Catégorisation rapide des scènes naturelles : L'objet, le contexte, et leurs interactions

Olivier R Joubert

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par *l'Université Toulouse III - Paul Sabatier*
Discipline ou spécialité : *Sciences Cognitives*

Présentée et soutenue par *Olivier R. Joubert*
Le *30 Septembre 2008*

Titre :
CATEGORISATION RAPIDE DES SCENES NATURELLES:
L'OBJET, LE CONTEXTE, ET LEURS INTERACTIONS

JURY

Pr. Aude Oliva – Brain and Cognitive Sciences – MIT, Boston	**Rapporteur**
Dr. Catherine Tallon-Baudry - LENA - CNRS, Paris	**Rapporteur**
Dr. Muriel Boucart - LNFP - CNRS, Lille	**Examinateur**
Dr. Jonathan Grainger – LPC - CNRS, Marseille	**Examinateur**
Pr. Jean-Michel Lassalle - CRCA - U. Paul Sabatier, Toulouse	**Examinateur**

Ecole doctorale : *CLESCO*
Unité de recherche : *CERCO - CNRS, UPS, UMR 5549*
Directeur(s) de Thèse : *Dr. Michèle Fabre-Thorpe*

« Partout où l'homme apporte son travail, il laisse aussi quelque chose de son cœur »
Henryk Sienkiewicz

Quand on entre en thèse, on ne sait pas forcément ce que l'on va rechercher… Quand on en sort, on ne peut qu'apprécier tout ce qu'on y a trouvé…Un épanouissement professionnel certes, mais également et surtout des rencontres humaines. Car ils ont été d'une manière ou d'une autre présents à mes cotés tout au long de cette grande et merveilleuse aventure, je tiens à exprimer mes remerciements et mon amitié…

A **Michèle Fabre-Thorpe,**
J'ai tant de raisons de te remercier…
Malgré tes responsabilités de directrice de laboratoire, tu as su rester disponible…
Tu as sculpté l'âme de chercheur latente au fond de moi en y gravant l'organisation, la minutie, la rigueur, la liberté, la patience, le courage, la soif d'apprendre…
J'ai tout autant apprécié nos enrichissantes discussions scientifiques que nos impétueux débats…
Mais tout cela n'est rien à côté du soutien et du réconfort que tu m'as apporté à tant d'occasions, aussi bien au niveau professionnel qu'extra-professionnel… et il y en a eu quelques unes !
Merci d'avoir été présente à mes côtés et jusqu'aux derniers mots de ce mémoire…

A **Denis Fize,**
Parfait complément de Michèle, cette thèse n'aurait pas eu la même saveur sans toi
Tu as été pour moi un encadrant à part entière et une aide précieuse.
Grâce à toi, nos amis Matlab et Fourier n'auront, je l'espère, bientôt plus de secrets pour moi…
Mais avant tout, tu t'es montré constamment à mon écoute, sincère, vrai…
Tu es et resteras pour moi un ami, un complice, un modèle de vie…

A **Guillaume Rousselet**,
Ô Cap'tain ! Mon capitaine de galère… Le gibbon ne se définit pas, il se vit !
En plongeant dans la thèse en compagnie de tant d'enthousiasme, il n'est pas surprenant d' en arriver à bout…
Merci de m'avoir initié à la recherche, merci de m'avoir fait voyager dans les méandres de la science, et finalement merci de m'avoir apporté tes conseils du bout du monde…

Aux rapporteurs **Aude Oliva** et **Catherine Tallon-Baudry,**
Pour votre sympathie et pour avoir accepté de relire mon travail malgré les contraintes temporelles.
Ainsi qu'à **Muriel Boucart**, **Jonathan Grainger** et **Jean-Michel Lassalle**,
Pour avoir accepté de faire partie des membres du jury de ma thèse.

A **ma famille**, **ma marraine** et plus particulièrement à **ma mère**,
Aucun mot ne pourra décrire tout ce qu'une mère peut apporter à son fils,
Si le sang nous lie, c'est l'amour qui nous réunit ,
Et surtout n'oublie pas… Quand tu diras à tout le monde combien tu es fière de ton fils,
Rappelle toi combien ton fils est fier de toi…

A **Magali**, **Marianne**, **Rudy**, **Rufin**,
Mes débuts au CerCo n'auraient pas été les mêmes sans vous quatre…
Partages quotidiens, évasions nocturnes, délires impromptus et soutiens en toute circonstances ont été les clés de ce fabuleux quintet. Si on a tendance à s'éparpiller autour du globe, c'est pas pour autant que je vous oublierai…vous, et toute l'aide que vous m'avez apporté !

A **Sébastien**,
Compagnon de bureau, compagnon de voyage, compagnon d'escapades, et ami jusqu'au cou !
Surtout reste comme tu es, car même tes quelques faiblesses (de tête en l'air) se transforment en qualité… Merci aussi à toi Tevy, la belle amazone, pour toute l'amitié que tu as su me porter…

A **Ludovic**,
On a partagé pas mal de choses, plus ou moins racontables, mais c'est dans ces moments là qu'on sait sur qui on peut compter. Et merci à Mylène de nous avoir supporté…

A **Nathalie**,
Tant de choses à te dire, pourtant inutiles quand on se comprend sans se parler…
Je me contenterai juste de te dire : Crois en tes rêves et fais tout ce que tu peux pour les réaliser…
Et finalement, merci d'avoir été là, tout simplement…

A **Catherine** et **Claire**,
Fidèles au poste, vous facilitez grandement la vie de tous les étudiants du laboratoire,
Et surtout de ceux (comme moi !) qui peuvent parfois bouder l'administratif et la recherche bibliographique…

A toutes les âmes du CerCo présentes, passées, et peut être futures,
et plus particulièrement à **Simon, Marc**, **Nadège**, **Leila**, **Vince**, **Julien**,
Zoe, **Yanica**, **Florence**, **Mathieu**, **Yves**, **Maxime**

A **Sophie**, **Lucille**, et **Charlène**,
Grâce à vous, j'ai beaucoup grandi, beaucoup muri, et beaucoup appris sur moi même…
Ce genre de relations n'a pas de prix.

A mes précieux amis de Bordeaux, et plus particulièrement

A **Pierre**.
En moins de 10 ans, tu es devenu d'un ami, un frère…
On a tout partagé, des soirées de fêtes aux instants plus tristes, et en chaque instant, j'ai su que je pouvais compter sur toi (et quelquefois sur ton papa). Merci pour tous ces moments partagés à tes côtés!

A **Mathieu.**
On a beaucoup de points en commun, et pas toujours les meilleurs. Mais on s'en tire à chaque fois avec le sourire. C'est ainsi qu'on apprend à affronter la vie… Bon d'accord, un peu de techno et d'alcool à l'occasion, ca aide aussi… Merci de rester fidèle à toi même

Sans oublier **Aude**, **Nancy**, **Racha**, **Jean-Baptiste**, **Julien**,
Vous m'avez chacun apporté des instants de bonheurs à votre façon. C'est toujours un réel plaisir de partager du temps avec vous, on ne se lasse jamais des personnalités hors du commun…

Liste de publications

1 – Articles Publiés

- Rousselet, G.A., **Joubert, O.R.**, & Fabre-Thorpe, M. (2005). How long to get to the “gist” of real-world natural scenes? Visual Cogn, 12 (6), 852-877.
- **Joubert, O.R.**, Rousselet, G.A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. Vision Res, 47 (26), 3286-3297.
- **Joubert, O.R.,** Fize, D., Rousselet, G.A., & Fabre-Thorpe, M. (2008). Early interference of context congruence on object processing in rapid visual categorization of natural scenes. (accepté pour publication)

2 – Articles en révision ou en préparation

- **Olivier R Joubert**, Guillaume A. Rousselet, Michèle Fabre-Thorpe and Denis Fize (en revision dans Journal of Vision). *“Rapid visual categorization of natural scene contexts with equalized power spectrum and increasing phase noise.”*
- Marc Macé, **Olivier R. Joubert** and Michèle Fabre-Thorpe (soumis). *”The animal before the bird: feature diagnosticity and categorization speed."*

3 – Résumés de conférences publiés

- **Olivier R. Joubert**, D. Fize, Guillaume A. Rousselet and Michèle Fabre-Thorpe. *“Categorization of natural scene : global context is extracted as fast as objects.”* European Conference on Visual Perception 2005, A Coruna, Spain. Perception suppl, volume 34, abstract 0375.
- **Olivier R. Joubert**, D. Fize, Guillaume A. Rousselet, and Michèle Fabre-Thorpe. *“What about background and context influences on ultra-rapid object categorisation?”* European Conference on Visual Perception 2006, St-Petersburg, Russia. Perception supplement, volume 35, abstract p.76.
- Marc J. Macé, **Olivier R. Joubert** and Michèle Fabre-Thorpe. *“Entry level at the superordinate level in visual categorization.”* Proceedings of the 9th International Conference on Cognitive and Neural Systems, 52.
- Marc J. Macé, Michèle Fabre-Thorpe and **Olivier R. Joubert**. *“Dog or animal? What comes first in vision?”* Perception, 34, suppl, 8.
- **Olivier R. Joubert**, Guillaume A. Rousselet, D. Fize, and Michèle Fabre-Thorpe. “*Rapid categorization of Natural or Man-made scene contexts: different effects with amplitude and phase alterations”.* Vision Science Society 2007.

Floride, 17 Mai 2007…

La conférence internationale Vision Sciences Society réunissant les plus grands chercheurs en neurosciences visuelles s'est conclue il y a quelques jours. 1500 scientifiques venus de tous les horizons pour présenter leurs travaux et se tenir au courant des nouvelles découvertes de par le monde… Tant de noms déjà rencontrés dans la littérature scientifique… Neuropsychologues, cogniticiens, physiologistes, modélisateurs, toutes les disciplines se sont mélangées pour partager leur savoir avec toujours le même objectif… Arracher ses derniers secrets au plus mystérieux des mondes… notre cerveau…

Parmi ces explorateurs d'un nouveau temps, 3 jeunes doctorants français…Pour eux aussi, la conférence a tenu ses promesses… de nouvelles rencontres, un savoir un peu plus conséquent, et surtout des idées plein la tête…A leur retour en France, la motivation sera d'autant plus grande à élucider de nouvelles questions, à mettre en place de nouvelles expériences dans leur champ de recherche respectif…Une recherche qui vous ravit autant qu'elle vous frustre… Sans fin, elle répond à vos questions par d'autres questions… Séductrice et tentatrice, elle se donne chaque jour un peu à vous, juste assez pour s'assurer que vous lui serez fidèle le jour suivant…Et si pour quelques jours, nos 3 jeunes étudiants lui étaient infidèles ? Et si pour quelques jours, nos 3 jeunes chercheurs en profitaient pour explorer un autre monde, la Floride… Terre de paradoxe, mêlant urbanisme démesuré et terres sauvages…

Les voilà filant en voiture sur la route 75 qui longe le Pacifique. La nuit est calme, chaude et humide comme toujours en cette saison. A la frontière des Everglades largement réputées pour être le plus grand marécage des Etats Unis, les 3 compères se dirigent vers Miami et les Keys à toute allure, espérant admirer les premiers rayons de soleil sur les mers du sud. Au travers des vitres, le paysage presque menaçant défile… Une végétation abondante, humide et dense, qui se laisse bercer au gré des courants d'air, abritant sans nul doute une faune silencieuse qu'on devine sans pour autant percevoir…Une flore sauvage qui menace de reprendre ses droits sur cette route interminable… L'homme ici encore a dominé la nature, signant son passage d'une encre de bitume…Et toujours aucune autre voiture que la leur, ils pourraient presque s'imaginer pionniers… Les phares avant de la berline concourent avec la lune pour éclairer leur progression au gré des kilomètres qui défilent… vite… peut être beaucoup trop vite…

Il s'en est fallu de quelques secondes pour éviter l'accident ! Cette minuscule surface sphérique réfléchissant la lumière des phares à l'extrémité d'une forme obscure, indéfinie, allongée sur l'asphalte… 2 ou 3 mètres d'écailles humides reflétant les rayons de la lune… Et combien de millisecondes pour identifier l'animal ? Combien de millisecondes pour que notre système cognitif puisse transcrire l'information cachée dans les photons frappant notre rétine ? Combien de temps pour que notre cerveau puisse détecter, catégoriser et reconnaître le danger sur la base des quelques indices perçus? On peut d'ailleurs se demander quelles caractéristiques de l'animal se sont avérées diagnostiques ? Sa forme, sa texture, sa couleur, sa position dans l'espace ? Aurait-on simplement détecté l'animal s'il n'avait pas été au beau milieu de la route ? Aurait-on eu besoin d'un temps supplémentaire si l'environnement avait été différent, si l'action s'était déroulée dans la cathédrale Saint Sernin, plutôt qu'au milieu des marécages de Floride ? Finalement, combien de millisecondes auront été suffisantes pour qu'une action salvatrice, ici la décision de freiner, soit déclenchée sur la base d'une information visuelle peut être encore incomplète ?

Quelles que soient les réponses à ces questions, nous pouvons affirmer que le cerveau s'est une nouvelle fois avéré réellement efficace et rapide dans une situation qui aurait pu mettre en danger la vie d'êtres humains. Brillant produit de l'évolution des espèces et du développement de l'individu, son efficacité semble en majeure partie reposer sur de simples décharges de neurones en accord avec les lois physiques et électrochimiques…Des neurones interconnectés, tissant une toile en 3D dans l'enceinte de notre crâne, alliant des systèmes responsables de différentes fonctionnalités telles que la vision, l'audition, la spatialisation, le langage, la mémoire, ou encore la conscience et les émotions…

C'est une infime parcelle de ce vaste monde cérébral que je vous propose d'explorer à mes côtés… A l'orée des études sur la perception visuelle rapide des objets, une nouvelle voie de recherche est en défrichage… un sentier battu sinueux, entre objets isolés et scènes naturelles intégrales que l'on appelle contexte… Un contexte porteur d'informations que vous utilisez de manière inconsciente en chaque instant, alors que seuls les objets vous semblent d'intérêt…

Mais si aucun contexte n'avait été décrit précédemment, si l'action ne s'était pas déroulée à la lisière des marécages de Floride, aurions-nous simplement pensé au même animal ?

PARTIE I

ETAT DE L'ART

What is mind ?
No Matter
What is matter ?
Never mind

René Descartes

1. Les scènes naturelles et leurs caractéristiques

Prenons un instant pour découvrir de nouveau ce monde qui chaque jour nous entoure… De nombreux objets participent à notre environnement. Ils ont des formes variées, des surfaces plus ou moins grandes, respectent des règles physiques qui leurs sont propres, et pourtant, ils ont tous un point commun : en absence totale de lumière, leurs existences visuelles cesseraient. On perçoit le monde qui nous entoure uniquement grâce à la présence d'une source de lumière, la plupart du temps extérieure aux objets. Cette source de lumière peut être naturelle comme le soleil, les flammes d'un feu de camp, les lucioles mais peut être également artificielle comme les lampes électriques. La nature de la lumière a cependant été longuement débattue. Jusqu'au 18ème siècle, en grande partie sous l'influence d'Isaac Newton, de nombreux scientifiques prêtent à la lumière une nature corpusculaire : la lumière serait constituée de particules élémentaires, qui seront appelées plus tard « photons ». Par la suite, diverses expériences menées par Thomas Young et Augustin Fresnel (19ème siècle) mettant en évidence les phénomènes d'interférences et de diffraction de la lumière supporteront grandement l'hypothèse d'une nature ondulatoire de la lumière. Cette théorie ondulatoire se verra d'ailleurs confirmée par les recherches de Maxwell et de Hertz à la fin du 19ème siècle. Il faudra attendre les travaux de Millikan, puis d'Einstein (1916) sur la loi de Planck traitant des rayonnements du corps noir pour redonner à la lumière sa nature bimodale, à la fois corpusculaire et ondulatoire. Il est désormais reconnu que les photons, particules élémentaires électriquement non chargées, sont constitutifs d'ondes électromagnétiques et peuvent être échangés lors de l'absorption ou de l'émission de lumière par la matière. De manière anecdotique, ces études sur la lumière auront eu le mérite d'apporter la reconnaissance à ses investigateurs puisque entre autres, Max Planck (1918), Albert Einstein (1921) et Robert A. Millikan (1923) reçurent le Prix Nobel pour leurs travaux respectifs.

S'il était scientifiquement important de définir physiquement et entièrement la lumière, sa nature ondulatoire s'avère dans notre domaine de recherche la plus intéressante. En effet, tandis que l'ensemble des ondes électromagnétiques se distribuent sur un spectre large dépassant les limites des infrarouges et des ultraviolets, seules les ondes dont la longueur d'onde se situe entre 380 nm (violet) à 780 nm (rouge) sont perceptibles par l'œil humain. C'est donc la proportion des différentes longueurs d'ondes de ces photons qui va coder pour

l'intensité lumineuse et les couleurs, tandis que les variations de proportions coderont pour les contrastes globaux et locaux, seront à la base des contours et des formes et à la base de notre percept visuel de l'environnement dans sa globalité ! Il était donc bien normal de rendre hommage à la lumière en ce début de mémoire…

1.1. Définir la notion de scènes naturelles…

Il n'existe pas de scène naturelle plus pure et plus détaillée que notre environnement. On estime qu'un seul exemplaire de scène naturelle équivaut à plus de 1000 mots (Friedman, 1979). Que l'on soit à l'intérieur d'une cuisine, allongé sur le bord de la plage, ou encore en ballade en forêt, le monde environnant rassemble un nombre de propriétés visuelles importantes que tout chercheur dans le domaine de la vision se doit de comprendre et si possible de quantifier. Cependant, si de nombreuses dimensions physiques visuelles plus ou moins complexes ont été définies, rien ne prouve que nous soyons au bout de nos découvertes. Une approche de la question consiste à prendre comme point de départ les propriétés même de la lumière reflétée par le monde alentours, et d'en extraire le maximum de variables. Etant donné que la valeur des variables divergent légèrement entre la scène naturelle même et sa capture photographique, nous traiterons de ces variables dans le cadre des photographies, stimuli utilisés dans notre sujet d'étude. Il faut cependant garder en tête que la photographie n'est qu'une image statique de la scène en un instant donné et dépourvue de profondeur stéréoscopique.

1.2. Caractéristiques physiques bas-niveau des photographies

1.2.1 Couleurs

Il existe dans notre monde deux types de lumières : les lumières monochromatiques et les lumières polychromatiques. Tandis que les rayonnements monochromatiques ne sont constitués que d'une longueur d'onde, les rayonnements polychromatiques renferment un ensemble de longueurs d'ondes appelé spectre. L'œil humain n'a cependant pas une physiologie lui permettant de capter l'ensemble du spectre, mais seulement les ondes

s'étendant de 380 à 780 nm, on parle de spectre visible. En effet, les cônes, photorécepteurs au sein de la rétine sont divisés en 3 catégories : les erythrolabes, les chlorolabes, et cyanolabes, respectivement appelés respectivement cônes rouge (564 nm), cônes vert (534 nm) et cônes bleu (420 nm).

Or si les objets reçoivent l'ensemble de la lumière naturelle, leur nature et leur texture vont définir quelles longueurs d'onde seront absorbées, lesquelles seront renvoyées, et ainsi leur couleur perçue. Dans une image électronique, le codage RVB est le plus souvent utilisé : les couleurs sont retranscrites sur 3 couches de couleurs : rouge, vert, bleu en fonction de l'intensité de chacune des 3 composantes en chaque point de l'image (Figure 1).

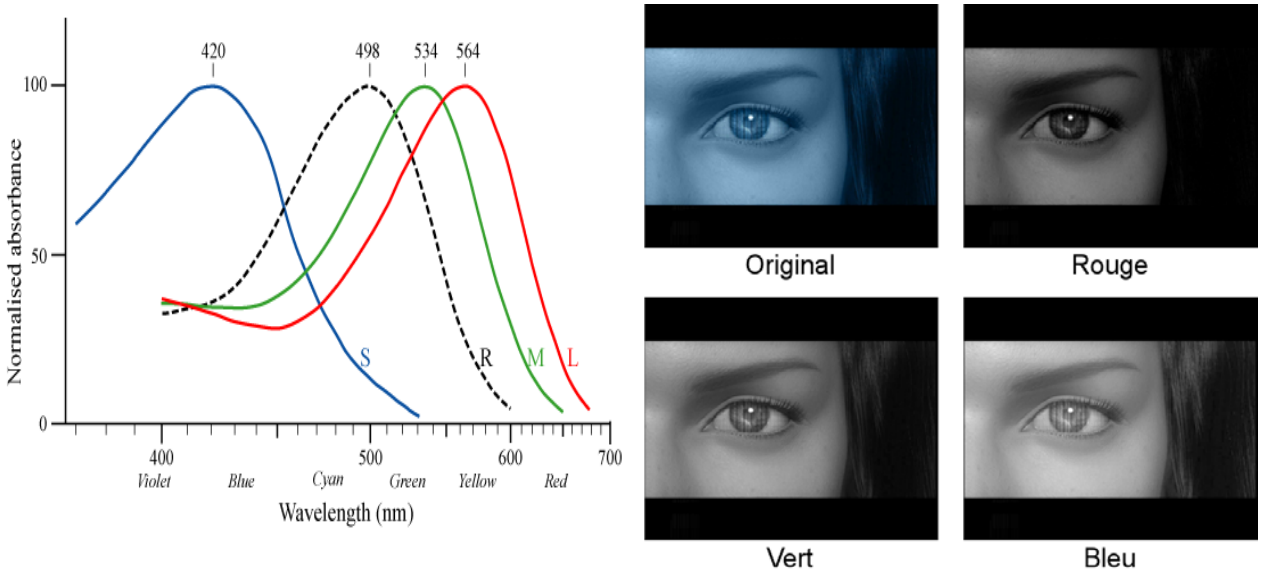

Figure n°1 : Chaque image polychromatique peut être décomposée en 3 canaux de couleurs rouge, vert, bleu. Au niveau de l'œil humain, les 3 familles de photorécepteurs traitent spécifiquement une des 3 composantes colorimétriques de l'image.

S'il est clair que la couleur est naturellement présente dans notre environnement, il est cependant intéressant d'évaluer son influence sur les traitements visuels nous permettant d'appréhender notre environnement et ainsi de donner un sens aux objets et à la scène globale. Des études menées sur la dénomination et la reconnaissance d'objet ont montré qu'au moins au niveau basique, nous étions aptes à reconnaître des objets isolés aussi rapidement et précisément en niveau de gris qu'en couleurs (Biederman, Blickle, Teitelbaum & Klatsky, 1988 , Davidoff & Ostergaard, 1988, Ostergaard & Davidoff, 1985). Les couleurs pourraient cependant s'avérer importantes dans les tâches de dénomination ou de catégorisation plus fine des objets. En réponse à cette interrogation, Biederman (Biederman et al., 1988) a démontré que même des objets caractérisés par une couleur hautement diagnostique (telle qu'une

orange) étaient aussi rapidement reconnus sur un dessin en noir et blanc que colorés. D'un autre côté, ils semblent que certaines catégories comme les fruits et légumes soient dénommés plus rapidement lorsqu'ils sont présentés en couleur (Davidoff & Ostergaard, 1988) que l'objet soit affiché clairement ou flou (Wurm, Legge, Isenberg & Luebker, 1993). De manière assez logique, les fruits et légumes sont moins rapidement dénommés quand des couleurs anormales leurs sont ajoutées (Price & Humphreys, 1989). Un léger avantage des stimuli colorés est également observé lors de la catégorisation rapide d'aliments présentés dans une scène complexe alors qu'elle est quasi-inexistante lors d'une catégorisation animal/non-animal (Delorme, Richard & Fabre-Thorpe, 2000) et cet avantage n'est pas enregistré sur les réponses les plus précoces.

Les informations colorimétriques pourraient donc être utilisées de manière tardive pour dénommer ou catégoriser finement un objet uniquement lorsqu'elles sont diagnostiques. Des traitements visuels de l'information achromatique pourraient donc être effectués rapidement par le système magnocellulaire de la voie ventrale, que je décrirai ultérieurement, afin de proposer une description grossière de l'image. Cette représentation rudimentaire de l'objet pourrait être par la suite complétée par l'intégration des informations chromatiques dépendantes du système parvocellulaire plus lent (Macé, Thorpe & Fabre-Thorpe, 2005), également décrit par la suite.

Nous verrons cependant tout au long de ce mémoire de thèse que les informations visuelles caractérisant le contexte et la représentation globale d'une scène divergent des informations visuelles nécessaires à la reconnaissance d'un objet. Il reste donc à déterminer si la couleur constitue une information essentielle pour la reconnaissance du contexte.

1.2.2 Luminance

Comme on l'a vu précédemment, les couleurs d'une photographie sont codées sur 3 couches respectivement pour les couleurs : rouge, vert, bleu. Dans mes travaux, lorsqu'il s'agira de s'affranchir de la composante chromatique, on va transformer l'image en niveau de gris. Chaque pixel sera un pixel gris codé par une seule valeur entre 0 et 255 correspondant à la luminance de ce point dans l'image. Cependant, la luminance de gris n'est pas égale à la luminance moyenne des 3 composantes de couleurs de l'image originale. Le C.I.E

(Commission Internationale de l'Eclairage) propose ainsi pour une image électronique visionnée sur un écran d'ordinateur la fonction de conversion suivante :

$$Gris = 0,299 \cdot Rouge + 0,587 \cdot Vert + 0,114 \cdot Bleu$$

Cette formule prend en compte la perception des 3 composantes colorimétriques par l'œil. On peut ainsi trouver une valeur d'intensité de gris pour chaque pixel de l'image et de là obtenir une distribution de luminance de l'image.

De nombreuses études contrôlent désormais le biais potentiel engendré par la luminance moyenne de stimuli variables. On peut en effet modifier la luminance de chaque image afin que tous les stimuli aient la même luminance moyenne. Il faut cependant savoir que cette opération peut entraîner une saturation des pixels ayant une luminosité extrême. Mais quelle est l'influence de la luminance globale des scènes naturelles sur les performances comportementales ? Peu d'études à ce jour ont étudié concrètement cette question. Les résultats d'une expérience de catégorisation rapide animal/non-animal menée chez l'homme montrent que la luminance moyenne des stimuli n'a qu'une influence très faible sur les performances (Macé, 2006). Tandis que les images étaient flashées 28 ms, une baisse de précision d'à peine 2% et une augmentation des temps de réaction moyens de 20 ms ont été enregistrées pour des images dont la luminance des pixels avaient été déplacée de +/- 48 sur l'axe de luminance (Figure 2).

Une autre étude cette fois basée sur l'enregistrement de saccades oculaires démontre une très faible sensibilité du système visuel à des changements globaux de luminance entre deux fixations oculaires (Henderson, Brockmole & Gajewski, 2008).

D'un point de vue plus local, la luminance d'une région donnée de la scène contribue entre autres à l'intégration des informations relatives aux couleurs, aux textures (Hanazawa & Komatsu, 2001), à la segmentation des surfaces (Fine, MacLeod & Boynton, 2003). Elle influence notre rapidité à détecter des visages (Lewis & Edmonds, 2003) et module la capture exogène de l'attention (Turatto & Galfano, 2000). De plus, le contraste résulte de la juxtaposition de pixels de luminances différentes.

Diverses études ont cependant montré que dans une scène naturelle, luminance et contraste locaux sont indépendants (Frazor & Geisler, 2006), tous deux capturés par des traitements visuels indépendants au sein du corps genouillé latéral (LGN, Mante, Frazor,

Bonin, Geisler & Carandini, 2005). Aucune étude n'a cependant encore précisé l'influence de la luminance globale sur la reconnaissance du contexte au sein d'une scène naturelle.

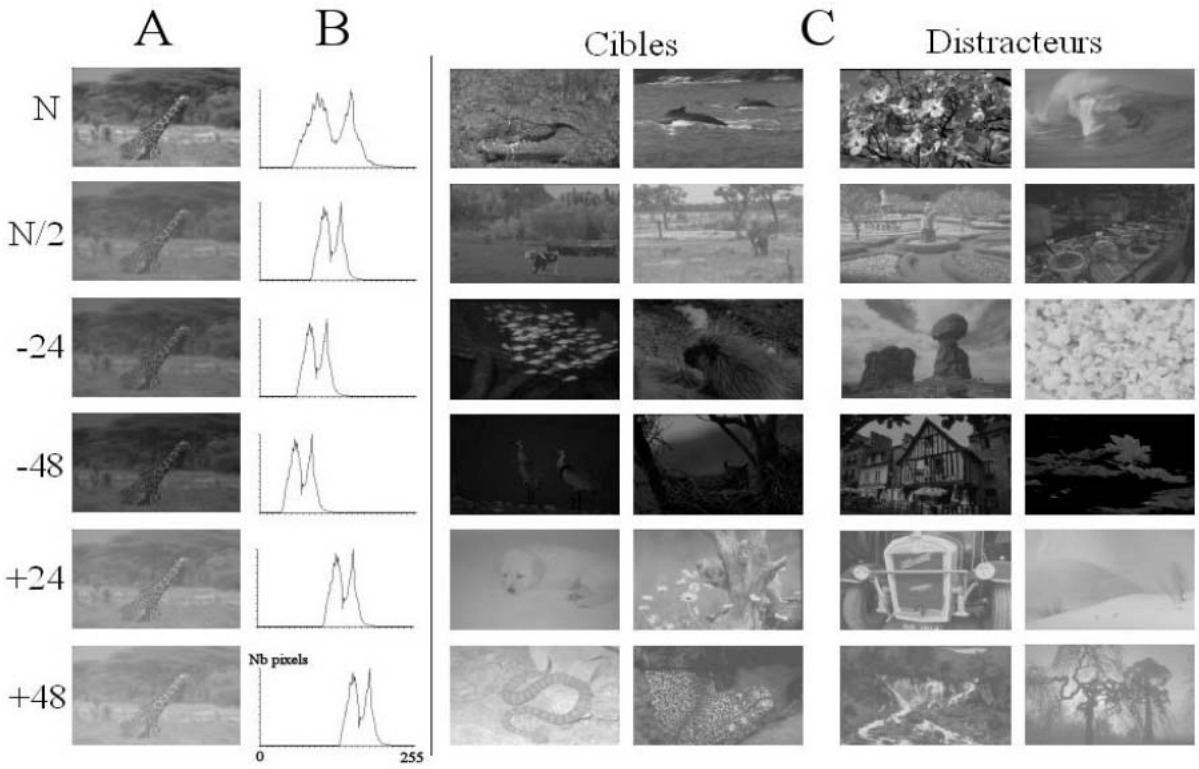

Figure n°2 : Figure et légende tirées de (Macé, 2006). A : Même image vue dans toutes les conditions de présentation. B : Histogramme de distribution des valeurs de luminance des pixels de l'image en A. Lorsqu'on divise le contraste par 2, l'écart de la luminance d'un pixel donné par rapport à la luminance moyenne de l'image est divisé par deux. Cela a pour effet de resserrer l'histogramme de distribution des luminances de tous les pixels d'une image autour de sa luminance moyenne (N/2) et de limiter la proportion de pixels saturés à 0 ou 255 dans les conditions -48 et +48. C : Exemples de stimuli utilisés dans la tâche animal/non animal.

1.2.3 Contraste global et local

On appelle contraste global la variance de la luminance, correspondant à l'écart-type de la distribution de luminance des pixels au sein d'une image. Tandis qu'une augmentation du contraste se traduit par une distribution de la luminance des pixels plus étalée, une réduction du contraste d'une image se traduit par un resserrement de la distribution. L'image apparaît alors moins nette comme si la photographie avait été prise dans le brouillard.

Le contraste local d'un pixel donné peut quant à lui être calculé selon la fonction suivante, Michelson contrast : (Lmax-Lmin)/(Lmax+Lmin). Un contraste local d'un pixel proche de 0 suggérera souvent l'appartenance du pixel à une surface relativement uniforme étant donné

que les pixels voisins sont de luminance similaire. Un contraste local proche de 1 suggérera l'existence d'un contour. Une analyse effectuée sur deux pixels voisins permettra de déterminer l'orientation privilégiée du contour. Dans tous les cas, le contraste local définit la netteté et l'intensité des contours.

Une fois de plus, peu d'études ont soulevé la question de l'influence du contraste sur la reconnaissance des scènes naturelles. Une première étude se basant sur une expérience de reconnaissance d'objets dessinés en noir et blanc montrent que les performances humaines restent bonnes pour des conditions de contrastes supérieures à 10% (Avidan, Harel, Hendler, Ben-Bashat, Zohary & Malach, 2002). De plus, la réduction du contraste global ne semble avoir qu'un faible effet sur notre faculté à détecter des visages (Lewis & Edmonds, 2003).

L'influence du contraste sur les traitements catégoriels a également été testée dans une tâche de catégorisation rapide plus proche de notre paradigme d'étude, au cours de laquelle les sujets devaient répondre le plus rapidement possible dès qu'ils apercevaient un animal dans des scènes flashées pendant 28 ms. Les photographies achromatiques avaient été au préalable altérées par une réduction de contraste plus ou moins importante. Les résultats démontrent des performances humaines robustes à la réduction du contraste puisque les sujets humains sont largement au dessus du niveau chance (70-80% correct) pour des conditions de contrastes réduit à 12/10% du contraste initial. Une augmentation maximale des temps de réaction de 60 ms est également constatée pour les images dont le contraste a été réduit à 8% du contraste initial (Mace et al., 2005, Figure 3).

A part dans l'étude précédente, l'importance du contraste local dans la perception des contextes de scènes naturelles n'a jamais été étudiée à ma connaissance. Une explication simpliste de cette apparente lacune réside dans le fait que l'étude des contrastes locaux est directement liée à l'étude des contours et surfaces. Si ce type d'étude est relativement aisé avec des stimuli simples, elle l'est beaucoup moins avec des scènes naturelles. Une approche alternative consiste à évaluer la capacité de notre système visuel à extraire l'information présente dans les fréquences spatiales d'une scène naturelle.

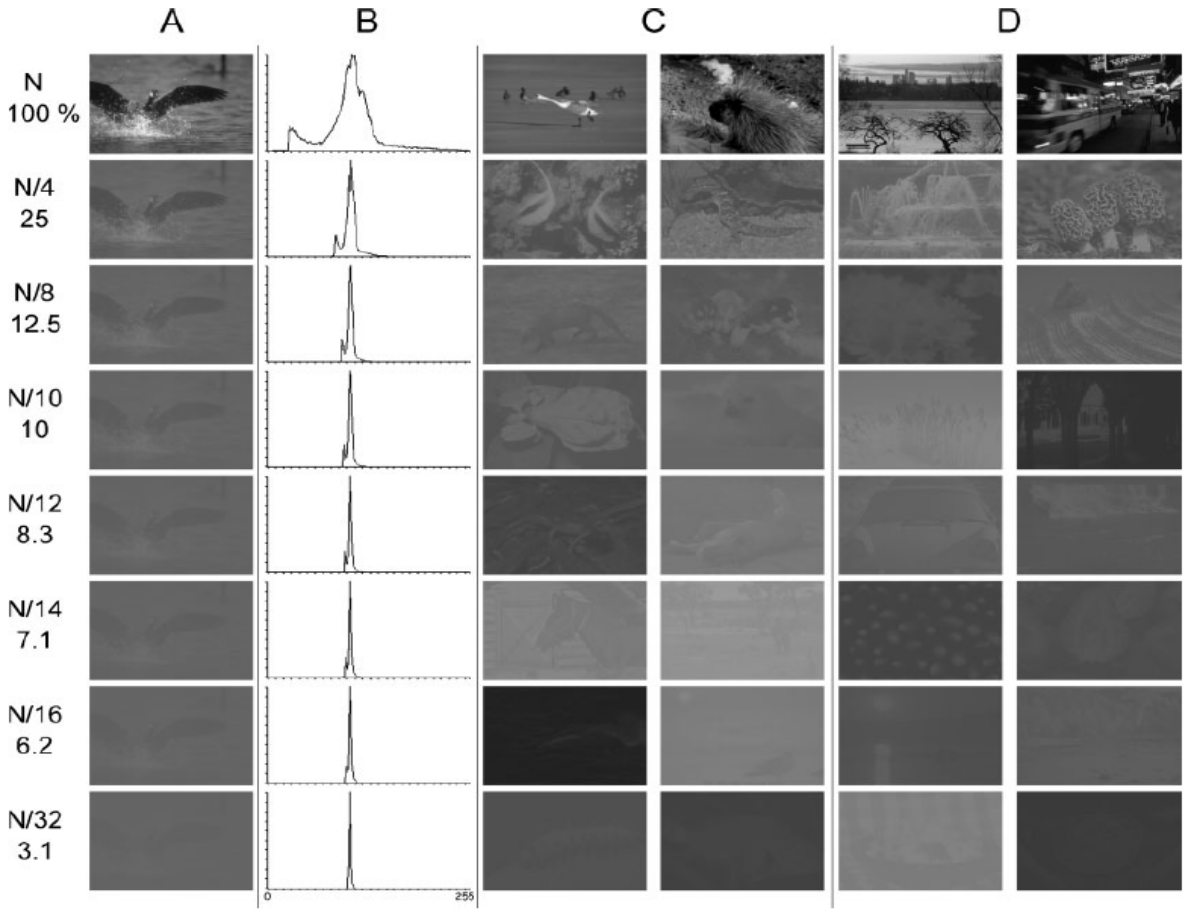

Figure n°3 : Tiré de (Mace et al., 2005). 1. Exemples de stimuli pour les 8 conditions de contrastes testées dans l'étude. Le contraste résiduel pour chaque condition (de N à N/32) est calculé en fonction du contraste de l'image originale (N) considéré à 100%. A. La même image présentée dans 8 conditions de contraste différentes. B. La distribution de luminance des pixels de l'image en A pour chaque condition, centrée sur la même luminance moyenne. C. et D. Des exemples de cibles (C) et de distracteurs (D) dans les 8 conditions de contrastes testées.

1.2.4 Fréquences spatiales

On définit une fréquence spatiale comme étant l'inverse d'une distance angulaire. On peut par exemple décomposer l'évolution de l'intensité du niveau de gris selon une orientation donnée. La fréquence spatiale correspond alors à l'inverse de la période d'une sinusoïdale selon laquelle le motif de l'image se répète.

Or, selon la théorie de Fourier, toute fonction périodique de fréquence f, peut se décomposer en une somme infinie de fonctions sinusoïdales de fréquences multiples de f (Figure 4).

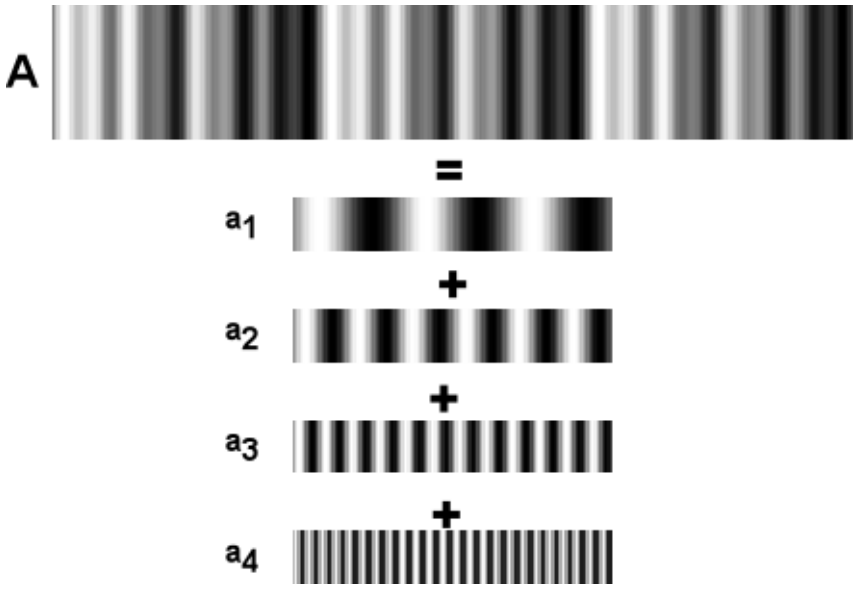

Figure n°4 :: La fonction périodique A peut se décomposer en une somme infinie de fonctions sinusoïdales de fréquences multiples de A. a_1, a_2, a_3, a_4 sont les 4 fonctions sinusoïdales majeures de A. Leur sommation suffit à recomposer grossièrement le signal A

En d'autres termes, en effectuant une sommation linéaire de plusieurs fonctions spatiales sinusoïdales, on obtient un signal spatial linéaire plus complexe. Bien sûr, toutes les fréquences spatiales ne vont pas contribuer de manière équivalente à la composition du signal, certaines fréquences spatiales peuvent être absentes du signal. La participation de chaque fonction est ainsi pondérée par un coefficient appelé coefficient de Fourier, correspondant à l'intensité de la fréquence spatiale.

Un exemple de décomposition d'un signal linéaire effectué sur deux scènes naturelles : un environnement manufacturé et un environnement naturel est présenté figure 5. En bas de la figure sont représentées les 3 fréquences spatiales majeures composant respectivement chaque scène.

En décomposant le signal de diverses images, on constate rapidement que les fréquences spatiales basses sont spécifiques de motifs larges (de larges surfaces comme la plage ou la mer…) tandis que les hautes fréquences spatiales sont spécifiques de motifs répétés de manière rapprochée (plutôt des contours comme ici, les contours verticaux des immeubles). Cette décomposition du signal effectuée dans l'exemple selon une unique orientation horizontale (0°) peut s'effectuer également sur l'ensemble des orientations (de 0 à 360°). On

évalue ainsi la contribution de chaque fréquence spatiale dans l'image distribuée sur l'ensemble des orientations. Procédant ainsi, on s'aperçoit que des scènes naturelles complexes et chargées en détails vont être en moyenne composées d'un plus grand nombre de hautes fréquences spatiales que des scènes naturelles ayant de larges surfaces homogènes. Finalement, la scène ayant un arrangement précis assurant la cohérence des objets et de la scène dans sa globalité, il est primordial de prendre en compte une variable définissant la position relative de l'information contenue dans les fréquences spatiales, c'est ce qu'on appelle la phase des fréquences spatiales.

Chaque photographie peut ainsi être représentée dans un nouvel espace dans lequel orientation, intensité et phase des fréquences spatiales sont codées. Cet espace appelé « Espace de Fourier » est un outil largement utilisé dans notre domaine de recherche pour mieux appréhender le système visuel souvent décrit comme un analyseur de Fourier (Marr, 1982 , Westheimer, 2001).

Un exemple d'application de la théorie de Fourier est le filtre passe-bas d'images.

Si tout signal complexe peut être décomposé en une infinité de composantes périodiques, il est également possible de reconstituer un signal complexe connaissant la contribution de chaque composante périodique. On peut donc décomposer une image en fréquences spatiales, puis recomposer l'image en ne préservant que ses fréquences spatiales basses par exemple.

Un des articles de ce mémoire évalue la contribution des dimensions de l'espace de Fourier dans les traitements catégoriels. Une présentation plus approfondie de l'espace de Fourier et un état de l'art spécifique de la question seront présentés dans un chapitre ultérieur.

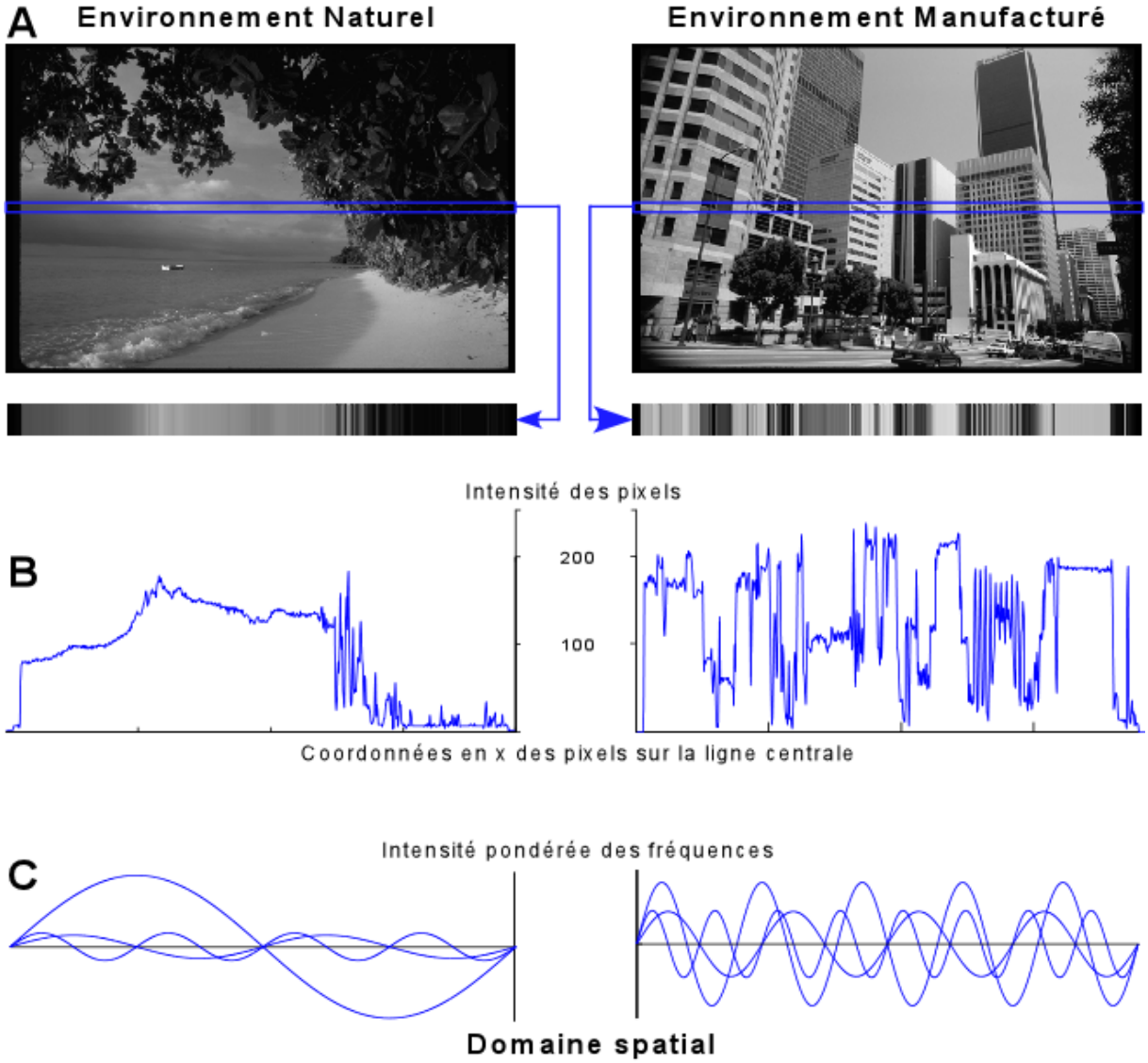

Figure n°5 : A. Deux exemplaires de scènes naturelles : un environnement naturel et un manufacturé dont on isole la ligne de pixels centrale. B. Transcription de l'information visuelle des lignes de pixels en intensité lumineuse. C. Les 3 fréquences spatiales pondérées par leur intensité majoritaires dans le signal lumineux : Les environnements manufacturés sont caractérisés par des fréquences spatiales hautes plus importantes.

1.2.5 Organisation spatiale

Les informations visuelles extraites de notre environnement permettent au système visuel de générer une représentation spatiale cohérente du monde alentours. La reconnaissance des objets, du contexte et la perception de notre corps dans l'espace serait impossible si les coordonnées spatiales de chaque information visuelle n'étaient pas relativement préservées. Relativement car on peut imaginer qu'une partie de l'information est traitée en termes de position relative et donc non absolue. En effet, les aires visuelles selon leur niveau dans

l'organisation pseudo-hiérarchique du système visuel ne semblent pas coder la position de la même manière. Au niveau de l'aire V1, les petits champs récepteurs des neurones préservent la localisation spatiale de l'information rétinienne selon une organisation rétinotopique. Au contraire, dans le cortex occipito-latéral (LOC), les champs récepteurs majoritairement invariant à la position absolue des stimuli, tendraient à coder une information de position plus relative mais suffisante à l'intégration de formes complexes cohérentes (Rousselet, Thorpe & Fabre-Thorpe, 2004).

Cependant, même dans cette aire haut-niveau, certains biais concernant la position des champs visuels persisteraient. En effet, les résultats d'une étude IRM tendent à laisser penser que la notion de champ visuel central vs. périphérique persisterait dans les aires haut-niveau et modulerait la représentation d'objets complexes. La représentation complexe d'objets nécessitant une prise d'information détaillée en région fovéale (tel que les visages) serait plutôt effectuée par des cellules du LOC fortement associée aux traitements de l'information centrale tandis que la représentation des objets considérés comme majoritairement périphériques (bâtiments, objets contextuels) serait fortement associée aux traitements de l'information périphérique (Levy, Hasson, Avidan, Hendler & Malach, 2001). Paradoxalement, une étude de catégorisation animal/non-animal effectuée avec des scènes naturelles présentées en périphérie démontrent que l'être humain est encore capable de détecter un animal affiché en périphérie lointaine à une excentricité de plus de 75° (Thorpe, Gegenfurtner, Fabre-Thorpe & Bulthoff, 2001). Dans ce sens, il serait intéressant de tester les performances des sujets dans une tâche de catégorisation de contexte périphérique.

1.3. Les photographies : des stimuli assez proches de la réalité…

Tout au long des études présentées dans ce mémoire, nous avons utilisé comme support expérimental les photographies, stimuli assez proches de la réalité sans en être une réplique exacte. En effet, contrairement à notre environnement réel, la photographie est dépourvue de la dimension de profondeur stéréoscopique. Quand nous observons le monde réel, nos deux yeux extraient des informations visuelles 2D similaires mais légèrement décalées spatialement (disparité verticale et horizontale) nous permettant ainsi d'appréhender la profondeur du monde par une vision 3D stéréoscopique. Cette vision 3D permet de porter un jugement sur la taille réelle et la forme des objets en fonction de leurs distances estimées. Les photographies

ont aussi le désavantage d'être statiques et donc dépourvues de dimension de mouvement. Tandis que les changements dans notre environnement apparaissent majoritairement par un côté de champ visuel, les photographies fournissent une information visuelle immobile, constante, immédiatement disponible.

Malgré ces quelques lacunes, la photographie d'une scène naturelle reste l'un des stimuli les plus complexes parmi la liste des stimuli utilisés dans le domaine de la vision. La majorité des études jusqu'à présent ont travaillé sur des stimuli plus simplifiés : « Random Dot kinetograms », gabors, formes non-figuratives, dessins faits main… La plupart des expériences de catégorisation d'objets ou de scènes ont jusque là été réalisées sur des représentations graphiques d'objets et de scènes dessinées à la main, accentuant grandement l'importance des contours. Contrairement aux photographies, ces dessins « faits main » fournissent peu ou pas le rendu des dégradés colorimétriques, de textures, ou encore d'ombrages. Certains s'affranchissent même des occlusions entre objets d'une même scène. Cet ensemble d'indices visuels est pourtant caractéristique d'une perception 3D monoculaire permettant d'estimer la taille réelle, la distance et l'orientation des objets.

Dans le cadre de nos études sur la reconnaissance de scènes à partir d'informations visuelles brièvement disponibles, les photographies contenant la gamme complète des informations chromatiques, lumineuses, fréquentielles, spatiales et 3D monoculaires semblent donc un atout majeur pour nos expériences. De plus, ces caractéristiques physiques sont distribuées sur l'ensemble de la scène naturelle tant pour l'objet que pour son contexte. C'est aussi pour ce type de stimuli que notre système visuel a été façonné par l'évolution.

1.4 Scènes naturelles et information visuelle bas-niveau

D'un point de vue écologique, on peut supposer que le fonctionnement de notre système visuel s'est construit à partir des régularités de l'environnement. Attneave (Attneave 1954) et Barlow (Barlow, 1961) furent les premiers à suggérer que notre système visuel aurait pu s'adapter au cours de l'évolution et de notre développement aux statistiques du monde environnant (« efficient coding »). Avec l'avancée de la technologie, de nombreuses études se sont déjà attachées à faire le lien entre les statistiques des scènes naturelles et les mécanismes visuels bas-niveau (Bell & Sejnowski, 1997, Field, 1987, Field, 1994, Olshausen, 2003, Olshausen & Field, 1996a, Simoncelli & Olshausen, 2001). Dans une économie d'énergie, un

des points forts du système visuel reposerait donc sur sa capacité à coder l'ensemble de l'information importante via l'intervention d'un nombre minimal de neurones (« sparse coding ») tout en limitant la redondance, pourtant hautement présente dans les scènes naturelles. Sur une photographie, la probabilité de corrélation entre deux pixels voisins par exemple est énorme : ils auront sûrement la même luminance, la même couleur... Si on peut donc apparenter la transcription de l'information visuelle au niveau des photorécepteurs à un codage pixel par pixel, il semble que dès V1, des fonctions de réponse indépendantes distribuées sur un ensemble de cellules de V1 seraient de mise (Olshausen & Field, 1996b).

Ces fonctions pourraient s'intégrer à un modèle visuel bas-niveau conçu tel un analyseur de Fourier (Marr, 1982 , Westheimer, 2001). Si l'hypothèse d'une transformée de Fourier sur l'ensemble du champ visuel (Westheimer, 2001) a été mise de côté, l'hypothèse d'une analyse par bandes de fréquences (Koenderink, 1984) ou par « ondelettes» (Mallat 1987) reste tout à fait envisageable. Elle présente d'ailleurs un avantage sur l'encodage pixel par pixel : elle permet d'encoder dans le même temps la dimension spatiale et le contenu fréquentiel (Daugman, 1984, Daugman, 1985). Cependant, un tel modèle est viable uniquement s'il se base sur le traitement d'un signal n'ayant subi jusque là que des opérations linéaires. Ce pré-requis est respecté au niveau physiologique. En effet, Enroth-Cugell et Robson (Enroth-Cugell & Robson, 1966) ont montré chez le chat que les cellules ganglionnaires « X-cells » sommaient de manière linéaire les stimuli sur l'ensemble du champ récepteur, que les régions du champ soient excitatrices ou inhibitrices. Cette sommation linéaire a également été retrouvée chez le singe par Hubel et Wiesel lorsqu'ils ont attribués aux cellules simples de V1 cette fonction de détecteur de barre (« bar-detector »). D'ailleurs, si les cellules de l'aire V1 répondent de manière optimale à des barres orientées de longueurs et largeurs diverses, cela signifie qu'elles sont sélectives aux dimensions de taille et d'orientation (Blakemore & Campbell, 1969), elles répondent donc de manière spécifique à des régions de l'espace 2-D de Fourier. Suite à ces travaux, Campbell et Robson ont également suggéré un modèle visuel constitué de canaux indépendants et spécifiques aux fréquences spatiales (Campbell & Robson, 1968).

Adelson et Bergen (Adelson & Bergen, 1991) ont tenté de préciser la nature de l'information visuelle bas-niveau en tentant de définir ce que le système visuel était capable d'extraire de son environnement. Quelles informations sont contenues dans la distribution de lumière d'une région de l'espace ? En réponse à cette question, ils ont proposé une fonction appelée « plenoptic function ». Cette fonction à la base indépendante d'un organisme

spécifique, peut s'appliquer au primate en prenant en compte 5 paramètres différents P = P(x,y,t,λ ,Vx). : les 2 coordonnées spatiales de la région spatiale considérée (x et y), la variable temps (t), les longueurs d'ondes caractéristiques de la couleur et intégrées par les 3 types de cônes chez l'humain (λ) et un dernier paramètre V prenant en compte la notion de disparité. Une bonne estimation de cette fonction peut être obtenue par le calcul de la moyenne de dérivés locales de cette fonction aboutissant à ce qu'Adelson et Bergen appelle des «low-order derivative operators ». Il est intéressant de noter que l'ensemble des opérateurs obtenus peut facilement s'intégrer à la structure des champs récepteurs de cellules simples. De plus, la description dans le domaine de Fourier de ces opérateurs dérivés correspond à un filtre de bande passante, et donc peut être assimilé à un canal de fréquences spatiales.

Ce type d'opérateurs peut représenter une solution pour l'intégration d'informations bas-niveau et servir de base à la construction de représentations haut-niveau interprétables.

2. Définition du contexte, de l'objet: une problématique en soi?

Lorsque les photons caractérisant le monde visuel environnant entrent en contact avec notre rétine, ils sont porteurs d'informations diverses et organisées spatialement nous permettant de percevoir un monde cohérent. Cependant, ils ne sont aucunement messagers d'informations permettant de caractériser l'entité qui les a renvoyés comme à un des objets présents dans la scène ou à son contexte. Le système visuel va donc être confronté à un problème complexe, celui de déterminer les informations visuelles relatives à un même objet ou à des objets différents (liage perceptif), ainsi que les informations relatives au contexte (ségrégation figure/fond).

2.1. Définition de l'objet

Comment définir la notion d'objet ? En allant chercher dans un dictionnaire, tel que le Petit Larousse (1993), on peut trouver la définition suivante :

Objet : n.m. (lat. objectum, chose placée devant).

1. Toute chose concrète, perceptible par la vue, le toucher
2. Chose solide considérée comme un tout, fabriquée par l'homme et destinée à un certain usage

Etendant la définition vers des aspects moins matériels

3. Ce sur quoi porte une activité, un sentiment, etc.
4. But d'une action, d'un comportement

Ainsi, l'objet est une entité concrète, visuellement cohérente sur laquelle on va pouvoir effectuer une action. Une chaise n'est chaise que si elle dispose de 4 pieds placés en dessous de l'assise, elle-même orthogonale au dossier de chaise.

Henderson & Hollingworth précisent la définition de l'objet en lui attribuant une dimension spatiale : « objects are smaller-scale discrete entities that are manipulable (e.g. can be moved) within the scene » (Henderson & Hollingworth, 1999). Ainsi, l'objet devrait être de petite taille et manipulable. On peut ici supposer un lien direct entre la taille et le fait qu'on

puisse le bouger. Un avion peut-il alors dans ce cas être considéré comme un objet ? L'avion est par exemple plus gros et lourd qu'un arbre, mais sa propriété mobile le rend plus « facilement » déplaçable qu'un arbre.

Finalement, l'étymologie du mot objet apporte une nouvelle information : l'objet est une chose placée devant. Il ne fait donc pas partie de la structure de la scène ou de l'arrière-plan qu'on peut apparenter au contexte.

2.2. Définition du contexte

On trouve également une définition du contexte dans le dictionnaire qui apparaît encore plus évasive.

Contexte : n.m. (lat. contexere, tisser ensemble).

1. Texte à l'intérieur duquel se situe un élément linguistique et dont il tire sa signification ou sa valeur.
2. Circonstances, situation globale où se situe un événement

Un peu plus loin, on peut trouver la définition d'un autre mot qui aurait peut être été plus à propos.

Contexture : n.f. (de contexte)

Façon dont sont assemblées les différentes parties d'un tout. Structure.

Le contexte consisterait donc en une organisation globale, à priori également cohérente et apporterait un supplément d'information, un cadre de référence, pour les éléments figurant en son sein. Cependant, contrairement aux objets dont les composantes sont disposées selon une configuration bien définie, les scènes semblent moins soumises à des règles spatiales précises. Une scène de plage sera toujours perçue comme une scène de plage, peu importe la position cohérente du sable, de l'eau, du ciel, des palmiers, ou encore des chaises longues.

Une fois de plus, Henderson et Hollingworth apportent des informations supplémentaires dans le domaine des recherches sur la perception visuelle : « Background elements are taken

to be larger-scale, immovable surfaces and structures, such as ground, walls, floors, and mountains » (Henderson & Hollingworth, 1999).

Le contexte, contrairement à l'objet, serait de plus composé de surfaces et structures figées de tailles importantes, tel que le sont les sols, les murs ou encore les montagnes…

Parallèlement à cette définition, un autre point de vue sur le contexte peut être envisagé. Le contexte d'une scène, son environnement, peut être considéré comme la généralisation spatiale et perceptuelle d'une configuration d'objets (Tversky & Hemenway, 1983).

Ces deux définitions du contexte ont l'avantage de mettre d'ores et déjà en avant une problématique importante dans le cadre de la recherche sur les scènes naturelles. Doit-on considérer le contexte comme une entité à part entière ou comme une dérivée de la collection d'objets présents.

2.3. Est-ce un objet ? Est-ce un contexte ?

Dans la définition de l'objet et du contexte, l'accent est donc mis sur les propriétés de mobilité et de taille. L'objet se veut mobile et de taille raisonnable tandis que les éléments du contexte sont immobiles et de tailles importantes.

Pourtant, les propriétés de spatialité dépendent en grande partie du point de vue de l'observateur. Ainsi, dans un souci de précision, je souhaite définir dès lors deux types de contexte sur lesquels je m'appuierai tout au long de ce mémoire : le « contexte relatif » et le « contexte absolu ».

Lorsqu'on porte son regard sur les deux scènes naturelles de la figure 6, on comprend quasi-immédiatement que cette photographie a été prise à l'intérieur d'un bureau selon un plan relativement large (A) ou selon un plan plus rapproché (B). Ainsi, on peut considérer que le bureau est le « contexte absolu ». Quel que soit le point de vue ou l'orientation des photographies, la scène représentera un bureau.

Figure n°6 : Selon notre position dans l'espace, selon les entités présentes dans notre champ visuel et selon la tâche à effectuer, nous ne considérons pas les mêmes objets comme objet d'intérêt (à gauche : la femme, à droite : le clavier par exemple). De plus, en avançant vers le bureau, ce dernier ayant jusque là le statut d'objet est devenu contexte.

Cependant, dans la scène A, notre attention est immédiatement capturée par la jeune femme qui travaille tandis que le clavier, l'écran d'ordinateur ou encore la table de réunion apparaissent de peu d'intérêt. Dans cette scène, la femme est considérée comme un objet biologiquement pertinent car animé (on peut interagir socialement avec elle, elle pourrait même nous agresser… ; New, Cosmides & Tooby, 2007), tandis que les autres objets de la scène sont peu pertinents, ils sont dès lors contextuels et composent le « contexte relatif » à la femme. Si maintenant on se désintéresse de la femme pour s'intéresser aux objet posés sur son bureau (B), le clavier est alors susceptible de devenir le principal objet d'intérêt. Dans ce cas, les pots à crayons situés juste derrière font partie du contexte relatif au clavier. Si au contraire, nous étions venus dans cette salle pour chercher de quoi écrire, les pots à crayon serait l'objet pertinent tandis que le clavier serait contextuel. Ainsi, selon des critères encore relativement méconnus, certains objets attirent ou capturent notre attention plus que d'autres.

La nature de l'objet pertinent et du contexte relatif à cet objet est donc modulée par le but visé, par ce que l'on cherche, par la relevance des objets à un instant donné. Le contexte relatif est de ce fait évolutif contrairement au contexte absolu qui est fixe.

Suite à cet exemple, il semble que les propriétés de mobilité ne soient pas essentielles à la caractérisation du contexte et de l'objet. Quant à la dimension spatiale, il y aura en effet toujours un élément de l'ensemble contextuel de taille bien supérieure à l'objet d'intérêt. Le contexte absolu constitue une structure spatiale de référence pour localiser et évaluer l'objet d'intérêt.

La composante attentionnelle modulant la nature de l'objet du contexte peut être également mise en évidence dans certaines images bi-stable telles que la désormais fameuse « Rubin's vase-faces ». Dans cette image bi-stable (Figure 7), il s'avère bien difficile de déterminer quel est l'élément objet et quel est l'élément contextuel.

Figure n°7 : Illustration d'un Rubin's vase-faces. En fonction d'où l'on porte l'attention, on distingue un vase ou des visages se faisant face.

Au premier regard, il est clair que le vase constitue l'objet d'étude. Mais dès que notre regard se porte sur les zones noires et que notre système visuel y discerne des visages, le vase capturant moins l'attention devient alors contexte. Dans un tel cas, la discrimination objet/contexte reviendrait à effectuer une ségrégation figure/fond.

2.4. Localiser chaque entité dans un espace en 3 dimensions

Chaque information visuelle de la scène peut à mon sens être placée dans un espace en 3 dimensions en fonction de son niveau d'intérêt, de la nature de son analyse et de son étape de traitement.

Un premier axe situerait l'information sur un continuum de pertinence, les informations caractéristiques de l'objet étant généralement des informations d'intérêt traitées en région fovéale ou capturant l'attention tandis que les informations relatives au contexte seraient traitées en région plutôt périphérique et serviraient de cadre de référence.

Un deuxième axe situerait l'information visuelle sur un continuum de résolution. Il permettrait ainsi de définir la nature du traitement perceptif, global ou local, mais également sa finesse.

Un troisième axe situerait l'information visuelle sur un continuum physico-sémantique. En fonction de l'étape de traitement de l'information, son intégration serait plus ou moins aboutie. Les deux extrêmes de l'axe sont respectivement les caractéristiques physiques de l'information et sa représentation sémantique.

2.5. Le contexte est-il fonctionnel ?

Comme on l'a vu, définir le contexte n'est pas une tâche aisée. La définition la plus aboutie qu'on a pu finalement mettre en place réside dans l'idée suivante. Le contexte serait une structure spatiale globale et cohérente de taille importante servant de référence pour l'appréhension des objets d'intérêt en son sein. Il n'est pas fonctionnel comme peuvent l'être les objets, et pourtant on peut se demander si les objets auraient une fonction sans contexte ? A quoi correspondrait un monde sans contexte ? Peut-on imaginer se trouver dans un espace incohérent et voir autour de nous une collection d'objets flottants, aléatoirement disposés, de taille non définie ? Les tableaux de René Magritte et la scène composée par Isabelle Bülthoff peuvent être envisagés comme les premiers pas vers un tel monde (Figure 8)…

Les objets présents dans ces tableaux ont sans nul doute une fonction qui leur est propre. Le peigne sert à coiffer, le verre à boire, la pipe à fumer, peu importe leurs taille et positions respectives.

Qu'en est-il de la fonctionnalité de l'objet ? Face à ces tableaux, nous serions tentés de dire qu'il n'a pas de fonction propre, il sert les objets, les organise, leur donne une référence spatiale, une cohérence, une existence... L'incohérence de la scène « Les valeurs personnelles » provient uniquement d'une échelle spatiale attribuée à chaque objet incongru avec le contexte. Dans « Golconde », l'incohérence est liée à l'inexistence d'un support pour chaque objet. Finalement, dans la scène des chaises d'Isabelle Bülthoff, nous sommes

capables de faire la part des choses entre les chaises réelles et les formes de chaises uniquement grâce au contexte. A noter que comme le disait René Magritte, tout objet peint dans un tableau n'est pas un objet mais uniquement sa représentation.

Figure n°8 : De gauche à droite et de haut en bas : « Les valeurs personnelles » de René Magritte, « Golconde » également de Magritte, une conception graphique de Isabelle Bülthoff (essayez donc de compter le nombre de chaises), et « Ceci n'est pas une pipe » de Magritte.

Dans le cadre de la théorie écologique de Gibson (Gibson, 1979), le contexte peut contenir des « affordances », c'est à dire des propriétés actionnables par l'individu. Par exemple la mer, considérée habituellement comme contexte, contient l' « affordance » « aller nager ». Doit-on alors considérer que l'eau devient objet ?

Les principales caractéristiques du contexte résident dans sa stabilité et sa régularité. Dans la vie de tous les jours, l'environnement est permanent. Il ne change pas à chaque instant, il est stable. Nous pourrions dire qu'il préexiste aux objets. Nous pouvons nous imaginer dans

un couloir dépourvu d'objets, seulement 2 murs, un plafond, un plancher, cela suffit à fournir un repère spatial cohérent. La perception visuelle de notre contexte module d'ailleurs la représentation 3D mise en place par notre système proprioceptif. La preuve en est des expériences effectuées sur des astronautes en mission dans l'espace qui continue à utiliser les repères « plancher »/ « plafond » pour discriminer le haut du bas, même si le plafond se trouve orienté à ce moment là vers la terre. Dans ce cadre de recherche dans l'espace, une étude démontre merveilleusement bien l'influence du contexte sur nos capacités perceptivo-motrices (McIntyre, Zago, Berthoz & Lacquaniti, 2001). Le mouvement de saisie d'une balle tombant du plafond a été évalué chez des astronautes alors qu'ils étaient soumis à une gravité de 0G (dans l'espace) et 1G (sur terre). Tandis qu'à 1G, la balle qui tombe subit une accélération due à la pesanteur, cette même balle tombe à vitesse constante à 0G. Les résultats montrent que les astronautes dans l'espace ont tendance à initier leur mouvement de saisie de balle précocement comme si la balle était soumise à l'accélération. Les auteurs suggèrent que malgré des preuves conscientes évidentes de l'absence de gravité (objets flottants, pression sur la peau…), la présence des murs, du plancher, du plafond, des lumières au-dessus de la tête des astronautes leurs confèrent la sensation d'être toujours dans un référentiel contextuel habituel soumis aux lois de la gravité.

Ces indices contextuels sont d'ailleurs valables uniquement parce que nous avons appris les régularités de notre environnement tout au long de notre vie. Nous avons ainsi pu créer des associations entre le contexte et l'objet comme en témoignent la liste des violations relationnelles décrites par Biederman : Violations de support, d'interposition, de probabilité d'apparition, de position, ou encore de taille (Biederman, Mezzanotte & Rabinowitz, 1982).

Ces régularités nous permettraient d'utiliser des représentations pré-activées de notre environnement permettant de restreindre la liste des objets probables tout en précisant leurs positions possibles dans l'espace. Nous reviendrons sur l'ensemble des interactions objet/contexte dans la partie « Des interactions objet/contexte précoces et bidirectionnelles ».

3. Le système visuel et ses dualités

3.1. Les aires visuelles bas-niveau

Comme nous l'avons vu préalablement, il ne pourrait exister de perception visuelle sans l'existence d'une source de lumière naturelle ou artificielle entraînant la génération et le mouvement d'un ensemble de photons constitutifs des ondes électromagnétiques. Pourtant, ce monde riche en couleurs nous resterait totalement inconnu si notre organisme était incapable d'intégrer ce signal ondulatoire. C'est donc par le biais de photorécepteurs spécifiques que notre cerveau, et plus précisément notre système visuel, va pouvoir transcrire, intégrer, et par la suite donner un sens au monde qui nous entoure.

Cônes et bâtonnets : la porte d'entrée d'un monde visuel

L'ensemble des photons passant la cornée et le cristallin vont traverser l'ensemble des couches de la rétine jusqu'à atteindre la couche la plus profonde contenant les photorécepteurs, cellules ciliées spécialisées. Environ 125 millions de photorécepteurs (120 millions de bâtonnets et 5 millions de cônes) vont alors assurer la première étape de traitement du signal lumineux : la transduction de l'onde électromagnétique en signal électrique.

Les bâtonnets, très sensibles à la variation de luminance même en conditions de faible luminosité, sont situés en périphérie de la rétine. Ils sont connectés à un plus grand nombre de cellules ganglionnaires que les cônes permettant ainsi un accroissement de la sensibilité à la lumière. En contrepartie, la résolution spatiale périphérique se voit diminuée. Les cônes plus petits et plus larges sont surtout sensibles à la longueur d'onde et donc aux variations de couleurs dans des conditions de forte luminosité. Ils sont majoritairement situés dans la partie fovéale de la rétine et assurent ainsi une très bonne acuité visuelle en vision centrale. L'ensemble de ces photorécepteurs projettent vers les neurones bipolaires dans lesquels le signal électrique va subir une première étape de traitement avant d'être envoyé vers les cellules ganglionnaires via les neurones bipolaires.

Cellules ganglionnaires

Une des premières découvertes capitales concernant la vision fut effectuée par Kuffler (Kuffler, 1953) en révélant la propriété des cellules ganglionnaires de la rétine à répondre majoritairement sur des points isolés. Il fut également suggéré que la plupart des cellules ganglionnaires ont soit des champs récepteurs de type centre ON – périphérie OFF, soit l'inverse (Barlow, 1953). Dans l'exemple d'une cellule centre ON – périphérie OFF, le centre du champ récepteur sera excitateur tandis que son pourtour sera inhibiteur. Ces neurones sont donc sensibles aux variations spatiales de contraste. De plus, au niveau de la région fovéale, d'autres cellules ganglionnaires apparaissent sélectives à des contrastes chromatiques spécifiques (Kuffler, 1953). Avant d'atteindre le corps genouillé latéral (LGN), les projections des cellules ganglionnaires vont passer par le chiasma optique : les informations visuelles de l'hémichamp temporal de chaque œil vont ainsi être transmises vers le LGN de l'hémisphère ipsilatéral tandis que les informations visuelles de l'hémichamp nasal de chaque œil seront transmises vers le LGN de l'hémisphère controlatéral. L'ensemble de l'hémichamp visuel gauche est ainsi traité par les aires visuelles de l'hémisphère droit, et inversement.

Si la majorité des cellules ganglionnaires projettent en direction du LGN, puis vers le cortex, il est intéressant de noter qu'une partie d'entre elles projettent vers d'autres structures sous-corticales telles que le colliculus supérieur (Schiller & Malpeli, 1977), le pulvinar, l'hypothalamus, le prétectum, ou encore le noyau thalamique latéral postérieur. Bien que n'ignorant pas le rôle important que ces voies visuelles sous corticales peuvent être amenées à jouer, je m'attacherai plutôt à décrire les voies visuelles corticales qui apparaissent indispensables dans le traitement de l'objet.

Le corps genouillé latéral

Si le LGN a longtemps été et reste encore souvent défini comme un simple relais thalamique, il est à noter que ses neurones sont activés par des contrastes de luminance. Le LGN, plus qu'un relais participe donc à l'intégration du signal visuel. Il pourrait, en fonction de l'état de veille de l'individu, transmettre ou non l'information en provenance des cellules ganglionnaires (Wiesel & Hubel, 1966). Nous verrons de plus dans un chapitre ultérieur que le

LGN reçoit de nombreuses connexions feedback provenant des aires visuelles supérieures suggérant un rôle bien plus conséquent.

Aire visuelle primaire V1

La majorité des connexions en provenance du LGN aboutissent dans l'aire visuelle primaire V1. A noter que parallèlement, des projections du LGN vers le colliculus supérieur et le pulvinar ont été mis en évidence et serviraient à guider l'attention et l'action (Grieve, Acuna & Cudeiro, 2000). Par diverses études menées chez le chat (aire 17) et chez le singe (V1), Hubel et Wiesel furent les premiers à caractériser de manière approfondie les neurones de la première aire visuelle corticale (Hubel & Wiesel, 1959, Hubel & Wiesel, 1962, Hubel & Wiesel, 1968) leurs donnant le nom de « bar-detectors ». Contrairement aux cellules de la rétine ou du LGN, les cellules de V1 répondent de manière sélective à la présentation dans leurs champs récepteurs, de barres orientées selon un angle spécifique. La majorité de ces neurones sont sensibles à la couleur, à l'orientation et aux fréquences spatiales (Bullier & Nowak, 1995, Leventhal, Thompson, Liu, Zhou & Ault, 1995). Ils sont binoculaires (et donc sensibles à la disparité), et leurs réponses pourraient être influencées par la direction du regard (Trotter & Celebrini, 1999). Leurs champs récepteurs de 1° en moyenne sont divisés en sous-régions excitatrices et inhibitrices. On constate au sein de V1 plusieurs types de cellules : les cellules simples et les cellules complexes. Alors que les cellules simples ont des champs récepteurs organisés en subdivision centre/pourtour et effectuent une sommation linéaire du signal, les cellules complexes quant à elles permettent une intégration non-linéaire du signal et sont les premières cellules à démontrer une invariance à la position du stimulus. La réponse des cellules de V1 est de plus influencée par les stimuli présents à l'extérieur du champ récepteur, incluant ainsi des effets de facilitation colinéaire et d'inhibition latérale (Allman, Miezin & McGuinness, 1985, Knierim & Van Essen, 1992, Sillito, Grieve, Jones, Cudeiro & Davis, 1995). De manière générale, des régions voisines d'une image contiennent des orientations colinéaires (Sigman, Cecchi, Gilbert & Magnasco, 2001). Ainsi, si deux champs récepteurs proches s'activent à la présence de traits physiques colinéaires, il y a de fortes chances qu'entre les deux régions de l'image traitée, il y ait également présence de traits colinéaires. Le contour des champs récepteurs en V1 pourrait donc recevoir des informations des connexions latérales allant dans ce sens (Gilbert, Sigman & Crist, 2001). Ce réseau de

connexions verticales et horizontales seraient en fait globalement organisé « en colonnes corticales » (Figure 9). Les neurones seraient regroupés en fonction de leur orientation préférée et en fonction de la position de leur champ récepteur dans le champ visuel. Entre ces colonnes d'orientation se trouveraient des « blobs » transmettant les informations de couleurs.

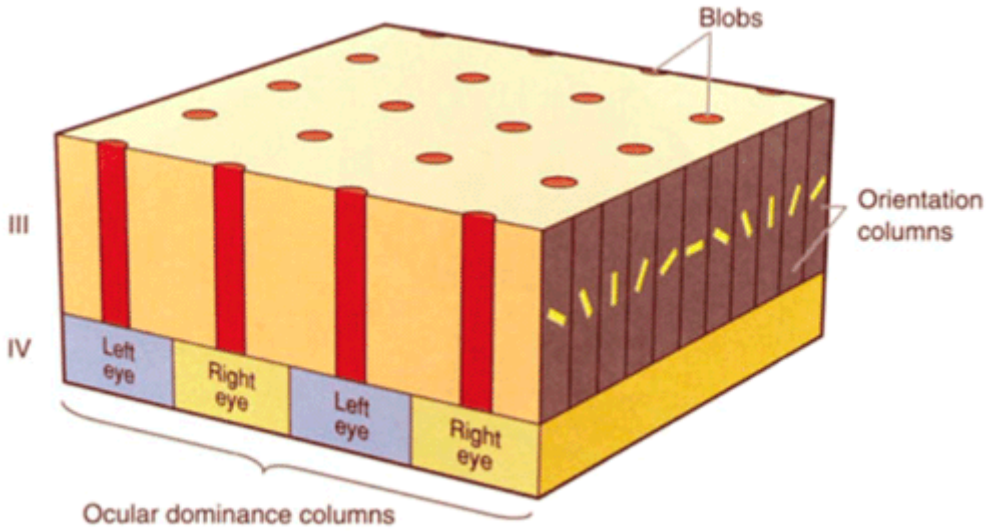

Figure n°9 : Représentation des colonnes d'orientations
Tiré de(Bear, Connors & Paradiso, 2002)

L'information visuelle traitée par les neurones de V1 va alors transiter majoritairement via les projections axonales en direction de l'aire hiérarchique supérieure suivante V2 (Cragg, 1969, Zeki, 1969). On note cependant des projections des cellules M de V1 (IVb) en direction de l'aire MT (Lund, Lund, Hendrickson, Bunt & Fuchs, 1975), de l'aire V3, et du colliculus supérieur (V ; Schiller, Malpeli & Schein, 1979), ainsi que des connexions feedback (VI) projetant sur le LGN (Lund, Henry, MacQueen & Harvey, 1979).

Aire visuelle secondaire V2

Les neurones de V2, première aire visuelle extra-striée, sont organisés en bandes corticales (fines, épaisses, et interbandes) et ont des champs récepteurs plus grands que les cellules de V1. Si la plupart montre une sélectivité à l'orientation similaire aux neurones de V1 (Zeki, 1978), d'autres sont déjà spécifiques à des formes locales comme des coins ou des lignes courbes.

L'aire visuelle V2 représente la dernière aire visuelle avant la séparation des voies ventrales et dorsales communément appelées voies « What » et « Where ». Les neurones de V2 vont ainsi projeter sur l'aire visuelle V4 et les aires V3 et MT, représentant respectivement les premières aires des voies ventrale et dorsale.

3.2. La dualité des voies ventrale et dorsale au sein du système cortical

Suggérée la première fois chez le singe par Van Essen et collaborateurs, puis confirmée par Ungerleider et Mishkin par la suite, la division du système visuel cortical en deux voies distinctes est communément acceptée (Mishkin & Ungerleider, 1982, Mishkin, Ungerleider & Macko, 1983, Van Essen & Zeki, 1978). Elle a d'ailleurs été démontrée comme toute aussi valable chez l'humain (Haxby, Grady, Horwitz, Ungerleider, Mishkin, Carson, Herscovitch, Schapiro & Rapoport, 1991, Ungerleider & Haxby, 1994). Tandis que la voie ventrale (V1-V2-V4-IT) serait la voie préférée des informations de formes et de couleurs dans un but ultime de reconnaissance d'objets (Desimone, Schein, Moran & Ungerleider, 1985), la voie dorsale (V1-V2-V3-MT) porterait des informations de spatialité et de mouvements pour permettre la localisation et le guidage de l'action, dans certains cas en direction des objets (Goodale & Humphrey, 1998, Goodale, Milner, Jakobson & Carey, 1991). Plus récemment, ces deux voies furent représentées selon une perspective différente. Remplaçant la dichotomie perception/action, il fut suggéré que la voie ventrale serait le siège de traitements perceptifs conscients tandis que la voie dorsale permettrait une perception inconsciente vers l'action (Goodale & Milner, 1992, Milner, 1995).

Certaines similarités subsistent cependant entre les deux voies, notamment au niveau physiologique. Dans un but d'intégration efficace de l'information, les traitements visuels des deux voies semblent devenir de plus en plus complexes au fur et à mesure qu'on progresse dans la hiérarchie du système visuel. Cela est d'ailleurs corrélé à des champs récepteurs de plus en plus complexes couvrant une région du champ visuel de plus en plus importante.

3.2.1 La voie ventrale ou voie « What »

L'aire visuelle V4

La voie ventrale débute dès les aires visuelles primaires V1, V2 mais c'est à partir de l'aire visuelle V4 qu'elle se sépare de la voie dorsale. Zeki fut l'un des premiers à la caractériser précisément (Zeki, 1971) et à définir la sélectivité de ses champs récepteurs. Il a ainsi démontré que les cellules de V4 s'avèrent sélectives à l'orientation et présentent une plus grande invariance à l'orientation que les neurones de V1. Elles sont également les premières cellules de la voie ventrale à être sensibles aux couleurs perçues (Bartels & Zeki, 2000, Hadjikhani, Liu, Dale, Cavanagh & Tootell, 1998, Zeki, 1980, Zeki, 1983), différentes des couleurs physiques que traitent V1 et V2 (mais aussi Schein, Marrocco & de Monasterio, 1982). Enfin, Il a été démontré que les champs récepteurs de V4 étaient sensibles à l'occurrence de courbes apparaissant à un endroit donné d'une forme (Pasupathy & Connor, 2001, Pasupathy & Connor, 2002). Une telle sélectivité démontre efficacement que des propriétés de formes sont également intégrées dans cette aire (Merigan & Pham, 1998). Ce traitement des formes de l'objet pourrait être facilité par le fait que les neurones sélectifs aux mêmes propriétés d'orientation et aux mêmes tailles sont regroupés en modules au sein de V4. Une telle organisation se retrouve uniquement dans la portion de V4 représentant le centre du champ visuel (Ghose & Ts'o, 1997).

L'aire visuelle V4 semble donc participer à la reconstruction de formes physiques intermédiaires (Gallant, Connor & Van Essen, 1998 , Kobatake & Tanaka, 1994) sans pour autant permettre la représentation d'un objet naturel abouti. Il semblerait pourtant que les neurones de V4, en plus de présenter une sensibilité au stimuli de complexité intermédiaire, puissent également représenter la réponse comportementale associée aux stimuli (Mirabella, Bertini, Samengo, Kilavik, Frilli, Della Libera & Chelazzi, 2007). Ses connexions avec les aires du cortex inféro-temporal (Desimone, Fleming & Gross, 1980, Rockland & Pandya, 1979) permettront une intégration supérieure du signal visuel par les neurones de IT.

Cortex inféro-temporal chez le singe (IT : TEO et TE) / Cortex occipito-latéral (LOC) chez l'homme

Une approche efficace pour mieux comprendre la fonctionnalité du cortex inféro-temporal passe par la caractérisation chez le singe de ses neurones et plus particulièrement de leur champ récepteur. Il est dans ce sens intéressant de noter que IT est la première région visuelle dont les champs récepteurs des neurones n'ont pas une organisation rétinotopique (Boussaoud, Desimone & Ungerleider, 1991). Leurs champs récepteurs sont larges (en moyenne 25° d'angle visuel), bilatéraux et comprendraient la plupart du temps la région fovéale (Desimone & Gross, 1979, Gross, Bruce, Desimone, Fleming & Gattass, 1981, Gross, Rocha-Miranda & Bender, 1972 , Ito, Tamura, Fujita & Tanaka, 1995). D'autres études ont cependant montré qu'il existait également au sein de IT des cellules dont le champ récepteur couvraient les régions périphériques ipsilatérales, et même controlatérales (Boussaoud et al., 1991) . De fait, certains neurones de IT démontrent une grande invariance à la position et à la taille (Gross, 1994, Logothetis & Pauls, 1995, Logothetis & Sheinberg, 1996, Op De Beeck & Vogels, 2000, Tanaka, 1996). Les champs récepteurs de l'aire antérieure de IT, siège de la représentation d'objets complexes, couvriraient une surface moyenne de 10° d'angle visuel (Op De Beeck & Vogels, 2000), avec des surfaces maximales observées de 70° (Rousselet et al., 2004). Cependant, il apparaît que même dans ces aires haut-niveau, certains champs récepteurs couvrent une petite surface. En effet, des cellules ayant des champs récepteurs de 2.8° ont également été découvertes au sein de IT (Op De Beeck & Vogels, 2000). Après avoir entraîné un singe à discriminer de petits stimuli (0.6° de large), les neurones de IT apparaissaient même sensibles à des changements de position des stimuli de 1.5° (DiCarlo & Maunsell, 2003).

L'aire visuelle IT peut donc en même temps mettre en place des représentations complexes d'objets ou de scènes étendues, et en parallèle, intégrer des représentations d'objets complexes de taille plus modeste sur la base d'informations purement ascendantes plus ou moins détaillées. Ainsi, les neurones de IT semblent rassembler les propriétés caractéristiques pour être impliqués dans la reconnaissance d'objets. Une preuve plus évidente de la capacité des neurones de IT à répondre sélectivement à une catégorie d'objet fut décrite par Gross en 1972. Alors qu'il enregistrait des réponses unitaires de neurones dans IT, il découvrit accidentellement l'existence de neurones de IT sélectifs aux mains (Gross et al.,

1972). Ces neurones déchargeaient quelle que soit la position ou l'angle de vue de la main. De manière anecdotique, une « hand cell » avait été découvert avant qu'on ne découvre les « face cells ». Depuis, ces résultats ont été de nombreuses fois répliqués, et on a ainsi révélé l'existence de neurones sélectifs aux visages (Bruce, Desimone & Gross, 1981 , Perrett, Rolls & Caan, 1982), aux arbres ou à certaines catégories d'arbres (Vogels, 1999a, Vogels, 1999b) mais également aux brosses de WC (Richmond, Wurtz & Sato, 1983)… Une autre preuve de la capacité des neurones à répondre à des catégories d'objets, et ici aux visages, fut apportée par Afraz et ses collaborateurs. Dans un premier temps, ils entraînèrent un singe à regarder à droite lorsqu'on lui présentait un visage et à regarder à gauche lorsqu'aucun visage n'était présent. Quand par la suite, ils présentèrent au singe des stimuli bruités et stimulèrent via des micro-stimulations électriques un groupe de neurones sélectifs au visage, le singe eut majoritairement tendance à regarder vers la droite comme s'il avait perçu un visage (Afraz, Kiani & Esteky, 2006).

A noter que certains neurones déchargent lorsque le visage est vu en entier, mais leurs réponses diminuent si on gomme une partie de l'illustration tel que les yeux ou le nez. D'autres encore ne répondent qu'aux visages perçus selon un angle précis. En fait, en se dirigeant vers la partie antérieure, les neurones deviendraient sélectifs à des objets de plus en plus complexes via une analyse probablement plus globale (Kobatake & Tanaka, 1994 , Tanaka, 1996). Il est de plus fortement envisagé que la représentation de certains objets dans IT soit distribuée sur plusieurs neurones d'une même population (Booth & Rolls, 1998 , Rolls, Treves & Tovee, 1997, Rolls, Treves, Tovee & Panzeri, 1997), les cellules déchargeant pour des traits physiques similaires étant regroupées en colonnes corticales à l'instar des colonnes d'orientations de V1 (Figure 10, Tanaka, 1996). Les représentations des objets dans IT ne seraient pas figées, mais au contraire plastiques et évolueraient avec l'entraînement (Logothetis, Pauls & Poggio, 1995, Logothetis & Sheinberg, 1996) . Dans une tâche de catégorisation, les réponses des neurones de IT pourraient évoluer en fonction de la diagnosticité des traits physiques présentés au cours de l'expérience (Sigala, 2004, Sigala, Gabbiani & Logothetis, 2002, Sigala & Logothetis, 2002). Ces changements de représentation s'accompagneraient d'une restructuration cellulaire au niveau des colonnes corticales spécifiques de l'objet (Tanaka, 1996).

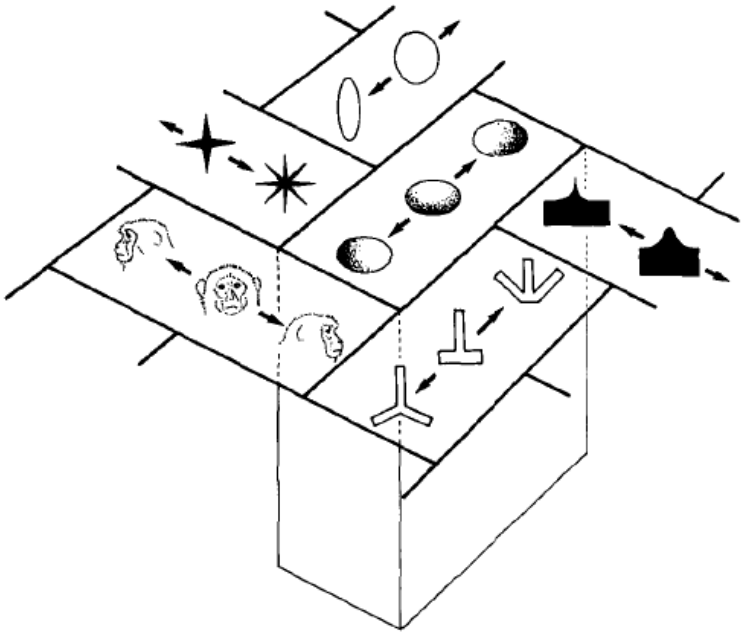

Figure n°10 : Tiré de (Tanaka, 1996). Représentation de colonnes corticales au sein desquelles les cellules déchargent à la présentation d'objets similaires.

Chez l'humain, des études menées sur la région du cortex occipito-latéral (LOC) démontrent des similarités avec l'aire IT du singe. Il semblerait notamment que les champs récepteurs des neurones dans le LOC soient également ipsi et controlatéral (Grill-Spector, Kushnir, Hendler, Edelman, Itzchak & Malach, 1998 , Tootell, Mendola, Hadjikhani, Liu & Dale, 1998). De plus, Kourtzi et al. ont montré que LOC serait cependant plus impliqué dans la reconnaissance de la forme des objets plutôt que de leurs contours (Kourtzi & Kanwisher, 2000). Ils sont par contre rarement sélectifs à des formes trop simples tel qu'un simple point lumineux et invariants au contraste, à la taille, ou à la position dans le champ visuel (Gross, 2008). Si il semble que les neurones du LOC soient plus activés lors de la perception d'objets en 3D (vs. formes 2D, Moore & Engel, 2001), d'autres études suggèrent que cette aire ne serait pas spécialement impliquée dans la perception de la profondeur (Kourtzi & Kanwisher, 2000). De plus, tous les neurones du LOC ne semblent pas sélectifs à une catégorie d'objets complexes. Les neurones de la partie postérieure de LOC coderaient pour des catégories d'objets relativement simples ou même des parties d'objets (Grill-Spector et al., 1998). En se dirigeant vers la partie antérieure, les neurones deviendraient sélectifs à des objets de plus en plus complexes (Bussey & Saksida, 2002, Lerner, Hendler, Ben-Bashat, Harel & Malach, 2001 , Murray & Bussey, 1999 , Tyler, Stamatakis, Bright, Acres, Abdallah, Rodd & Moss, 2004). Selon Tyler et Moss (2001), les concepts d'objets émergeraient ainsi de patterns

d'activations de populations neuronales codant diverses propriétés sémantiques au sein d'un système distribué. D'ailleurs, deux régions, la FFA (« fusiform face area », Kanwisher, McDermott & Chun, 1997) et la PPA (« Parahippocampal Place Area », Aguirre, Zarahn & D'Esposito, 1998, Epstein, Harris, Stanley & Kanwisher, 1999, Epstein & Kanwisher, 1998) situées dans des régions très proches du LOC apparaissent respectivement comme spécifiques de deux catégories de stimuli : les visages et les lieux.

De nombreuses expériences (pour revue : Kanwisher, 2001) ayant eu pour sujet d'étude principal la FFA ont démontré une activation de la FFA lors de la présentation de stimuli divers contenant des visages : photographies de visages, dessins fait-main de visages, têtes de chats, visages de dessins animés ou encore représentations de visages retournées. S'il semble que la FFA soit impliquée dans la détection et la reconnaissance des visages (Kanwisher, 2000), il semble également que la FFA soit également légèrement impliquée dans la catégorisation d'autres objets sujets à l'expertise (Grill-Spector, Knouf & Kanwisher, 2004).

Dans le cadre des travaux présentés au cours de ce mémoire, la PPA (Aguirre et al., 1998, Epstein et al., 1999, Epstein & Kanwisher, 1998) présente un intérêt encore plus conséquent. En effet, cette région corticale s'activerait sélectivement lors de la présentation de stimuli représentatifs de scènes tels que des bâtiments, des maisons, ou encore des constructions LEGO et serait impliquée dans le traitement de la configuration spatiale des scènes (Epstein, Graham & Downing, 2003, Epstein & Kanwisher, 1998). Son activité globale modulable par une composante attentionnelle ne semble pas influencée par la nature de la tâche à effectuer (O'Craven, Downing & Kanwisher, 1999). De façon intéressante, une étude neurophysiologique révéla l'existence d'un patient avec lésion de la PPA dont la perception des lieux était préservée. Les auteurs de cette étude suggèrent à ce propos que la PPA serait plus impliquée dans l'encodage mnésique des scènes que dans leur perception. Enfin, une récente étude de Moshe Bar, ayant montré des activations de la PPA lors de la présentation de visages familiers impliquant de fortes associations contextuelles, défend l'idée que la PPA ne traiterait pas uniquement les informations spatiales de lieu mais les associations contextuelles en général (Bar, Aminoff & Ishai, 2008). D'autres aires seraient néanmoins impliquées de près ou de loin dans la reconnaissance des scènes naturelles indépendamment de la reconnaissance d'objets, notamment au niveau du cortex pariétal, du gyrus lingual.

Dans tous les cas, ces abondantes recherches montrent que les neurones de IT codent une information visuelle qui transcende le traitement et la caractérisation physique des stimuli perçus. Cependant, plusieurs études menées sur l'aire IT n'ont pu encore démontrer une

réponse des neurones de IT spécifique de la réponse comportementale associée à la catégorie des stimuli (Baker, Behrmann & Olson, 2002, Op de Beeck, Wagemans & Vogels, 2001b).

Aires supérieures à IT et LOC

IT est l'une des ultimes aires perceptives de la voie ventrale à avoir été sujet de nombreuses expériences dans le domaine de la reconnaissance visuelle. Pourtant, quelques études montrent que IT n'est pas l'étape finale de la reconnaissance des objets. Des aires supérieures, davantage associatives que visuelles, ont ainsi été mises en évidence. Par exemple, dans le lobe temporal médian (MLT), Quiroga et al. ont révélé l'existence de cellules dont les réponses à l'objet (célébrités, monuments célèbres, animaux) mettent en jeu de multiples représentations (Quiroga, Reddy, Kreiman, Koch & Fried, 2005). En effet, de façon remarquable, des neurones de MLT déchargent sélectivement à des présentations de l'actrice « Halle Berry » que cette dernière soit une photographie ou un croquis de son visage, une capture de film d'Halle Berry déguisée en catwoman (Batman), ou même à son nom écrit !! L'explication de ce remarquable résultat reposerait dans la fonction même de MLT. Il semblerait en effet que MLT soit largement impliquée dans la consolidation mnésique des multiples informations liées à un même objet (Murray & Bussey, 2001). L'aire MLT située dans le cortex périrhinal serait donc une région simultanément perceptive, associative et mnésique.

Les cellules de IT projeteraient parallèlement vers l'amygdale potentiellement impliquée dans la « lecture » émotionnelle des stimuli, ainsi que vers le cortex préfrontal duquel émergerait le choix d'une réponse comportementale appropriée. Dans ce sens, des neurones du cortex préfrontal ont été enregistrés chez le singe lors d'une tâche de catégorisation chien/chat et déchargeaient sélectivement aux images appartenant à une des catégories (Freedman, Riesenhuber, Poggio & Miller, 2001). Les interactions entre cortex préfrontal et IT pourraient sous-tendre le stockage et le rappel mnésique d'informations visuelles et associatives (Tomita, Ohbayashi, Nakahara, Hasegawa & Miyashita, 1999). Elles pourraient également permettre de structurer la sélectivité des cellules de IT en fonction des caractéristiques pertinentes permettant de réaliser la tâche (Sigala et Logothetis, 2002).

3.2.2 La voie dorsale ou voie « Where »

L'aire visuelle V3

Si les cellules de l'aire V2 projettent vers l'aire V4 de la voie ventrale, elles projettent dans le même temps vers l'aire V3 appartenant à la voie dorsale. Gegenfurtner et ses collaborateurs se sont cependant attachés à décrire fonctionnellement les propriétés des cellules de V3. Grâce à des enregistrements unitaires de neurones chez le macaque anesthésié, ils ont pu caractériser une sélectivité des neurones de V3 à l'orientation assez similaire à celles des neurones de V2. Ils ont également montré que ces neurones répondent préférentiellement aux fréquences spatiales basses et hautes fréquences temporelles. Plus sensibles au contraste que les neurones de V2, ils montrent également une sensibilité à la direction et à la couleur des stimuli. Les auteurs suggèrent que V3 pourrait être un aire d'interaction entre les traitements de la couleur et ceux du mouvement (Gegenfurtner, Kiper & Levitt, 1997, Felleman & Van Essen, 1987), d'autant plus que des projections vers les aires V4 et MT ont été révélées.

MT ou V5

Les neurones de MT, peu sélectifs à l'orientation et la couleur, sont par contre sensibles à la direction d'un mouvement, à sa vitesse et à sa disparité (Albright, Desimone & Gross, 1984 , Dubner & Zeki, 1971 , Maunsell & Van Essen, 1983b , Van Essen, Maunsell & Bixby, 1981, Watson, Myers, Frackowiak, Hajnal, Woods, Mazziotta, Shipp & Zeki, 1993 , Zeki, 1974). Malgré une taille conséquente des champs récepteurs, certaines cellules traiteraient une information locale de mouvement (Majaj, Carandini & Movshon, 2007). De plus, certains neurones de MT coderaient le mouvement perçu plutôt que la direction physique du stimulus (Britten, Newsome, Shadlen, Celebrini & Movshon, 1996). La majeure partie de ses cellules projettent vers MST et le cortex pariétal, très peu vers le cortex temporal (Maunsell & van Essen, 1983a).

MST et Cortex pariétal

Les neurones de MST et du cortex pariétal ont des champs récepteurs plus larges que MT, répartis sur l'ensemble du champ visuel, certains n'incluant pas la fovéa (Motter & Mountcastle, 1981, Robinson, Goldberg & Stanton, 1978). Ils démontrent une plus faible sélectivité à la couleur que les neurones de IT (Robinson et al., 1978) et sont comme MT très sensibles aux propriétés de mouvement (Motter & Mountcastle, 1981), même tactile (Beauchamp, Yasar, Kishan & Ro, 2007). MST et le cortex pariétal, permettant une intégration du mouvement plus complexe que MT, seraient les étapes ultimes d'une perception du mouvement orientée vers l'action. En effet, les neurones de ces aires seraient sensibles aux flux optiques (Merchant, Battaglia-Mayer & Georgopoulos, 2001), composants reconnus pour aider à la perception du mouvement de notre propre corps (Duffy, 1998). De plus, elles seraient impliquées dans le contrôle de l'attention préalablement à des actions motrices (Quraishi, Heider & Siegel, 2007), dans le contrôle des saccades oculaires (Andersen, Bracewell, Barash, Gnadt & Fogassi, 1990, Bremmer, Distler & Hoffmann, 1997), et dans la planification de l'action (MacKay & Riehle, 1991, MacKay & Riehle, 1992).

Cette voie visuelle dorsale semble donc avoir deux rôles majeurs. Elle permet l'analyse spatiale de l'environnement qui s'offre à nous, et permettrait de ce fait d'aider à localiser les objets traités par la voie ventrale. Elle préparerait le mouvement du corps lors d'action à accomplir telle que la saisie d'un objet. Elle représente donc une réelle interface perception-action. De plus, les informations spatiales intégrées tout au long de la voie dorsale sont rapidement transmises au cortex entorhinal, une aire perceptivo-mnésique au sein duquel on peut trouver des neurones sélectifs à des objets complexes (Suzuki, Miller & Desimone, 1997). Cette sélectivité pourrait s'expliquer par la présence de connexions entre cortex périrhinal et entorhinal. Dans ce sens, les informations spatiales de la voie dorsale plus rapidement intégrées que les informations de formes de la voie ventrale pourraient influencer la reconnaissance des objets.

3.3. La dualité des systèmes parvocellulaire et magnocellulaire

En supplément de la dualité des voies ventrale et dorsale existe une autre dualité entre systèmes parvocellulaire et magnocellulaire. En effet, de nombreuses études ont démontré que les cellules ganglionnaires pouvaient être divisées en 3 classes distinctes, que ce soit chez le chat (Enroth-Cugell & Robson, 1966) ou chez les primates (Sherman, Wilson, Kaas & Webb, 1976) :

Les petites cellules ganglionnaires de type P (pour parvus, petit en latin, X-like chez le chat) représentent environ 90% de la population totale de cellules ganglionnaires et sont situées majoritairement en région fovéale (Dacey & Petersen, 1992). Elles présentent des réponses de type tonique, soutenues aussi longtemps que le stimulus est présent dans leur champ récepteur. Elles sont de plus très sensibles aux propriétés chromatiques, peu sensibles au contraste, et leurs petits champs récepteurs permettent un traitement de l'information plus détaillée caractérisée par de hautes fréquences spatiales.

Les cellules de type M (pour magnus, grand en latin, Y-like chez le chat ; Shapley & Perry, 1986) distribuées uniformément sur la rétine ont quant à elles de plus grands champs récepteurs et constituent environ 5% de la population des cellules ganglionnaires. Achromatiques, très sensibles au contraste et aux basses fréquences spatiales, elles répondent de manière phasique au changement de luminosité ce qui en fait de parfaits détecteurs de mouvement. En comparaison des cellules P, le diamètre des axones des cellules M est plus important, la propagation des potentiels d'action vers les aires ultérieures est donc plus rapide.

Les 5% (environ 150000 neurones !) de cellules ganglionnaires restantes sont des cellules dites koniocellulaires autour desquelles de nombreuses découvertes restent à faire. On sait cependant que leurs axones présentent une faible rapidité de propagation du signal, et qu'une partie d'entre elles projettent vers le colliculus supérieur (Schiller & Malpeli, 1977). Nous avons déjà vu que les cellules ganglionnaires projetaient en grande partie vers le LGN.

Organisé en 6 couches distinctes, le LGN reçoit les afférences des cellules magnocellulaires au niveau des 2 couches inférieures et les afférences parvocellulaires au niveau des 4 couches supérieures (Malpeli & Baker, 1975, Wilson, Rowe & Stone, 1976). Les afférences des cellules koniocellulaires aboutissent quant à elles dans les zones interlaminaires. (Irvin, Norton, Sesma & Casagrande, 1986). Par la suite, la majorité des connexions magnocellulaires en provenance du LGN aboutissent dans l'aire visuelle primaire

V1 au niveau de la couche 4Cα. Les connexions parvocellulaires quant à elles projettent sur les couches 4Cβ et 4A.

Une description plus détaillée des connexions au sein des couches de V1 est présentée sur la figure 11.

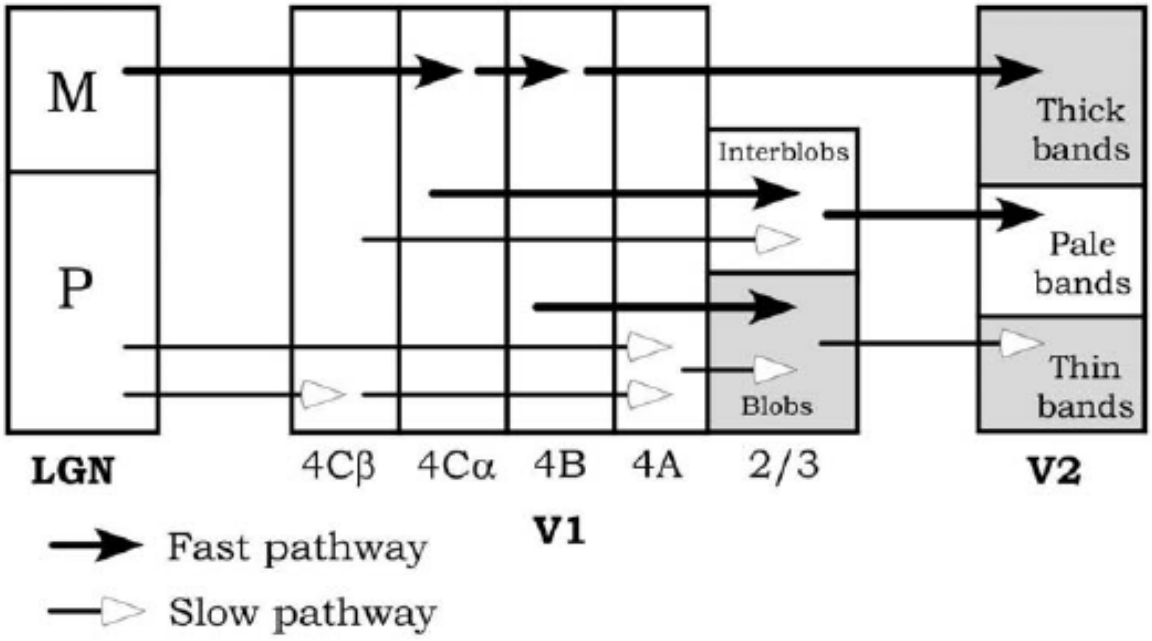

Figure n°11 : Tiré de (Munk, Nowak, Girard, Chounlamountri & Bullier, 1995). Illustration des connexionscorticales pour les systèmes magnocellulaire et parvocellulaires au sein de l'air V1.

Il est important de noter que dès V1, les cellules M possédant des axones de plus gros diamètre, déchargent en moyenne 10 ms plus tôt que les cellules P (Maunsell, 1992). Cet avantage des cellules M augmente en moyenne de 10 ms au niveau de l'aire V2. Or, s'il a toujours été reconnu que la voie dorsale impliquait très majoritairement le système magnocellulaire, on ne sait que depuis récemment que la voie ventrale repose sur la participation du système parvocellulaire, mais également magnocellulaire. (Allison, Melzer, Ding, Bonds & Casagrande, 2000, Ferrera, Nealey & Maunsell, 1992, Nealey & Maunsell, 1994). D'un point de vue physiologique, il semble donc évident (1) que les informations visuelles circulant dans la voie dorsale sont plus rapidement intégrées que l'ensemble des informations circulant dans la voie ventrale, (2) que les informations magnocellulaires sont plus rapidement intégrées que les informations parvocellulaires au sein même de la voie ventrale. De fait, une représentation spatiale et grossière des objets pourraient être construite assez rapidement sur la base des informations magnocellulaires dorsales et ventrales,

représentation qui serait progressivement affinée par l'intégration des informations parvocellulaires de la voie ventrale (Bullier, 2001a, Bullier, 2001b).

3.4. La dualité des traitements visuels ascendants et descendants

3.4.1 Traitements ascendants

Un autre aspect important du système visuel émerge lorsqu'on s'intéresse de plus près aux contraintes temporelles imposées au système visuel d'un point de vue comportemental et physiologique. D'un point de vue comportemental, il est désormais prouvé que nous sommes capables de dénommer un objet familier au sein d'une image en 600 ms (Intraub, 1979). Nous sommes en outre capables d'extraire le sens d'une scène présentée au sein d'une séquence d'images successives affichées à une fréquence de 113 ms par image (Potter, 1975). Enfin, nous sommes encore capables de dénommer ou de catégoriser une scène non masquée flashée 26, voir 10 ms dans les cas extrêmes (Delorme et al., 2000, Fabre-Thorpe, Delorme, Marlot & Thorpe, 2001, Fabre-Thorpe, Richard & Thorpe, 1998, Joubert, Rousselet, Fize & Fabre-Thorpe, 2007, Rousselet, Joubert & Fabre-Thorpe, 2005, Rousselet, Mace & Fabre-Thorpe, 2003, Thorpe, Fize & Marlot, 1996, VanRullen & Thorpe, 2001a) et (Intraub, 1979). D'un point de vue électrophysiologique, des études menées chez le singe éveillé montrent que des populations neuronales au sein de l'hypothalamus latéral et sélectives à des aliments ou à des stimuli aversifs déchargent à des latences de 150-200 ms après l'apparition du stimuli (Rolls, Sanghera & Roper-Hall, 1979). D'autres neurones sélectifs aux aliments et déchargeant à des latences de 100 ms ont également été enregistrés au niveau du cortex orbito-frontal (Thorpe, Rolls & Maddison, 1983). Chez l'homme, des études EEG ont révélé un signal spécifique du statut des photographies à catégoriser apparaissant en région frontale dès 150 ms alors que les stimuli n'étaient flashés que 20 ms (Thorpe et al., 1996). Enfin, des neurones sélectifs aux visages dans le lobe temporal du singe ont été enregistrés déchargeant à des latences de 80-100 ms (Oram & Perrett, 1992). Ces impressionnantes données temporelles sont encore plus intéressantes quand on fait le rapprochement avec des données physiologiques. Comme le mettent en évidence Thorpe et Imbert, on peut estimer à au moins 10 le nombre de synapses présentes entre les cellules rétiniennes et les neurones sélectifs aux visages du cortex inféro-

temporal (Figure 12, Thorpe & Imbert, 1989) Si on considère que ces 10 étapes de traitement sont nécessaires à la construction d'une représentation grossière d'une scène ou d'un objet et qu'aucune autre voie de « raccourci » ne peut être envisagée, cela suggère que chaque étape de traitement prend en moyenne 10 ms.

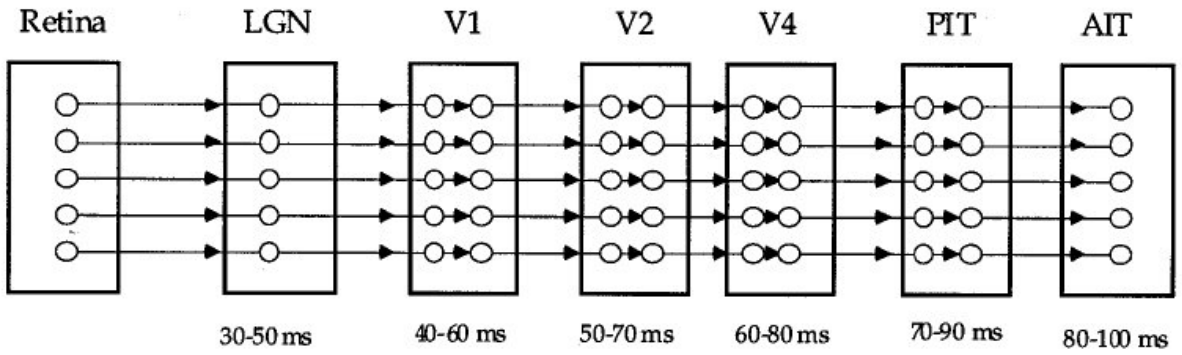

Figure n°12 : Tirée de (Thorpe & Gautrais, 1997). La voie visuelle purement ascendante jusqu'au cortex inférotemporal antérieur seraient composée d'une dizaine d'étapes de traitements. Les latences approximatives des neurones au niveau de différentes aires visuelles sont indiquées au-dessous de chaque aire.

Cette démonstration a deux implications majeures ! La première implication est que de telles contraintes temporelles de traitement ne laissent que peu de place à l'intervention de boucles de rétro-action ou de connexions feed-back. Ainsi la première représentation grossière d'un objet ou d'une scène, suffisant à une catégorisation au moins superordonnée, reposerait sur des traitements visuels purement ascendants !

A noter que d'autres modèles de reconnaissance d'objets prennent en comptent la nécessité de traitements purement ascendants de l'information à travers la hiérarchie des aires visuelles (Fukushima & Miyake, 1982, Poggio & Edelman, 1990, Riesenhuber & Poggio, 2000, Wallis & Rolls, 1997). A ma connaissance, un seul autre modèle purement ascendant est cependant compatible avec les contraintes temporelles imposées par les observations biologiques décrites dans ce paragraphe. Une description schématique de ce modèle proposé par Serre, Oliva et Poggio (Serre, Oliva & Poggio, 2007) est illustrée dans la figure 13.

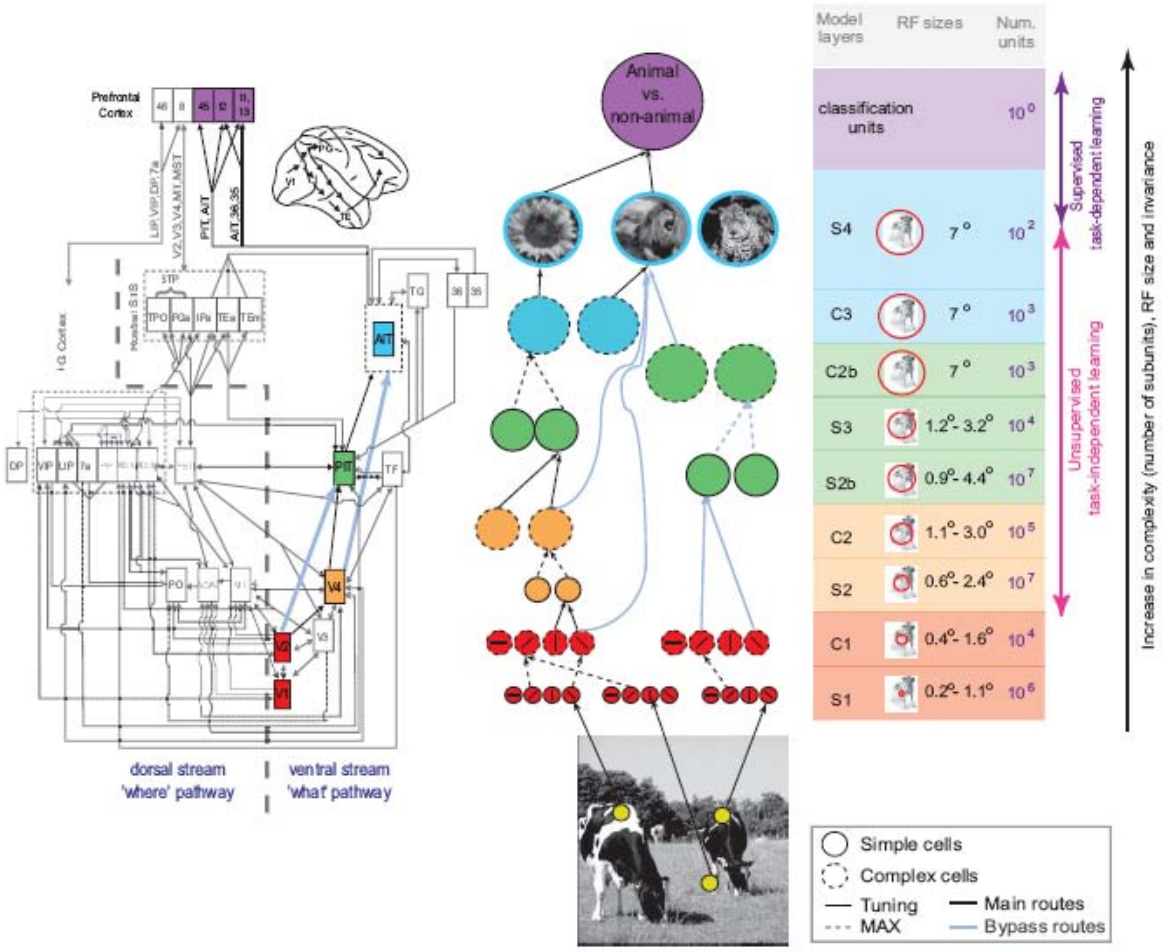

Figure n°13 : Tiré de (Serre et al., 2007). Illustration au centre du modèle de reconnaissance d'objet proposé par Serre, Oliva & Poggio reposant sur un flux d'informations purement ascendant et compatible avec les contraintes temporelles imposées par les observations expérimentales. A gauche sont précisées les aires anatomiques analogues aux couches du modèle computationnel. Dans le tableau de droite sont indiquées les surfaces des champs récepteurs et le nombre de neurones au sein de chaque aire.

3.3.2 Traitements descendants

Jusque là a été présentée une information visuelle ascendante transitant de manière unidirectionnelle vers les aires supérieures des voies visuelles et codant pour des propriétés du stimulus toujours plus complexes. Mais des études neuro-anatomiques ont démontré l'existence de connexions feedback massives et donc descendantes pouvant influencer les réponses des neurones à toutes les étapes de traitements des voies visuelles. Il convient cependant de distinguer deux types d'informations descendantes.

Certains traitements descendants pourraient en effet sous-tendre la pré-activation de populations neuronales spécifiques codant des informations diagnostiques de la tâche à

effectuer. D'autres traitements descendants seraient en complément partie intégrante de boucles itératives de rétro-action tout au long des voies visuelles, aires visuelles bas-niveau incluses (Bullier, 2001a, Bullier, 2001b, Ferster, 2000 #60, Lamme & Roelfsema, 2000), Figure 14), et permettraient d'affiner progressivement l'intégration ascendante des percepts visuels. Ces connexions feedback pourraient être porteuses d'informations spécifiques permettant de préciser la nature de l'objet observé ou encore de discriminer l'objet de son contexte au sein de la voie ventrale.

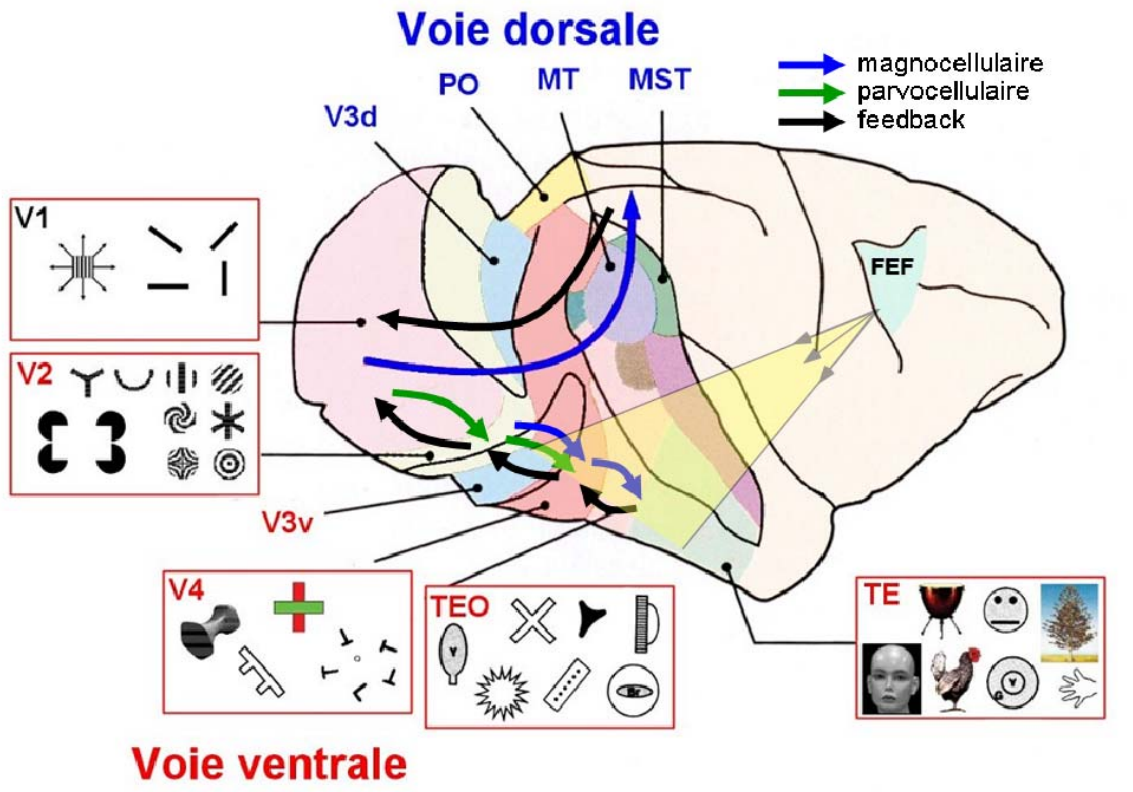

Figure n°14 : Reproduit d'après (Bullier, 2001b), modifié par Nadège Bacon-Macé, et inspiré d'une production de Marc Macé. Les informations magnocellulaires (en bleu) seraient les premières à activer les aires visuelles au sein de la voie ventrale et dorsale. Sur la base d'une information grossière ainsi intégrée, des flux descendants d'informations (en noir) pourraient alors être générer afin de préciser les informations contenues dans les flux parvocellulaires plus tardifs (en vert).

Lamme fut de ceux suggérant une modulation des réponses des neurones de V1 par des aires hiérarchiquement supérieures (Lamme, 1995). Alors qu'un singe effectuait une tâche de reconnaissance en saccade oculaire nécessitant une segmentation figure/fond, il a enregistré la réponse des neurones de V1. Il a démontré que non seulement les neurones de V1

déchargeaient lorsque la figure cible était présente dans le champ récepteur des neurones, mais a en plus observé une augmentation de l'activité de ces mêmes neurones 30 à 40 ms après l'apparition des premières réponses. Il défend ainsi l'idée d'une modulation des réponses des neurones de V1 suivant l'évaluation complète du contexte. Cette modulation pourrait également impliquer les connexions horizontales et les aires extrastriées (Bringuier, Chavane, Glaeser & Fregnac, 1999, Lamme, Super & Spekreijse, 1998). Cependant, d'autres études s'inspirant du même protocole n'ont pas réussi à répliquer ces résultats (Rossi, Desimone & Ungerleider, 2001).

Une autre modulation des réponses de V1 via des connexions feedback en provenance de MT a été également mise en évidence dans une tâche de discrimination de mouvement d'une barre entourée d'un contexte au contraste variable. Utilisant une méthode d'inactivation réversible de l'aire MT, Hupé et ses collaborateurs ont révélé que les connexions feedback facilitaient les réponses des champs récepteurs à la barre-cible, renforçaient l'inhibition du contexte, et que ces effets étaient d'autant plus importants que l'objet cible était peu saillant (Hupe, James, Payne, Lomber, Girard & Bullier, 1998).

Un autre exemple de rôle supposé des connexions feedback réside dans le lien qu'elles établissent entre le cortex visuel et le LGN. Sachant que 80% des entrées du noyau thalamique sont des connexions en provenance d'aires visuelles supérieures, il serait peu écologique de penser que ces dernières sont inutiles. Quelques auteurs ont ainsi prêté à ses connexions un rôle dans le développement de l'attention spatiale par renforcement de l'activité des champs récepteurs dans les régions d'intérêts (Crick, 1984) ou par inhibition de l'activité des champs récepteurs voisins (Vanduffel, Tootell & Orban, 2000). D'autres études d'enregistrements unitaires ou d'imagerie sont venues confirmer l'existence d'une influence attentionnelle sur l'activité des aires visuelles bas niveau tel que V1 (Gandhi, Heeger & Boynton, 1999, Motter & Mountcastle, 1981, Roelfsema, Lamme & Spekreijse, 1998 , Sengpiel & Hubener, 1999) et sur les étapes de traitement plus précoces (Hillyard, Vogel & Luck, 1998, Treue & Maunsell, 1996). Il est finalement intéressant de noter que cette modulation attentionnelle intervient non pas sur les premières activations des aires visuelles bas-niveau mais sur leurs recrutements plus tardifs (Martinez, Anllo-Vento, Sereno, Frank, Buxton, Dubowitz, Wong, Hinrichs, Heinze & Hillyard, 1999, Nobre, Allison & McCarthy, 1998).

Ainsi de manière globale, les connexions feedback permettraient de moduler la nature et la quantité d'information envoyée vers les aires de plus haut-niveau permettant ainsi une

meilleure intégration et résolution des objets d'intérêts. Ces mécanismes permettraient de fait, de résoudre précocement les interactions compétitives entre deux objets proches susceptibles d'appartenir aux mêmes champs récepteurs haut-niveau (Deco & Rolls, 2004, Reynolds, Chelazzi & Desimone, 1999).

Cet ensemble de résultats fut intégré dans un modèle de perception visuelle appelé « Reverse Hierarchy Theory » mettant en avant le rôle des informations visuelles descendantes (Hochstein & Ahissar, 2002). Pour ses auteurs, l'appréhension d'une nouvelle scène ou d'un nouveau stimulus se ferait en deux temps.

Dans un premier temps, une chaîne de traitement purement ascendante de traitements automatiques et implicites construirait une représentation globale de la scène ou du stimulus sur la base de traitements physiques globaux. On peut dire dans ce sens que le modèle proposé par Thorpe et ses collaborateurs correspond parfaitement à cette première vague ascendante d'informations (Thorpe & Gautrais, 1997). Cette représentation grossière intégrée par les aires visuelles haut-niveau serait suffisante pour réussir des tâches ne nécessitant pas d'informations physiques locales détaillées.

Dans un deuxième temps, une perception visuelle consciente serait rendue possible par des traitements visuels redescendant aussi bas que nécessaire. Le système visuel aboutirait ainsi à une représentation détaillée d'une partie du champ visuel.

Cette dichotomie entre traitements visuels ascendants et descendants expliquerait certains phénomènes psychophysiques.

- En recherche visuelle : la représentation globale permettrait le traitement parallèle de traits physiques simples qui « pop-out » mais ne permettrait pas de repérer une conjonction de propriétés physiques. Une seconde stratégie de recherche visuelle sérielle basée sur les traitements visuels descendants affinerait alors successivement l'information locale dans différents champs récepteurs jusqu'à détection de la cible. Ces deux étapes de traitement expliquerait des temps de réactions stables dans les conditions pop-out, et des temps de réaction augmentant avec le nombre de distracteurs dans les conditions de recherche sérielle (Treisman & Gelade, 1980).

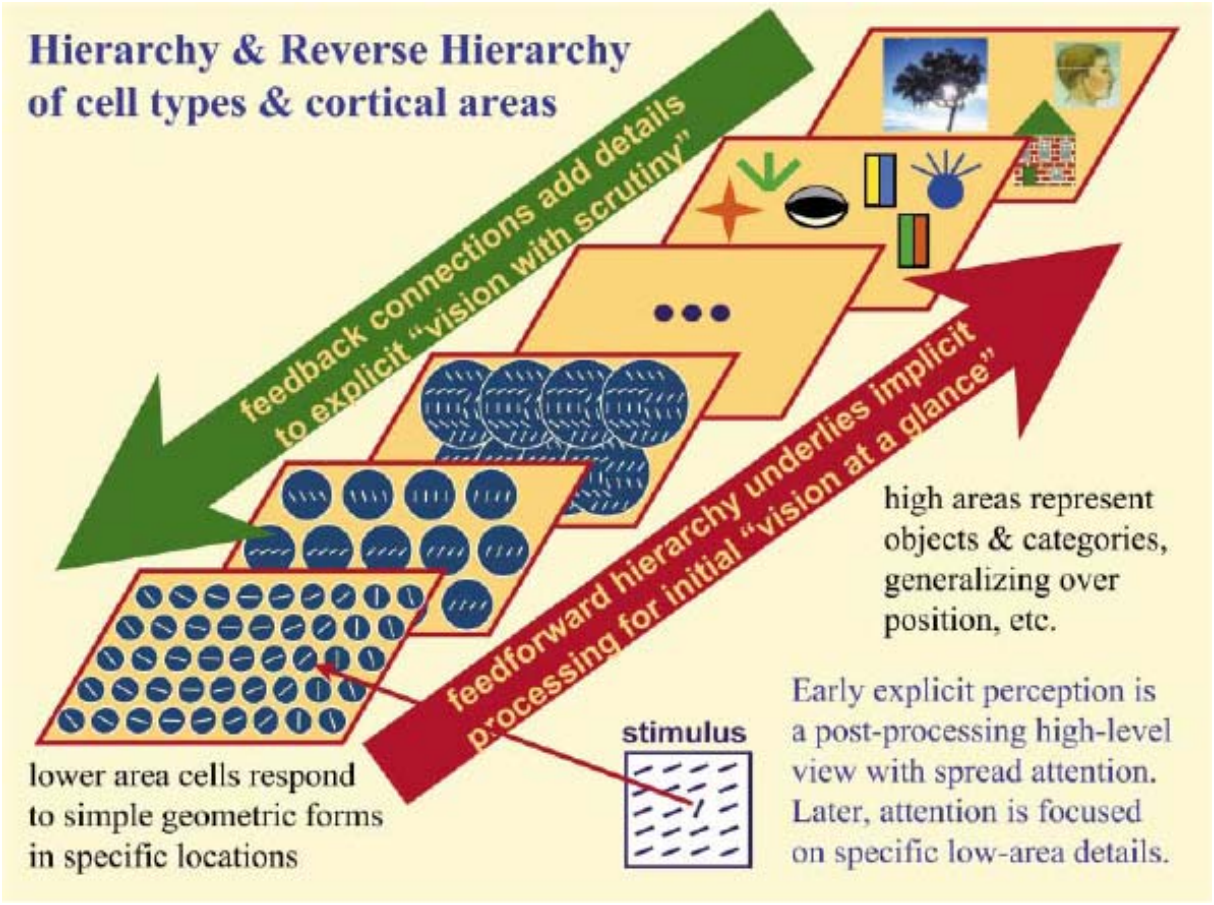

Figure n°15 : Tiré de (Hochstein & Ahissar, 2002). Au sein de la voie ventrale, le flux d'informations purement ascendant pourrait suffire pour établir en IT une représentation globale et grossière d'un objet ou d'une scène et ainsi catégoriser un stimulus à un niveau superordonné. Par la suite, un supplément d'information serait apportée par la mise en jeu de connexions feedback permettant un traitement plus affiné du stimulus.

Dans le domaine de la catégorisation des scènes naturelles, d'autres phénomènes pourraient être pris en compte par la « Reverse Hierarchy Theory » :

- Niveaux de Catégorisation : Selon les auteurs, la représentation de haut-niveau suffirait à catégoriser des scènes à un niveau basique. Une catégorisation plus fine serait obtenue à la fin des flux descendants après avoir réuni une plus grande quantité d'informations locales diagnostiques. Les études présentées dans les articles 2 et 3 de ce mémoire corroborent l'idée d'une catégorisation grossière effectuée plus rapidement qu'une catégorisation affinée. De manière intéressante, Wolfe et ses collaborateurs ont suggéré que les propriétés physiques qui pop-out en recherche visuelle seraient mieux décrites en termes catégoriels qu'en termes de dimensions quantifiables (Wolfe, Friedman-Hill, Stewart & O'Connell, 1992).

- Scène vs. objet : En se basant sur le paradigme RSVP, Potter en 1976 montrait que lorsqu'un ensemble d'images était flashé à une fréquence de 10 images par seconde, les sujets étaient encore capables d'effectuer une tâche de catégorisation sur ces images mais ne détectaient pas la répétition d'un même objet sur plusieurs images de la séquence (Potter, 1976). De plus, l'être humain peut négliger un changement progressif dans une scène tant que le *gist* de la scène reste le même (change blindness, Rensink, O'Regan & Clark, 1997). Il semble également que nous soyons plus efficaces a discriminer la taille moyenne d'un ensemble de 4, 8, 10, ou 15 ronds présenté 500 ms que la taille individuelle de chaque rond (Ariely, 2001). Le cerveau est donc capable d'extraire les traits globaux d'une scène tout en restant insensible, inattentif, aux objets et dimensions locales qui la composent. Nous verrons que l'article n°3 de ce mémoire présente des résultats divergents, l'objet pouvant dans certains cas interférer sur la reconnaissance du contexte (Part. 3, Chap.3).

La « Reverse Hierarchy Theory » suggère donc que la représentation globale de la scène serait disponible suite aux traitements visuels ascendants, tandis que des informations plus détaillées seraient uniquement disponibles après un certain nombre de traitements visuels descendants requérant potentiellement une composante attentionnelle. A noter que cette théorie n'est pas la première à prendre en compte les voies visuelles descendantes. D'autres modèles par le passé ont été suggérés dans la même optique de recherche tel que le modèle proposé par (Ullman, 1995) décrivant la participation de deux voies parallèles opposées ascendantes et descendantes qui évalueraient à chaque nœud du système visuel si représentation physique et mentales coïncident.

Finalement, l'influence des flux descendants dépendant de la tâche à effectuer et résultant en une pré-activation des populations neuronales spécifiques n'est pas prise en compte dans ce modèle. Il serait intéressant de comprendre comment ces derniers s'intègrent au modèle.

4. Des traitements visuels spécifiques de la tâche : la catégorisation

4.1. La catégorisation facilite notre appréhension du monde

A chaque instant de notre vie, nous catégorisons des objets. Face à la très grande diversité du monde environnant, notre cerveau n'est pas capable d'encoder en mémoire une description détaillée de chaque exemplaire des objets que nous rencontrons. Il existe par exemple une quantité infinie de lampes, de tailles, de couleurs et de formes diverses, ayant chacune un fonctionnement spécifique. Pourtant, nous sommes à chaque fois capable d'associer le concept de lampe et la fonctionnalité « éclairage » à l'ensemble des percepts de lampes existants. Cela est rendu possible par un apprentissage en partie inconscient des traits diagnostiques de chaque famille d'objets. Ayant au préalable encodé une représentation d'une lampe A, notre perception d'une lampe B différente de A entraînera tout de même l'activation de la représentation « lampe » par des mécanismes reflétant une association comparative, c'est à dire une estimation des distances perceptuelles, ressemblances et différences, entre l'objet d'intérêt et l'ensemble des catégories déjà encodées. « Categorization is the fundamental process whereby highly variable perceptual imputs are progressively reduced to a smaller number of equivalence classes (called 'categories') whose memory representations (called 'concept') mediate thinking and adaptive action» (Schyns, 1997). Ce processus s'appuyant sur notre capacité à extraire rapidement les différences et similarités entre objets, nous permet de ne pas encoder les objets environnants au cas par cas ce qui soulage grandement la charge mnésique de notre cerveau. Cela est d'autant plus avéré que plusieurs attributs coexistent habituellement au sein d'une catégorie. Dans le monde, ce qui a un bec a la plupart du temps des plumes. La seule exception me venant à l'esprit est l'ornithorynque qui de ce fait est très atypique. De par cette redondance entre attributs, la reconnaissance de certains objets serait facilitée. La catégorisation nous permet également d'associer les différentes vues d'un même objet à une unique entité. « Bien que différentes vues de votre propre visage dans le miroir produisent des motifs de stimulation rétinienne différents, elles ne sont pas pour autant catégorisées comme des visages différents, mais plutôt comme des vues différentes d'un même visage, qui dans ce cas présent, est reconnu comme le vôtre » (Schyns, 1997). Ainsi, notre capacité à reconnaître et catégoriser les objets serait invariante au point de vue de l'objet (Biederman, 1987). Cette capacité de notre cerveau se voit confirmée par la découverte de neurones dans le cortex inféro-temporal répondant

sélectivement à un objet présenté sous différents points de vue (Booth & Rolls, 1998, Logothetis & Pauls, 1995 , Vogels & Orban, 1996). A noter cependant qu'une grande quantité d'objets semble être plus efficacement reconnue selon un point de vue spécifique dit « canonique » plus représentatif (Bulthoff & Edelman, 1992, Edelman & Bulthoff, 1992, Tarr, Williams, Hayward & Gauthier, 1998, voir également Hayward & Williams, 2000).

Via la catégorisation, le cerveau s'affranchit ainsi de l'existence d'un objet réel pour privilégier le travail sur une représentation d'un objet et donc d'un concept, intégrant ainsi sa description physique, sa fonctionnalité, son label ou encore ses probabilités d'occurrence. Si ce mécanisme simplifie nos possibilités d'interactions avec les objets, il peut également présenter des limites plutôt perverses en aboutissant à des catégorisations excessives.

En effet, assimiler les objets à des catégories tend à accentuer les ressemblances entre stimuli de même catégorie (biais d'assimilation) tout en accentuant également les différences entre stimuli de catégories différentes (biais de contraste). Ainsi, Tajfel et Wilkes ont montré en 1963 que face à deux groupes de barres verticales A et B, les barres A étant supérieures aux barres B, les sujets à qui on demandait d'estimer quantitativement la taille des barres avaient tendance à surestimer les différences entre les deux groupes (Tajfel & Wilkes, 1963). Si ce problème paraît anecdotique lorsqu'on s'intéresse à des objets inanimés, il devient fondamental appliqué dans le domaine social, pouvant être en partie la source de ces fléaux que sont le racisme et l'intolérance.

Il est donc plus que pertinent d'avoir la capacité de moduler ces mécanismes tendancieux, de spécifier les catégorisations d'intérêts ainsi que les traits physiques diagnostiques pour la tâche à accomplir.

4.2. Nature de l'information visuelle haut-niveau

Si de nombreux modèles de reconnaissance d'objets s'appuient sur l'intégration de percepts plus ou moins complexes (Biederman, 1987, Edelman, 1997, Riesenhuber & Poggio, 1999), tous s'accordent à penser que ces informations ne peuvent être utilisées en tant que telles pour y trouver un sens. Poggio, Edelman et ses collègues ont dans ce but proposé un cadre de travail pour expliquer la transition d'une représentation des propriétés physiques d'un stimulus vers la représentation d'objets complexes (Edelman & Duvdevani-Bar, 1997, Poggio & Edelman, 1990). Ils proposent l'utilisation d'une classe spécifique de fonctions, les

RBFs (pour « Radial basis functions ») qui permettraient de caractériser les réponses des neurones de haut-niveau. En se plaçant dans un espace de représentation défini par des dimensions de « significations psychologiques », les RBFs répondraient à un stimulus optimal. Leurs courbes de réponses dépendraient ainsi de la distance entre la représentation du stimulus présenté et la représentation d'un stimulus optimal. Chaque fonction serait ainsi caractérisée par un paramètre correspondant à la position du stimulus optimal dans l'espace de représentation. Des données neurophysiologiques enregistrées en IT chez le macaque appuient cette idée en démontrant une forte relation existante entre la distance de représentation et les profils de réponse des neurones (Op de Beeck, Wagemans & Vogels, 2001a).

Ce type de modèle est d'ailleurs très similaire à la « Template Theory » proposée par (Tarr & Bulthoff, 1998) qui suggère une reconnaissance de l'objet basée sur une comparaison entre la représentation des objets stockés en mémoire, et la représentation du stimulus perçu. Ils proposent de plus que la représentation du stimulus perçu pourrait être modifiée en termes de taille ou de point de vue.

Enfin, d'autres auteurs tels que Biederman ou Marr proposent l'existence d'un « alphabet » de composantes qui pourraient être utilisé pour décrire l'objet perçu (Biederman, 1987, Marr & Nishihara, 1978). Ainsi, notre perception d'une sphère reposant sur la base d'un cône nous permettrait d'identifier l'objet en tant que glace. Ces hypothèses trouvent appui dans le fait que des neurones de IT seraient sélectifs aux « composantes primitives » ou géons utilisées pour construire leur « alphabet » (Biederman & Bar, 1999, Vogels, Biederman, Bar & Lorincz, 2001). Il est à noter d'ailleurs que si les neurones de IT sont sensibles à des objets dans leur globalité et à des composantes primitives, leurs réponses peuvent aussi s'avérer spécifiques de fragments d'objets et même de fragments de composantes quelle que soit la taille du stimulus. Sur cette base, Edelman et Intrator (Edelman & Intrator, 2003) suggèrent que ces neurones répondant à des fragments d'objets pourraient également jouer un rôle dans l'évaluation de la distance entre représentations en mémoire et stimuli perçus.

Dans tous les cas, on peut se demander si la construction des représentations d'objets dans les aires haut-niveau est la première étape nécessaire à la catégorisation de ces mêmes objets. A mon sens, on peut dire que la construction de la représentation d'un objet est sa catégorisation. En effet, si la construction d'une représentation d'objet défini repose sur l'intégration de ses différents traits diagnostiques, cela sous-entend que l'ensemble des indices nécessaires à la catégorisation sont d'ores et déjà disponibles. Il semblerait peu

écologique d'utiliser les traits diagnostiques pour la construction d'une représentation, puis d'identifier l'objet, et enfin d'évaluer la quantité de traits en commun entre la représentation et les différentes catégories d'objets stockées en mémoire. Ainsi, la catégorisation pourrait être une étape préalable à la représentation aboutie de l'objet. Cette question sur la chronologie des différentes étapes de traitements d'un objet – détection, catégorisation, identification – a d'ailleurs été débattue dans une étude récente menée par Grill-Spector et Kanwisher intitulée « As soon as you know it is there, you know what it is » (Grill-Spector & Kanwisher, 2005). Dans leur expérience n°3, les auteurs ont testé leurs sujets sur 3 tâches de choix forcé différentes au sein desquelles ils faisaient varier le temps d'exposition des images test non masquée : une tâche de détection d'objet (vs. texture randomisée), une tâche selon eux de catégorisation basique d'objet et une tâche de catégorisation subordonnée. De manière surprenante, les sujets obtiennent des performances similaires en terme de précision et de temps de réaction dans la tâche de détection et dans la tâche de catégorisation basique. Les auteurs, sur la base de ces résultats et d'autres résultats préalables défendent alors l'idée que détection et catégorisation basique reposeraient sur les mêmes traitements visuels contrairement à l'idée communément acceptée que la détection de l'objet est préalable à sa catégorisation. Il est néanmoins important de noter comme le soulève Bowers et Jones (Bowers & Jones, 2008) que les sujets impliqués dans leur tâche de catégorisation basique devaient par exemple répondre « oui » lorsqu'ils apercevaient une voiture tandis que les distracteurs correspondaient à d'autres objets de catégories superordonnées différentes. Dans ce cas, une simple détection des traits physiques diagnostiques de la catégorie suffit à effectuer correctement la tâche de catégorisation qui devient alors une simple tâche de détection. Désireux de contrôler plus amplement ces résultats, les auteurs ont répliqué les travaux de Grill-Spector et Kanwisher en utilisant pour la tâche de catégorisation des cibles et des distracteurs de même catégorie superordonnée : des chats et des chiens par exemple. Avec ces nouvelles conditions scientifiquement plus valides vu les hypothèses testées, il apparaît que la tâche de détection est plus facilement réalisée que la tâche de catégorisation basique démontrant que détection et catégorisation résultent d'étapes de traitements visuels différentes. D'autres preuves de la dissociation entre ces deux traitements ont d'ailleurs depuis été apportées par Mack et collaborateurs qui ont révélé que lors de la présentation de stimuli dégradés, les sujets étaient meilleurs à détecter les objets qu'à les catégoriser (Mack, Gauthier, Sadr & Palmeri, 2008). Le chapitre suivant vise à préciser les connaissances déjà

acquises sur la catégorisation, aussi bien dans le domaine de la linguistique que de la neuropsychologie, et plus spécifiquement, dans le domaine de la perception visuelle.

4.3. Un même stimulus, des statuts différents… L'influence de la tâche

De nombreuses tâches expérimentales impliquent la reconnaissance d'indices diagnostiques, l'étude des traitements sous-jacents représente un cadre de recherche à part entière (Schyns, 1998). Selon la tâche visuelle que nous avons à effectuer, le travail de notre système visuel ne va pas être dirigé vers la recherche des mêmes caractéristiques physiques, mais des flux d'informations descendantes vont permettre la pré-activation de populations neuronales privilégiant la détection de l'information la plus diagnostique. En outre, il arrive souvent qu'une partie de l'information diagnostique ne soit pas disponible à l'extraction, par exemple lorsqu'une partie de l'objet cherché est masquée ou encore lorsque les stimuli sont partiellement altérés. Dans ce cas, notre capacité à catégoriser un objet dépendra de notre capacité à déterminer les meilleurs indices diagnostiques, indices variant en fonction du protocole expérimental, de la consigne, de l'information exploitable dans les stimuli utilisés, et de la stratégie expérimentale adoptée. Cette diagnosticité n'est d'ailleurs pas figée mais peut évoluer avec l'expérience. Ainsi, la reconnaissance diagnostique ne se limite pas à identifier les indices diagnostiques et à comprendre comment ils sont intégrés dans l'objet à catégoriser Cela est d'autant plus vrai que bien souvent, le fait même de lier un objet à une catégorie modifie notre perception et notre analyse de ses traits diagnostiques (Goldstone, 1994, Schyns & Rodet, 1997, Tajfel & Wilkes, 1963). De plus, on peut également se demander ce qu'il advient du traitement des traits physiques non-diagnostiques (Schyns, 1998).

Il est maintenant reconnu que la réponse des neurones à différents niveaux de la voie visuelle peut être modifiée selon la tâche que le sujet doit effectuer. Le sujet est capable en fonction de la consigne de concentrer son attention sur la recherche de certains traits physiques diagnostiques et d'ignorer les autres. Ces observations comportementales s'accompagnent au niveau neurophysiologique d'une modification des réponses et de la sélectivité des neurones impliqués (Desimone & Duncan, 1995, Sigala & Logothetis, 2002, Treue, 2001). Op de Beck et collaborateurs ont de plus montré que notre système visuel peut traiter deux dimensions d'un même objet différemment selon la tâche (Op de Beeck,

Wagemans & Vogels, 2003). Sous l'influence de la composante attentionnelle, les sujets étaient capables d'effectuer une tâche de catégorisation sur les courbes d'un objet indépendamment de sa taille, puis d'effectuer une tâche relative à la taille de l'objet indépendamment des courbes présentes. Ils démontrent donc que certaines dimensions s'avèrent séparables (comme les courbes et la taille de l'objet) tandis que d'autres dimensions sont considérées comme intégrales et ne peuvent être traitées indépendamment.

La catégorisation n'est pas un mécanisme uni-directionnel mais un couplage entre traitements ascendants et descendants liant percept et concepts. A noter qu'ici les traitements descendants ne correspondent pas à des boucles de rétro-action (« feedback ») mais à une pré-activation des représentations catégorielles en provenance des aires plutôt frontales permettant la sélection des caractéristiques physiques diagnostiques à rechercher. Ainsi, en fonction de la tâche, la diagnosticité des percepts serait modulée par la tâche à accomplir (Schyns, 1997).

Dans une tâche visuelle de catégorisation d'animaux mammifères/non-mammifères, nous aurions par exemple tendance à utiliser abusivement la caractéristique diagnostiques « 4 pattes », et de ce fait pourrions générer des erreurs lors de la présentation d'une photographie contenant un dauphin ou une chaise !

4.4. Un même item, des niveaux de catégorisation différents

Les labels de catégorie d'objets encodés dans notre mémoire sont loin d'être qualitativement égaux. Un père se promenant avec son petit garçon dans un parc et croisant un aveugle guidé par son labrador aura plutôt tendance à parler de l'animal à son enfant en utilisant le terme « chien » plutôt que le terme « labrador », « animal », ou encore « mammifère ». De cette observation ressortent plusieurs phénomènes intéressants. Tout d'abord, nous disposons d'un lexique important pour dénommer et caractériser un objet unique : le labrador est un chien, un canidé, un mammifère, un animal, un être vivant, mais également un animal de compagnie et un guide pour aveugle. Si l'enfant connaît personnellement l'animal, le père pourra même utiliser son nom, par exemple, « Rantanplan ». Nous privilégions donc l'utilisation d'un label en fonction de la situation, selon le but de notre dialogue et selon la personne à qui s'adresse le discours. Il serait assez inadéquat de parler de mammifère à un enfant, ce dernier risquant de ne pas comprendre le

sujet de la discussion. De plus, selon la thématique du dialogue, nous nous attacherons plutôt à l'aspect descriptif de l'objet (un chien) ou à son aspect fonctionnel (un guide, un animal de compagnie). Le choix du label est donc un compromis instantané effectué dans le but de donner le plus d'informations exploitables possible.

En 1976, une étude déterminante de Rosch et al. (Rosch, Mervis, Gray, Johnson & Boyes-Braem, 1976) précisa l'importance d'un niveau de catégorisation appelé « niveau de base » (Rosch et al., 1976), « accès principal » (Biederman, 1987) ou encore « point d'entrée » (Jolicoeur, Gluck & Kosslyn, 1984). Ce niveau de catégorisation privilégié serait le plus spontanément utilisé pour dénommer les objets visuels apparaissant dans notre champ visuel (Jolicoeur et al., 1984, Rosch et al., 1976). Dans l'exemple précédent, le niveau de base correspond à la catégorie « chien ». De même, si on fait entendre au sujet le nom d'une catégorie d'objet 0.5 seconde avant d'afficher une image d'un objet appartenant ou non à cette catégorie, les sujets sont plus rapides à effectuer une tâche de vérification lorsque le mot prononcé correspond au niveau de catégorisation basique (expérience 7, Rosch et al., 1976). Dans cette même expérience, les catégories dites superordonnée (plus générales et abstraites, animal par exemple) et subordonnées (plus fine et stricte, labrador par exemple) entraînaient des temps de traitement plus longs.

C'est par le biais de cette impressionnante étude (décrivant 12 expériences différentes !) que Rosch et al. ont suggéré une hiérarchie des niveaux de catégorisation reflétant l'inclusion des catégories les unes dans les autres. On distinguerait ainsi un niveau d'entrée, le niveau basique privilégié dans toutes les tâches sémantiques (telle que la dénomination) comprenant des catégories telles que « voitures », «chiens », « chats », «chaises »… Ces catégories pourraient être incluses dans des catégories plus générales et abstraites, les catégories superordonnées : « véhicules », « animaux », « meubles »… Elles incluraient dans le même temps les catégories subordonnées plus fines et limitées telles que « peugeot 205 », « labrador », « chaise longue »… Le niveau de base étant le niveau d'entrée, la catégorisation d'images selon leur niveau superordonné ou leur niveau subordonné nécessiterait un temps de traitement supplémentaire. Alors que des temps de réactions similaires sont enregistrées pour des images présentées 75 et 250 ms lors d'une tâche de catégorisation à un niveau basique, une baisse de performance dans la condition 75 ms est démontrée lorsque les sujets sont impliqués dans une tâche de catégorisation subordonnée (Jolicoeur et al., 1984). L'accès à un niveau de catégorisation subordonné nécessiterait des traitements cognitifs supplémentaires

dans le but d'accentuer de légères différences et ainsi préciser une catégorie beaucoup plus fine. Cependant, ces temps de traitements plus longs constatés lors d'une catégorisation subordonnée furent discutés par Murphy et al.. Les auteurs suggèrent que le label plus long des catégories serait susceptible de rendre plus complexe la compréhension du label et pourrait être ainsi la cause d'une décision moins rapide (Murphy & Smith, 1982). L'accès à une catégorie superordonnée dans une tâche de dénomination nécessiterait également des traitements cognitifs supplémentaires cette fois dans le but de gommer les différences entre catégories afin de préciser une catégorie plus générale mais plus abstraite. Les objets d'une même catégorie superordonnée partageraient ainsi de nombreuses caractéristiques fonctionnelles et au contraire peu de caractéristiques perceptuelles. Cependant, lors de la perception d'une scène contenant plusieurs objets non-isolés d'une même catégorie superordonnée, l'avantage de la dénomination à un niveau basique par rapport au niveau superordonné serait plus faible. En outre, lorsque le contexte de la scène est inapproprié avec les objets d'intérêts, de plus grandes interférences surviendraient lors de l'identification de ces objets à un niveau superordonnée (Murphy & Wisniewski, 1989).

Les catégories basiques seraient quant à elles les premières catégories apprises par l'enfant et les plus fréquemment utilisées (Murphy & Smith, 1982). Elles correspondraient aux catégories les plus générales contenant des objets de formes similaires interagissant fonctionnellement de manière semblable. L'étude menée par Rosch et al. a en effet montré que les sujets impliqués dans une description libre des objets avaient tendance à utiliser les mêmes attributs pour des objets de même catégorie de base. A la prise de connaissance de ces résultats, Tversky et al. notèrent qu'une grande partie de ces attributs correspondaient à des parties de l'objet partageant aussi bien des propriétés physiques que fonctionnelles (le manche d'un tournevis est adapté à la forme de main et permet de saisir l'outil), ce qui expliquerait leur grande valeur informative (Tversky & Hemenway, 1984). Cependant, d'autres études défendent la valeur critique des caractéristiques perceptuelles uniquement (Murphy & Smith, 1982) et démontrent l'importance d'attributs ne correspondant pas à des parties d'objets (Murphy, 1991). Finalement, catégoriser un objet pourrait correspondre à l'évaluation de la distance entre la représentation de l'objet et les concepts stéréotypiques des catégories déjà encodées en mémoire. Un argument allant dans ce sens est le fait que l'être humain est aussi rapide à catégoriser une autruche en tant qu '« oiseau » qu'en tant qu' « autruche ». Ainsi dans le cas d'objet trop atypique, le niveau d'entrée pourrait être le niveau subordonné (Jolicoeur et al., 1984). Il est également important de noter que le niveau d'entrée pourrait être

fortement modulé par le contexte. Au rayon outillage d'un grand magasin de bricolage, on tendra spontanément à utiliser le niveau de catégorisation subordonnée pour désigner un objet d'intérêt. On cherchera plutôt un tournevis cruciforme qu'un simple tournevis.

4.5. Une même dénomination, des représentations inter-individuelles différentes

La bibliothèque des traits physiques diagnostiques disponibles et des analyses perceptives utilisables par les mécanismes de catégorisation n'est pas fixe, mais au contraire se construit progressivement au cours de notre historique, modulée par la découverte d'objets nouveaux (Schyns & Rodet, 1997). Ainsi, les adultes utiliseraient une bibliothèque de traits diagnostiques plus abondante que les jeunes enfants (Smith, Carey & Wiser, 1985). De cette idée pourrait découler également le phénomène d'expertise, les experts dans une catégorie (les ornithologues par exemple) étant plus à même de déterminer quels indices sont diagnostiques, et développant également une capacité à extraire plus d'informations (Tanaka & Taylor, 1991). Une fois développées, les règles de reconnaissance sous-tendant cette expertise sont automatiquement appliquées et difficilement transmissibles à un novice comme en témoignent ces études comparatives entre radiologistes novices et experts (Norman, Brooks, Coblentz & Babcook, 1992) ou encore entre « vétérinaires » novices et experts (Biederman & Shiffrar, 1987).

De plus, même à âge et expertise comparable, nos répertoires d'indices diagnostiques restent différents (Schyns & Rodet, 1997). Tandis que des sujets ont été informés qu'ils allaient devoir apprendre à catégoriser des « cellules martiennes » inventées, ils ont suivi un apprentissage des règles de catégorisation suivant un ordre différent. L'ensemble des cellules pouvaient être classées en 3 catégories (Figure 16) : cellules X, Y et XY contenant respectivement des formes génotypiques x, y et une conjonction des deux xy. A noter que dans les cellules xy, aucune séparation claire n'était présente entre le « génotype » x et le « génotype » y les composant. De plus, des formes non-diagnostiques figuraient en plus des formes génotypiques dans l'ensemble des cellules. Tandis qu'un groupe de sujets A a appris à reconnaître d'abord les cellules X et Y, puis les cellules XY, l'autre groupe de sujet B a appris d'abord à repérer les cellules XY puis les X et Y. Les auteurs ont ainsi démontré que contrairement aux sujets A, les sujets B avaient appris la configuration XY sans se rendre compte que xy n'étaient que la conjonction des traits x et y. Cette étude met bien en évidence

que nos représentations conceptuelles dépendent grandement de nos confrontations passées avec notre environnement.

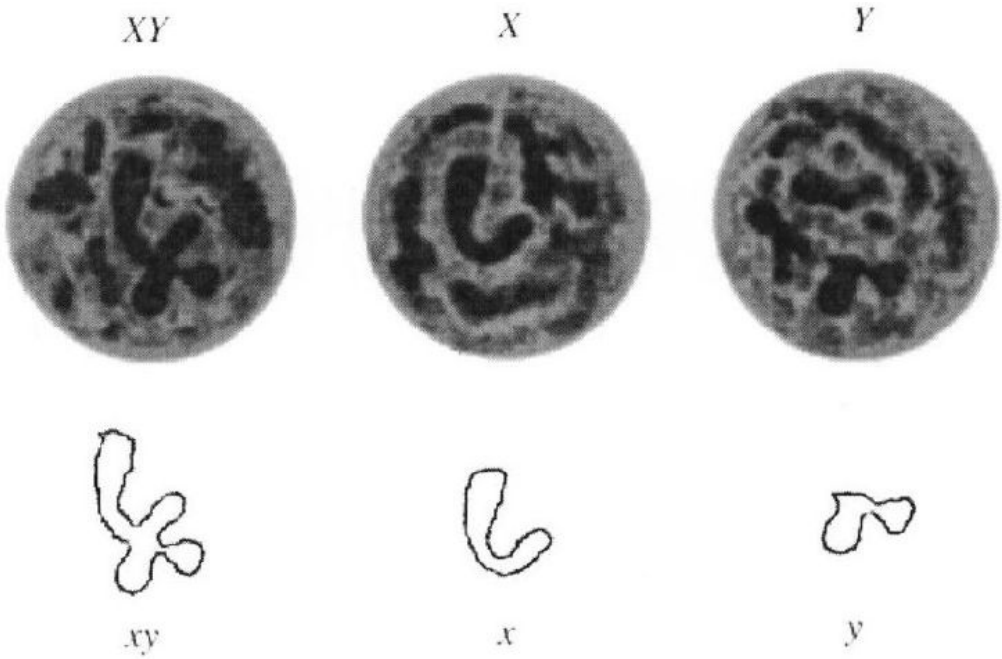

Figure n°16 : Tiré de (Schyns & Rodet, 1997). Exemples de stimuli « cellules martiennes ». Les cellules XY présentent un génotype assimilable à la conjonction des traits diagnostiques de X et de Y. Les sujets entrainés à catégoriser ces cellules martiennes selon une chronologie différente ne vont pas développer le même répertoire de traits diagnostiques. La tâche influence donc la diagnosticité du percept.

4.6. A la base de la reconnaissance sémantique, la mémoire…

"if a fragment of a stimulus categorizes objects (distinguishes membres from non-members), the fragment is instantiated as a unit in the representational code of object concepts (p.310, Schyns & Murphy, 1994):" Cette citation illustre parfaitement la théorie du « Principe de Fonctionnalité » (Functionality Principle**)** et sous-tend l'idée d'un « lexique » de concepts stockés en mémoire organisant l'ensemble des représentations mentales d'objets et leurs caractéristiques associées. Cette idée trouve assez bien son équivalent dans la notion de *frame* (Friedman, 1979) définissant à la base une structure mnésique dont le parallèle physiologique pourrait correspondre aux cellules grand-mère. Une *frame* contient la description stéréotypée d'un évènement en associant des connaissances physiques, sémantiques, relationnelles, relatives à cet évènement. Ces *frames* seraient regroupées les unes aux autres au sein d'une hiérarchie conceptuelle, l'ampleur et la nature de leur contenu dépendant de leur niveau dans la hiérarchie. De plus, les *frames*, comme les catégories taxonomiques préalablement décrites,

pourraient s'enchâsser les unes dans les autres. La *frame* « tête » contiendrait la *frame* « nez » mais appartiendrait dans le même temps à la *frame* « humain ». Dans tous les cas, plus une *frame* serait considérée comme abstraite, plus ses arguments seraient globaux et conceptuels. Ces *frames* pourraient se construire au fur et à mesure de notre expérience de vie, sur la base d'un savoir en partie inné, nous permettant d'inférer de nouvelles connaissances. Leurs attributs ne seraient pas nécessairement précis mais pourraient représenter des gammes de valeurs (la couleur des fleurs…). Ces gammes de valeurs pourraient d'ailleurs être modifiées, complétées avec le temps aussi bien d'un point de vue phylogénétique qu'ontogénétique. Ainsi, une *frame* pourrait avoir une certaine invariance. D'un point de vue phylogénétique, il est évident que le règne animal a évolué avec le temps soumis aux lois de la sélection naturelle. Les animaux les plus avancés dont l'humain fait partie, ont dû au cours de l'évolution optimiser leurs représentations des catégories d'objets afin d'être toujours plus précis et toujours plus rapide à réagir. Cela pourrait d'ailleurs être la cause de la spécificité de certaines catégories d'objets tels les « animaux » ou les « visages », objets animés capables d'interagir indépendamment de notre propre volonté (New et al., 2007). D'un point de vue ontogénétique, lorsqu'un nouvel évènement survient dans notre vie, nous construirions des attentes fondées sur le savoir déjà acquis. Si ces attentes s'avèrent fausses, nous n'allons pas pour autant complètement remettre en cause notre savoir mais plutôt le compléter si cela est opportun.

Les *frames* fonctionneraient comme des analyseurs ou des détecteurs de patterns sémantiques en fonction de notre familiarité et des attentes dans une situation donnée. Face à une situation familière, des *frames* pré-activées limiteraient la gamme d'objets probablement attendus rendant ainsi possible une perception quasi-automatique de chaque objet dans la scène. La quantité d'analyses physiques locales à effectuer sur l'objet dépendrait alors de la quantité d'information à accumuler pour distinguer cet objet des autres objets contenus dans une *frame* plus ou moins vaste. Il serait alors plus rapide de distinguer un objet faisant partie d'une *frame* simple que d'une *frame* complexe (la *frame* désert par exemple contiendrait moins d'éléments qu'une cuisine habituellement plus chargée et dense en éléments, (Palmer, 1975). Allant dans le même sens, un objet attendu serait automatiquement détecté tandis qu'un objet inattendu nécessiterait des traitements ascendants et descendants interactifs. Si cette théorie des *frames* est susceptible d'illustrer l'idée d'un registre mnésique stockant l'ensemble des informations relatives à un nombre divers de concepts, elle est donc également en faveur d'une catégorisation basée sur l'interaction de traitement ascendants et descendants.

Il est également intéressant de noter que cette étude fait mention de *frame* de scènes, telle qu'une *frame* « cuisine » qui ne serait activée qu'après détection de plusieurs items représentatifs comme un four, une réfrigérateur ou une machine à café. Elle privilégie donc l'idée d'une reconnaissance de scènes naturelles basée sur la détection des objets en son sein.

4.7. Et la catégorisation du contexte ? Un aspect encore méconnu

La majorité des études sur la catégorisation du contexte absolu s'attachent à caractériser les environnements selon des mesures quantitatives multidimensionnelles susceptibles de servir de base à une catégorisation efficace. Plus rares sont les études abordant la question d'un point de vue taxonomique toute aussi importante. Dans la vie de tous les jours, lorsque nous rencontrons quelqu'un dans la rue ou encore au téléphone, il nous arrive souvent de demander « ou es-tu ? » ou encore « D'où viens-tu ? ». L'intérêt de la question, mise à part le fait qu'elle puisse combler le vide d'une discussion mal engagée, réside dans le fait que la réponse permet de préciser aussi bien la position géographique de l'interlocuteur que son occupation du moment. S'il est à la piscine, on peut deviner la structure spatiale de l'environnement, tout en s'imaginant que la personne y est allée pour bronzer ou nager. L'ensemble de ces possibilités correspond aux attributs physiques et fonctionnels de la catégorie environnementale. Dans ce sens, une seule étude à ma connaissance répliqua les travaux de Rosch à propos des niveaux de catégorisation des objets en les appliquant aux environnements de scènes naturelles. Partant des catégories superordonnées arbitrairement définies « Intérieurs » et « Extérieurs », Tversky et Hemenway (Tversky & Hemenway, 1983) ont après enquête sélectionné et organisé diverses catégories de contextes selon la hiérarchie présentée figure 17.

Suite à la mise en place de cette hiérarchie, les auteurs demandèrent à leurs sujets de définir pour chaque catégorie de chaque niveau de catégorisation un maximum d'attributs perceptifs (pour la forêt par exemple : sombre, présence d'arbres et d'animaux…), d'activités liées (chasser, camper, monter dans les arbres…), et de parties de l'environnement (arbres, arbustes, herbe, vie sauvage…). Le but de cette opération était de faire l'analogie avec les travaux de Rosch (Rosch et al., 1976) démontrant une augmentation de la liste des attributs perceptifs et fonctionnels au fur et à mesure que l'on descend dans les niveaux de

catégorisation, le niveau basique étant le meilleur compromis entre quantité d'informations et généralisation.

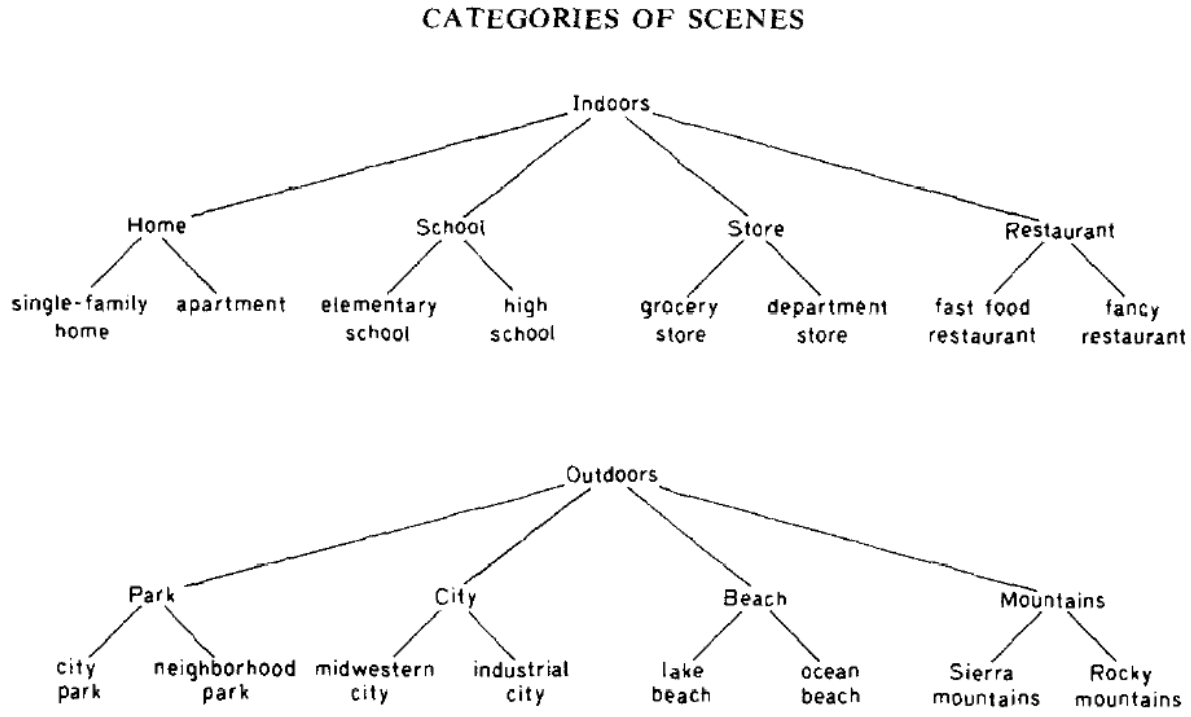

Figure n°17 : Tiré de Tversky et Hemenway, 1983. Comme pour les catégories d'objets, il est possible de dresser une hiérarchie des catégories de scènes et de déterminer différents niveaux de catégorisation : superordonné, basique, et subordonné.

Après vérification des attributs cités par les sujets, il apparait que l'augmentation de la liste des caractéristiques entre catégorie superordonnée et catégorie de base est supérieure à l'augmentation de la liste des caractéristiques entre catégorie de base et catégorie subordonnée. Cette constatation vaut pour les attributs perceptifs autant que pour les activités ou encore les parties de scènes. Les catégories considérées de base correspondent de plus parfaitement à la définition donnée par Rosch, c'est-à-dire des catégories partageant de nombreux attributs et activités communes, tout en possédant d'autres caractéristiques exclusives. De plus, les catégories superordonnées de scènes sont plus générales et abstraites, les catégories subordonnées plus fines et définies. La cooccurrence est aussi de mise pour les scènes, un environnement au sein duquel se trouve une balançoire contiendra aussi probablement un toboggan, des enfants, de l'herbe. Ainsi, comme pour les objets, les parties et leurs fonctions définissent le tout.

I.4. Des traitements visuels spécifiques de la tâche : la catégorisation

Dans deux ultimes expériences, les auteurs confièrent à leurs sujets la tâche de dénommer des photographies de scènes et de compléter des phrases avec le nom d'une catégorie (le sens de la phrase menant bien sûr à la dénomination d'une catégorie) en utilisant le label le plus simple et le plus commun possible. Cette consigne est à mon avis discutable tant elle biaise le choix des sujets vers les labels que les auteurs ont sélectionnés comme basiques. L'analyse des résultats montre que les sujets privilégient quasi-exclusivement le nom de la catégorie de base pour dénommer les photos et choisissent uniquement la catégorie de base pour compléter les « blancs » dans les phrases.

Cette étude appliquée aux scènes naturelles réplique remarquablement bien les résultats de l'étude initiale menée sur la catégorisation d'objet. Elle démontre l'existence des niveaux de catégorisation superordonnés, basiques, et subordonnés organisant les scènes naturelles au moins d'un point de vue sémantique.

Plusieurs articles présentés dans ce mémoire décrivent des expériences dans lesquels il est demandé aux sujets de catégoriser des scènes naturelles à un niveau superordonné « Environnement manufacturé vs. naturel » ainsi qu'à un niveau de base « Mer / Montagne / Intérieur urbain / Scènes de rues ». Nous verrons que cette organisation hiérarchique bien établie dans des tâches lexicales (dénomination / description) n'est peut être pas identique dans des tâches visuelles.

5. La reconnaissance d'une scène naturelle d'un seul regard…

Lorsque nous évoluons dans notre monde, ce que nous pouvons apercevoir autour de nous est habituellement stable, continu, sans changements majeurs. De plus les changements survenant ne sont pas brutaux. Ce sont généralement des objets qui évoluent indépendamment de nous selon des mouvements ayant une vitesse et une orientation donnée. Pourtant dans certaines situations, quand nous survolons le contenu des chaînes à la télévision, quand nous allumons la lumière dans une salle jusque là non éclairée, ou encore quand nous apprécions la bande annonce d'un film au cinéma, les scènes apparaissent brutalement en un fragment de seconde. Malgré ces apparitions brutales, nous sommes capables très rapidement d'extraire le sens des nouvelles scènes. Etant donné qu'aucune attente sur les scènes naissantes n'est possible, ces situations mettent en évidence la capacité de notre système visuel à extraire et intégrer immédiatement une information visuelle complexe sans l'influence de traitements cognitifs descendants préalables.

Ce chapitre a pour but de préciser la nature de l'information visuelle extraite et les traitements rapides sous-jacents. Je tâcherai dans un premier temps de mettre en avant les différences entre le concept et le percept d'une scène, puis j'effectuerai une revue de la littérature portant sur la quantité d'information extraite d'une analyse globale de scènes naturelles. Je présenterai finalement différents modèles de reconnaissance de scènes naturelles s'inscrivant dans deux courants distincts. Le premier défend l'hypothèse d'une précédence de l'analyse locale sur l'analyse globale. Le second courant repose sur l'idée qu'une quantité d'informations visuelles disponibles après une analyse globale et immédiate de la scène suffirait à donner un premier sens et un cadre de référence à notre perception.

5.1. Une représentation sémantique rapidement disponible : le gist conceptuel

De manière assez stupéfiante, notre système visuel est capable de détecter la présence d'un animal, d'un véhicule, d'un visage dans des scènes naturelles flashées pendant 20 ms seulement. (Delorme et al., 2000, Fabre-Thorpe et al., 2001, Fabre-Thorpe et al., 1998, Rousselet et al., 2003, Thorpe et al., 1996, VanRullen & Thorpe, 2001a). Nous arrivons également à détecter une image cible incluse dans une séquence d'images flashées toutes les 125 ms par image (Rapid Serial Visual Presentation), que la description de la cible ait été

visuelle ou verbale (Potter, 1975). Cependant, un temps d'apparition de l'image de 300 ms serait le minimum nécessaire pour encoder en mémoire l'image (Potter, 1976). En dessous, la représentation de l'image serait trop volatile car encore non consolidée. Cette représentation rapidement formée sur la base de la présentation d'une image correspond au « gist » de l'image, c'est à dire son « essence ». Sur la base de ces premiers résultats, une autre étude plus récente a tenté de préciser la nature des informations extraites d'un seul regard, le gist, en faisant varier le temps de présentation des scènes naturelles. Les sujets ayant reçu la consigne de décrire librement les images masquées qu'ils avaient perçues durant un temps variant de 27 à 500 ms tendent à caractériser des attributs physiques pour les conditions de 27 à 53 ms. A partir de 53 ms, des qualificatifs correspondant à des objets et des scènes d'extérieurs sont cités. Il faut atteindre un temps de présentation de 80 ms pour que les concepts liés aux scènes d'intérieurs soit plus largement mentionnés que de simples descriptions physiques (Fei-Fei, Iyer, Koch & Perona, 2007). Dans tous les cas, l'avantage dans la détection de traits physiques distribués laisse envisager l'idée de traitements physiques globaux bas-niveau survenant avant l'identification locale d'objets. Cette dernière idée va de paire avec un autre résultat de l'étude montrant que les niveaux superordonnés des catégories auraient tendance à être dénommés avant les niveaux de base. Si l'antériorité d'une analyse globale des traits physiques est à ce niveau là encore discutable, il semble cependant que les caractéristiques physiques et le gist sémantique des scènes ne soient pas complètement corrélés. En effet, dans la cadre d'une expérience s'appuyant sur le phénomène de cécité attentionnelle au changement (Change blindness), Rensink et al. montrent que certains changements (de position, de couleurs, d'apparition/disparition) appliqués sur de larges régions d'une scène restent souvent ignorés (Rensink et al., 1997). Cela est d'autant plus vérifié que la zone de changement est dépourvue d'intérêt. D'un point de vue statistique, la surface des régions marginales modifiées est pourtant plus importante que la surface des régions d'intérêts. Le phénomène de cécité attentionnelle est donc largement dépendant du gist (les objets le composant tendraient à capturer l'attention) et peu dépendant des traits physiques.

5.2. Une représentation perceptuelle rapidement disponible : Le *gist* perceptuel

Les études décrites précédemment visaient à mieux comprendre l'intégration d'informations sémantiques au sein de scènes visuelles plutôt que l'extraction d'un percept physique proprement dit. Ainsi, la première notion de *gist* et la plus généralement utilisée correspond à l'essentiel sémantique d'une scène, dont la représentation dépendrait majoritairement des aires visuelles de haut-niveau. Plus récemment, Oliva proposa de préciser la nature du *gist* en suggérant l'idée d'un *gist* perceptif et d'un *gist* conceptuel, chacun dépendant d'étapes de traitements hiérarchiques différentes lors de la perception de scènes (Oliva, 2005). Dans le cadre de la recherche sur les scènes naturelles, les notions de *gist* et de *layout* sont utilisées de manière croissante. Si chacun de ces termes ont pour origine des études différentes, ils sont néanmoins très complémentaires et désormais communément utilisés. Il est de ce fait important de définir à quoi se réfère chacun d'entre eux pour éviter de possibles confusions, d'autant que certaines différences peuvent apparaitre subtiles. Au sein d'une scène naturelle, les informations visuelles n'ont pas la même valeur et leurs traitements sont soumis à des décours temporels différents. Dans tous les cas, l'information initiale correspond à un percept physique dont découle progressivement la construction d'une représentation sémantique indispensable à la compréhension du monde. Ce percept physique ou *gist* perceptuel est sous-tendu par la structure spatiale de la scène (ou *spatial layout*)

Le gist percepuel vs. gist conceptuel

Le *gist* perceptuel reposerait sur l'extraction précoce puis l'intégration d'informations physiques globales diagnostiques de la scène. Le *gist* perceptuel ainsi construit serait suffisant pour construire une représentation globale de la scène ne correspondant pas obligatoirement à l'essentiel sémantique de cette même scène. Tandis que le *gist* conceptuel de la photographie Figure 18.A correspond à « une personne déguisée en ours dans un carnaval », un *gist* perceptuel du type Figure 18.B induirait plutôt la représentation « scènes de rues ». Il serait l'aboutissement d'une intégration globale bas-niveau des contours des bâtiments, de la surface de la route, et de l'espace de ciel, éléments diagnostiques de la structure spatiale. Des scènes schématiques mettant en avant les caractéristiques globales de la scène sont d'ailleurs plus précisément et plus rapidement identifiés que des photos détaillées (Ryan & Schwartz, 1956).

Ces caractéristiques globales feraient partie intégrante de la structure spatiale de la scène intégrée différemment selon les modèles et auteurs.

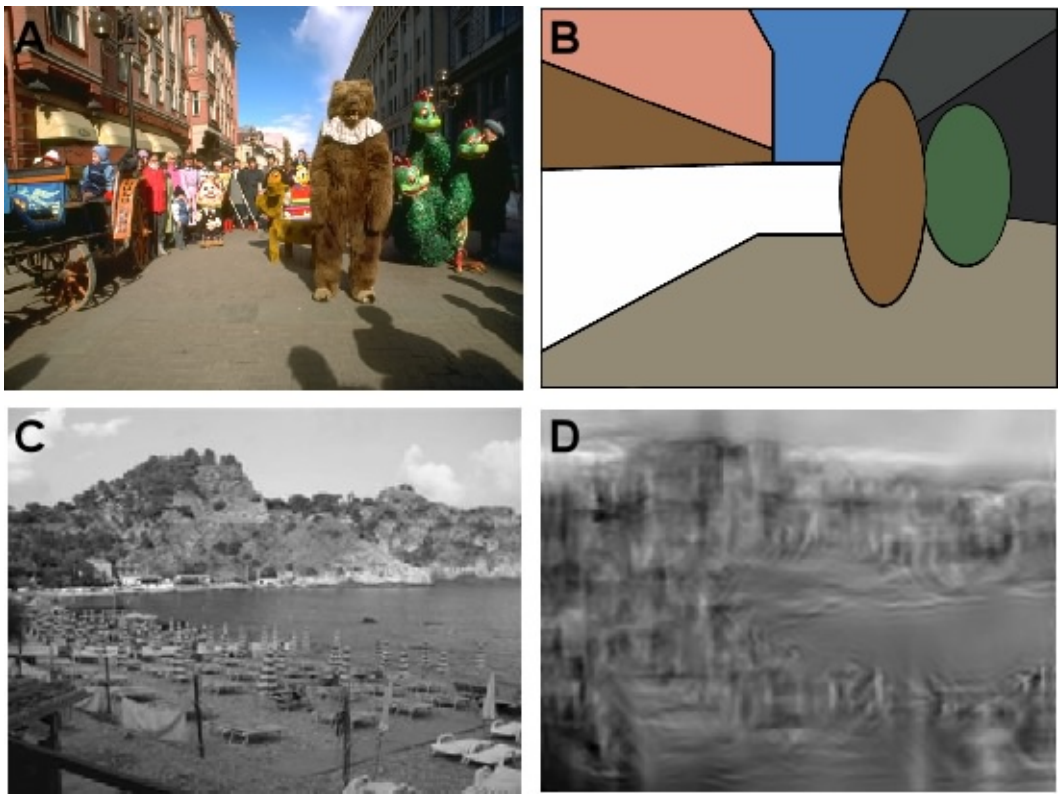

Figure n°18 : Tandis que le *gist* conceptuel de la scène A pourrait correspondre à une scène de carnaval, son *gist* perceptuel activerait la représentation scène de rue. La structure spatiale en partie porteuse du *gist* perceptuel pourrait dépendre des fréquences spatiales et des textures des images C et D tirées de (Oliva, 2005).

La structure spatiale d'une scène

Au sein d'une scène naturelle, les informations physiques globales organisées spatialement sur l'ensemble de la scène, vont constituer la structure spatiale de la scène, en anglais « spatial layout ». Cette structure spatiale s'appuie sur les caractéristiques physiques de l'ensemble de la scène, aussi bien les objets que le contexte. Evidemment, ce sont les entités de surface importante qui vont majoritairement participer à la structure spatiale. La nature de l'information participant à la structure spatiale est cependant encore incertaine. Certaines études suggèrent la contribution de volumes en 3 dimensions organisés, les géons (Biederman, 1995), d'autres proposent un arrangement de formes de couleurs et contrastes spécifiques (Oliva & Schyns, 2000, Schyns & Oliva, 1994) pouvant dépendre des

informations physiques globales bas-niveau telles que les fréquences spatiales et les textures (Figure 18 B, Oliva, 2005, Oliva & Torralba, 2001, Schyns & Oliva, 1994 , Torralba & Oliva, 2002).

Cette structure spatiale constituerait un cadre de référence suffisant pour se repérer dans la scène et pour faciliter des tâches visuelles purement physiques telle qu'une estimation de la profondeur des objets (Sanocki & Epstein, 1997). En effet, utilisant un protocole d'amorçage (« priming »), Sanocki & Epstein ont présenté à des sujets des scènes comportant à chaque fois deux objets identiques côte à côte mais décalés en profondeur, la tâche des sujets étant de déterminer le plus rapidement possible quel objet (celui de gauche ou celui de droite) était le plus proche. La structure spatiale de l'amorce visuelle pouvait être soit identique, soit différente de celle de la cible, ou encore être absente. De plus, les scènes d'amorçage étaient plus ou moins complexes (dessin à la main ou photographie), certaines contenant une information sémantique et d'autres non. Les auteurs ont ainsi montré que les sujets répondaient plus rapidement à la tâche lorsque amorce et cible avaient la même structure spatiale indépendamment de la présence d'informations sémantiques. Ils démontrent ainsi que la structure spatiale a une utilité propre, au moins dans des tâches de décision purement perceptives. Il a également été démontré que mélanger les portions d'images constitutives d'une scène entraînait une baisse de performance dans le choix forcé du label de cette scène (Biederman, Rabinowitz, Glass & Stacy, 1974). Cependant, un tel réarrangement porte atteinte aussi bien à la structure spatiale qu'au *gist* conceptuel, il est donc peu évident de faire la part des choses en ce qui concerne la participation de chacun dans une telle tâche.

Dans le cadre des modèles suggérant une analyse globale de la scène précédant les analyses locales, cette structure spatiale des scènes pourrait constituer la première information physique intégrée exploitable par notre système perceptif. Selon l'hypothèse des schémas de scènes (« scene schema hypothesis » ; Antes, Penland & Metzger, 1981, Biederman, 1981, De Graef, De Troy & D'Ydewalle, 1992, Henderson, 1992), le *gist* perceptuel rapidement intégré pourrait être comparé avec les représentations de scènes stockées en mémoire afin de guider les traitements visuels descendants vers les informations visuelles à affiner. Le prochain chapitre décrivant différents modèles de reconnaissance de scènes naturelles, précisera comment le *gist* perceptuel d'une image est différemment construit selon les défenseurs d'une reconnaissance des scènes naturelles basée principalement sur une analyse globale préalable à l'analyse locale.

5.3. Les modèles de reconnaissance des scènes

5.3.1. La reconnaissance d'une scène en tant qu'ensemble d'objets

Un courant majeur décrit la reconnaissance visuelle globale des scènes comme une suite de traitements ascendants de plus en plus complexes permettant l'intégration progressive de traits physiques locaux en représentations perceptives puis sémantiques des différents objets constituant la scène (Barrow & Tenenbaum, 1981, Biederman, 1987, Marr, 1982). De l'identité des objets découlerait la reconnaissance de la scène dans son ensemble, son contexte absolu, son environnement. Selon le paradigme des *frames* défendu par Friedman, certaines *frames* (structures mnésiques) de scènes pourraient comporter des concepts d'objets indispensables à leur activation. Ainsi dès qu'un objet obligatoire serait détecté, la *frame* de la scène serait automatiquement activée (Friedman, 1979). La reconnaissance d'une scène naturelle reposerait donc sur l'identification sérielle réussie de quelques uns des objets spécifiques de la scène. De plus, dès qu'un objet serait reconnu, il activerait non seulement la *frame* de contexte à laquelle il appartient, les *frames* d'objets susceptibles de faire également partie de la scène mais aussi leurs orientations, leurs tailles, et leurs positions probables. La reconnaissance des autres objets de la scène serait ainsi facilitée. Allant dans le même sens, une étude menée par Antes en 1977 révèle une meilleure performance des sujets à reconnaître l'appartenance à une scène d'une portion grandement informative comparée à une portion peu informative. Etant donné que les régions informatives comportent la plupart du temps un objet, il en déduit que les objets sont reconnus avant la scène globale (Antes, 1977). Loftus confirme l'idée en démontrant que les régions les plus informatives sont traitées en premier (Loftus & Mackworth, 1978). Cependant, Metzger répliquera cette étude en insérant dans le design expérimental la reconnaissance de portions de scènes moyennement informatives qu'il estimait à son sens plus caractéristiques du contexte, et défendra l'idée d'une influence contextuelle (Metzger & Antes, 1983).

Une autre voie d'exploration privilégiant l'utilisation d'informations locales fut de centrer le débat sur les relations inter-objets. En 1987, Henderson pointa le fait que des objets sémantiquement reliés tendent à souvent apparaître dans la même scène. Sur la base de cette observation, il suggéra que l'effet contextuel révélé dans certaines études pourrait en fait être un effet d'amorçage entre objets de la scène. La présence d'un chien dans la scène pourrait ainsi faciliter la détection d'un chat dans la même scène (Henderson, Pollatsek & Rayner,

1987). De même, la nature même des relations spatiales entre les objets pourrait être caractéristique de l'identité de la scène (De Graef, Christiaens & d'Ydewalle, 1990). Enfin, l'ensemble des relations entre objets de la scène pourrait constituer une unité permettant d'extraire le sens global de la scène sans que l'identification de chacun des objets ne soit nécessaire.

Selon ces hypothèses, les objets seraient systématiquement catégorisés avant les scènes (Biederman, 1987, Riesenhuber & Poggio, 2000).

Mais de nombreux travaux basés sur la présentation brève de scènes naturelles suggèrent une alternative à ces deux théories. En effet, Biederman d'abord, puis Intraub, et Oliva et Schyns ont montré qu'il est possible de reconnaître la catégorie à laquelle appartient une scène alors que l'image est présentée de façon très brève, tout en atteignant une bonne précision avec des temps de réaction très faibles, (Biederman, 1972, Biederman, Mezzanotte & Rabinowitz, 1982, Intraub, 1997, Intraub, 1999, Oliva & Schyns, 1997, Oliva & Schyns, 2000, Potter, 1975, Schyns & Oliva, 1994) ce qui va à l'encontre des théories précédentes. Dans ce sens, l'hypothèse d'une reconnaissance globale de la scène basée sur des indices visuels qui lui serait propre est préférable.

5.3.2. La reconnaissance de scène sur la base de ses caractéristiques propres

La majorité des études récentes défendent désormais l'hypothèse que les traitements précoces de la scène seraient basés sur une analyse plutôt globale que locale (Antes et al., 1981, Loftus, Nelson & Kallman, 1983, Metzger & Antes, 1983, Schyns & Oliva, 1994) de caractéristiques physiques bas-niveau telles que des primitives 3D (Biederman, 1981, Biederman, 1995), des orientations locales (Guerin-Dugue & Oliva, 2000) ou encore des bandes de fréquences spatiales (Oliva & Schyns, 1997, Oliva & Torralba, 2001, Schyns & Oliva, 1994). Cependant, la majorité de ces études utilisent pour stimuli des dessins faits main, composés uniquement de contours noir sur blanc. Comme on l'a vu dans le premier paragraphe, il existe dans notre environnement naturel de nombreuses variables physiques à prendre en compte qui amenuisent grandement l'importance des contours d'objets et qui constituent une importante quantité d'informations supplémentaires à prendre en compte pouvant moduler l'importance des objets et du contexte. D'un autre côté, les quelques études déjà menées sur le contexte des scènes naturelles se sont peu, voire pas du tout intéressées au

décours temporel des traitements sous-jacents. Il est ainsi important de combler cette lacune en caractérisant les décours temporels des traitements de l'objet et du contexte au sein des scènes naturels, un des principaux objectifs des travaux de ce mémoire.

5.3.2.1 Hypothèse de Biederman : les géons des scènes naturelles

Biederman proposa un modèle de reconnaissance de scène ressemblant à celui des objets (Biederman, 1981, Biederman, 1995)… Dans son étude de 81, Biederman décrit les travaux de Robert Mezzanotte démontrant qu'une interprétation sémantique de la scène peut être construite sur la base de l'agencement de primitives 3D, des « géons », respectant la taille relative des objets et dépourvus de sens propre (Figure 19). Des primitives avec une échelle spatiale plus importante que celles servant à la reconnaissance d'objet pourraient ainsi représenter des informations visuelles spécifiques des scènes indépendamment des informations relatives aux objets. De manière intéressante, ces scènes schématiques seraient reconnues assez rapidement pour interférer sur l'identification d'objets intacts inappropriés à la scène.

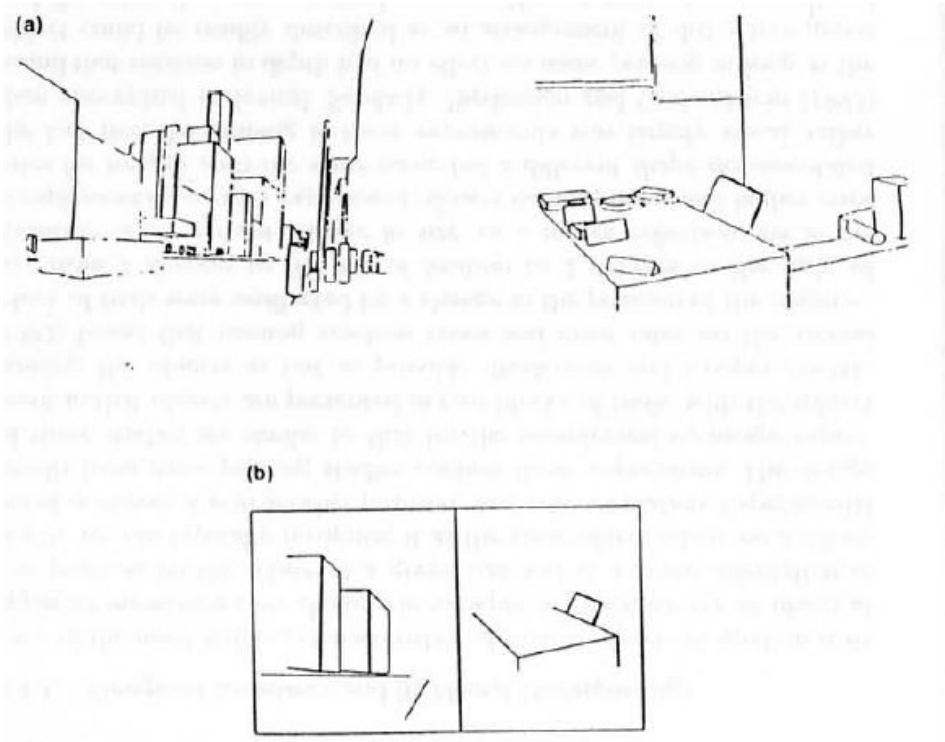

Figure n°19 : Tiré de Biederman, 1981 #255. Des primitives en 3 dimensions respectant la taille relative de l'objet et dépourvus de sens pourraient suffire à permettre une catégorisation de la scène. Dans la figure, les « géons » en (b) préciseraient grandement le contexte dans lequel on se trouve (a).

Cependant, les modèles par description structurale ne sont pas possibles biologiquement (Rolls, Stringer & Trappenberg, 2002) tant la quantité d'information à encoder et stocker serait importante. De plus, tandis qu'un objet tend à posséder un arrangement spatial bien défini des parties le composant, une scène au contraire est moins contrainte du point de vue des relations spatiales entre les objets qu'elle contient (Henderson, 1992, Henderson & Hollingworth, 1999, Hollingworth & Henderson, 1998). La configuration spatiale des scènes est davantage distribuée au sein du champ visuel et encodée différemment des objets. Sans compter le fait que des études en neuropsychologie montrent que les relations spatiales intra et inter-objets seraient représentées différemment, respectivement au sein des aires fronto-pariétales et occipito-pariétales (Humphreys & Riddoch, 1994, Humphreys & Riddoch, 1995) au niveau du système visuel. Finalement, rappelons qu'il a été proposé que l'identification d'objet et l'identification de scènes auraient lieu dans des aires corticales différentes (Epstein & Kanwisher, 1998) : pour les objets et les visages respectivement au sein du LOC et de la FFA, pour les scènes au sein de la PPA. .

5.3.2.2. Hypothèse d'une analyse « Coarse to fine »

L'hypothèse « Coarse to fine » s'ancre dans le courant de pensée d'une analyse grossière préalable aux traitements fins d'une image suffisant à la construction d'un *gist* perceptuel (Oliva & Schyns, 1997, Schyns & Oliva, 1994). Contrairement à certaines hypothèses, cette dernière bénéficie d'une argumentation importante aussi bien d'un point de vue expérimental que computationnel. Elle s'appuie sur la caractérisation des scènes naturelles sur le plan des fréquences spatiales et de l'espace de Fourier. En effet, tout signal en deux dimensions peut être décomposé en une somme de sinusoïdes d'amplitude et de phase différentes. En rajoutant une variable angulaire (orientation), il est alors possible de coder l'intégralité d'une scène naturelle selon un nombre théoriquement infini d'échelles spatiales. D'un point de vue pratique, le nombre d'échelles spatiales prises en compte est restreint car dépendant de la résolution d'affichage de l'image (sur un ordinateur, le nombre de pixels). Dans ce cadre de travail, il est donc possible après décomposition d'une image en fréquences spatiales de reconstruire une image 2D ne comportant que des fréquences spatiales en dessous d'une fréquence de coupure spécifique, on applique alors dans ce cas là un filtre passe-bas. On peut au contraire reconstruire une image 2D en gardant uniquement l'information présente dans les

fréquences spatiales supérieures à une fréquence de coupure spécifique, on parle alors de filtre passe-haut. On peut finalement ne garder que les fréquences spatiales comprises entre deux zones de coupure, on parle alors de filtre passe-bande. Cette façon de représenter le signal au sein d'une image n'est pas aussi artificielle que cela pourrait paraitre à première vue. En effet, des canaux visuels ascendant traiteraient séparément différentes bandes de fréquences spatiales (Campbell, Nachmias & Jukes, 1970, Campbell & Robson, 1968, Marr & Hildreth, 1980 , Shulman, Sullivan, Gish & Sakoda, 1986). Dans ce sens, les canaux de fréquences spatiales basses coderaient plutôt des surfaces étalées et floues de la scène (« blobs ») possiblement représentatives de la structure spatiale. De façon complémentaire, les canaux de fréquences spatiales hautes coderaient majoritairement les contours des objets (Biederman & Ju, 1988). On peut enfin imaginer que les textures seraient codées par des canaux spécifiques des bandes de fréquences spatiales moyennes. D'un point de vue anatomique, il est intéressant de faire le parallèle avec les systèmes magnocellulaires et parvocellulaires qui pourraient respectivement permettre l'intégration d'une représentation grossière (fréquences spatiales basses) et d'une représentation spatiale affinée d'une même entité. Afin d'évaluer l'importance des différentes bandes de fréquences spatiales dans la reconnaissance des scènes naturelles, Schyns et Oliva développèrent d'astucieux stimuli hybrides : des scènes recomposées à partir des fréquences spatiales basses d'une scène et des fréquences spatiales hautes d'une seconde scène (Figure 20, Oliva & Schyns, 1997, Schyns & Oliva, 1994).

Figure n°20 : Tiré de (Oliva, 2005). Deux exemples d'images hybrides complémentaires. Tandis que l'image de gauche est composée des hautes fréquences spatiales d'une scène de couloir couplées aux basses fréquences spatiales d'une scène de ville, l'image de droite est quant à elle composée des hautes fréquences spatiales d'une scène de ville couplées aux basses fréquences spatiales d'une scène de couloir. En « bridant » les yeux, vous pouvez plus aisément distinguer la scène représentée dans les basses fréquences.

Dans une première expérience (Schyns & Oliva, 1994), les auteurs impliquèrent leurs sujets dans une tâche de vérification. Chaque essai consistait en une première image affichée 30 ou 150 ms, suivie d'un masque, puis de l'image cible. Les sujets devaient alors faire un choix forcé et décider si la première image et l'image cible représentaient la même scène. Les premières images de l'essai étaient soient des images normales (N), soit des images passe-bas (LF), soit des images passe-haut (HF), soit des images hybrides. Les images hybrides pouvaient soit être composées des fréquences spatiales hautes de l'image cible couplées aux fréquences spatiales basses d'une autre scène (hybride HF), soit être composées des fréquences spatiales basses de l'image cible couplées aux fréquences spatiales basses d'une autre scène (hybride LF). Les résultats montrent que lorsque des images hybrides sont présentées, la décision des sujets s'appuie de façon privilégiée sur l'information contenue dans les fréquences spatiales basses lorsque le temps de présentation de l'image est de 30 ms. Au contraire, lorsque la durée d'affichage est de 150 ms, les sujets privilégient l'information contenue dans les fréquences spatiales hautes. Cette préférence n'est cependant pas liée à un problème de perception physique puisque les images HF et LF sont bien reconnues dans les deux conditions de présentation.

Afin de tester l'existence d'une stratégie préférée de traitement lors de la catégorisation de scènes naturelles, ils sélectionnèrent des images appartenant à 4 catégories différentes (« salon », « chambre », « ville », « périphérique »). Ils formèrent par la suite des couples d'images de catégories différentes sur la base desquels ils construisirent deux images hybrides. Les deux images de chaque couple furent alors présentées successivement à une vitesse de 45 ms par image hybride selon les deux chronologies possibles. Les traits physiques d'une scène apparaissaient ainsi soit selon un ordre *Coarse-to-fine* (CtF : l'image hybride LF de la scène suivie de l'image hybride HF) ou *Fine-to-Coarse* (FtC : la HF suivie de la LF). Deux réponses parmi les 4 possibles étaient donc avérées, puisqu'à chaque essai, les traits physiques de deux catégories différentes étaient mixés. Les résultats démontrent une préférence pour catégoriser les scènes naturelles dont les traits physiques apparaissent selon une chronologie CtF (67%). Cependant, 29% des réponses sont effectuées à partir des informations contenues dans les traits physiques FtC. Si une analyse *coarse-to-fine* des scènes est globalement privilégiées dans les tâches de catégorisation, elle n'exclue pas pour autant la possibilité d'analyse *fine-to-coarse*. Le choix de la stratégie pourrait dépendre de la diagnosticité des éléments au sein de la scène.

I.5. La reconnaissance d'une scène naturelle d'un seul regard…

Suite à ces deux premières expériences, les auteurs s'attachèrent à étudier la composante attentionnelle impliquée dans la sélection de l'échelle « optimale » la plus diagnostique. Entrainant les sujets à catégoriser des scènes contenant de l'information uniquement dans les fréquences spatiales basses ou uniquement dans les fréquences spatiales hautes, ils démontrèrent que lors du test composé d'images hybrides, les sujets catégorisaient uniquement les scènes en fonction du contenu des fréquences spatiales contenant une information durant l'entraînement. La quasi-totalité des sujets n'avaient d'ailleurs pas remarqué que les informations diagnostiques d'une autre scène étaient présentes dans les stimuli. A l'aide de deux expériences ultimes s'appuyant sur le paradigme d'amorçage et sur l'utilisation d'images hybrides, les auteurs démontrèrent que les informations contenues dans les fréquences spatiales non sélectionnées par l'attention étaient tout de même traitées sans pour autant permettre une reconnaissance de la scène.

Ainsi, selon la tâche à effectuer et les informations diagnostiques à extraire, le système visuel privilégierait l'extraction et l'intégration des informations contenues dans les fréquences spatiales basses ou hautes, respectivement des informations grossières ou fines. Pour une catégorisation superordonnée, les informations grossières pourraient être les plus diagnostiques (« scènes de rues par exemple »). Au contraire, pour une catégorisation subordonnée, par exemple (« Le Louvre »), les informations fines pourraient s'avérer plus diagnostiques. Ainsi, le choix de l'échelle spatiale de l'image à analyser serait contraint par la nature de la tâche (Schyns & Oliva, 1994) et sélectionné par une composante attentionnelle (Shulman & Wilson, 1987). La détermination inconsciente de l'échelle spatiale d'intérêt pourrait également dépendre de la prise de vue générale du set de stimuli utilisé. Nous pouvons en effet penser que les contours d'un objet en plan large et du même objet en gros plan ne sont pas discriminés de façon optimale au sein de la même échelle spatiale. Il n'existerait donc pas d'échelle spatiale optimale et préférée pour tout type de tâches. Dans ce sens, l'idée d'un traitement privilégiant le traitement grossier avant un traitement affiné *coarse-to-fine* (Oliva & Schyns, 1997 , Schyns & Oliva, 1994) constitue un bon compromis. Une description grossière de l'image serait construite avant d'en extraire des informations visuelles détaillées.

Comme le précise à juste titre Oliva et al. (1997), il est important de ne pas confondre traitements *coarse-to-fine* et traitements « global-to-local ». La figure 21 illustre parfaitement les différences entre ces deux types de traitements. Il est tout à fait possible d'imaginer une analyse globale de la scène à une échelle spatiale (ou résolution) importante, dans la figure, le

F formé de L. Il est de même possible d'effectuer un traitement visuel local à une échelle spatiale faible, les petits L formant le C. Cette idée de traitements globaux précédant les traitements locaux découle en grande partie des travaux de Navon (Navon, 1983), expérience 2). Affichant pendant 40 ms des images masquées contenant une grosse lettre composé de petites lettres identiques ou différentes, par exemple un grand « H » composé de petits « H » ou de petits « S », il impliqua ses sujets dans une tâche de reconnaissance des lettres globales ou locales. Navon démontra ainsi que la reconnaissance des lettres globales n'étaient pas influencée par l'identité des lettres locales tandis que les temps de réaction dans la tâche de reconnaissance des lettres locales augmentaient lorsque la lettre globale était différente des lettres locales. Différentes stratégies de traitements perceptifs pourraient être mises en place en fonction de la tâche à effectuer, privilégiant l'analyse locale ou globale indépendamment de la préférence de résolution grossière ou fine. La catégorisation de scènes naturelles à un niveau superordonné pourrait de fait s'appuyer sur une analyse globale et grossière de la scène. La résolution des traitements globaux pourraient s'affiner pour une catégorisation à un niveau inférieur. La recherche d'objets pourrait quant à elle s'appuyer sur des traitements globaux de la scène afin de guider ultérieurement des traitements plus locaux et plus discriminatoires.

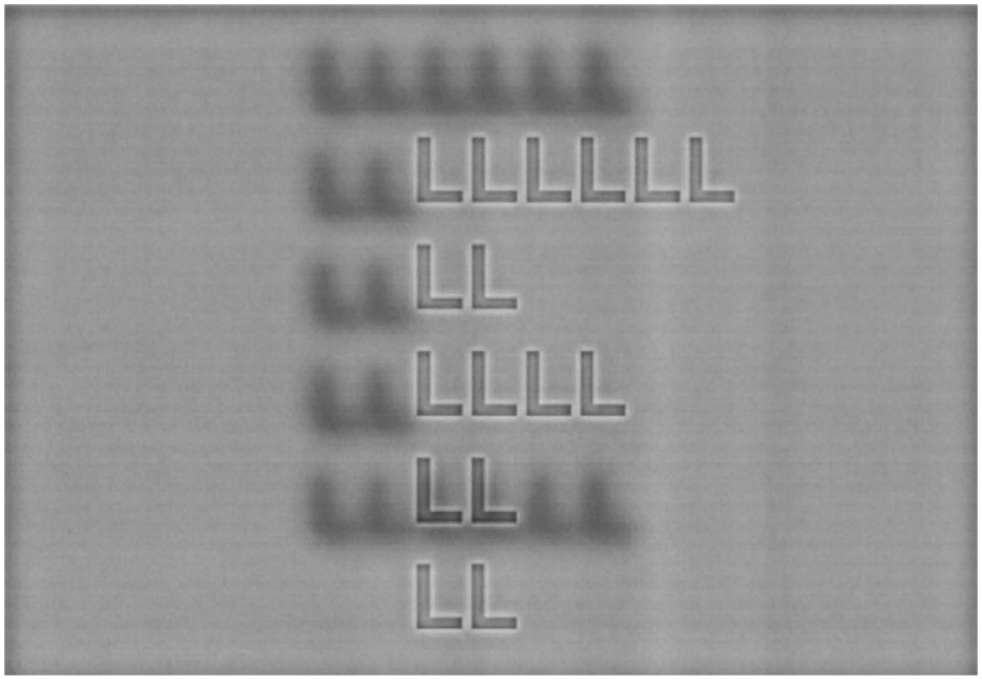

Figure n°21 : Tirée de (Oliva & Schyns, 1997). Un parfait exemple mettant en évidence le fait que « coarse » et « fine » n'équivalent pas à « global » et « local ». Les petits « L » flous composant le C sont par exemple des lettres « coarse » locales.

5.3.2.3. Modèle d'Oliva et Torralba

Les fréquences spatiales semblant porteuses d'une information suffisante pour extraire le *gist* perceptuel d'une scène, Oliva et Torralba ont proposé un modèle computationnel de reconnaissance des scènes naturelles s'appuyant sur le contenu spectral de l'image, représentable dans l'espace de Fourier (Oliva & Torralba, 2001). Transposé dans ce nouvel espace, les scènes pourraient tout comme les objets être considérées comme unitaires et disposeraient d'une forme propre. Différentes scènes appartenant à la même catégorie auraient alors des formes et des structures spatiales similaires (Torralba & Oliva, 2003). Si nous étions amenés à comparer deux photos de scènes différentes, l'une prise sur les plage de Tahiti et l'autre sur les plages de Normandie, nous pourrions constater qu'elles partagent de nombreux points communs. Certes, il y aura peut être plus de soleil et de palmiers sur la photo de Tahiti. Cependant, prises d'un angle de vue similaire, les photographies figureront toutes deux des scènes homogènes, peu complexes et dépourvues de textures manufacturées. Une surface de texture sableuse ou caillouteuse se trouvera en bas de l'image, attenante à une étendue de mer touchant la zone de ciel au niveau de l'horizon. Cette relation entre les contours, les surfaces et diverses autres propriétés bas-niveau de la scène constitue ce qu'Oliva et Torralba ont appelé l' « enveloppe spatiale » de la scène. Afin d'isoler les caractéristiques physiques de l'enveloppe spatiale porteuse de sens, les auteurs ont demandé à 17 observateurs de classer un ensemble de photographies en 8 groupes différents sur la base des traits physiques uniquement, excluant ainsi un classement s'appuyant sur la présence et l'identité d'objets. Après avoir demandé à chacun les critères qu'ils avaient utilisé pour faire cette classification, ils sélectionnèrent 5 propriétés de l'enveloppe spatiale semblant importantes dans la classification des images : en anglais : « Naturalness », « Openness », « Roughness », « Expansion » et « Ruggedness ». La clé du modèle réside dans la possibilité de dériver ces 5 propriétés de statistiques de second ordre des images : le spectre d'énergie, et le spectrogramme (spectre d'énergie localisé). Le spectre d'énergie est fonction de la contribution de chaque fréquence spatiale sur l'ensemble de l'image. Le spectrogramme est quant à lui fonction de la contribution des fréquences spatiales en 16 régions délimitées de la scène. Les 5 propriétés estimées de la scène permettent de projeter cette dernière dans un espace multidimensionnel à 5 dimensions au sein duquel les scènes de même catégorie sémantique se retrouvent à proximité les unes des autres.

Après test, ce modèle démontre de bonnes performances de catégorisation aux alentours de 80% pour la catégorisation de scènes naturelles à un niveau superordonné et à un niveau basique. Ses performances sont assez similaires voire supérieures à d'autres modèles de catégorisation de scènes s'appuyant en parallèle sur les attributs globaux et locaux de la scène (Vailaya, Jain & Zhang, 1998 , Vogel & Schiele, 2004). L'idée d'un traitement global de la scène précédant l'analyse locale des objets a d'ailleurs été intégré dans des modèles d'interactions contexte/objet, comme par exemple le modèle proposé par Moshe Bar (Bar, 2004, Bar, 2005) que nous décrirons dans le chapitre sur les interactions entre objet et contexte.

De nombreux autres modèles ont été développés afin de toujours mieux comprendre la nature des traitements rendant possible la reconnaissance des scènes naturelles par le système visuel. L'ensemble des articles de ce mémoire défende l'hypothèse d'une reconnaissance globale de la scène au moins aussi rapide que la reconnaissance des objets. Elle peut s'appuyer sur des informations qui lui sont propres, comme en témoigne la merveilleuse capacité de notre système visuel à reconnaître une toute nouvelle scène apparaissant brutalement dans notre champ visuel. Cependant, il reste encore beaucoup à découvrir, aussi bien d'un point de vue psychologique, physiologique ou computationnel. Le prochain chapitre a pour objectif de préciser les objectifs poursuivis tout au long de mes travaux et le choix d'un protocole adapté.

6. Spécificité de la catégorisation rapide dans un paradigme go/no-go

Lorsque l'on perçoit une scène naturelle, l'ensemble de l'information parvenant à notre rétine peut être divisé en deux types… Les informations relatives à l'objet et les informations relatives au contexte. S'il est avisé de penser que ces attributions ne sont pas fixes mais au contraire modulables, un objet pouvant devenir contexte, il est difficile d'imaginer une information qui ne participerait ni à la composition d'un objet, ni à la composition de son contexte… Il est tout aussi difficile d'imaginer que les informations du contexte et de l'objet soient complètement indépendantes… La taille d'un objet est relativisée par son contexte, la perception de sa luminance ou encore sa saillance tout autant... Même la sémantique d'une forme ambigüe est dépendante du contexte…

Pourtant, il existe un gros déséquilibre entre contexte et objet… L'objet d'une scène étant souvent centre d'intérêt, il semble qu'une grande partie du monde scientifique en ait oublié de s'intéresser au contexte jusqu'à récemment…

Ainsi les lacunes à combler dans le domaine sont importantes. Quelle est la nature et le décours temporel des traitements sous-tendant la reconnaissance du contexte ? Quelle est la nature de l'information extraite ? Quelle influence a le contexte sur la reconnaissance des objets d'intérêts ?

Afin d'étudier ces questions, j'ai inscrit mon travail dans le cadre de recherche de la catégorisation visuelle rapide des scènes naturelles tout en privilégiant l'utilisation d'un paradigme go/no-go. Quelle argumentation pour ce choix ?

6.1. Les scènes naturelles, une approche écologique…

Lorsqu'on veut comprendre comment le système visuel fonctionne, deux approches opposées sont envisageables concernant la nature des stimuli utilisés. La première approche consiste à choisir les stimuli les plus simples possibles nous permettant de répondre à la question spécifique que l'on se pose. L'avantage est un contrôle parfait des différentes variables physiques du stimulus et de ce fait l'absence de biais influençant les résultats. Il est ainsi possible de caractériser de manière optimale l'ensemble des dimensions du percept. Cependant, le désavantage est conséquent. On étudie une situation définie que le cerveau n'a jamais à gérer dans la vie quotidienne, ou alors de manière très simplifiée. On espère ainsi

partager un problème scientifique majeur, la densité d'informations dans notre champ visuel, en sous-questions plus faciles à aborder.

La seconde approche consiste à reproduire au mieux les conditions écologiques tout en gardant en tête les questions posées. Alors que la majorité des études ayant traité de la catégorisation d'objets et de contextes ont utilisé des dessins, peu ont traités de la catégorisation des scènes naturelles via l'utilisation de photographies. Les photographies de scènes naturelles ont pourtant l'avantage d'être des stimuli complexes et aboutis proposant une information visuelle de dimensions variées telle qu'on peut la trouver dans un environnement réel. Elles sont pourtant dépourvues de mouvements et de disparité 3D, mais restent un bon compromis pour préciser la nature des traitements catégoriels du contexte. Les photographies ont l'avantage de leurs inconvénients. Plus réelles, elles sont également plus complexes, et donc moins aisément paramétrables et contrôlables. La finesse consiste à multiplier le nombre et la diversité des photos dans chaque condition pour moyenner les informations d'intérêts tout en gommant les effets liés aux informations de bruit.

Le premier objectif principal de ce mémoire sera de préciser les informations physiques bas-niveau à la base de la reconnaissance du contexte en modulant le contenu spectral des photographies et en étudiant les effets de ces modulations sur les performances comportementales. (Article n°1).

6.2. Catégorisation rapide et paradigme go/no-go

La catégorisation rapide et le paradigme go/no-go sont deux caractéristiques de mes travaux marquant une volonté d'étudier des traitements visuels purement ascendants avec un minimum de boucle de contrôle au moment de la prise de décision. Pour mieux comprendre l'accent mis sur la rapidité de traitement, il est important de s'intéresser à la description du protocole.

Les différentes études décrites dans les articles de ce mémoire testent notre capacité à catégoriser des scènes naturelles au cours d'essais organisés de la façon suivante. Une croix de fixation (0,1° d'angle visuel) apparait au centre de l'écran pendant un laps de temps aléatoire de 300 à 900 ms afin d'éviter les anticipations. La croix est immédiatement suivie de la présentation d'un stimulus pendant 26 ms (soit la durée de deux rafraîchissements d'écran)

toujours au centre de l'écran. L'apparition du stimulus est assez brève pour empêcher l'exploration oculaire de la scène.

Si l'image appartient à la catégorie des distracteurs, le sujet doit garder le doigt appuyé sur le boîtier de réponse (réponse no-go). Si l'image appartient à la catégorie cible, le sujet doit relever le doigt le plus rapidement et le plus précisément possible, sa réponse étant détectée par des diodes infrarouges (réponse go). Passé le délai d'1 seconde, la réponse est considérée comme une réponse « no-go ». L'affichage du stimulus est suivi d'un écran noir pendant 300 ms avant l'apparition de la croix de fixation de l'essai suivant. Un essai dure donc entre 1600 et 2200 ms.

Comme on peut le voir, l'apparition brève du stimulus pendant 26 ms oblige à traiter la scène d'un seul regard. La seule information exploitable est une information extraite instantanément sans possibilité d'exploration oculaire, de vérification ou de guidage de l'attention. Une fois le système visuel pré-activé pour une recherche optimale des informations diagnostiques de la catégorie, les traitements visuels de catégorisation se limitent à une vague d'information à priori purement ascendante de part l'encouragement à répondre le plus rapidement possible et les temps de réaction minimaux enregistrés.

Le second objectif de ce mémoire sera d'estimer le décours temporel des traitements visuels de reconnaissance de scènes naturelles à différents niveaux de catégorisation superordonnés « Environnement Naturel » vs. « Environnement Manufacturé » et à différents niveaux de base : « Mer », « Montagne », « Scène de rue », « Intérieur urbain ». Pour les études concernant l'influence du contexte sur la reconnaissance des objets, le niveau de catégorisation basique d'objet utilisé sera « Animal » (Article n°2 et n° 3).

6.3. Des travaux de référence en catégorisation d'objet

Un autre avantage majeur dans le choix de ce paradigme pour étudier la reconnaissance du contexte absolu dans les scènes naturelles est lié au fait que de nombreux travaux sur la reconnaissance d'objets ont déjà été publiés avec le même paradigme.

Lors de travaux antérieurs menés au CerCo, la très grande rapidité du système visuel pour catégoriser différents types d'objets cibles tels que des animaux, des aliments ou encore des moyens de transport a été mise en évidence. Pour des images flashées seulement 20 ms, un sujet humain est capable de catégoriser des animaux dans 94% des cas, avec un temps de

réaction moyen de 400 ms. Les réponses les plus précoces surviennent dès 250 ms (Delorme et al., 2000, Fabre-Thorpe et al., 2001, Fabre-Thorpe et al., 1998, Thorpe et al., 1996). Ces travaux étendus à d'autres catégories d'objets tel les moyens de transport (VanRullen & Thorpe, 2001a) ou les visages animaux et humains (Rousselet et al., 2003) ont rapporté des résultats équivalents. Ces temps de réaction incluent à la fois le traitement de l'image ainsi que le temps nécessaire pour générer une réponse motrice. En analysant les enregistrements des potentiels évoqués pendant la tâche, on peut s'affranchir de la réponse motrice et situer le temps minimal pour catégoriser une image aux alentours de 150 ms.

Le troisième objectif de ce mémoire sera la réalisation d'une étude comparative des décours temporels des traitements relatifs à l'objet et au contexte dans des situations expérimentales grandement similaires.

6.4. Des interactions précoces entre objet et contexte

Au cours d'une étude à laquelle j'ai participé (Macé, Fabre Thorpe & Joubert, , Macé, Joubert & Fabre Thorpe, 2006), nous avons montré à l'aide du paradigme go/no-go décrit précédemment, que les sujets étaient capables de catégoriser les objets à un niveau basique en 450 ms environ. A un niveau superordonné, ces mêmes objets sont catégorisés avec des temps de réaction plus courts de 50 à 60 ms. Ce résultat est paradoxal puisqu'il est communément admis que c'est la catégorie de base qui est la plus facilement et rapidement accessible (Rosch et al., 1976). Nous verrons cependant que les études n°2 et 3 de ce mémoire reproduisent ces résultats pour les catégorisations de contexte et nous discuterons bien sur cet apparent paradoxe. Tandis qu'un environnement est catégorisé à un niveau basique en 450 ms, le même contexte à un niveau superordonné est catégorisé en moins de 400 ms. Les décours temporels des traitements catégoriels d'objet et de contexte à ces deux niveaux hiérarchiques semblent donc très similaires, et dans tous les cas se chevauchent.

Enfin, d'autres résultats obtenus dans l'équipe au CERCO, ont démontré que le système visuel était capable de traiter deux scènes en parallèle (Rousselet, Fabre-Thorpe & Thorpe, 2002). Cet ensemble de données suggèrent donc qu'objet et contexte pourraient être traités en parallèle selon des décours temporels similaires, et rend envisageable l'existence d'interactions entre traitements catégoriels d'objet et de contexte.

I.6.Spécificité de la catégorisation rapide dans un paradigme go/no-go

Le quatrième et ultime objectif de ce mémoire sera l'analyse des interactions entre objets et contexte au sein des scènes naturelles. Une première approche évaluera l'influence d'un objet sur la catégorisation du contexte à un niveau superordonné (article n°3). Une seconde approche testera et caractérisera l'influence du contexte sur la catégorisation des objets, également à un niveau superordonné (article n°4).

Ce mémoire pris dans sa globalité défend l'hypothèse d'interactions précoces et bi-directionnelles entre les traitements catégoriels de l'objet et du contexte au sein des scènes naturelles.

PARTIE II

CATEGORISATION VISUELLE DU CONTEXTE

«*Tout ce que nous voyons cache quelque chose d'autre*.»
René Magritte (Le faux mirroir)

1. Quelles informations physiques bas-niveau mises en jeu ?

1.1. Un cadre d'étude basé sur le spectre de Fourier.

Il a été démontré que notre système visuel pourrait se comporter tel un analyseur de Fourier (Marr, 1982, Westheimer, 2001). Puisque les informations physiques de premier ordre (luminance, contraste…) ne suffisent pas à diagnostiquer la catégorie des scènes naturelles, il est important de définir et comprendre la nature des informations d'ordres supérieurs présentes dans l'espace de Fourrier.

Comme nous l'avons vu en début de mémoire, toute image codée sous forme matricielle peut également être transcrite par une transformée de Fourrier dans un espace dit « Espace de Fourrier ». Cet espace se décompose en deux spectres. Le spectre d'amplitude code l'intensité des fréquences spatiales de l'image en fonction de leur orientation. Le spectre de phase code la phase des fréquences spatiales en fonction de leur orientation, c'est-à-dire la position relative de chaque sinusoïde spatiale au sein de l'image.

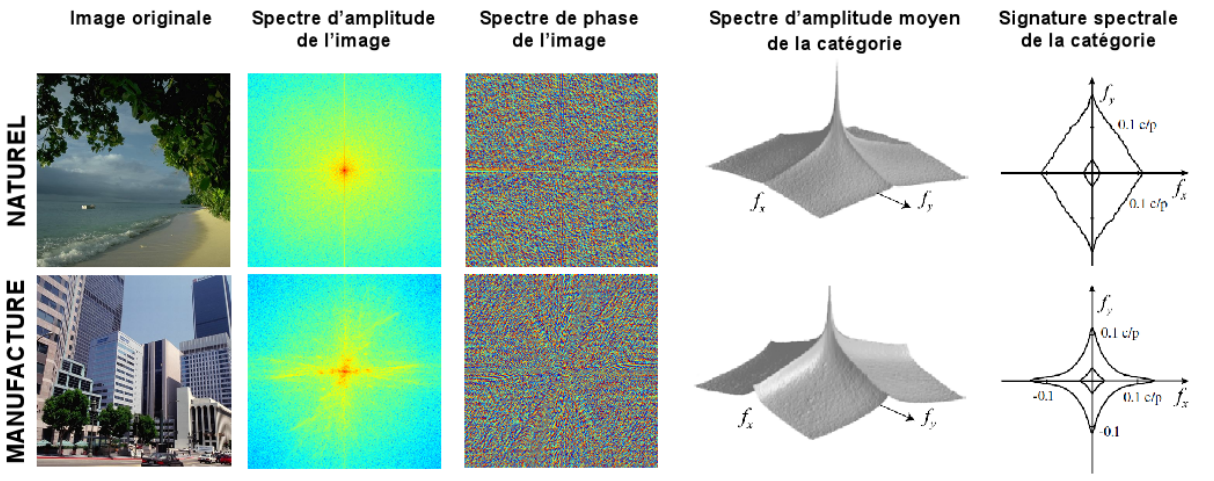

Figure n°22 : Pour chaque catégorie de contexte superordonnée : « Env. Nat » (en haut) et « Env. Man » (en bas), de gauche à droite : un exemple de photographie originale, son spectre d'amplitude, son spectre de phase, le spectre d'amplitude moyen de la catégorie et la signature spectre de la catégorie.

Au sein du spectre d'amplitude, sur l'axe f_x horizontal sont représentés du centre aux extrêmes, les contours verticaux participant à un continuum des fréquences spatiales basses

aux fréquences spatiales hautes. De la même manière, sur l'axe f_y vertical sont représentés les contours horizontaux selon un même continuum (Figure 22).

Une première constatation concernant le spectre d'amplitude des scènes naturelles fut une plus grande intensité des fréquences spatiales basses comparées aux fréquences spatiales hautes. Selon Field, la forme du spectre d'amplitude, l'intensité des fréquences spatiales suivrait une loi en 1/f (Field, 1987), plus tard affinée en $1/f^{\alpha}$, α variant selon les images utilisées et selon leurs catégories. Enfin, cette loi peut être davantage affinée en prenant en compte le facteur orientation (Torralba & Oliva, 2003). Il faut cependant garder en tête que si cette loi semble caractéristique des scènes naturelles en général, et peut être précisée pour des catégories de scènes particulières, elle n'est pas pour autant forcément avérée pour des exemplaires uniques de scènes.

Une seconde constatation importante fut la prépondérance des horizontales et verticales sur les obliques. Quand on regarde le spectre d'amplitude de chaque image, même si la grande croix centrale est un artefact induit par les bordures de l'image, les axes verticaux et horizontaux du spectre ressortent d'avantage que les obliques. Il en est de même quand on regarde le spectre d'amplitude moyen d'une catégorie ou encore sa signature spectrale. Cela illustre une généralité physique des images : les contours horizontaux et verticaux ont tendance à être plus présents que les contours obliques au sein des scènes naturelles (Baddeley, 1997, Oliva & Torralba, 2001 , van Hateren & van der Schaaf, 1998). Cela est d'autant plus intéressant qu'à un niveau physiologique, le nombre de cellules du cortex primaire codant pour les horizontales et les verticales serait aussi important que le nombre de cellules codant pour des obliques (De Valois & De Valois, 1988). De plus, à même champ de profondeur, les spectres des environnements naturels (Env. Nat.) ont tendance à être davantage isotropiques que les environnements manufacturés (Env. Man., Torralba & Oliva, 2003). Cette différence est une nouvelle fois due à une répartition différente des fréquences spatiales sur l'ensemble des orientations.

Enfin, à chaque catégorie de scènes correspondrait une signature spectrale assez stéréotypique (Figure 23).

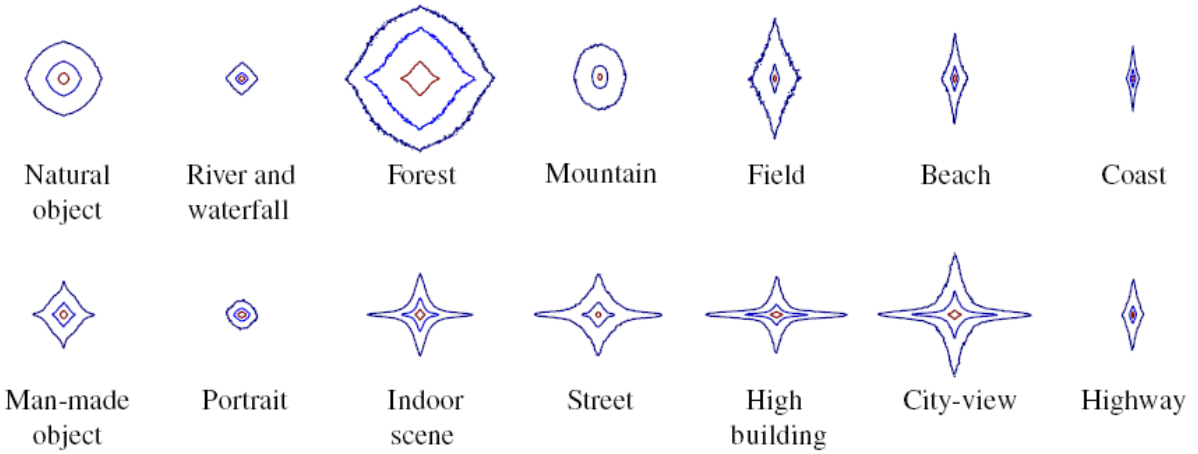

Figure n°23 : Tiré de (Torralba & Oliva, 2003). Signature spectrale de diverses catégories d'environnements contextes et d'objets.

Ces différences spectrales peuvent-elles alors suffire à effectuer une catégorisation des scènes naturelles ?

Si l'on observe 3 reconstructions d'une même image (Figure 24) : l'image originale (A), une reconstruction à partir de son spectre d'amplitude et d'un spectre de phase aléatoire (B), et une reconstruction à partir de son spectre de phase et d'un spectre d'amplitude aléatoire (C), il semble d'un seul coup d'œil que l'information porteuse de sens se situe dans le spectre de phase, et que l'on soit incapable de reconnaître une scène à partir de l'information contenue dans le spectre d'amplitude uniquement.

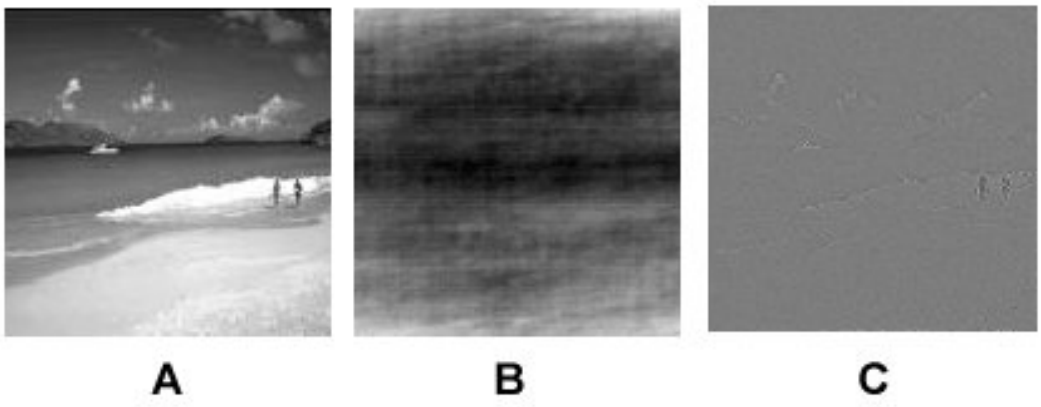

Figure n°24 : Tiré de (Guyader, Chauvin, Peyrin, Herault & Marendaz, 2004). Images amorces pouvant contenir (A) l'information de l'image en intégralité, (B) uniquement l'information d'amplitude de l'image, ou (C) uniquement son information de phase

A ma connaissance, deux études seulement ont suggéré notre capacité à utiliser les informations du spectre d'amplitude dans une tâche de catégorisation. Dans une première

étude (Guyader et al., 2004), les sujets impliqués dans une tâche de catégorisation go/no-go « scènes de ville » / « scènes de plages », percevaient une image amorce pouvant contenir (1) l'information de l'image en intégralité, (2) uniquement l'information d'amplitude de l'image, ou (3) uniquement son information de phase (Figure 24). De plus, les amorces pouvaient être de la même catégorie que la cible, ou de la catégorie opposée. La précision des sujets étant très bonne, les auteurs ont analysé uniquement les temps de réaction dans chaque condition et ont démontré que la facilitation liée à l'amorçage consistant reposait uniquement sur l'information contenue dans le spectre d'amplitude. Il est important cependant de noter que les auteurs avaient choisi deux catégories de scènes au spectre d'amplitude moyen très différent, les plages étant majoritairement composées de contours horizontaux et les villes de contours verticaux. Si cette étude donne un premier argument en faveur de la nature catégorielle de l'information d'amplitude, elle est loin d'être suffisante et nécessiterait d'être répliquée avec un plus grand nombre de catégorie.

Dans une seconde étude (Kaping, Tzvetanov & Treue, 2007), les auteurs ont montré que le système visuel humain impliqué dans une tâche de catégorisation d'Env. Man. vs. Env. Nat à l'endroit ou à l'envers, était biaisé par des images d'adaptation simulant les propriétés physiques bas-niveau non-localisées des catégories. Dans le cadre de travail de l'espace de Fourier, ces résultats suggèrent l'implication des informations du spectre d'amplitude dans la catégorisation des scènes, au moins à un niveau superordonné.

D'un point de vue computationnel, d'autres évidences sur le contenu catégoriel du spectre d'amplitude sont apportées. Les catégories des scènes peuvent être extraites des informations contenues dans le spectre d'amplitude (Guerin-Dugue & Oliva, 2000 , Guyader et al., 2004, Herault, Oliva & Guerin-Dugue, 1997, Oliva & Torralba, 2001, Schyns & Oliva, 1994, Torralba & Oliva, 2003). Ces informations pourraient d'ailleurs être utilisées comme support de l'exploration oculaire (Gautrais & Thorpe, 1998).

Néanmoins, la majorité de ces modèles ne se base pas uniquement sur l'information du spectre dans sa globalité mais s'appuie sur des analyses plus locales. Cette approche mérite d'ailleurs d'être mise en parallèle avec une caractéristique des scènes : la catégorie de la scène ne varie pas avec sa résolution ou avec la distance de la prise de vue. On accumule de ce fait plus d'information sur la catégorie d'une scène en combinant à l'analyse globale des analyses plus locales. Un autre parallèle réside dans les travaux menés par Morgan démontrant que l'importance supérieure de l'information de phase sur l'information d'amplitude diminue

d'autant plus qu'on multiplie les régions locales d'analyse pour devenir moins informative que l'information d'amplitude (Morgan, Ross & Hayes, 1991).

A l'opposé, d'autres études défendent l'importance d'une information localisée (phase) pour effectuer convenablement une catégorisation des scènes naturelles. Dans deux études menées par Loschky, les sujets devaient effectuer un choix forcé par la mise en correspondance de la catégorie d'une scène naturelle reconstruite avec un label de catégorie apparaissant en fin d'essai. Les auteurs démontrent ainsi que les scènes naturelles reconstruites à partir de l'intégralité de leur information d'amplitude et de moins de 50 % de leurs informations de phase originale étaient globalement mal catégorisées (Loschky & Larson, 2008, Loschky, Sethi, Simons, Pydimarri, Ochs & Corbeille, 2007). Elles démontrent de plus l'importance des informations non-localisées aussi bien pour une catégorisation basique que pour une catégorisation superordonnée. Ces résultats contradictoires avec les études précédentes pourraient être liées à l'utilisation d'un plus grand nombre de catégories environnementales rendant la tâche plus complexe, mais aussi plus écologique. Les informations non-localisées auraient ainsi un simple rôle facilitateur. Finalement, les auteurs constatent un biais des sujets à répondre Env. Nat. Pouvant sous-tendre l'utilisation des contours locaux comme indices diagnostiques, indices progressivement détruits avec l'augmentation du bruit dans la phase (Wichmann, Braun & Gegenfurtner, 2006). A noter que l'importance d'une information localisée a également été démontré dans une tâche de détection d'animaux (Wichmann et al., 2006).

Si les travaux de Loschky et ses collaborateurs démontrent l'importance prépondérante des informations de phase dans la catégorisation d'environnements à différents niveaux, ils ne posent cependant pas directement la question de l'influence des informations d'amplitude dans la tâche. En effet, les stimuli utilisés dans leurs expériences avaient conservé leurs informations d'amplitude dans leur intégralité, informations pouvant interagir avec les informations de phase pour une meilleure catégorisation. De plus, aucune analyse des décours temporels des traitements perceptifs n'a pu être effectuée de part le mode de réponse choisi, les sujets devant répondre si l'image cible correspondait ou non à un nom de catégorie présenté après l'image.

1.2. Influences des informations de phase et d'amplitude dans la catégorisation rapide du contexte à un niveau superordonné.

1.2.2. Objectifs et protocoles

Cet article, actuellement en révision dans « Journal of Vision », s'inscrit directement dans ce cadre de recherche. Le principal objectif de cette étude était d'évaluer jusqu'à quel point les traitements catégoriels des scènes naturelles à un niveau superordonné pouvaient reposer sur les statistiques bas-niveau de la scène, et plus particulièrement sur les informations contenues dans le spectre de phase. L'accent a été mis sur l'analyse des temps de réaction dont on peut déduire les décours temporels des traitements impliqués et sur le contrôle des biais potentiels, ce qui n'avait pas été fait dans les études décrites précédemment.

Les deux expériences présentées dans cet article consistent en des tâches de catégorisation go/no-go au cours desquelles les sujets devaient relever le doigt le plus vite et le plus précisément possible à chaque fois qu'une image cible apparaissait, Env. Man ou Env. Nat selon la tâche. La nature des réponses et les temps de réaction étaient enregistrés pour chaque essai.

1.2.2. Les informations d'amplitude facilitent la catégorisation des scènes naturelles

Dans la première expérience, 3 types de stimuli par catégories ont été utilisés : (1) des scènes naturelles achromatiques originales (O), (2) des scènes naturelles achromatiques dont la luminance et le contraste global avaient été égalisés à travers l'ensemble des scènes des deux catégories (EL), et (3) des scènes achromatiques dont la luminance, le contraste et le spectre d'amplitude avaient été égalisés à travers l'ensemble des scènes des deux catégories (ELA). Cette première expérience visait à estimer les conséquences d'une suppression de la diagnosticité des informations d'amplitude sur les performances tout en contrôlant les biais bas-niveau de premier ordre.

II.1. Quelles informations physiques bas-niveau mises en jeu ?

Résultats

Les sujets humains s'avèrent très efficaces à effectuer une catégorisation de contextes achromatiques à un niveau superordonné puisqu'ils atteignent en moyenne 96% de bonnes réponses avec un temps de réaction médian de 410 ms environ, que les informations de luminance soient égalisées ou non. Lorsque les stimuli sont dépourvus d'informations diagnostiques d'amplitude, les performances des sujets chutent d'environ 6% et leurs temps de réaction augmentent d'environ 20 ms. En outre, l'analyse du décours temporel via le calcul du d' au cours du temps révèle que ce délai de traitement peut aller jusqu'à 40 ms pour les réponses les plus précoces.

Ainsi la diagnosticité des informations d'amplitude ne serait pas essentielle à la catégorisation des scènes naturelles, mais elles faciliterait et accélérerait leurs traitements.

1.2.3. Diagnosticité supérieure des informations de phase « Env. Man »

Dans la seconde et principale expérience, nous avons utilisé les stimuli ELA de la première expérience pour concevoir de nouveaux stimuli au sein desquels l'information de phase avait été bruitée de 0 à 100% par pas de 11% afin d'obtenir 10 conditions de phase par image et par catégorie. Nous avons ainsi pu tracer les courbes psychophysiques relatives, ainsi que l'évolution du d' au cours du temps pour chaque condition de phase alors qu'aucune information diagnostique d'amplitude n'était disponible. Cette expérience fut répliquée dans un protocole de choix forcé Env. Man vs. Env. Nat afin de s'assurer que le biais des réponses enregistrées n'étaient pas liés au mode de réponse.

Finalement, nous avons profité de ce protocole expérimental bien spécifique au cours duquel les sujets effectuent deux tâches croisées pour calculer une nouvelle mesure exploratoire, que nous avons choisi d'appeler DT-value (D-T pour Distractor-Target).

II.1. Quelles informations physiques bas-niveau mises en jeu ?

Résultats

La seconde expérience testant la capacité des sujets à catégoriser les scènes uniquement sur la base des informations diagnostiques de phase nous a quant à elle permis de dresser des courbes psychophysiques s'apparentant à des sigmoïdes avec un point d'inflexion aux alentours de 50% d'informations de phase résiduelle. Alors que leurs performances se situent au dessus de 80% de réussite tant que les images contiennent plus de 50 % de la phase originale, leurs performances chutent brutalement au niveau de la chance lorsque les informations de phase sont davantage bruitées. De plus, alors que les Env. Man sont mieux catégorisés que les Env. Nat. lorsque la plus de 50% des informations de la phase originale est présente, ils semblent également les plus sensibles au bruitage de phase lorsque celui-ci dépasse 50%, laissant apparaître un biais vers les réponses « Env. Man ». Cette différence entre les deux catégories a été répliquée avec un protocole de choix forcé et n'est donc pas liée au mode de réponse mais aux traitements perceptifs sous-tendant la catégorisation
L'altération de moins de 50% des informations de phase n'a que très peu de conséquences sur les performances de catégorisation. Au delà, les informations diagnostiques des Env. Man. semblent plus touchées que celles des Env. Nat.

Si aucune différence significative entre conditions de phase n'est à noter concernant les temps de réactions, l'analyse des distributions des réponses au cours du temps ainsi que les courbes d' suggèrent une sensibilité plus précoce des traitements catégoriels du contexte aux informations d'Env. Man qu'aux informations Env. Nat. De plus, l'analyse de la DT value laisse présager que dans les conditions de phase fortement bruitée, nous baserions notre décision sur la présence ou l'absence des informations diagnostiques des Env. Man.

Si des informations diagnostiques des Env. Man. sont immédiatement disponibles, nous privilégions la réponse Env. Man, dans le cas contraire, nous répondons plus tardivement et comme par défaut Env. Nat que l'image présentée soit un « Env. Man » fortement bruitée un « Env. Nat. ».

1.2.4. Précision sur la D-T Value

La mesure du d' calculée dans de nombreuses expériences psychophysiques permet d'évaluer les performances des sujets, et donc la quantité d'information diagnostique extraite des stimuli indépendamment des stratégies individuelle liées à toute prise de décision incertaine. Le d' correspond à la distance perceptive interne permettant de discriminer le signal du bruit, dans notre expérience le signal perceptif de la catégorie cible du signal perceptif de la catégorie distracteur. On le calcule selon l'équation $d' = z(pHits_{task}) - z(pFA_{task})$ où z représente la fonction inverse de la distribution normale et $pHits_{task\ et}\ pFA_{task}$ respectivement la probabilité de réponses go correctes et de fausses alarmes dans une même tâche. En calculant son évolution au cours du temps, on peut ainsi avoir une estimation de la quantité moyenne d'information diagnostique accumulée au cours du temps depuis l'apparition du stimulus jusqu'à la prise de décision. En fonction du critère de décision choisi, privilégiant plutôt la précision ou la vitesse d'exécution, on fera plus ou moins de fausses alarmes et on manquera plus ou moins de cibles.

Néanmoins, s'il est facile d'apparenter cette mesure lorsqu'on parle de capteur électronique tel qu'un radar, il semble moins évident de comprendre exactement à quoi cette mesure correspond au sein du cerveau humain. En catégorisation, on suppose qu'elle pourrait représenter la distance diagnostique entre les deux catégories à discriminer. Elle prendrait donc en compte aussi bien les traitements perceptifs que les traitements décisionnels.

Au cours de notre expérience, les sujets réalisaient deux tâches de catégorisation croisées Env. Nat. et Env. Man sur le même ensemble d'image. Les mêmes stimuli étaient donc cibles dans une tâche de catégorisation et distracteurs dans l'autre, protocole ayant déjà permis en analyse de potentiels évoqués la mesure d'une activité différentielle isolant les composantes décisionnelles (Thorpe et al., 1996, VanRullen & Thorpe, 2001a, VanRullen & Thorpe, 2001b). La mesure de la DT value s'inspire grandement de ce calcul d'activité différentielle décisionnelle et de la mesure du d'. Elle tend à isoler une distance dans un espace restreint à des mesures décisionnelles, s'affranchissant des composantes perceptives. Elle se calcule par l'équation $DT = z(pHits_{task\ 1}) - z(pFA_{task\ 2})$, voir figure 25. Pour illustration, dans notre expérience, elle permettrait d'évaluer une distance décisionnelle séparant les traitements décisionnels engendrés par les images Env. Nat. cibles dans la tâche de catégorisation Env. Nat des traitements décisionnels engendrés par les images Env. Nat. distractrices dans la tâche de catégorisation Env. Man. Cette mesure est encore à l'état

exploratoire, elle suppose une étape cognitive décisionnelle différente des étapes perceptives. Elle a permis de mettre en évidence l'utilisation précoce d'indices spécifiques des « Env. Man. ».

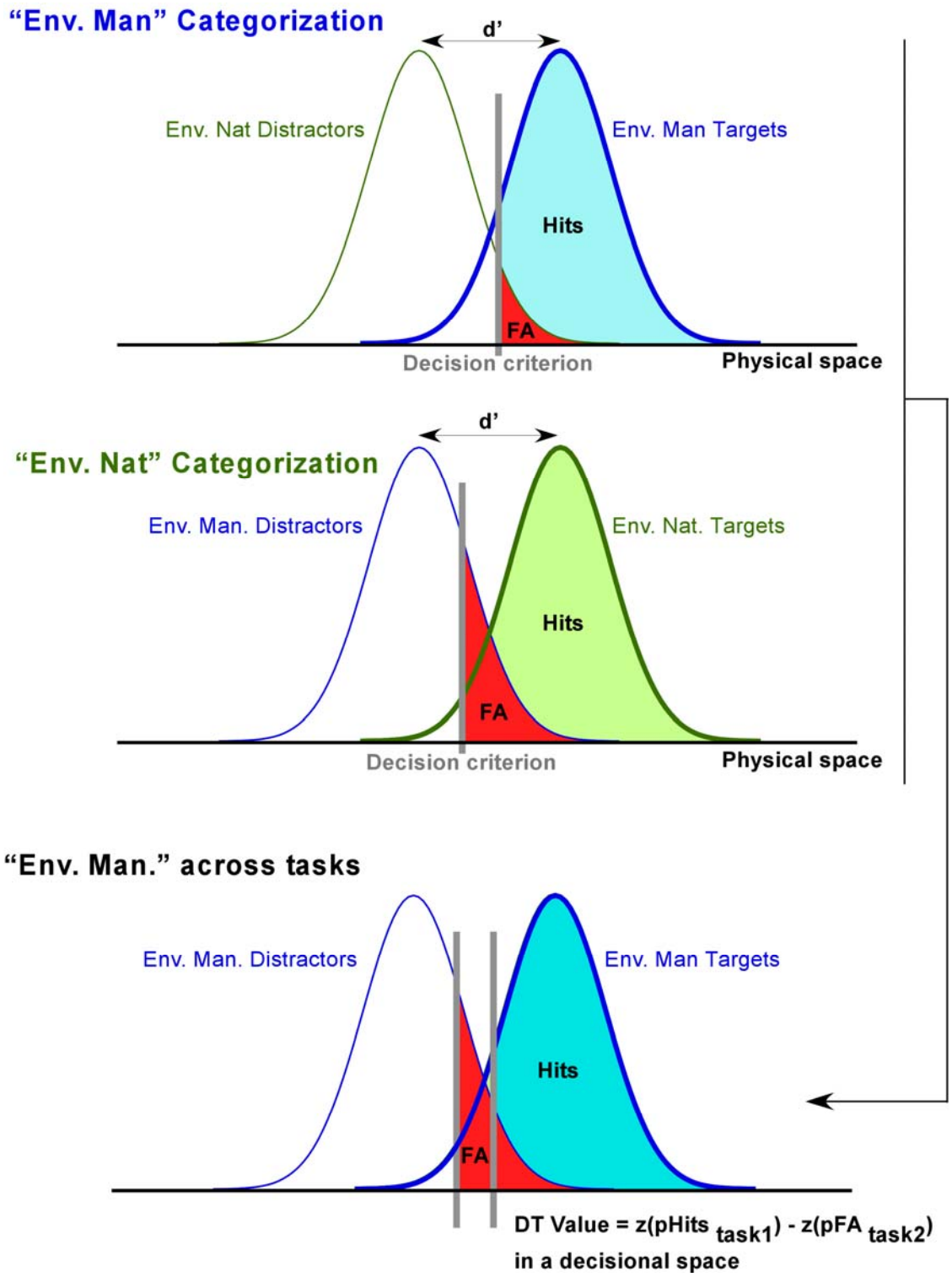

Figure n°25 : Illustration de la DT-Value.

Article n°1

Rapid visual categorization of natural scene contexts with equalized power spectrum and increasing phase noise

---

en revision dans Journal of Vision

Olivier R. Joubert, Guillaume A. Rousselet, Michèle Fabre-Thorpe & Denis Fize

Rapid visual categorization of natural scene contexts with equalized amplitude spectrum and increasing phase noise

Olivier R. Joubert [a,b], Guillaume A. Rousselet [c], Michèle Fabre-Thorpe [a,b] and Denis Fize [a,b,*]

a Université de Toulouse, CerCo, UPS

b CNRS, UMR 5549, Faculté de Médecine de Rangueil, Toulouse, France

c Centre for Cognitive Neuroimaging (CCNi), Department of Psychology, University of Glasgow, UK

* corresponding author

Centre de Recherche Cerveau et Cognition, CerCo, UMR 5549, Faculté de Médecine de Rangueil,

133 route de Narbonne 31062 Toulouse, cedex 9 France

Email : denis.fize@cerco.ups-tlse.fr

Abstract

This study aimed to determine the extent to which rapid visual context categorization relies on global scene statistics, such as diagnostic amplitude spectrum information. We measured performance in a Natural vs. Man-made context categorization task using a set of achromatic photographs of natural scenes equalized in average luminance, global contrast, and spectral energy. Results show that the visual system uses amplitude spectrum characteristics of the scenes to speed up context categorization processes. In a second experiment, we measured performance impairments with a parametric degradation of phase information applied to power spectrum averaged scenes. Results showed that performance accuracy was virtually unaffected up to 50% of phase blurring, but then rapidly fell to chance level following a sharp sigmoid curve. Response time analysis showed that subjects tended to make their fastest responses based on the presence of diagnostic man-made information; if no man-made characteristics enable to reach rapidly a decision threshold, because of a natural scene display or a high level of noise, the alternative decision for a natural response became increasingly favoured. This two phase strategy could maximize categorization performance if the diagnostic features of man-made environments tolerate higher levels of noise than natural features, as proposed recently.

Introduction

Our rapid understanding of complex visual scenes (Potter & Faulconer, 1975; Schyns & Oliva, 1994; Thorpe, Fize, & Marlot, 1996) suggests that a surprisingly large amount of information can be captured within a glance. Such rapidly extracted information leads to an image *gist* (Potter, 1976; Friedman, 1979; Oliva, 2005) that not only provides information about the spatial layout of the scene, but also information about a few objects and their surface characteristics (Rensink, 2000). This information is sufficient to assign a semantic label to the scene, its category or function (Biederman, Rabinowitz, Glass, & Stacy, 1974; Biederman, 1981; G.A. Rousselet, Joubert, & Fabre-Thorpe, 2005; Fei-Fei, Iyer, Koch, & Perona, 2007; Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007). Experimental evidence has shown that categorical information about scenes or objects can involve visual processes that do not require focused attention and are largely parallel (Li, VanRullen, Koch, & Perona, 2002; Rousselet, Fabre-Thorpe, & Thorpe, 2002; Rousselet, Thorpe, & Fabre-Thorpe, 2004; VanRullen, Reddy, & Koch, 2004; but see Evans & Treisman, 2005; Walker, Stafford & Davis, 2008). However, beyond the question of their possible feed-forward nature (Thorpe et al., 1996; Thorpe, Delorme, & VanRullen, 2001; Vogel, Schwaninger, Wallraven, & Bulthoff, 2006; Serre, Oliva, & Poggio, 2007; VanRullen, 2007), the kinds of information extracted by these fast visual processes are still unclear. To account for the speed of complex scene categorization, Oliva and Torralba proposed that diagnostic information was present at different image scales, including in the global statistical properties described by the Fourier spectral signature of the scene (Oliva & Torralba, 2001; Oliva, 2005). .

Following Marr (1982), the visual system could behave as a Fourier analyzer. The first steps of visual processing have often been described as filtering operations (Campbell & Robson, 1968; Field, 1999, Westheimer, 2001). In particular, some theories of sensory coding have suggested that one aim of the early visual system is to reduce redundancy in the coding of natural scenes, thus providing a more compact description of natural stimuli (Barlow, 1961). For illustration, given that the amplitude spectrum of natural scenes falls with increasing spatial frequency according to a 1/f rule, the compression of the visual information can be optimised by suppressing its high frequency content as can be seen in the visual system outside the fovea (Field 1999). Fourier analysis captures a large amount of image redundancy since the power spectrum of a picture (the square of its amplitude spectrum) provides a direct measure of its autocorrelation. The power spectrum of natural scenes can be modified, if the frequency fall-off is preserved, without seriously impairing our understanding of these scenes – probably because these modifications mainly affect redundant information. Therefore, averaging the power spectrum of a set of visual stimuli (while keeping the original phase information of each stimuli) is potentially a very good way to discard low-level biases among experimental conditions without necessarily critically deteriorating the informative content. This method strictly equalizes global

luminance and, rather importantly, averages the spatial frequency contents of all stimuli at each scale and orientation.

On the other hand, scene redundancies could have typical characteristics that could be of importance for the visual system, for example by supporting fast scene categorization processes. Oliva and Torralba proposed that the visual system could construct a meaningful representation of scene gist directly from such global Fourier based low-level features (Oliva & Torralba, 2001; Torralba & Oliva, 2003). Their results showed that complex scenes exhibit different Fourier spectral signatures that can be used by computational models to infer scene categories: for example, natural scenes and man-made environments were accurately distinguished using the distribution of orientation within the amplitude spectrum (Baddeley, 1997; Oliva & Torralba, 2001).

Recent experiments have provided evidence that the human visual system exploits these low-level image statistics for performing categorization. In particular, Kaping, Tzvetanov & Treue (2007) showed that the visual system adaptation to the amplitude spectrum of Natural or Man-made images biases rapid scene categorization toward the opposite category of the adapter. Similar results were obtained by Guyader, Chauvin, Peyrin, Herault & Marendaz (2004) for basic level categories. However, other studies that manipulated phase information have concluded that higher order statistics are needed to perceive scenes as natural because performance decreased following a sigmoid curve that characterizes categorical processes (Einhauser et al., 2006). Recently, Loschky and coll. obtained similar results using a Natural vs. Man-made scene categorization task in which stimuli were equalized for average luminance and global contrast (Loschky & Larson, 2008). But this sigmoidal performance was not reported in an Animal vs. Non-animal categorization task in which phase information was also systematically randomized (Wichmann, Braun, & Gegenfurtner, 2006). However, none of these studies provided a reaction time analysis, thus leaving open the issue of the early contribution of low-level spectral signatures to the speed of scene categorization.

In the present study, we aimed to determine to what extent rapid scene categorization depends on Fourier spectrum signatures, and how much performance speed would be affected when this information is no longer diagnostic. Using a Natural vs. Man-made task already used to investigate context categorization (Joubert et al., 2007), we measured categorization speed in several conditions. First, we used an original set of achromatic photographs of natural scenes to establish a benchmark level of performance. Second, the same photographs were equalized in luminance and contrast. Third, luminance and contrast equalized images were also equalized in spectral energy by applying the same averaged power spectrum to each image. Finally, we measured performance impairments when the images were further altered by mixing phase noise with their original phase spectra.

Experiment I. Amplitude spectrum equalization

The first experiment was designed to evaluate the impact of low-level image statistics equalization on context categorization performance. Setting the power spectrum of each image to the average across the entire image set eliminates amplitude spectrum diagnostic information that distinguishes man-made from natural scenes (Oliva & Torralba, 2001). We also characterized performance in an intermediate condition in which photographs were just equalized for global luminance and contrast.

Methods

Subjects and Tasks

Twelve volunteers (9 men, mean age 25, range 21-30, 4 of them left handed, normal or corrected to normal vision) performed 2 rapid visual go/no-go context categorization tasks both using man-made and natural scenes stimuli: one task used natural scenes as the target category, the other task used man-made scenes as target category. Subjects were asked to lift their finger as quickly and as accurately as possible (go responses) in response to a target picture, and to withhold their response (no-go responses) otherwise. Subjects gave their written informed consent; all of them had normal or corrected to normal vision.

The experiment consisted of two blocks, one for each of the natural and man-made categorization tasks. Each block was preceded by a training series of 96 trials. Each block was composed of 6 series of 96 trials containing an equal proportion of the three different stimulus conditions: original grey-level photographs (O), stimuli with luminance and contrast equalized (EL) and amplitude spectrum equalized stimuli (ELA). In each series, target and distractor trials were equally likely. The order in which tasks were performed (man-made and natural), the successive conditions tested (O, EL, ELA), and the stimuli used, were all counterbalanced among subjects. A given stimulus was seen by 4 subjects (twice as a target and twice as a distractor), but was never seen more than once (either as target or distractor) by a given subject.

Stimuli

The set of 1152 original scenes (O, see examples in Figure 1.O) was selected from a large commercial CD-ROM library (Corel Stock Photo Libraries) and included 576 man-made scenes and 576 natural scenes. Natural scenes were composed of sea scenes, mountain scenes, desert, iceberg, forest, and field scenes. Man-made scenes contained images of street scenes (with or without pedestrians), indoor scenes like kitchens, museums, churches, and aerial views of cities. None of the man-made scenes contained mountains or sea views, and none of the natural scenes contained buildings. Within each category, images were as diverse as possible. Images were grey-level coded, jpeg 8-bit format, with a size of 512 x 512 pixels (about 5 to 6° of visual angle).

Each of the 1152 stimuli of EL subset (Figure 1.EL) was set to the same global luminance and the same global (RMS) contrast, computed by taking the average luminance and RMS contrast of the 1152 original scenes (O).

Each of the 1152 stimuli of ELA subset (Figure 1.ELA) had the same amplitude spectrum, computed by averaging the power spectra across the 1152 EL stimuli. The ELA stimuli were constructed by inverse Fourier transforms using the square root of the averaged power spectrum and the original phase of each EL stimulus.

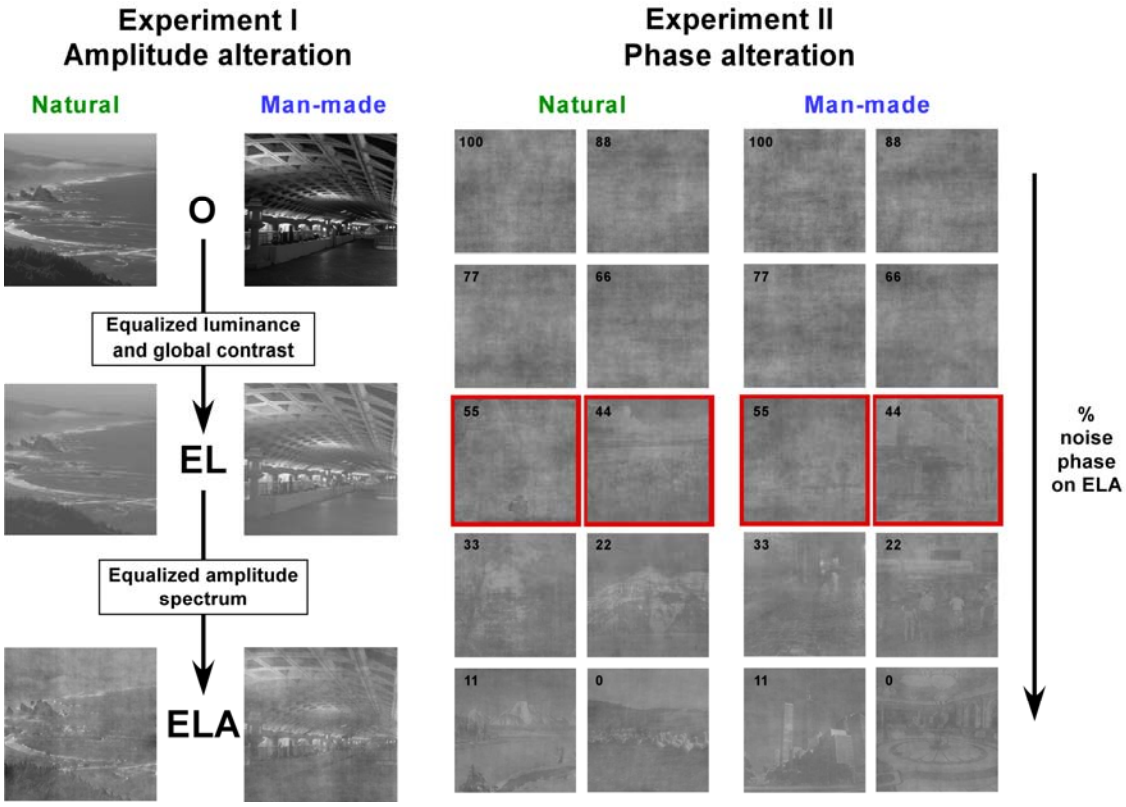

Figure 1: *Examples of stimuli used in experiments I and II.*
All stimuli were computed from gray-level pictures of man-made and natural scenes. In both experiments, subjects performed a go/no-go task in which man-made and natural scenes were alternatively targets or distractors. In experiment I, stimuli were divided in different subsets: original scenes (O), global contrast and luminance equalized scenes (EL), amplitude spectrum equalized scenes (ELA). In experiment II, stimuli were ELA scenes that included a variable percentage of phase noise, from 0 to 100 % with increments of 11%. In red the two conditions at which subjects' performance dropped to chance level.

Procedure

Subjects sat in a dimly lit room, 1 m away from a computer screen (resolution 1024 x 768, vertical refresh 75 Hz) connected to a PC computer. Stimulus display and behavioral response measurements were carried out using Presentation software (NeuroBehavioral Systems, http://nbs.neuro-bs.com/).

Each trial started with a fixation cross (1° of visual angle) displayed at the centre of a black screen for 300-900 ms randomly, immediately followed by the stimulus for two frames (26 ms), also in the middle of the screen. These brief presentations prevented exploratory eye movements and constrained the time available for information uptake. To start stimulus display, subjects had to place their fingers on a response pad equipped with infrared diodes that allowed timing with micro-second precision. For each target-image, subjects had to respond (finger lift) in less than 1000 ms; longer reaction times (RT) were considered as no-go responses. RT were computed between stimulus onset and finger lift. Following this 1 sec period, a black screen was displayed during 300 ms before the next trial started. A trial lasted between 1600 and 2200 ms.

Evaluation of performance

Performance was recorded in terms of accuracy and response speed. For each subject, task and condition, global accuracy, correct-go and false alarms rates, and median RT were computed. In order to evaluate statistical differences across subjects among the three conditions and the two tasks, we used the non-parametric two-way repeated measure Friedman test. When Friedman test showed statistical differences, a paired Wilcoxon test, Bonferonni corrected, was used to assess pairwise comparisons (in the text, only Wilcoxon p-values are indicated). Correct-go and false alarm responses were pooled for all subjects and expressed over time using 20 ms time bins in order to plot RT distributions. Minimal RT was determined as the first 10 ms bin for which correct responses significantly exceed errors using chi-square tests (target and distractor stimuli were equally likely). We used $d' = z_{hits} - z_{FA}$, where z is the inverse of the normal distribution function (Macmillan & Creelman, 2005), as a sensitivity measure to compare conditions and tasks. In order to plot d' as a function of time (d' curve), we computed the cumulative proportions of hits and false alarms using 20 ms bins. The d' curves describe performance time courses and provide estimates of the processing dynamics for the entire subject population (Rousselet et al., 2003, 2005).

Results

Results for man-made and natural categorization tasks are summarized in Figure 2 for the three conditions (O, EL, ELA). Categorization time-courses are illustrated by RT distributions and d' curves in Figure 3.

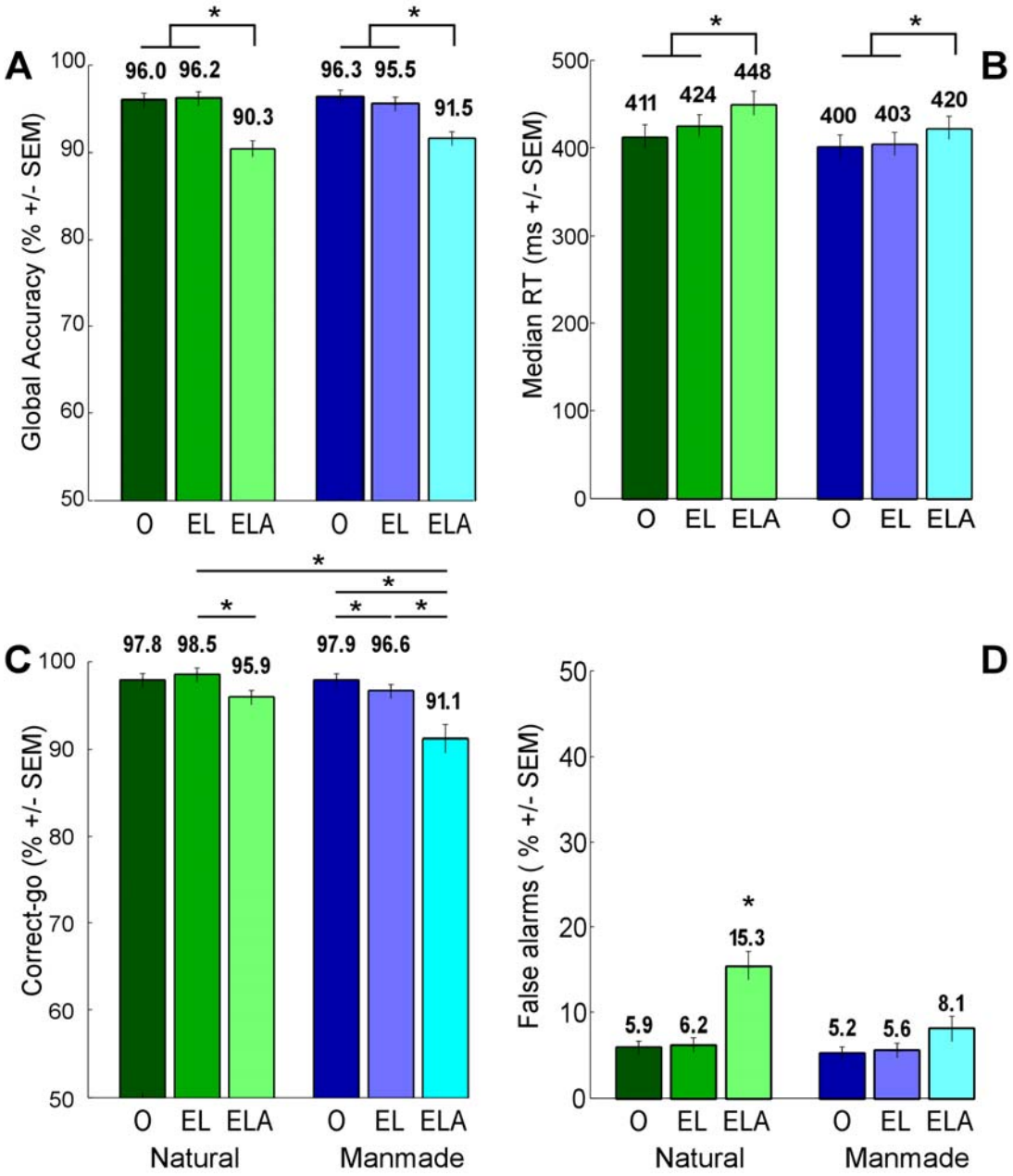

Figure 2: *Average performance in experiment I.*
For the three conditions O, EL, ELA described in figure 1, green bars indicate average subject performance in the natural categorization task (targets: natural scenes, distractors: man-made scenes), and blue bars indicate average performance in the man-made categorization task (targets: man-made scenes, distractors: natural scenes) in the three conditions O, EL, ELA described in figure 1. Error bars indicate the standard errors of the mean across subjects. **A** Average global accuracy for the 12 subjects. **B** Average median RT for correct-go responses (ms). **C** Average accuracy for correct-go responses (in %). **D** Average false alarm rate. Asterisks indicate statistically significant differences between conditions. Note that the ELA condition was associated with a systematic accuracy drop, longer reaction times in both categorization tasks. Moreover, in the ELA condition, subjects tend to bias their responses towards the natural category.

Accuracy

Subjects were very efficient at categorizing scene contexts, but equalizing amplitude spectra significantly impaired performance accuracy. With original scenes, subjects reached an averaged global accuracy of 96% and 96.3% (Wilcoxon test, n.s.) respectively for natural and man-made categorization tasks (Fig. 2A). Equalizing global luminance and contrast did not significantly modify global accuracy (natural task: 96.2%, man-made task: 95.5%, both Wilcoxon tests between O and EL: n.s.). For ELA condition, results showed a significant accuracy decrease in both tasks (5.9 and 4%, respectively; Wilcoxon test between EL and ELA for natural categorization task: $p = .003$, man-made task: $p = .004$). No significant difference was found between the two ELA conditions for global accuracy. The ELA manipulation affected differently hits and false alarms. In the ELA condition, hits (Fig. 2C) decreased by 1.9% and 2.6% relatively to the O and EL conditions in the natural categorization task, and by 6.8% and 5.5% respectively in the man-made task (for all comparisons, Wilcoxon tests: $p < .016$). False alarms (Fig. 2D) increased by 9.4% and 9.1% relatively to O and EL conditions in natural categorization task, whereas the increase was smaller, 2.9% and 2.5%, in the man-made categorization task. This higher error rate on distractors was only significant in the natural categorization task (ELA_{Nat} vs. EL_{Nat}Wilcoxon test: $p = .002$).

The effect of amplitude spectrum equalization was not identical in the two tasks. Subjects tended to make more correct go responses to natural targets (95.9%) than with man-made targets (91.1%, Wilcoxon test: $p = .007$). They also incorrectly categorized man-made distractors as natural scenes in the natural task more often than they categorized natural scenes as man-made opposite in the man-made task (15.3% vs. 8.1% of false alarms, Wilcoxon test: $p = .008$). Thus, when using stimuli of equal amplitude spectrum, there was a clear bias toward categorizing scenes as "natural".

Categorization processing time-course

Context categorization was very fast with similar median reaction times regardless of the target category: about 411 ms in the natural categorization task and 400 ms in the man-made task (Fig. 2B, Wilcoxon test, n.s.). Categorization time-courses were also very similar in both tasks as illustrated by the virtually superimposed RT distributions and d' curves (Figure 3A). Minimal RT in the O condition was 280 ms in both tasks. A Wilcoxon test revealed no significant difference between O and EL median RT in both tasks (EL, natural task: 424 ms, man-made task: 403 ms). EL and O correct-go RT distributions were almost superimposed in both tasks (Fig. 3B,C), but the minimal RT was slightly delayed (natural task: 310 ms, man-made task: 290 ms). This delay between O and EL conditions was thus probably due to a larger number of fast false alarms in the EL condition. A similar shift towards longer latencies was

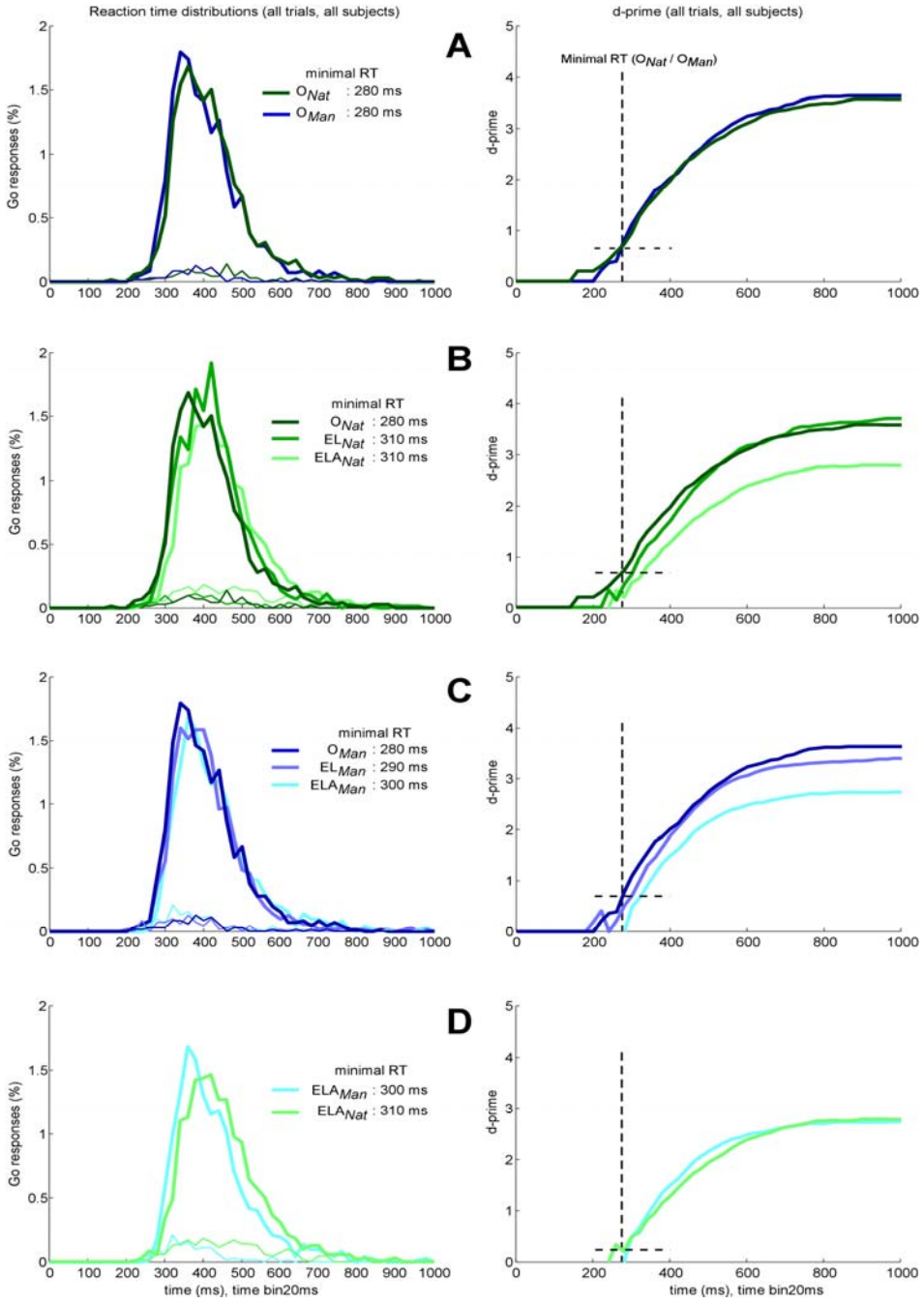

Figure 3: *Categorization time-course in experiment I.*
Left column: RT distributions for correct go-responses (thick curves) and false alarms (thin curves) expressed over time using 20 ms bins and RT pooled across all subjects. Minimal RTs were determined as the first 10 ms bin for which correct responses significantly exceeded errors (targets and distractors were equally likely). **Right column:** d' expressed over time using 20 ms cumulative bins. Dash lines indicate the d' reached at the minimal RT computed for the O condition (280ms) and is used as a reference. **A** Comparison of natural and man-made tasks using original scenes (O condition). **B** Comparison of O, EL and ELA conditions in the natural task. **C** Comparison of O, EL and ELA conditions in the man-made task. **D** Comparison of natural and man-made tasks using amplitude spectrum equalized scenes (ELA condition). With amplitude spectrum equalized stimuli, subjects remain able to perform context man-made and natural categorization tasks but a processing shift toward longer latencies can be observed in both tasks. Stimulus detectability at 280 ms is clearly lower for ELA condition than for O or EL conditions.

observed for the d' curves; the shift between conditions was defined as the time delay necessary to reach a threshold value. We defined that threshold as the d' level reached at the minimal RT of 280 ms in the O condition (Fig. 3B,C horizontal and vertical dashed lines). This delay was about 20-30 ms in the natural task and 10-20 ms in the man-made task. In summary, equalizing global luminance and contrast introduced a 10-30 ms processing speed delay in discriminating targets from distractors in both categorisation tasks, without further effects on median RT.

In contrast, equalizing amplitude spectra produced a more pronounced effect on response latencies and the delay was visible on median RT (natural task: 448 ms, man-made task: 420 ms, Fig 2B). This 24 ms and 17 ms delay in median RT between EL and ELA conditions reached significance (Wilcoxon tests, natural task: $p = .003$, man-made task: $p = .012$), and was significantly longer for the natural categorization task (Wilcoxon test: $p = .028$). Compared to O and EL conditions, the ELA d' curves were clearly delayed in both tasks (about 40ms at the threshold d' value defined as above, Fig. 3B,C horizontal and vertical dashed lines). This latency shift was also observed on the RT distributions for correct-go responses at early latencies. The effect was clearly visible for natural categorization task and less visible but nevertheless present in the man-made categorization task. To summarize, equalizing the amplitude spectrum of man-made and natural scene stimuli introduced a clear median RT increase of about 20 ms for both categorization tasks. This delay could reach 40 ms when assessed using the d' sensitivity measure for early response latencies.

As with accuracy, the effect of amplitude spectrum equalization on reaction times was not the same in the two tasks. We have already mentioned that the latency shift between EL and ELA correct-go RT distributions was less pronounced for man-made than natural targets at early latencies and the same effect was seen for median RT. Figure 3D directly compares the two tasks in the ELA condition. Whereas the two distributions were almost superimposed with O stimuli, a complete shift of the natural correct ELA go response distribution toward longer latencies is clearly visible when compared to the man-made condition correct ELA go responses (Fig. 3D). We also observed a more subtle difference in false alarm RT distributions: in the man-made categorization task, most of the false alarm responses tended to be produced before 450 ms, while false alarms in the natural categorization task were more equally distributed. Thus, there was a tendency to produce faster responses on man-made targets, with an earlier distribution of false alarms in the man-made task. Such time-course characteristics will be further discussed in terms of processing strategy, since they suggest that man-made information could accumulate faster than natural one, as observed in the second experiment.

Experiment II. Phase Alteration

Experiment II was designed to evaluate the additional effect of phase spectrum alteration on context categorization when other features such as global luminance, global contrast and amplitude spectrum were equalized across all stimuli. The higher order cues present in the phase spectrum were progressively blurred by adding phase noise. Behavioral performance was evaluated on both man-made and natural scene categorization tasks.

Method

Subjects and Task

Ten volunteers (6 men, mean age 27, range 24-31, 2 of them left handed, normal or corrected to normal vision) performed the same rapid visual go/no-go context categorization tasks described in experiment 1. This experiment was designed in two task blocks (natural categorization task, man-made categorization task), each composed of 7 series of 100 trials, 50 targets and 50 distractors in a random order. In each series, 10 conditions of phase alteration (Figure 1) were equally likely among target and distractor stimuli. Each task block was preceded by a training series of 90 trials. Each subject categorized each image in only one noise condition; across subjects, every image was seen in all noise conditions. Task order (natural, man-made) was counterbalanced among subjects.

Stimuli

In order to counterbalance efficiently all noise conditions across scenes and subjects, we completed the image set from Experiment I with 248 extra pictures selected to contained a high diversity of scenes as described earlier. A total of 1400 pictures from the Corel Stock Photo Libraries were used; they included 700 man-made scenes and 700 natural scenes. We Fourier transformed each grey-level picture and computed an average power spectrum. Then, we constructed every stimulus by inverse Fourier transform using the average power spectrum, and using a phase spectrum that contained a variable percentage of the phase of the original image and an inversely proportional percentage of the phase of a white noise image. Thus, from one original image, 10 stimuli were created that contained 0%, 11%, 22%, 33%, 44%, 55%, 66%, 77%, 88% or 100% of phase noise. Thus, the complete set of 14,000 stimuli had the same global luminance and the same amplitude spectrum. Images were encoded in jpeg 8-bit format, with a size of 512 x 512 pixels (approximately 5 to 6° of visual angle).

Control experiment

Subjects performing go/no-go tasks that include difficult test conditions could bias their behaviour towards no-go responses even when instructed to respond on about 50% of the trials. In order to control for such a possible bias, we replicated the experiment using a 2-alternative forced-choice paradigm and 10 other subjects (6 men, mean age 27, range 23-32, 2 of them left handed, normal or corrected to normal vision). Stimuli, display and procedure were identical to the above description except for the response mode. At the beginning of the session, subjects associated left or right hand with natural or man-made category; to start a series, subjects had to place their middle fingers simultaneously on the two "control" keyboard buttons. They further responded to each stimulus by lifting up the finger of the correct hand from the key; no feedback was provided. For half of the subjects, the left and right control keys corresponded respectively to the natural and man-made categories, and the opposite applied to the other half.

Results

Results on man-made and natural categorization tasks using phase spectrum alteration are summarized in Figure 4. Categorization time-courses are illustrated by RT distributions and d' curves in Figure 5.

Accuracy

Subjects' accuracy was measured for each of the 11% steps in phase noise levels (Fig.4A). For both tasks, accuracy followed a sigmoid curve that started with a high plateau at above 80% accuracy for the lowest percentage of phase noise, presented an inflexion point around 50% of phase noise and quickly reached chance level for the highest noise levels. This accuracy pattern was highly similar for man-made and natural categorization tasks; it was also similar to the accuracy pattern obtained in the control experiment that used forced-choice responses. An ANOVA comparing the accuracy in both go/no-go tasks showed no significant task effect. There were no significant difference between 11%, 22%, 33% and 44% noise conditions, and neither were there significant differences between the 55%, 66%, 77% and 88% noise conditions. To simplify subsequent analyses, we merged conditions from 11% to 44% of phase noise (low noise stimuli), and those from 55 to 88% (high noise stimuli), leaving apart the condition with original phases preserved and the condition with complete random phases. Using these merged conditions, a paired t-test did not show any difference between man-made and natural categorization tasks (low noise: $p = .59$, high noise: $p = .38$).

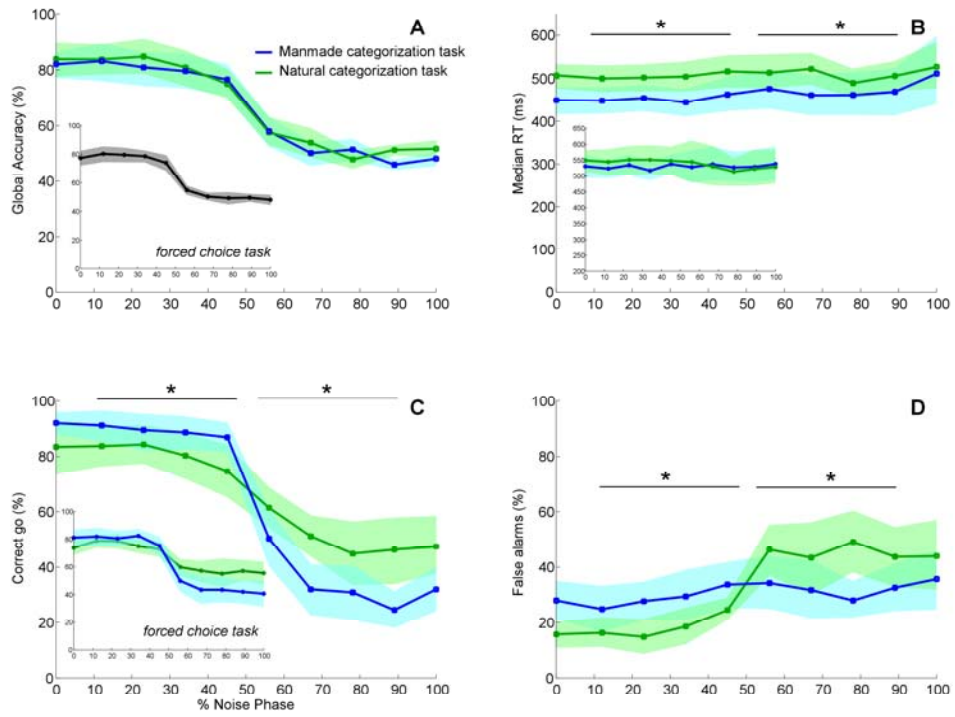

Figure 4: ***Average performance in experiment II***

Lines indicate subjects' average performance as a function of percentage of phase noise. The area around each line is the 95% confidence interval computed using a percentile bootstrap technique with 2000 resamples. **A** Average accuracy; **B** average median RT for correct-go responses (ms); **C** Average accuracy for correct-go responses; **D** Average false alarm rate. Green (blue) lines indicate performance in the natural (man-made) go/no-go categorization task. Subplots in (A,B,C) illustrate average performance recorded in the forced-choice control task using the same stimuli than the go/no-go tasks; for the forced-choice task, green and blue lines stand thus for performance on natural and man-made *stimuli*, respectively. Asterisks indicate statistically significant differences between go/no-go tasks. Global accuracies on man-made and natural categorization tasks are highly similar across phase noise levels. However, subjects' responses were strongly biased toward natural category when more than 55% of phase information was altered, while the opposite response bias was observed for less than 44% of phase noise stimuli. Reaction times were 60 ms longer in the natural categorization task.

On the other hand, the separate analysis of correct-go responses and false alarms (Fig. 4C,D), revealed a different effect of phase degradation depending on the target category. Correct-go responses were significantly more accurate for man-made targets than for natural targets for low phase noise levels ($p < .002$), and significantly less accurate for high noise levels ($p < .0001$). More false alarms were recorded during man-made categorization task for low phase noise levels than during natural task ($p = .006$), and conversely fewer false alarms were observed during the man-made categorization task with low phase noise levels ($p = .031$). Thus, when most of the original phase was preserved, subjects were biased toward man-made responses but when little of the original phase information remained, subjects were biased toward natural responses. This pattern was also observed in the forced-choice task.

Categorization time-course

Median RT was almost constant across the phase noise conditions in both tasks (Fig. 4B). Median RT during the man-made categorization task (464 ms) was about 60 ms shorter than during the natural categorization task (527 ms), irrespective of the phase noise level (paired t-test, low noise condition: p = .0012, high noise condition: .045). In the forced-choice task, a longer median RT was observed (540 ms); but again, response speed was also almost constant across phase noise levels.

Categorization time-courses were also found to be remarkably similar across phase noise levels, as assessed by the RT distributions of correct-go responses. Figure 5A compares correct-go RT distributions for low phase noise stimuli to the one for the original phase noise condition (0% phase noise). For each categorization task, man-made and natural, the correct-go RT distributions with low phase noise and original phase noise conditions were very similar. Figure 5 C shows the RT distributions for the high phase noise conditions. The lower hit accuracy in the man-made categorization task compared to the natural task is clearly observed, but with high or low phase noise level, within each task RT varied in the same range, i.e. if normalized, RT distributions would appear very similar.

However, we observed differences in timing between the man-made and the natural categorization tasks. Man-made correct-go RT distributions always deviated from zero earlier than with the RT distributions for natural targets and a similar delay was observed for false alarm RT distributions (Fig. 5A-C). A second observation can be made concerning the false alarm RT distributions (Fig. 5B): as in experiment I, the percentage of false alarms in the man-made categorization task was higher for short latency responses than for late ones, while false alarms in the natural categorization task were more evenly distributed along the overall response time range. Using d' curves, we further explored the different timing characteristics between man-made and natural categorization tasks (Fig 5D). In addition to the decrease in sensitivity associated with increasing amount of phase noise, d' curves deviated earlier from zero with the natural categorization task than for man-made categorization task, unlike correct-go RT distributions. However, after a short plateau followed by a rebound from 350-360 ms for the natural categorization task, d' curves reached similar values in both tasks. To better understand how this d' pattern could be related to man-made and natural categories, we further explored categorization time courses using hit and false alarm data from both tasks by performing new analyses using the D-T value described below.

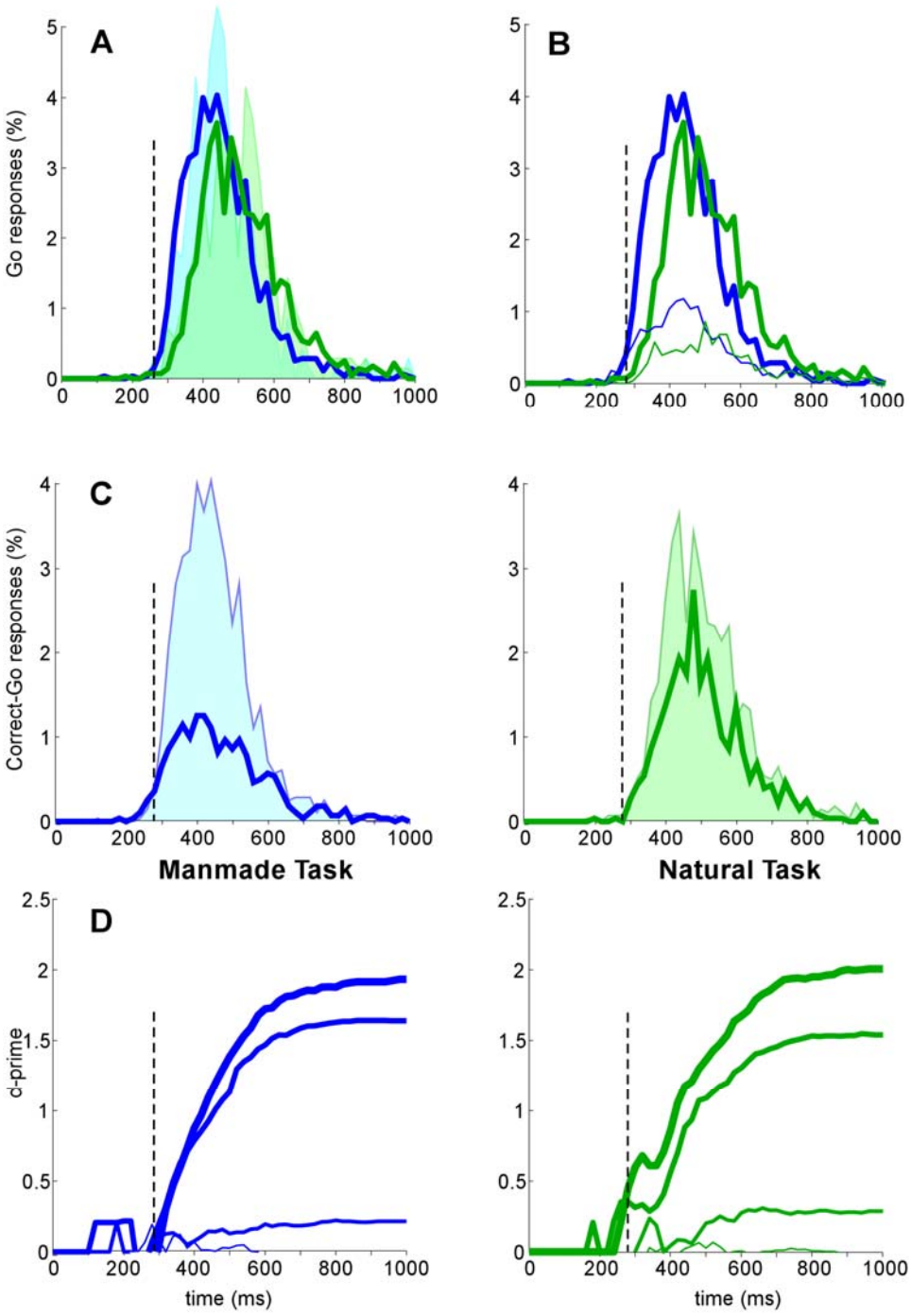

Figure 5: *Categorization time-course in experiment II*.
Green (blue) lines indicate performance in the natural (man-made) go/no-go categorization task, as a function of time using RT pooled across subjects. **A** RT distributions in the natural and man-made tasks for all low phase noise level stimuli pooled together (11-44%, thick line) compared to original phase stimuli (0%, filled histogram), using 20 ms bins. **B** RT distributions, in the natural and man-made tasks, of correct-go responses (thick curves) and false alarms (thin curves), for low phase noise level stimuli (11-44%), using 20 ms bin. **C** Comparison between correct-go response RT distributions (20 ms bins) for stimuli with low phase noise level (11-44%, same as B thick curves) shown with filled histograms and the stimuli with high phase noise level (55-88%, thick curves). **D** d' curves computed using 20m bin for, from the thickest to the thinnest lines, stimuli with phase noise levels 11-22%, 33-44%, 55-66% and 77-88%.

Categorization time-course using D-T value

By analogy to d' sensitivity measure, we designed a new measure (D-T value) that describes how a particular stimulus set was detected with respect to its task status in alternated crossed tasks. D-prime is a sensitivity measure based on the signal detection theory framework, which provides a measure of performance that is independent of decision bias like strategy choice or speed-accuracy trade off. By using hit and false alarm data, d' measures how sensitive subjects are at distinguishing between two different stimulus sets. D-prime is often interpreted as a 'physical' distance between stimuli: in the analyses described above, d' between man-made scene targets and natural scene distractors could be related for instance to the number of visual cues that were characteristic of man-made and natural scenes and that were preserved from noise blurring.

In the current experiment, the two stimulus sets alternated between target and distractor status in the two tasks employed. We took advantage of this paradigm to determine how the two stimulus sets could be differentiated using only their target and distractor status. As an attempt to better describe perceptual processes, we measured how sensitive subjects were to make a target/distractor decision on *the same* stimulus set. In other words, by using hit and false alarm responses to the same stimuli, we removed the 'physical' distance in order to enhance a 'decision' factor that could characterize the current context go/no-go categorization task.

We thus designed a sensitivity measure by computing $DT = z(pHits_{task\ 1}) - z(pFA_{task\ 2})$ where z is the inverse of the normal distribution function, $pHits_{task\ 1}$ being the proportion of hits for task 1, and $pFA_{task\ 2}$ being the proportion of false alarms for task 2, where tasks 1 and 2 are the two crossed tasks using both the same two stimulus sets but alternating their status of target and distractor (Fig. 6A). In order to plot D-T value as a function of time (D-T curve), the cumulative proportions of hits and false alarms was computed using 10 ms bins. We computed D-T curves separately for man-made category stimuli and for natural category stimuli using the behavioural performance recorded in the two categorization tasks (Fig 6B). The D-T measure allows us to describe the time-course of processing for a stimulus set belonging to the same category when processed either as target or as distractor.

Several observations can be made from these data. First, deviations from zero were observed much earlier for man-made stimuli than natural ones (around 250 ms for man-made stimuli, around 380 ms for natural stimuli). Second, man-made scenes reached a higher D-T plateau than natural scenes when the phase noise was below 50%. Finally, while D-T values increased with RT response times for all noise conditions with natural category stimuli, the D-T values computed for the highest noise conditions with man-made stimuli quickly decreases to zero baseline starting at about 360 ms. This return to baseline coincided with the rebound observed for man-made scenes and the increase in D-T value observed for natural scenes (Fig. 6C).

Thus, the earliest correct responses were primarily observed when man-made stimuli were shown. When no rapid decision was made about man-made category, subjects increasingly biased their choice toward "natural" responses from around 360 ms.

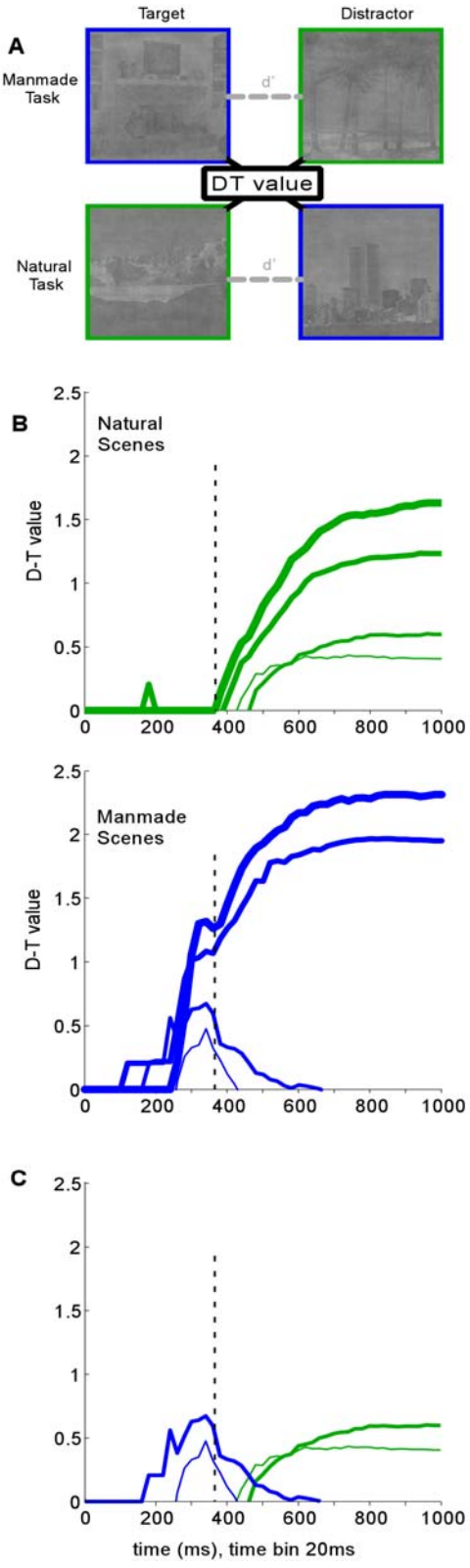

Figure 6: *Illustration of the D-T value computed on man-made and natural stimuli for several noise levels.*

We introduced the D-T value (see Result section of experiment II) in order to provide a sensitivity measure for the same stimuli considered alternatively as target and distractor in crossed go/no-go tasks. **A** Illustration of the crossed go/no-go tasks in which the same stimuli are processed as target or distractor depending on the task. Hits and false alarms recorded in both tasks were used to compute the D-T value of man-made and natural stimuli. Hits and false alarms were taken from the same scene category, only their task status (target or distractor) were different. **B** D-T curves for natural stimuli (green) and man-made stimuli (blue), for noise levels 11-22%, 33-44%, 55-66% and 77-88%, from the thickest to the thinnest lines. Natural scenes sensitivity, as measured by D-T value, takes about 150 ms longer to reach a similar level than the one for man-made scenes. For high phase noise levels, sensitivity decreases sharply for man-made scenes. It might be that, in certain scenes, some remaining original phase information that is diagnostic for man-made scenes is used immediately by subjects to decide for a man-made response; when remaining phase information is not sufficient nor diagnostic for man-made scenes, no man-made response is produced after 400 or 600ms. **C** Direct comparison of D-T curves for natural stimuli (green) and man-made stimuli (blue), using average performance of 55-66% and 77-88% phase noise stimuli.

Discussion

Context categorization performance was analyzed using both average luminance and global contrast equalized scenes, and scenes in which the Fourier's amplitude spectrum had been equalized. Results from experiment I suggest that the visual system uses amplitude spectrum characteristics of man-made and natural scenes to speed up context categorization processes. Equalizing the stimuli's Fourier amplitude resulted in an accuracy drop of about 6% and a median RT increase of about 20ms. Furthermore, analyses of d' showed a substantial 40 ms delay at the level of the earliest responses, indicating that fast processing of context was delayed by amplitude spectrum equalization.

When the specific amplitude spectral signatures of natural and man-made environments are no longer available as diagnostic cues, context categorization processes has to rely on finer details captured by higher order statistics. Experiment II showed that performance accuracy was insensitive to phase alteration until a threshold of 50% of phase noise, but then rapidly fell to chance level following a sigmoid curve. Despite these large accuracy differences, response time-courses showed highly similar timing characteristics for all noise conditions.

In the two experiments, the time course of both correct-go responses and false alarms in the man-made and natural tasks suggested that processing of context categorization may be sensitive earlier to man-made than to natural information. We tested this hypothesis by introducing the D-T value that took advantage of the crossed tasks design to provide a decision sensitivity measure for each context category. Analyses using this D-T value showed that when the information necessary for categorizing a scene as man-made was available, this information was extracted very early. When no decision toward a man-made response could be rapidly made, subjects appeared to opt for natural response as a default late response.

Accessing context category from low-level features

In their modelling work, Oliva and Torralba found that the amplitude spectrum of pictures can be used to classify them as natural scenes and man-made environments (Oliva & Torralba, 2001). They used the fact that pictures of man-made environments tend to have an amplitude spectrum that is elongated on the horizontal and vertical frequencies, while pictures of natural scenes have a more uniform frequency content (Baddeley, 1997; Torralba & Oliva, 2003). This kind of global description of a scene that captures an intrinsic property of the visual world could thus provide information to help categorization and recognition processes.

Averaging Fourier power spectrum equates a large amount of global information among stimuli. In addition to equate average luminance and global RMS contrast, we thus also equated for all stimuli the energy in all directions and frequency bands; for example global properties such as a symmetry in luminance distribution would be equal in each stimulus, although the specific localization of these

luminance patterns could change depending on the particular scene considered. Under these conditions in experiment I, any characteristics of the amplitude spectrum that could systematically help differentiating between man-made and natural environments was removed.

The results from experiment I suggest that the global information contained in the amplitude spectrum of scene pictures was indeed used by subjects for early discrimination between natural and man-made environments. Furthermore, it suggests that the use of this sort of visual information can be particularly fast, since with amplitude spectrum equalized stimuli, the whole reaction time distribution was shifted toward longer latencies, starting with the fastest responses at about 300 ms. These visual processes could take place from the initial part of the sensory encoding, that relies on parallel processing common across a large range of tasks (Bacon-Mace, Kirchner, Fabre-Thorpe, & Thorpe, 2007).

Indeed, context categorization can be performed as fast as other natural scene categorization tasks. The median RT observed in current experiments were in the same range than those observed in previous context categorization tasks with original color photographs(Rousselet et al., 2005; Joubert et al., 2007), tor in object categorization tasks (animals, humans, faces, means of transport, or food objects). Such experiments imply that the processing of context and object can progress in parallel and results in similar time-course and performance levels. They also imply that substantial interactions between object and context processing are perfectly possible (Davenport & Potter, 2004; Joubert et al., 2007; Joubert, Fize, Rousselet & Fabre-Thorpe, in revision).

One could thus suggest that object and context processes could share similar spectral information to determine object and scene categories. Such quickly available information used in both processing streams could account for the fast interference between context and object in a rapid coarse processing of a scene, possibly through the dorsal magnocellular pathway as proposed by Bar and collaborators (Bar, 2004; Bar et al., 2006), or using the magnocellular information processed within the ventral pathway, in accordance with previous results from our group (Macé, Thorpe, & Fabre-Thorpe, 2005). The scene gist provided by a coarse blurred representation of the contextual frame might be sufficient to guide both context and object processing. The good resistance of context categorization processes to phase noise demonstrated in experiment II reinforces this proposal.

On the other hand, the fact that amplitude spectrum equalization significantly impairs fast context categorization supports computational theories that claim that a large amount of information can be rapidly extracted from a visual scene without the need to access elaborated high-level representations. Indeed, the present results show that this rapidly extracted information includes image features that can be described using global image statistics. Furthermore, the results from experiment II further suggest that part of the phase information, that encodes edges and contours, is also rapidly used to enrich an early contextual representation used for categorization. These results are in keeping with the spatial envelope theory (Oliva & Torralba, 2001) that proposes a computational approach “that

constructs a meaningful representation of scene gist directly from the low-level features pool, without binding contours to form surfaces, or surfaces to form objects" (Oliva, 2005, p. 256).

What other cues than low-order statistics are used in context categorization?

Using images with equalized amplitude spectra, we showed that the visual system encoded the global information of the amplitude spectrum as a cue that can be used to determine rapidly context category. However, despite these unusual visual stimuli, subjects were still able to perform well; the information encoded by the amplitude spectrum cannot be the only source of information for performing the task. Other scene properties that take into account edges and contours, encoded in the phase spectrum, were also involved in rapid context categorization.

Experiment II was designed to estimate how context categorization performance relied on phase when amplitude spectra were equalized. Under these conditions performance accuracy was remarkably insensitive to phase alteration of less than 50%, and did not follow the amount of original phase that decreased linearly, but rather followed a sharp sigmoid curve that characterizes categorical processes. This latter observation is consistent with the results from a recent study on scene perception, in which subjects had to decide whether a scene was natural in the presence of phase noise (Einhauser et al., 2006).

A similar sigmoid function that depicts categorization processes was also observed in a recent experiment that used context categorization of natural and man-made scenes blurred with phase noise (Loschky & Larson, 2008). In their experiment, Loschky et al. reported a greater sensitivity of performance accuracy to phase noise than seen in the present experiments. Their results showed substantial performance impairments with 40% of phase noise, much more than the mild decrease observed in the present results for the same noise level. This discrepancy cannot be explained by the force-choice task modality they used, since in the present study the control forced-choice task gave similar results than the go/no-go one. Rather, the difference could be related to the stimuli used. In their study, the man-made and natural stimuli sets were not amplitude spectrum equalized, and this might well enhance the effect of phase degradation: when original phase was predominant, reconstruction using the original amplitude spectrum would be expected to lead to high accuracy performance, which was indeed the case (95% correct vs. our 82%). When phase noise was predominant, stimuli were both degraded by the loss of original phase and the absence of matching between this phase and the original scene amplitude (this can be observed in their Figure 1). Despite this probable overestimate of the role of phase information in context categorization tasks, their results are qualitatively compatible with the accuracy levels observed in the present study.

In contrast, Wichmann and collaborators reported a gradual decrease of performance with phase noise in an Animal/Non-animal rapid categorization task (Wichmann et al., 2006). The

performance accuracy they reported could be described more precisely as a sigmoid curve with an inflexion point around 130° of phase randomization, that corresponds to 75% of the maximum noise level (their Figure 5). Their results thus show a much stronger resistance to phase degradation in the case of object categorization than the present context categorization task; a possible explanation could be that the diagnostic information for object categorization, compared to the one of context, could be encoded in a wider range of contour orientations or spatial frequencies. It would be interesting to compare directly, using the same phase alteration technique, the resistance of both context and object categorisations to the addition of phase noise.

Finally, in an interesting attempt to characterize what information could be diagnostic in the phase spectrum to categorize basic-level scenes, McCotter, Gosselin, Sowden, & Schyns, 2005 used a technique of selective alteration of stimulus phase, which isolated the regions of the phase spectrum that maximized categorization accuracy in their observers. Their results showed that the diagnostic bandwidths of the scene spectra for all categories occurred at relatively low spatial frequency, and coincided with the characteristic amplitude spectral signatures reported by Oliva & Torralba, 2001. They showed that different regions of the phase spectra are diagnostic for different scene categories, with a low level of overlap, suggesting that the local structures and edges described by the diagnostic phase spectra of a scene category are not common to other scene categories. They also mentioned that some visual features, contours or edges that were poorly redundant across the entire image (a coastline separating two different textures for example), could be particularly sensitive to disruption of amplitude and as a consequence could not tolerate the effect of phase noise to the same extent than other ones. These results stress the fact that scene redundancies or spectral signature provide diagnostic information on scene category, and are encoded by both amplitude and phase spectra. The amplitude spectrum encodes first and second order statistics, and represents luminance variations at each orientation and scale. The phase spectrum, that encodes higher order scene statistics and represents outlines, boundaries and edges, can also be divided in separated regions of spatial frequencies that are diagnostic for a limited number of scene categories. Despite these significant advances, more computational and experimental work is needed to quantify the visual features that are effectively used by scene categorization processes.

The time-course of context categorization

The results of experiment II provide some insight into the time-course of context categorization. We introduced the D-T value to provide a decision sensitivity measure for natural and man-made scene categories, taking advantage of the task design that switched the target-distractor status of all stimuli between tasks. D-T value was computed by using hit and false alarm responses to the same stimuli in the two tasks, thus removing any effects of stimulus differences on trials with hits and false alarms. The D-T curve analysis showed that the earliest correct responses were primarily observed following the

presentation of man-made stimuli. When no rapid decision was made, subjects increasingly biased their choice towards a "natural" response after about 350 ms. While the interpretation and the significance of the D-T value will need further investigation, this measure succeeded in underlining the stimulus specificity of subject responses.

The results suggest that when the information necessary for categorizing a scene as man-made was present, the decision threshold was reached quickly. On the other hand, if no diagnostic man-made features could be detected either because a natural stimulus was displayed, or because of a high noise level, the resulting slower information accumulation process favours a natural response decision. This two phase categorization time-course could be specific to the particular context categorization task or the stimulus set used. However, it nevertheless suggests the existence of a categorization processing strategy based on stimulus information uptake, that could be driven by preparatory top-down signals.

Subjects might bias their fast responses toward the category for which diagnostic elements are more readily available across noise levels. This bias could account for the different performance observed in experiment I and II with the same ELA stimuli. As the hardest experimental condition in experiment I subjects tended to respond with long response times with a bias toward "natural" responses whereas, as the easiest condition in experiment II subjects tended to produce fast responses a bias toward "man-made" responses. Biases toward natural response have been reported in a few other studies on image perception that manipulated various experimental conditions (Rousselet et al., 2005; Fei-Fei et al., 2007; Loschky & Larson, 2008). Stimulus difficulty could be accounted for by the results of McCotter et al., 2005 reported earlier, who observed that some visual features were more sensitive to phase noise than others. Particularly, they mentioned that artificial scenes often contain well-defined and redundant edges that are more resistant to high levels phase noise, compared to more distributed diagnostic features of some natural scenes like textures and smoother edges.

Acknowledgments

This work was supported by the CNRS (Centre National de la Recherche Scientifique), by the French Government (Ministère de la Recherche et de l'Enseignement Supérieur) and by the Fondation pour la Recherche Médicale.

References

Bacon-Mace, N., Kirchner, H., Fabre-Thorpe, M., & Thorpe, S. J. (2007). Effects of task requirements on rapid natural scene processing: From common sensory encoding to distinct decisional mechanisms. *Journal of Experimental Psychology: Human Perception and Performance, 33*(5), 1013-1026.

Baddeley, R. J. (1997). The correlational structure of natural images and the calibration of spatial representations. *Cognitive Sciences, 21*(3), 351-372.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience, 5*(8), 617-629.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmidt, A. M., Dale, A. M., et al. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Acadademy of Sciences of the United States of America, 103*(2), 449-454.

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In W. A. Rosenblith (Ed.), *Sensory Communication* (pp. 217-236): John Wiley & Sons.

Biederman, I. (1981). On the semantics of a glance at a scene. In M. K. J. R. Pomerantz (Ed.), *Perceptual Organization* (pp. 213 -254). Hillsdale: Lawrence Erlbaum.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review, 94*(2), 115-147.

Biederman, I., Rabinowitz, J. C., Glass, A. L., & Stacy, E. W., Jr. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology, 103*(3), 597-600.

Campbell, F. W., & Robson, J. G. (1968). Application of Fourier analysis to the visibility of gratings. *Journal of Physiology, 197*(3), 551-566.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Sciences, 15*(8), 559-564.

Einhauser, W., Rutishauser, U., Frady, E. P., Nadler, S., Konig, P., & Koch, C. (2006). The relation of phase noise and luminance contrast to overt attention in complex visual stimuli. *Journal of Vision, 6*(11), 1148-1158.

Evans, K. K., & Treisman, A. (2005). Perception of objects in natural scenes: is it really attention free? *Journal of Experimental Psychology: Human Perception and Performance, 31*(6), 1476-1492.

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision, 7*(1), 10.

Field, D. J. (1999). Wavelets, Vision and the Statistics of Natural Scenes. *Philosophical Transactions of the Royal Society of London A, 357*, 2527-2542.

Friedman, A. (1979). Framing pictures: the role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General, 108*(3), 316-355.

Guyader, N., Chauvin, A., Peyrin, C., Herault, J., & Marendaz, C. (2004). Image phase or amplitude? Rapid scene categorization is an amplitude-based process. *Comptes Rendus de l'Academie des Sciences: Biologies, 327*(4), 313-318.

Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research, 47*(26), 3286-3297.

Joubert, O. R., Fize, D., Rousselet, G. A., & Fabre-Thorpe, M. (2008). Early interference of context congruence on object processing in rapid visual categorization of natural scenes. *Journal of Vision* (in revision)

Kaping, D., Tzvetanov, T., & Treue, S. (2007). Adaptation to statistical properties of visual scenes biases rapid categorization. *Visual Cognition, 15*(1), 12-19.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Acadademy of Sciences of the United States of America, 99*(14), 9596-9601.

Loschky, L. C., & Larson, A. M. (2008). Localized information is necessary for scene categorization, including the Natural/Man-made distinction. *Journal of Vision, 8(1)*(4), 1-9.

Mace, M. J., Thorpe, S. J., & Fabre-Thorpe, M. (2005). Rapid categorization of achromatic natural scenes: how robust at very low contrasts? *European Journal of Neuroscience, 21*(7), 2007-2018.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory A user's Guide*: Routledge.

Marr D. (1982). *Vision* (New York): W. H. Freeman

McCotter, M., Gosselin, F., Sowden, P., & Schyns, P. G. (2005). The use of visual information in natural scenes. *Visual Cognition, 12*(6), 938-953.

Oliva, A. (2005). Gist of the Scene. In L. Itti, G. Rees & J. K. Tsotsos (Eds.), *Neurobiology of Attention* (pp. 251-256): Elsevier.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42*(3), 145-175.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2*(5), 509-522.

Potter, M. C., & Faulconer, B. A. (1975). Time to understand pictures and words. *Nature, 253*, 437-438.

Rainer, G., Augath, M., Trinath, T., & Logothetis, N. K. (2001). Nonmonotonic noise tuning of BOLD fMRI signal to natural images in the visual cortex of the anesthetized monkey. *Current Biology, 11*(11), 846-854.

Rensink, R. A. (2000). Seeing, sensing, and scrutinizing. *Vision Research, 40*(10-12), 1469-1487.

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neurosciences, 5*(7), 629-630.

Rousselet, G. A., Joubert, O. R., & Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition, 12*(6), 852-877.

Rousselet, G. A., Thorpe, S. J., & Fabre-Thorpe, M. (2004). Processing of one, two or four natural scenes in humans: the limits of parallelism. *Vision Research, 44*(9), 877-894.

Sawamura, H., Georgieva, S., Vogels, R., Vanduffel, W., & Orban, G. A. (2005). Using functional magnetic resonance imaging to assess adaptation and size invariance of shape processing by humans and monkeys. *Journal of Neuroscience, 25*(17), 4294-4306.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: evidence for time- and spatial-scale-dependent scene recognition. *Psychological Sciences, 5*(4), 195-200.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Acadademy of Sciences of the United States of America, 104*(15), 6424-6429.

Thorpe, S., Delorme, A., & VanRullen, R. (2001). Spike-based strategies for rapid processing. *Neural Networks, 14*(6-7), 715-725.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520-522.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network, 14*(3), 391-412.

VanRullen, R. (2007). The power of the feed-forward sweep. *Advances in Cognitive Psychology, 3*(1-2), 167-176.

VanRullen, R., Reddy, L., & Koch, C. (2004). Visual search and dual tasks reveal two distinct attentional resources. *Journal of Cognitive Neuroscience, 16*(1), 4-14.

Vogel, J., Schwaninger, A., Wallraven, C., & Bulthoff, W. H. (2006). *Categorization of natural scenes: local vs. global information.* Paper presented at the 3rd symposium on Applied perception in graphics and visualization table of contents, Boston, Massachusetts.

Westheimer, G. (2001). The Fourier theory of vision. *Perception* 30(5), 531-541.

Wichmann, F. A., Braun, D. I., & Gegenfurtner, K. R. (2006). Phase noise and the classification of natural images. *Vision Research, 46*(8-9), 1520-1529.

2. A quelle vitesse peut-on catégoriser un contexte ?

2.1. Objectifs

Dans le chapitre précédent, nous avons pu davantage préciser les caractéristiques physiques diagnostiques des catégories de scènes naturelles « Env. Nat. » et « Env. Man. » par l'utilisation de stimuli dégradés.

Les deux articles présentés dans ce chapitre ont pour objectif d'estimer le temps nécessaire au système visuel pour catégoriser le contexte d'une scène naturelle non dégradée, ainsi que les décours temporels des traitements sous-jacents à différents niveaux de catégorisation : basique et superordonné.

Comme décrit précédemment dans le mémoire (Part I. Chap. 4.7), l'entrée privilégiée dans la hiérarchie des catégories d'objets se ferait au niveau de la catégorie basique qui serait accessible avant les catégories subordonnées et superordonnées (Rosch et al., 1976). Cette entrée privilégiée au niveau basique a été étendue aux contextes de scènes par Tversky et Hemenway (1983) qui suggèrent que dans des tâches verbales, les sujets accèderaient plus facilement aux catégories basiques (Mer, Montagne) de la scène qu'à leur catégorie superordonnée (Env. Naturel par exemple). En est-il de même au sein dans nos tâches de catégorisation visuelle où la réponse motrice rapide remplace la réponse verbale ?

Le premier article publié dans Visual Cognition en 2005 décrit une étude portant sur la catégorisation go/no-go du contexte au niveau basique et sur l'influence de la couleur dans ce type de tâche.

Le second article publié dans Vision Research en 2007 évalue à l'aide du même protocole notre rapidité à catégoriser le même ensemble de scène, cette fois-ci à un niveau superordonné. Au cours de cette étude, nous avons en outre réalisé une analyse post-hoc révélant l'influence des objets sur la catégorisation du contexte qui sera discuté dans la troisième partie de ce mémoire.

Ces deux études ont été mises en place afin de répondre à plusieurs questions:

(1) Peut-on reconnaître assez rapidement le contexte d'une scène naturelle pour envisager une influence du contexte sur les premières étapes de traitement menant à la catégorisation puis à la reconnaissance des objets ?

(2) Quel est –en termes temporel- le niveau de catégorie, basique ou superordonné, le plus rapidement accessible lorsque l'on s'affranchit (comme dans nos protocoles) de l'aspect "verbal" de la tâche ?

(3) Doit-on envisager une hiérarchie des niveaux de catégorisations visuelles des contextes telle qu'elle a été définie par Rosch pour les objets et par Tversky pour les contextes à un niveau taxonomique (Rosch et al., 1976, Tversky & Hemenway, 1983)? Ou doit-on penser que le niveau d'entrée des catégories visuelles diffère de celui des catégories taxonomiques comme cela a déjà été suggéré par notre équipe pour les catégories d'objets (Macé et al., 2006) ?

(4) Dans les catégorisations rapides de contexte à un niveau superordonné, les informations de couleurs sont-elles peu influentes comme c'est le cas pour les catégorisations d'objets ?

2.2. Accès à la catégorie visuelle basique d'un contexte en 450 ms.

Protocole

Dans les deux études, les sujets furent impliqués dans une tâche de catégorisation rapide de contexte de scènes naturelles selon le paradigme go/no-go déjà préalablement décrit. Dans la première étude, 4 catégories basiques cibles ont été successivement proposées aux sujets : 2 catégories dites « naturelles », « Mer » et « Montagne », et 2 catégories dites « manufacturées », « Intérieurs » et « Scènes de rue ». Alors que les scènes étaient flashées 26

ms, les sujets devaient répondre le plus précisément et le plus rapidement possible lorsqu'une image cible apparaissait tout en évitant de répondre sur les images appartenant aux trois autres catégories distractrices. Une première expérience ne comportait que des scènes naturelles en couleur tandis qu'une seconde expérience testait l'influence de la couleur dans la tâche par l'introduction de 50 % de scènes achromatiques.

Résultats résumés

Les sujets humains se sont montrés très efficaces à effectuer la tâche de catégorisation de contexte à un niveau basique atteignant des performances de plus de 90 % de réussite. Lors de la catégorisation des contextes à un niveau basique, la majorité des erreurs était commise faites sur la catégorie basique distractrice appartenant à la même catégorie superordonnée que la catégorie cible.

De plus, les sujets répondaient rapidement, les temps de réaction médians variant de 400 à 460 ms selon la catégorie avec un avantage pour les catégories basiques naturelles. La couleur semble une information peu cruciale pour la catégorisation rapide de scènes comme déjà démontré pour l'objet (Delorme et al., 2000), même si son influence semble légèrement plus importante pour certaines catégories telle que « Mer ». La couleur pourrait avoir un rôle facilitateur (notamment pour les décisions prises autour des temps de réaction moyens) pouvant accélérer la prise de décision sans jouer pour autant un rôle fondamental.

Discussion

De par ces résultats, il apparaît évident que les catégories visuelles de contexte sont également organisées selon une hiérarchie définie. Les analyses sur la précision des réponses montrent en effet que les quelques erreurs enregistrées durant la tâche de catégorisation à un niveau basique sont majoritairement effectuées lors de la présentation de scènes appartenant à la même catégorie superordonnée. Les catégories visuelles basiques naturelles sont donc perceptivement plus proches l'une de l'autre qu'elles ne le sont des catégories visuelles basiques manufacturées. L'inverse est d'ailleurs tout aussi vrai. Les catégories basiques

manufacturées partagent plus de points communs entre elles qu'avec les catégories basiques naturelles. Cette différence perceptive entre catégories peut d'ailleurs expliquer le biais observé au niveau des temps de réaction à l'avantage des scènes naturelles. On peut en effet penser qu'il existe un recouvrement perceptif moins important entre les deux catégories « naturelles » qu'entre les deux catégories « manufacturés ». Les catégories basiques naturelles possèderaient des traits physiques diagnostiques non partagés par l'autre catégorie naturelle permettant ainsi une catégorisation optimale. Ces informations diagnostiques pourraient d'ailleurs inclure le facteur couleur qui semble avantager plus particulièrement les scènes de « Mer ». Par exemple, une surface bleuâtre dans la moitié inférieure de l'image a 99% de chance d'indiquer une scène de « Mer ». Au contraire, les catégories basiques manufacturées plus proches perceptivement et donc contenant moins d'informations diagnostiques spécifiques à chacune nécessiteraient des traitements perceptifs plus importants pour être reconnues. Pour autant, la couleur ne constitue pas un facteur diagnostique majeur puisque les scènes naturelles achromatiques sont tout aussi précisément catégorisées.

2.3. Accès à la catégorie visuelle superordonnée d'un contexte en moins de 400 ms

Protocole

Dans la seconde étude, l'ensemble des scènes naturelles utilisé précédemment a été complété afin de couvrir une plus grande variété d'environnements. Ainsi ont été ajoutées des photographies de « champs », de « banquises », de « forêts », ou encore des « vues aériennes de ville ». Cette fois-ci, les sujets devaient effectuer une tâche de catégorisation à un niveau superordonné «Environnements naturels » vs. « manufacturés », toujours selon le même protocole expérimental.

Résultats

Alors que la précision enregistrée est très similaire à l'expérience précédente (>90%), des temps de réaction médians plus courts sont cependant obtenus dans la tâche de catégorisation de contexte à un niveau superordonné ! En effet, les sujets sont capables de

déterminer si une scène est naturelle ou manufacturée en moins de 400 ms avec des réponses précoces enregistrées dès 220 ms pour la catégorie « Env. Man. »

Discussion

Alors que les latences des réponses moyennes enregistrées dans la tâche de catégorisation basique se situent aux alentours de 450 ms, les latences de réponses les plus courtes sont obtenues lorsque ces mêmes scènes chromatiques sont catégorisées à un niveau superordonné : en moins de 400 ms. Le niveau superordonné apparaît donc comme le niveau d'accès privilégié dans les tâches visuelles ne nécessitant pas une décision verbale ou lexicale ultérieure à la présentation du stimulus. La catégorisation à un niveau basique nécessiterait des traitements perceptifs supplémentaires afin d'accumuler une information diagnostique plus affinée, permettant la reconnaissance de catégories perceptivement plus complexes ou la discrimination de catégories perceptivement plus proches. Cette hypothèse va dans le sens d'une analyse de la scène « coarse-to-fine » comme l'ont suggéré les études de Schyns et Oliva (Oliva & Schyns, 1997, Oliva & Torralba, 2001, Schyns & Oliva, 1994). Il serait dans ce sens intéressant d'évaluer le décours temporel des traitements catégoriels lorsque la reconnaissance d'une catégorie subordonnée est demandée. Dans tous les cas, il semble évident que la définition du niveau basique en tant que niveau d'entrée ne peut être élargie à tout type de tâches de catégorisation. Cette évidence apportée par les résultats dans ces deux tâches de catégorisation rapide de contexte avec réponse motrice, a également été fournie dans le cadre de la catégorisation d'objets. Dans un article actuellement en soumission auquel j'ai participé (Macé et al., 2006), nous avons pu en effet démontrer qu'une catégorisation animal/non-animal de photographies flashées était effectuée en 400 ms environ, et qu'une catégorisation de ces mêmes animaux à un niveau basique (chien ou oiseaux) nécessitait 40 à 70 ms supplémentaires. Cette différence entre nos travaux et les travaux de Rosch ou de Tversky (Rosch et al., 1976, Tversky & Hemenway, 1983) pourrait s'expliquer par la nature différente des traitements impliqués en fonction des tâches également différentes. Dans leurs travaux, les tâches effectuées par les sujets sont toutes de nature verbale (dénomination d'objet, caractérisation verbale d'attributs) et font donc appel au lexique alors que nos expériences impliquent majoritairement des traitements visuels perceptifs. Il est très possible que l'accès visuel plus tardif (une cinquantaine de ms) aux représentations catégorielles

basiques soit masqué et même inversé dès que les tâches nécessitent un accès au lexique par un accès verbal beaucoup plus rapide aux catégories de base. Le niveau d'accès privilégié aux catégories ne serait donc pas fixe mais pourrait donc varier en fonction du protocole utilisé selon qu'il nécessite ou non un accès au lexique.

De manière intéressante, ces latences obtenues lors d'une catégorisation d'objets ne sont pas sans rappeler les latences de réponses obtenues dans la présente tâche de catégorisation de contexte. Un recensement des temps de réaction enregistrés au cours de l'ensemble des études effectuées au CerCo, et présenté dans l'article n°3, nous a permis d'établir certaines correspondances entre la catégorisation d'objets et de contexte. Il apparaît que les traitements catégoriels de l'objet et du contexte ont des décours temporels très similaires. Les distributions des réponses au cours du temps obtenues dans les tâches de catégorisation d'objets et de contexte à un même niveau (basique ou superordonné) présentent un chevauchement très important. Notre système visuel pourrait donc en moyenne catégoriser un contexte ou un objet à un niveau superordonné en 400 ms environ, l'accès au niveau basique demandant en moyenne 50 ms supplémentaires, que ce soit pour les catégories d'objets ou les catégories contextuelles.

Ces latences similaires laissent envisager la possibilité d'interactions entre les traitements catégoriels relatifs à l'objet et au contexte, rendant possible une influence contextuelle sur la reconnaissance des objets. C'est cette possibilité qui va être explorée dans l'article suivant.

Article n°2

2005

How long to get to the "gist" of real-world natural scenes?

Visual Cognition, 12 (6) 852-877

Guillaume A. Rousselet, **Olivier R. Joubert,** & Michèle Fabre-Thorpe

VISUAL COGNITION, 2005, 12 (6), 852–877

Psychology Press
Taylor & Francis Group

How long to get to the ''gist'' of real-world natural scenes?

Guillaume A. Rousselet, Olivier R. Joubert, and Michèle Fabre-Thorpe

Centre de Recherche Cerveau et Cognition, UMR 5549 (CNRS-UPS), Toulouse, France

This study aimed at assessing the processing time of a natural scene in a fast categorization task of its context or ''gist''. In Experiment 1, human subjects performed 4 go/no-go categorization tasks in succession with colour pictures of real-world scenes belonging to 2 natural categories: ''Sea'' and ''mountain'', and 2 artificial categories: ''Indoor'' and ''urban''. Experiment 2 used colour and grey-level scenes in the same tasks to assess the role of colour cues on performance. Pictures were flashed for 26 ms. Both experiments showed that the gist of real-world scenes can be extracted with high accuracy (>90%), short median RT (400–460 ms) and early responses triggered with latencies as short as 260–300 ms. Natural scenes were processed faster than artificial scenes. Categories for which colour could have a diagnostic value were processed faster in colour than in grey. Finally, processing speed is compared for scene and object categorization tasks.

Natural scenes are more than a simple collection of objects. However, much of the research on scene processing has been devoted to the understanding of object processing in scenes, leaving aside the question of how we process the whole scene itself.

This issue is important given that we do not only process objects but we also analyse the context in which they appear. Global coarse information about a scene (mainly its category, or *gist*, and its spatial structure, or *layout*) is relatively crucial in memory-free models of scene perception in which little information is integrated across saccades. According to this idea, perception is constructed by integrating abstract scene representations with volatile object

Please address all correspondence to: Guillaume A. Rousselet, McMaster University, Department of Psychology, Hamilton, L8S4K1, ON, Canada. Email: rousseg@mcmaster.ca

This work was supported by the CNRS, the ACI ''Integrative and Computational Neurosciences'', and the Cognitique grant no. IC2. Financial support was provided to G. A. Rousselet by a PhD grant from the French government. We thank Anne-Sophie Paroissien for her very valuable help running the experimental sessions in Experiment 1.

http://www.tandf.co.uk/journals/pp/13506285.html DOI:10.1080/13506280444000553

representations available at the locus of attention (O'Regan, 1992; Rensink, 2000, 2002; Wolfe, 1999; but see Henderson & Hollingworth, 2003; Hollingworth, 2003; Hollingworth & Henderson, 2002; Simons, Chabris, Schnur, & Levin, 2002). More generally, visual context has been shown to guide attention toward potential target objects (for a review see Chun, 2000). The influence of the background of a scene on object identification is still controversial (see reviews in Henderson, 1992; Henderson & Hollingworth, 1999), with evidence both in favour and against such a view (see among many others Biederman, Mezzanotte, & Rabinowitz, 1982; Boyce, Pollatsek, & Rayner, 1989; but see Ganis & Kutas, 2003; Hollingworth & Henderson, 1998, 1999). But our knowledge about the global processing of a scene is severely limited compared to the knowledge accumulated about object processing. In particular, for scene context to influence object identification, one fundamental constraint is the speed at which scene context can be extracted.

What do we know exactly about scene processing? Only recently this topic has received more attention from cognitive neuroscience researchers, revealing a set of cortical areas involved in different aspects of scene processing, that include the parahippocampal and parietal cortices (Nakamura et al., 2000; Sato et al., 1999). More specifically, the parahippocampal area has been attributed functions such as the processing of the spatial layout of the scene (Epstein, Graham, & Downing, 2003; Epstein & Kanwisher, 1998) and the learning and recognition of buildings and landscapes (Maguire, Frith, & Cipolotti, 2001). It is also thought to mediate, in conjunction with the retrosplenial cortex, both spatial and nonspatial contextual processing (Bar & Aminoff, 2003). Finally, the right lingual sulcus has been implicated in the perception of buildings (Aguirre, Zarahn, & D'Esposito, 1998). This distributed system, largely separated from the object system, might lead to the hypothesis that scenes are processed very efficiently, perhaps fast enough to be able to influence object processing.

Scene categorization is often regarded as the ultimate representation generated along the ventral pathway in which a scene would be reconstructed progressively by integrating local contrasts (Marr, 1982). Following such a view, objects would be processed almost systematically before the scene (Biederman, 1987; Riesenhuber & Poggio, 2000).

Alternatively, many studies have suggested that scene categorization can be performed very efficiently from very brief visual presentations (Biederman et al., 1982; Intraub, 1997; Oliva & Schyns, 1997, 2000; Potter, 1975, 1976; Potter & Levy, 1969; Schyns & Oliva, 1994). The fact that scene categorization is possible with very brief presentations is often taken as an evidence for fast underlying mechanisms. Thus, scene categorization could be performed simultaneously or even precede object identification. Indeed, although some theories have suggested that scene categorization might result from the identification of the component objects (e.g., Friedman, 1979), other theories have emphasized that scenes might also be identified from scene-specific cues. For instance,

Biederman (1988) suggested that his original structural model of object recognition using ''geons'' (3-D primitives) might be extended to scene recognition. He proposed that primitives with a larger spatial scale than those used to represent objects could represent scene specific information independently of object information. Although this proposal has not been tested empirically, Henderson and Hollingworth (1999) have suggested that, given the lack of strong constraints on their structural organization, scenes are not likely to be represented as large objects. However, it remains possible that the spatial organization of a scene (even at a coarse level) might mediate its identification. Sanocki and Epstein (1997, p. 378) suggested that the representations of the spatial layout ''may include information about the extent and location of the ground plane and other reference objects and surfaces, as well as size and distance relations within the scene''. This information about the spatial structure of the scene might be used to extract its meaning. Indeed, the gist of a scene could be extracted from low spatial frequency versions of photographs preserving coarse spatial layouts, or spatially arranged colour blobs, but in which information to categorize component objects was not preserved (Oliva & Schyns, 2000; Schyns & Oliva, 1994). This strengthens the idea that object and scene categorization might be mediated by distinct visual cues. Finally, computational evidence suggests that scene-based visual filters derived from the combination of a restricted set of low-level filters might be sufficient to perform most of the visual discrimination needed to categorize the context of our environment (Oliva & Torralba, 2001; Torralba & Oliva, 2003).

Given such a reduced dictionary of scene-based physical properties and the known capacity of the visual system to dynamically adjust its strategies as a function of task constraints to pick the most adequate image features (Schyns, 1998), it is plausible that scenes might be identified very fast, probably as fast or even faster than objects. So far, the evidence for fast scene processing comes from experiments using brief visual presentations (see above). Unfortunately, brief visual presentations and particularly RSVP sequences (Rapid Sequential Visual Presentations; e.g., Intraub, 1997; Potter, 1975, 1976; Potter & Levy, 1969) provide a rate of visual processing rather than an absolute evaluation of the visual processing time. Furthermore, experiments like those performed by Oliva and Schyns (1997, 2000) used vocal responses or involved a matching task between a written name and a visual scene with a two-choice response. Using 16 different categories in the matching task, they showed that the averaged mean reaction times (RT) were largely distributed from 476 ms (*city* category) to 631 ms (*valley* category).

This mean RT can be compared to the 400 ms mean RT that has been commonly found for object categorization in various studies from our group. These studies employed a go/no-go animal categorization task in which human subjects were required to respond as fast and as accurately as possible each time a natural photograph, that was flashed for the first time and for only 20–40 ms,

contained an animal (Delorme, Richard, & Fabre-Thorpe, 2000; Fabre-Thorpe, Delorme, Marlot, & Thorpe, 2001; Thorpe, Fize, & Marlot, 1996). This finding has been extended to other object categories such as means of transport (Van-Rullen & Thorpe, 2001), human faces and animal faces (Rousselet, Macé, & Fabre-Thorpe, 2003), although food objects might take up to 30 ms longer (Delorme et al., 2000). Compared with Schyns and Oliva's studies, objects might thus be identified before extraction of scene context. But the difference in processing time might also originate in the motor response required in both tasks as our go/no-go task relies only on a single motor output, whereas their matching task requires a choice of response. In keeping with this idea, Goffaux, Jacques, Mouraux, Oliva, Schyns, and Rossion (2005) report an averaged median reaction time of 428 ms in a go/no-go verification task.

In this study, we have assessed the time course of the categorization of the scene context, or its ''gist'', with the same go/no-go visual categorization task previously used to study object categorization. We selected four categories of colour pictures, two of them were natural, ''sea'' and ''mountain'', and the two others were artificial, ''indoor and urban'' scenes. In a first experiment, subjects were asked to perform four go/no-go categorization tasks, one per category. A second experiment was designed to assess the effect of colour on gist processing speed.

EXPERIMENT 1

The aim was to provide an estimate of the temporal constraints in the visual processing of the gist of a natural scene, for four scene categories representative of our environment (sea, mountain, urban, and indoor). These categories were relatively coarsely defined in order to present subjects with pictures that were as varied as possible (see Stimuli and Figure 1).

Method

Participants. Twenty-four adults (12 women and 12 men, mean age 30, ranging from 19 to 51, three of them left handed), volunteered in this study and gave their informed written consent. All participants had normal or corrected to normal vision.

Stimuli. We used 24-bit (16 million colours) photographs of natural scenes (768 $\times$ 512 pixels, sustaining a visual angle of about 15.6° $\times$ 10.5°) taken from a large commercial CD-ROM library (Corel Stock Photo Libraries). From this databank, we selected 384 images for each of the four environmental categories. For each category, half of them were horizontal photographs, the other half were vertical photographs. They were all chosen to be as varied as possible, representing the four types of scenes from a large range of viewpoints and

Figure 1. Tasks and stimuli in Experiments 1 and 2. A. Tasks: While performing one of the four scene categorization tasks (sea, mountain, indoor, urban), nontargets belonged equally to the three other categories. Note the variety of stimuli used in this experiment. B. Pixel by pixel average picture for each stimulus category presented separately for horizontal and vertical images. For each category, the top two images represent the raw average pictures and the two bottom images are the equalized versions obtained using the ''equalize'' function in Photoshop 5.5. For each colour channel and the luminance channel, the function attributes a ''black'' value to the darkest pixel and a ''white'' value to the brightest one. It then redistributes regularly the intermediate pixel values of the distribution between these two extremes. C.. For each category, the nine target pictures associated with the fastest reaction times of individual subjects in Experiment 1 are presented. In Experiment 1, all stimuli were in colour while in Experiment 2 half were in colour and half in grey levels. To view this figure in colour please see the online issue.

perspectives (Figure 1). Each image was seen only once by a given subject to prevent learning.

Sea pictures were composed of various coast scenes (including beach scenes, cliff scenes, or showing various rocks, icebergs, etc.) as well as ''full sea'' pictures with boats, sailboards, and surfboards. In all cases the sea was clearly visible on the pictures. The mountain category contained pictures that showed large mountain backgrounds at different distances in all seasons as well as various photographs taken from the point of view of mountain hikers. Urban pictures were almost exclusively taken from the point of view of someone walking in towns ranging from small villages to large cities. Photographs depicted streets, buildings, houses, public squares, etc., from many places around the world. Indoor scenes were photographs taken from inside various urban constructions like houses, apartments, churches, museums, and stores, and so on.

There was no overlap in the pictures from the four target categories: Sea scenes did not contain harbours or mountains in the background; mountain scenes did not contain villages; street scenes did not include streets constructed along the sea, etc.

Procedure. Image presentation and behavioural response measurement were carried out using the software Presentation (NeuroBehavioral Systems; http://nbs.neuro-bs.com/). Subjects sat in a dimly lit room at 100 cm from a computer screen (horizontal resolution = 1024 pixels, vertical resolution = 768 pixels, vertical refresh rate: 75 Hz) piloted by a PC computer. To start a block of trials, they had to place their finger on an infrared response pad for 1 s. A trial was organized as following: A fixation cross (0.1° of visual angle) appeared for 300–900 ms and was immediately followed by the stimulus presented for two frames, i.e., about 26 ms, in the middle of the screen. These brief presentations prevented any exploratory eye movements. Participants had to lift their finger as quickly and as accurately as possible (go response) each time a target scene was presented and to withhold their response (no-go response) when the photograph did not belong to the target category. Responses were detected using infrared diodes. Subjects had 1000 ms to respond; longer reaction times were considered as no-go responses. This maximum response time delay was followed by a 300 ms black screen, before the fixation point was presented again for a variable duration, resulting in a random 1600–2200 ms intertrial interval.

Subjects were tested in two experimental sessions on two different days. A given picture category was the target category for four consecutive series. In each session they performed two categorization tasks for a total of eight blocks of 96 trials with target and nontarget trials being equiprobable. This led to a total of 3072 trials per subject. The order in which the subjects performed the four category tasks was counterbalanced across subjects. In a given task, the 48 nontarget images belonged equally to the three other environmental categories.

Thus, when performing the sea categorization task, a 96 trial series contained 48 target sea pictures, 16 nontarget mountain scenes, 16 nontarget urban scenes, and 16 nontarget indoor scenes. To avoid any bias, the design was organized so that across subjects each of the 384 pictures of a given category was seen 12 times as a target and 12 times as a distractor. Furthermore, when seen as a distractor, each image appeared the same number of times in the three different categorization tasks. Subjects had one training block of 48 images before starting the four series of a given categorization task. Training pictures were not used during testing.

Performance was evaluated by determining the percentage of correct trials and the latency (computed between stimuli onset and finger lift) at which subjects triggered their finger movement response. When repeated measures ANOVA were used, a Greenhouse-Geisser correction for nonsphericity was applied.

Results

Performance in the four tasks was evaluated by analysing separately accuracy and reaction times (RT). A summary of the results can be found in Table 1.

Accuracy. Subjects performed remarkably well in the four tasks. Mean accuracy was 96.2% correct in the sea categorization task; 95.6% in the mountain task; 95.5% in the indoor task; and 95.1% in the urban task. A one-way analysis on ranks was performed on these data and showed no significant effect, Friedman test: $\chi^2(3\ df) = 6.8$, n.s. We then analysed separately go and no-go responses.

In all four tasks subjects were better at responding on target scenes (97.4% correct) than at withholding their response on nontarget trials (93.8% correct). Correct go responses reached 98.1% in the sea task; 97.5% in the mountain task; 97.0% in the indoor task; and 97.1% in the urban task. These results were not homogenous, Friedman test: $\chi^2(3\ df) = 10.4$, $p = .016$, subjects scored better with the sea targets. Planned post hoc Wilcoxon tests showed that this higher accuracy with sea scenes reached significance when compared to indoor and mountain targets (both $Z < -2.3$, both $p < .02$). All other comparisons failed to reach significance.

Correct no-go responses reached 94.5% in the sea task; 93.6% in the mountain task; 93.9% in the indoor task; and 93.4% in the urban task. Sea and mountain scenes both belonged to natural categories, whereas indoor and urban scenes belonged to artificial categories. This means that, in each categorization task, one third of the distractors had a very strong relationship with the target category. Thus, the performance on distractors was studied separately depending on whether distractors were ''related'' or ''unrelated'' to the target category. Data were analysed using repeated measures ANOVA with category (four

TABLE 1
Experiment 1: Summary of results. Correct no-go related (or unrelated) accuracy refers to correctly categorized distractor images that belonged (or did not belong) to the same high-level category (natural vs. artificial scenes) as the target images

	Sea	*Mountain*	*Indoor*	*Urban*
Accuracy (%)				
Mean	96.2 (1.9) [91.1–99.7]	95.6 (2.6) [87.8–98.4]	95.5 (2.6) [88.0–98.4]	95.1 (2.7) [88.0–99.2]
Correct go	98.1 (2.4) [88.5–100]	97.5 (1.8) [92.2–99.5]	97.0 (2.6) [91.1–100]	97.1 (3.1) [86.0–100]
Correct no-go	94.5 (2.7) [88.5–99.5]	93.6 (4.5) [81.8–98.4]	93.9 (3.7) [84.9–99.0]	93.4 (4.8) [80.7–99.0]
Correct no-go related	85.9 (7.9)	83.5 (10)	84.3 (9.7)	83.8 (13)
Correct no-go unrelated	98.8 (1.2)	98.7 (2.0)	98.7 (1.2)	98.2 (1.7)
d'	3.9 (0.6) [2.7–5.7]	3.7 (0.4) [2.4–4.3]	3.7 (0.6) [2.4–4.6]	3.6 (0.6) [2.5–5.3]
RT (ms)				
Mean	422 (37) [348–494]	444 (46) [358–535]	466 (50) [373–555]	482 (45) [403–569]
Median	405 (37) [332–477]	425 (45) [334–499]	448 (47) [359–523]	463 (42) [384–538]
Minimal RT (ms)				
Overall data	260	290	300	300
Individual data	331 (28) [280–370]	346 (35) [290–420]	363 (32) [310–420]	372 (36) [310-460]

Standard deviation is indicated in brackets. Range of individual responses is indicated in square brackets [min-max].

levels) and related/unrelated (two levels) as within-subject factors. The analysis showed that subjects performed equally well at ignoring distractors regardless of the task, category factor: $F(2.6, 23) = 1.0$, n.s. However, correct no-go responses were strongly modulated by the categorical relationship between distractors and targets. Accuracy was significantly worse with distractors that belonged to the related category (84.4%) compared to the two others (98.6%), related/unrelated effect: $F(1, 23) = 82.5$, $p < .0001$. This was true for all categories (no interaction with the category factor; further confirmed by separate Wilcoxon tests on each category), all $Z < -4.1$, all $p < .0001$.

Reaction times. Although the analysis of accuracy did not reveal major differences between tasks, speed of processing measured by mean and median reaction times (RT) differed strongly between the four categorization tasks, Friedman tests: Both $\chi^2(3\ df) > 42$, both $p < .0001$. Mean and median RT were respectively 422/405 ms in the sea task; 444/425 ms in the mountain task; 466/448 ms in the indoor task; and 482/463 ms in the urban task. All two by two comparisons on mean and median RT were significant, Wilcoxon tests: All $Z < -2.6$, all $p < .01$, except the comparisons between the urban task and the indoor task (Figure 2A). Thus the four tasks were ranked according to processing speed as follows: (1) Sea, (2) mountain, (3) indoor = urban. These differences in processing speed can be observed in the RT distributions of Figure 3. Speed of processing was thus faster for the sea context and was also less variable (Figure 3, A, B, C, and D) as shown by a narrower RT distribution in the sea task compared to the mountain task (standard deviation: Sea = 37 ms, mountain = 46 ms, indoor = 50 ms, urban = 45 ms). Moreover this faster processing speed for sea pictures and, to a lesser extent, for mountain pictures, could be seen even on the fastest responses produced by the subjects. Thus, a complete shift of the RT distributions towards shorter latencies could be seen for sea and mountain pictures (Figure 3E). Expressing performance as cumulative d' curves as a function of time revealed that discriminative information was available earlier in the sea task than in the three other tasks and accumulated faster to reach a higher value (Figure 3F).

We then assessed more directly whether this average processing speed ranking of the four tasks was also true for the earliest responses that could be triggered. As target and distractor trials were equally likely, a random behaviour should equalize hits and false alarms; hence the minimal behavioural processing time was determined by the latency at which correct go responses started to significantly outnumber incorrect go responses, $\chi^2(1\ df)$, $p < .001$, using a noncumulated RT histogram with 10 ms time bins. Such early correct go responses cannot be considered as anticipations. The analyses were performed both on the overall data (pooling together all trials from all subjects), and for each subject separately. With the overall data set, the minimal processing time was 260 ms in the sea task, 290 ms in the mountain task, and 300 ms in both the

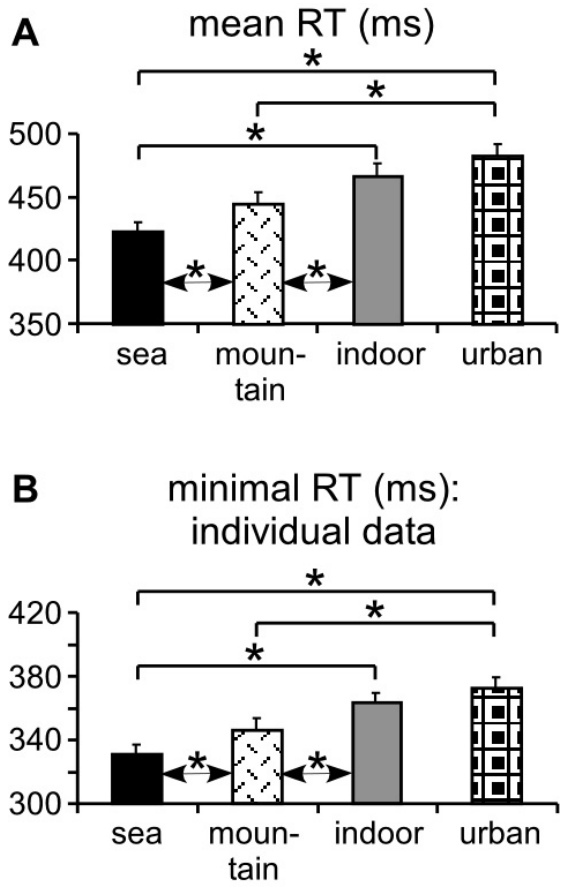

Figure 2. Mean and minimal reaction time with associated standard errors obtained for each of the four scene categorization tasks in Experiment 1. Asterisks indicate statistically significant differences (see text for details). In B, asterisks correspond to Wilcoxon tests where all $Z < -2.4$ and all $p < .02$.

indoor and the street task. Individual data (computed using cumulated RT histograms with 10 ms time bins) confirmed this tendency with a mean individual minimal processing speed of 331, 346, 363, and 372 ms respectively for sea, mountain, indoor, and urban target photographs (Figure 2B).

In conclusion, natural environments could be classified faster than artificial environments and this was true for the whole range of responses produced, from the earliest to the latest responses. Furthermore, among natural environments, sea scenes presented a clear processing speed advantage over mountain scenes. Although the accuracy performance was not very different between the four

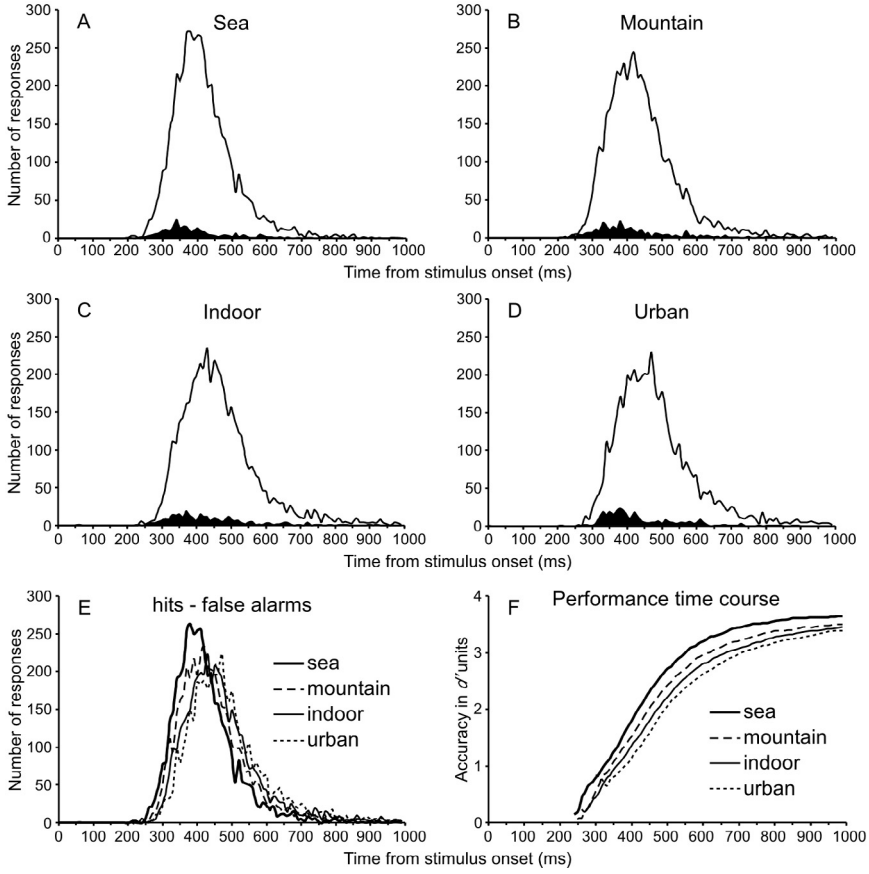

Figure 3. Time course of visual processing in Experiment 1. From A to D, RT distributions on correct (upper curve) and incorrect go responses (black histogram) are presented with the number of responses expressed over time, with 10 ms time bins. In E, false alarm distributions have been subtracted from hit distributions to allow a more direct comparison of the four tasks. In F, average performance accuracy (in *d'* units) is plotted as a function of processing time with 10 ms time bins. Cumulative numbers of responses were used. The *d'* was calculated from the formula $d' = z_n - z_s$, where zn is chosen such that the area of the normal distribution above that value is equal to the false-alarm rate, and where zs is chosen to match the hit rate. Note that the *d'* calculated here is not presumed to represent the actual distributions of signal and noise that underlie performance in the response time task. By taking into account the hit and FA rates in a single value at each time point, this time course of performance gives an estimation of the processing dynamics for the entire subject population. The plateau values correspond to the *d'* calculated from the overall accuracy results.

tasks, the rate of information processing was higher in the natural scenes (especially the sea scenes) compared to the artificial scenes.

Comparison with an object categorization task. Results presented above show that the gist of a natural scene flashed for 26 ms can be extracted both very efficiently and very quickly. But how fast is that processing compared to the categorization of objects in natural scenes? Previous studies from our laboratory have extensively assessed the performance of human subjects with the same go/no-go categorization task using animal as target category. A recent study (Rousselet et al., 2003) is particularly appropriate to compare the present human performance on global scene categorization with animal categorization because the same number of subjects were tested ($n = 24$) with the same number of trials per category, an identical set-up, the same image data bank, the same number of trials per category, and the same behavioural procedures (subjects had to alternate between the animal categorization task and a human face categorization task). A similar level of accuracy was also reached in the animal task (96.3%, n.s.) but the speed of processing was faster than with scenes. This faster processing was seen when using the median RT, which was significantly shorter in the animal task (371 ms) than in any of the four scene categorization tasks used here; two by two comparisons using Mann-Whitney tests: All $U < -2.8$, all $p < .005$. The fastest discriminative responses were found at the same latency than in the sea categorization task; thus earlier than for any other scene context, Mann-Whitney tests: All $U < -2.6$, all $p < .01$, an effect that was true for both the overall data set (260 ms in both tasks) and the individual data. On the other hand, performance accuracy increased more rapidly in the animal task than in the sea task. This can be seen on the RT distribution and even more clearly when accuracy performance is expressed in function of time by a *d'* curve (Figure 4).

Discussion

This first experiment confirmed that the general meaning of a visual scene can be extracted both very rapidly and highly efficiently with only brief visual presentations (e.g., Biederman, 1972; Potter, 1975, 1976). It also sets a minimal and an average processing time, respectively in the range of 260–300 ms and 400–460 ms, to extract this meaning from natural photographs. These figures are not very different from those obtained for object categorization in a variety of studies performed in our group. However a difference clearly emerged between natural scenes (sea, mountain) and artificial scenes (indoor and urban), natural scenes being categorized faster.

Among many properties that might be used to categorize natural scenes, the most obvious one, and probably the easiest to test, is colour (e.g., Torralba & Oliva, 2003). For instance, it is very plausible that colour was used as a

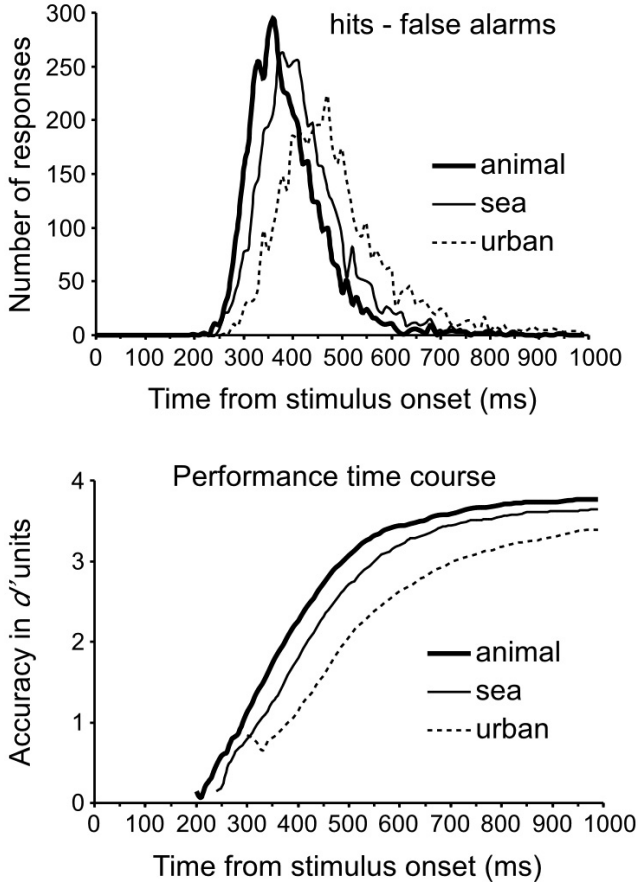

Figure 4. Comparison between scene and object categorization. Top panel: The RT distribution obtained in a preceding study (Rousselet et al., 2003) on an object categorization task (target = animal) is compared to the RT distributions obtained with the fastest (sea task) and the slowest (urban task) gist categorization tasks (false-alarm distributions have been subtracted from hit distributions to allow a more direct comparison). Bottom panel: The d' curves show that signal discrimination started earlier in the animal task compared to the sea task, a processing speed advantage that was maintained over the entire range of response latencies. For more details see text and caption of Figure 3.

diagnostic cue to categorize sea scenes. The influence of colour would be maximal in the sea task, because of large blue textured surfaces in the lower part of the picture and might explained the faster speed of processing in this task. The importance of colour cues depends on whether they constitute diagnostic features of the target category (Oliva & Schyns, 2000; Tanaka & Presnell, 1999). Colour contrasts may also be used to improve image segmentation, accelerating

image analysis (Gegenfurtner, 2003). For example, the importance of colour has already been demonstrated in a recognition test using natural scenes (Gegenfurtner & Rieger, 2000). Indeed, such diagnostic cues could allow the presetting of specific groups of ''diagnostic'' neurons and speed up the processing of the expected visual information (Delorme, Rousselet, Macé, & Fabre-Thorpe, 2004).

The use of colour cues in the rapid categorization of objects has been investigated and it was unexpectedly shown that removing chromatic information had little effect on average accuracy and speed of processing as well as on minimal processing speed (Delorme et al., 2000). But colour might not be such an efficient diagnostic cue in animal categorization as in global scene categorization. Experiment 2 was thus designed to test the effect of removing colour cues on the categorization of the gist of natural scenes.

EXPERIMENT 2

In this second experiment, we wanted to assess more specifically how the removing of colour information from natural scenes would impair performance and whether colour could be considered as one of the cues mediating the fast processing of natural scenes revealed in the first experiment. Thus, we tested another group of subjects using the same paradigm and the same set of images employed in Experiment 1. The difference between the two experiments was that half of the images were presented in colour and half in black and white (BW pictures; 256 grey levels). Colour and BW images were mixed at random in the series of stimulation in order to prevent subjects from relying on different strategies when categorizing colour and BW pictures and to allow a more direct comparison of human performance in these two conditions.

Method

Participants. Twenty-four adults (12 women and 12 men, mean age 30, ranging from 20 to 52, three of them left handed) volunteered in this study and gave their informed written consent. All participants had normal or corrected to normal vision. None of them had been tested in the first experiment.

Stimuli. The same set of natural scenes photographs used in Experiment 1 served as stimuli in Experiment 2. For each 24-bit (16 million colours) photograph, an 8-bit version (256 grey levels) was generated using Photoshop 5.5.

Procedure. The design of Experiment 2 was identical to the one of Experiment 1 except on two points. First, subjects were tested in a single session. Second, subjects were presented with 50% colour and 50% BW photographs. Thus, each experimental condition was subdivided into two colour

conditions. The design was counterbalanced so that across subjects each image was seen the same number of times in colour and in black and white.

Results

In Experiment 2, the mean accuracy for colour images was not significantly different from the one reached in the first experiment (95.1% and 95.6% correct respectively; between-subject analysis on ranks), Mann-Whitney test: $U = 281$, $Z = -0.4$, n.s. There was also a nonreliable tendency in favour of faster responses in the first experiment compared to the second (mean RT were 454 ms and 476 ms respectively), $U = 216$, $Z = -1.5$, n.s. As in Experiment 1, we analysed separately accuracy and reaction time data. A summary of the results can be found in Table 2.

Accuracy. Global accuracy was very good in Experiment 2. Data were entered in a repeated measure ANOVA with category (four levels) and colour (two levels) as within-subject factors. This analysis showed that the levels of accuracy reached with colour and BW images were not significantly different (colour = 95.1%; BW = 94.7%), $F(1, 23) = 3.8$, n.s. This was true for all four categories of natural scenes as there was no significant interaction between category and colour factors.

Contrary to Experiment 1, accuracy on go and no-go responses did not differ significantly from one another (go = 95.4%; no-go = 94.5%), $F(1, 23) = 1.2$, n.s. Separate analysis showed that the removal of colour cues had no significant effect on either go responses (colour = 95.5%; BW = 95.2%), $F(1, 23) = 0.7$, n.s., and no-go responses (colour = 94.8%, BW = 94.1%), $F(1, 23) = 3.3$, n.s. In addition, there was no difference in accuracy between the four tasks for both go responses, $F(2.2, 51) = 0.2$, n.s., and no-go responses, $F(2.4, 55) = 2.2$, n.s. There was no significant interaction between colour and category factors.

As previously found in Experiment 1, no-go responses were made more frequently toward distractors that belonged to the same higher level category as the targets (natural versus artificial). In other words, subjects proved much better at categorizing distractors that were unrelated (98.6%) than related (86.0%) to the target category, $F(1, 23) = 92.6$, $p < .0001$. The only effect induced by the removal of colour cue was seen on related distractors in the urban task: Indoor pictures were correctly ignored with a higher accuracy when presented in colour (88.2%) than in BW (82.9%), $Z = 2.3$, $p = .02$.

Reaction times. An ANOVA analysis showed that reaction times were affected both by the category of the target scene (category effect on both median and mean RT, both $F > 26$, $p < .0001$) and by the availability of colour cues (mean RT): $F(1, 23) = 9.1$, $p = .006$; median RT: $F(1, 23) = 5.2$, $p = .03$, so that

TABLE 2
Experiment 2: Summary of results. For each condition, the results are indicated separately for colour pictures (colour) and for grey-level pictures (BW)

	Sea		*Mountain*		*Indoor*		*Urban*	
	Colour	*BW*	*Colour*	*BW*	*Colour*	*BW*	*Colour*	*BW*
Accuracy (%)								
Mean	95.1 (2.6)	95.6 (2.8)	95.1 (3.4)	94.5 (2.8)	95.2 (4.3)	94.5 (3.5)	95.0 (3.5)	94.2 (3.1)
	[89–99]	[86–99]	[83–99]	[84–98]	[77–98]	[82–99]	[84–99]	[87–98]
Correct go	95.0 (3.2)	95.6 (3.4)	95.1 (5.2)	95.1 (4.5)	96.1 (7.1)	95.1 (5.3)	95.7 (4.8)	95.1 (4.5)
	[86–100]	[88–100]	[74–100]	[78–100]	[64–100]	[75–100]	[81–100]	[82–100]
Correct no-go	95.2 (3.8)	95.5 (3.4)	95.1 (3.3)	93.9 (3.2)	94.4 (3.7)	93.9 (4.0)	94.4 (4.6)	93.2 (4.7)
	[85–99]	[85–100]	[86–100]	[88–99]	[86–99]	[86–100]	[80–100]	[83–100]
Correct no-go related	87.2 (10.8)	88.8 (9.2)	87.1 (8.8)	83.7 (8.1)	86.2 (8.7)	84.1 (10.3)	88.2 (7.4)	82.9 (11.3)
Correct no-go unrelated	99.2 (1.5)	98.8 (1.5)	99.0 (1.9)	99.0 (1.6)	98.5 (1.9)	98.8 (2.0)	97.5 (3.8)	98.4 (1.9)
d'	3.5 (0.6)	3.6 (0.6)	3.5 (0.6)	3.5 (0.6)	3.7 (0.6)	3.5 (0.8)	3.6 (0.6)	3.4 (0.6)
	[2.4–5.2]	[2.2–4.5]	[2.1–4.9]	[2.0–4.9]	[1.7–4.3]	[1.9–5.0]	[2.0–5.2]	[2.3–4.6]
RT (ms)								
Mean	443 (56)	461 (60)	471 (58)	462 (63)	493 (70)	503 (65)	498 (55)	499 (58)
	[300–566]	[323–586]	[359–603]	[332–595]	[344–640]	[352–632]	[347–599]	[360–636]
Median	429 (56)	444 (61)	452 (59)	443 (61)	475 (70)	485 (67)	479 (57)	478 (61)
	[288–549]	[310–589]	[336–572]	[320–572]	[314–619]	[324–625]	[318–580]	[322–615]
Minimal RT (ms)								
Overall data	290	300	310	310	320	340	330	310
Individual data	388 (45)	398 (50)	401 (50)	400 (54)	425 (64)	434 (55)	429 (48)	432 (53)
	[280–490]	[280–520]	[300–500]	[290–510]	[300–600]	[320–560]	[310–520]	[310–580]

For other details see Table 1 caption.

speed of performance was analysed separately on colour and BW pictures for each categorization task (Figure 5).

However the results were not consistent from one category to another. Whereas sea and indoor pictures were categorized faster in colour than in BW—about 15 ms and 10 ms faster respectively; Wilcoxon tests: both $Z < -3.2$, both

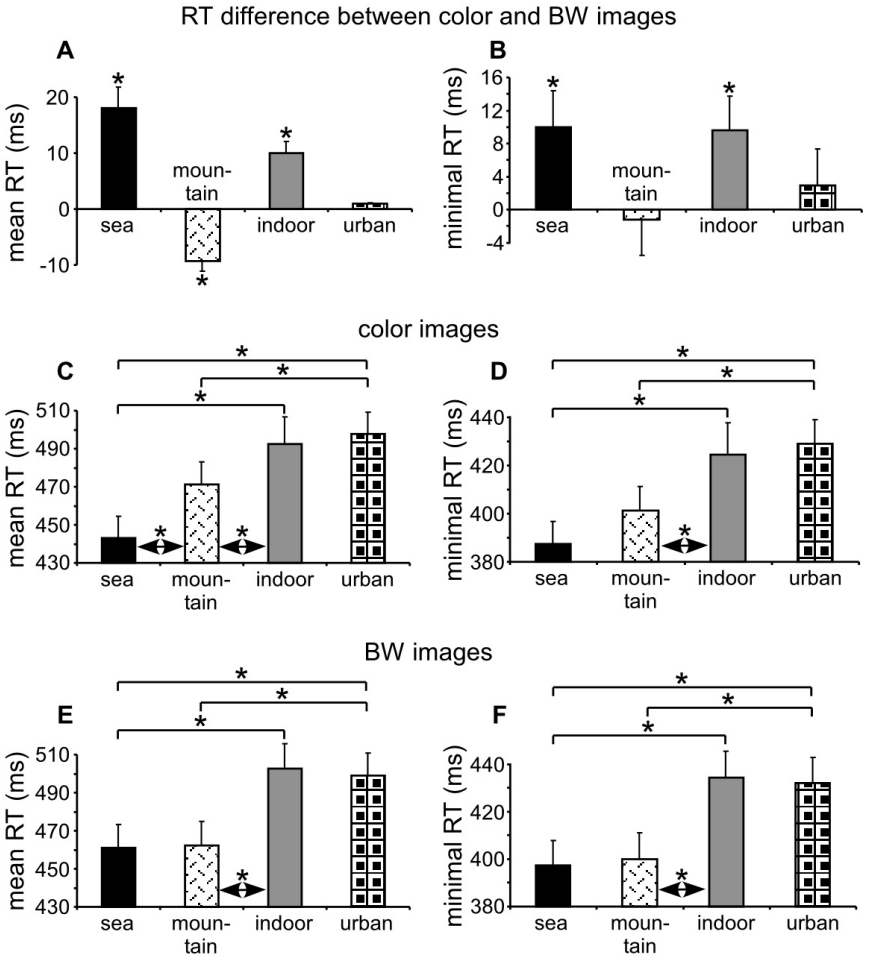

Figure 5. Speed of processing in Experiment 2: Mean RT (left column) and minimal RT (right column). The top two graphs illustrate the colour processing speed advantage by subtracting, for each of the four categorization tasks, the value obtained with colour images from the value obtained with BW images. Mean reaction times for each categorization task and associated standard errors are shown for colour images (middle panel) and for BW images (bottom panel). An asterisk shows statistically significant between-category effects.

$p < .001$, for both mean and median RT—urban scenes showed no effect of colour cue removal—mean and median RT: Both $Z > -0.7$, both n.s.—and, surprisingly, mountain pictures showed a tendency to be categorized with a slower speed in colour (9 ms slower) than in BW. This tendency reached significance for mean RT, $Z = -2.2$, $p = .03$, but not for median RT which presented only a borderline effect, $Z = -2.0$, $p = .05$.

Regarding differences in processing speed between the four categories, the results obtained with colour images tested separately showed the robustness of the results obtained in Experiment 1. Indeed, as in this first experiment, all two by two comparisons were significant for both mean and median RT, all $Z < -5$, all $p < .02$, except between indoor and urban scenes. Thus the two artificial categories were categorized at about the same speed.

When BW pictures were analysed separately, the same pattern of results appeared again for both mean and median RT, all $Z < -3.5$, all $p < .0001$, with the exception that the two natural categories (mountain and sea pictures) were categorized at the same speed.

The limited effect on performance speed linked to the removal of colour information when extracting the gist of natural scenes can be seen in RT distributions (Figure 6). Colour and BW RT distributions were virtually superimposed in the case of urban pictures and the amplitude of the shift towards shorter RT (for sea and indoor scenes) or towards longer RT (for mountain scenes) was indeed limited. The time course of performance (Figure 6, insets) again shows a small effect of colour removal on the subjects' capacity to discriminate between targets and distractors. Cumulated d' curves were virtually superimposed from the earliest responses to the plateau, with very similar slopes, indicating that information accumulated at a similar speed for colour and BW pictures. Regarding the small differences in plateau value, two by two one-way analyses on ranks revealed only one significant difference, namely that signal detection was slightly higher with colour images compared to BW images in the urban task (colour = 3.4; BW = 3.3; $Z = -2.1$, $p = .03$; other comparisons were not significant).

Discussion

Two main points have been stressed by the results obtained in these two studies. First, human subjects are both very accurate and very fast at categorizing the gist of real-world scenes with a processing speed advantage in favour of natural scenes compared to artificial scenes. Second, colour cues do not appear to be essential to trigger fast and accurate behavioural responses (some subjects were even unable to notice that half of the scenes were presented in BW); however, for certain categories where colour has a diagnostic value, there can be a small speed-up that even affect the fastest responses. Moreover, comparing the present

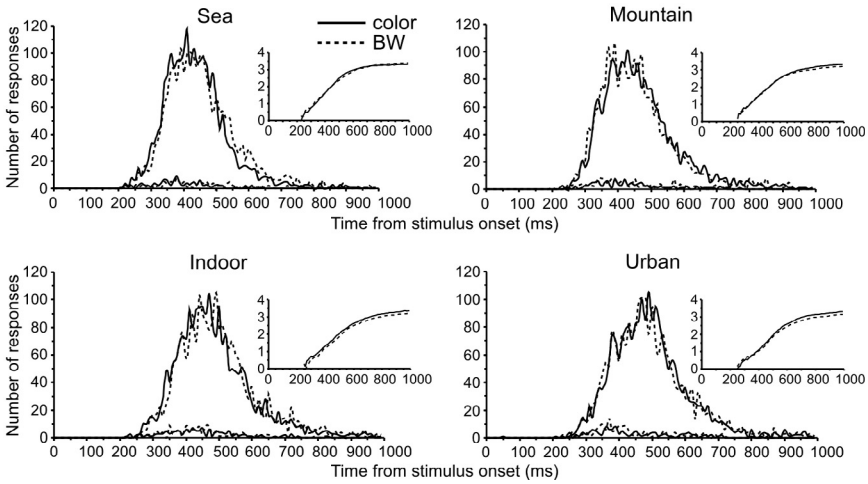

Figure 6. Reaction time distributions for correct hits (top two lines) and for false alarms (bottom two lines) for the four scene categorization tasks in Experiment 2. The graphs compare for each target category the reaction time distributions associated with colour (solid line) and BW images (dotted line). Each time, the insets show the d' computed from the cumulative number of responses in the RT histograms in both conditions. For details see caption of Figure 3.

results with those obtained for object categorization using the same task and the same set-up provides information on the possible interactions between the processing of objects and of the context in which they are presented.

Scene categorization is accurate and fast, but faster for natural scenes. Experiment 2 confirmed that subjects could extract very rapidly the global context of a natural scene. It also strengthened the finding that natural scenes can be processed faster than artificial scenes, and confirmed a slight processing speed advantage of sea images over mountain images.

Overall, Experiments 1 and 2 showed that human subjects can differentiate complex scene categories with median RT ranging from about 400 to 460 ms but with early responses observed at latencies that can be as short as 260–300 ms. This processing time is remarkably short compared to RT observed when subjects are simply asked to detect the appearance of a natural scene on the screen. Indeed, human subjects are able to detect that a scene has appeared, whatever its category, with a mean RT of about 230 ms (VanRullen & Thorpe, 2001) or even shorter (Rousselet, unpublished data: 20 subjects, mean RT = 211 ms). Thus, using a mean RT of 230 ms in such detection task as a reference, the additional cost to realize a complex gist categorization task is on average 170–230 ms, but can be as low as 30–70 ms for early responses. These strong temporal constraints

are in favour of models of visual processing relying essentially on coarse, feedforward, and massively parallel mechanisms to achieve scene recognition (Oliva & Torralba, 2001; Torralba & Oliva, 2003; VanRullen & Thorpe, 2002).

Both experiments also revealed a clear processing speed advantage in favour of natural scenes compared to artificial scenes. This advantage was not mediated by colour cues, given that the same was true with BW pictures. Aside from colour, several combinations of low-level scene-based properties could potentially explain how subjects classified pictures in these two experiments (Oliva & Torralba, 2001; Torralba & Oliva, 2003). But the bias towards natural scenes might find its explanation in a stronger variability in the physical properties of artificial environments compared to the natural environments used in these experiments. Indeed, although we used very varied natural scenes, they were still limited to sea and mountain pictures. On the other hand, indoor and outdoor artificial scenes covered almost the whole collection of imaginable artificial scenes. This resulted in a more constrained dictionary of physical properties for natural compared to artificial scenes, which could recruit more specifically preset groups of ''diagnostic'' neurons responding to target properties in the natural scene images. The stronger variability in artificial scenes compared to natural scenes can be appreciated from the average scenes and the examples in Figure 1 (B and C respectively). This in turn might allow faster decision making, on the basis of limited need for evidence, in the case of natural scenes compared to artificial scenes. Future experiments will be needed to evaluate more carefully how gist processing speed is affected by target diagnosticity, for instance by varying the physical similarities among targets and between target and distractor scenes.

Influence of colour cues in scene categorization. The second experiment provided evidence about the influence of colour information in task performance for the four scene contexts studied here. In the first place, it is surprising to note that colour had virtually no effect on subjects' capacity to detect targets. Thus, colour is not necessary to categorize real-world pictures in this fast go/no-go paradigm. Concerning the speed of processing of urban scenes, the lack of effect of removing the colour cues is in agreement with one previous report from Oliva and Schyns (2000). However, in the present study we found that colour helped the subjects at correctly ignoring nontarget indoor pictures in the urban task (artificial category-related distractors) as they scored better when such scenes were presented in colour than in BW. Colour also speeded up the processing of indoor scenes as subjects were also faster to make a decision on coloured than on BW indoor target pictures. This effect was not found by Oliva and Schyns (2000), and might have been strengthened by our protocol. The two artificial scene categories used in this experiment might share many low-level properties, forcing subjects to rely on distinctive features to discriminate between those two categories. One of these features might have been the diagnostic yellowish/

brownish colour present in many indoor scenes because of artificial lighting (see the colour version of Figure 1B online). Note that this small colour advantage was also seen on the earliest responses produced. Thus, in this task, colour information was used very early during the course of visual processing, in keeping with other reports (Delorme et al., 2004; Edwards, Xiao, Keysers, Foldiak, & Perrett, 2003; Gegenfurtner & Rieger, 2000; Goffaux et al., 2005).

Processing speed was also faster (15 ms on average) when colour was available in the sea task. This is not surprising, given that the blue of the sea was largely predictive of the category. However, colour alone does not appear to be able to explain the good performance in this task. First, BW sea pictures were also categorized very efficiently. Second, the blue feature was not a specific attribute of the sea category; it was present in large proportions in the skies of the mountain task and some urban pictures. Thus, it seems that the use of a single strategy relying on the diagnostic use of blue would not be sufficient to perform the task. But subjects could also rely on a slightly more sophisticated strategy based on the detection of a blue surface situated in the lower part of the scene (Delorme et al., 2004).

The use of blue textured surfaces as a cue in the sea task might also explain the unexpected negative effect of colour in the mountain task in which large surfaces of blue sky were often present and could induce ambiguity as a common feature with some of the distractors. Anyway, colour effects in this task (about 9 ms), contrary to the indoor and the sea task, did not affect the earliest responses, but essentially behavioural responses triggered with average RT larger than the median RT (>450ms).

The magnitude of the colour effects reported here is relatively small compared to the average 50 ms advantage in naming RT reported for colour pictures over BW pictures by Oliva and Schyns (2000). The category that is the more related to our *sea* category was their *coastline* category, which presented a colour effect of about 50 ms. This effect is much larger than the one reported here (about 15 ms). A very plausible interpretation for this discrepancy relies on the use of two different tasks, a go/no-go task in the present study and a naming task in Oliva and Schyns. It is possible that the necessity to name pictures, which was also associated with an additional 200–300 ms RT, forced subjects to rely on different representations, more sensitive to the colour factor. Another more plausible explanations stems in the fact that in the Oliva and Schyns experiment, subjects had to name a picture belonging to eight possible categories on every trial, while in our experiment subjects had a unique target for 384 consecutive trials, a protocol used to allow subjects to respond as fast as they could. This, in addition with the intermixed appearance of BW and colour pictures, might have biased subjects' strategy towards the use of noncolour properties like spatial frequencies, textures, depth of field, and other properties that have been shown to constitute valid cues for scene categorization (Torralba & Oliva, 2003). On the other hand, Goffaux et al. (2005) report a colour advantage of about 30 ms using a task very similar to the one we used here. Two main differences might

explain the discrepancy between the two studies. First, subjects in the Goffaux et al. study were presented with colour, grey, as well as abnormally coloured images, which might have affected their response strategy. It would be interesting to see if the 30 ms colour effect can be replicated when subjects are only presented with normal colour and grey-level images. Second, Goffaux et al. used four categories known to be colour diagnostic (deserts, forests, canyons, and coastlines), while our stimuli were not as well controlled in this respect. Thus, both targets and distractors could be categorized efficiently by relying on colour cues in their task, while only two of our categories were potentially colour diagnostic. The different results between the two studies are compatible with the idea that diagnostic cues are used whenever they are available and that diagnosticity in itself cannot be determined a priori but depends on such factors as the physical differences between targets and distractors (Schyns, 1998). In keeping with this idea, Delorme et al. (2000) showed virtually no effect of colour on the fast categorization of animals in natural scenes whereas a significant effect on accuracy was found with food objects.

Overall, there is now sufficient evidence showing that colour can be extracted very rapidly during the course of visual processing and can be used to improve performance when it has diagnostic value for the target scene category (Edwards et al., 2003; Gegenfurtner & Rieger, 2000; Goffaux et al., 2005; and the present study). However, our study also demonstrates that at least in some circumstances, colour is not a crucial feature for processing speed in a fast go/no-go categorization task of real-world scenes. More general conclusions about the importance of colour cues on scene categorization will need a more systematic investigation than the one presented in this paper.

Object categorization versus scene categorization. In Experiment 1, we performed a tentative comparison between object and scene categorization performances. This analysis revealed that although subjects were fast in categorizing a whole scene, this fast processing of scene gist was actually on average 30–90 ms slower than the processing of objects in natural scenes, like animals, but also like humans, human and animal faces, and means of transport (Delorme et al., 2000; Rousselet et al., 2003; VanRullen & Thorpe, 2001). This object advantage might very plausibly find its origin in the weaker structural constraints found in natural scenes. Indeed, the same gist can be assigned to scenes with relatively different low-level features and spatial arrangements. It is probably this relatively loosely defined structure of the scenes compared to component objects (like animals, vehicles, faces, etc.) that can explain their slower processing speed. However, this does not mean that scene categorization relies on higher level representations than object categorization. First, the fact that subjects made systematically more errors on distractors that belonged to the same higher level category as the target of the task (‘‘natural’’ vs. ‘‘artificial’’ categories) might be taken as an evidence for the use of relatively low-level cues in these tasks. Second, as we have argued recently (Rousselet et al., 2003), the

rapid categorization of objects like faces and animals in natural scenes might depend on coarsely defined features of intermediate complexity rather than on high-level complete descriptions (see also Ullman, Vidal-Naquet, & Sali, 2002). Thus, the slower processing of scenes compared to objects might find its explanation in the need to integrate in parallel a larger conjunction of relatively low-level features in order to reach a decision level in the processing of a natural context. The rapid categorization of objects in natural scenes would rely on the conjunction of a more limited number of features than the categorization of gist. Hence, because a given object category like animals can often be defined on the basis of a local combination of features, neurons coding for target objects in the ventral pathway would benefit from a finer task related top-down presetting in the animal task compared to the scene tasks. It is thus predicted that in a task where scene categories can be discriminated on the basis of a very limited set of basic features (e.g., natural vs. artificial scenes), processing speed could be even faster for scenes than objects, a hypothesis that we are currently testing. In addition, the difference in processing speed between objects and scenes could reflect the limitations of our visual system to process natural scenes in parallel. Indeed, although we recently demonstrated that two scenes can be processed in parallel (Rousselet, Fabre-Thorpe, & Thorpe, 2002), we have now evidence that this capacity is limited to certain conditions (Rousselet, Thorpe, & Fabre-Thorpe, 2004; VanRullen, Reddy, & Koch, 2004). Constraints on parallel processing would be even stronger on gist categorization because, to make a decision, it requires integrating low-level features over a wider spatial scale than for object detection.

However, these results should not be considered as an argument in favour of models postulating that there is no early interaction between scene and object representations (e.g., Henderson & Hollingworth, 1999). Indeed, in everyday situations, the gist of a scene is not changing abruptly from one image to the next but is much stable over time, probably allowing predictive hypotheses about possible objects to build up, as proposed by modern interactive frameworks (Bullier, 2001; Rao & Ballard, 1999; Ullman, 1995). But even here, using very short presentations, a considerable overlap was observed between the RT distributions of object and scene categorization. This overlap shows that the processing of an object might benefit from the simultaneous processing of the context in which it appears. The activation of congruent populations of neurons would probably allow a faster identification of a cow in a field than in a church. This issue clearly deserves further investigation.

CONCLUSIONS

Confirming earlier studies that have used brief visual presentations, data from the two experiments reported here showed that the gist of real-world scenes could be extracted with a high accuracy and with short RT in a fast go/no-go

visual categorization task. Furthermore, it was shown that natural scenes used in these experiments were processed faster than artificial scenes, probably because features of natural scenes might be more diagnostic than those of artificial scenes, allowing a stronger top-down presetting. In addition, we showed that colour information does not appear the most crucial feature used by human subjects to perform the fast go/no-go categorization tasks of real-world scenes studied here. Even if under certain conditions where colour is diagnostic of a specific scene category it can be used to produce a small performance increase, the fact that subjects can do so well with grey scale images shows that form-based information is typically self-sufficient.

REFERENCES

Aguirre, G. K., Zarahn, E., & D'Esposito, M. (1998). An area within human ventral cortex sensitive to ''building'' stimuli: Evidence and implications. *Neuron*, *21*(2), 373–383.

Bar, M., & Aminoff, E. (2003). Cortical analysis of visual context. *Neuron*, *38*(2), 347–358.

Biederman, I. (1972). Perceiving real-world scenes. *Science*, *177*(43), 77–80.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147.

Biederman, I. (1988). Aspects and extensions of a theory of human image understanding. In Z. W. Pylyshyn (Ed.), *Computational processes in human vision: An interdisciplinary perspective* (pp. 370–428). Norwood, NJ: Ablex.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*(2), 143–177.

Boyce, S. J., Pollatsek, A., & Rayner, K. (1989). Effect of background information on object identification. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 556–566.

Bullier, J. (2001). Integrated model of visual processing. *Brain Research Reviews*, *36*(2–3), 96–107.

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences*, *4*(5), 170–178.

Delorme, A., Richard, G., & Fabre-Thorpe, M. (2000). Ultra-rapid categorisation of natural scenes does not rely on colour cues: A study in monkeys and humans. *Vision Research*, *40*(16), 2187–2200.

Delorme, A., Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (2004). Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Cognitive Brain Research*, *19*(2), 103–113.

Edwards, R., Xiao, D., Keysers, C., Foldiak, P., & Perrett, D. (2003). Color sensitivity of cells responsive to complex stimuli in the temporal cortex. *Journal of Neurophysiology*, *90*(2), 1245–1256.

Epstein, R., Graham, K. S., & Downing, P. E. (2003). Viewpoint-specific scene representations in human parahippocampal cortex. *Neuron*, *37*(5), 865–876.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601.

Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience*, *13*(2), 171–180.

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, *108*(3), 316–355.

Ganis, G., & Kutas, M. (2003). An electrophysiological study of scene effects on object identification. *Cognitive Brain Research*, *16*(2), 123–144.

Gegenfurtner, K. R. (2003). Cortical mechanisms of colour vision. *Nature Reviews Neuroscience*, *4*(7), 563–572.

Gegenfurtner, K. R., & Rieger, J. (2000). Sensory and cognitive contributions of color to the recognition of natural scenes. *Current Biology*, *10*(13), 805–808.

Goffaux, V., Jacques, C., Mouraux, A., Oliva, A., Schyns, P. G., & Rossion, B. (2005). Diagnostic colours contribute to the early stages of scene categorization: Behavioural and neurophysiological evidence. *Visual Cognition*, *12*, 878–892.

Henderson, J. M. (1992). Object identification in context: The visual processing of natural scenes. *Canadian Journal of Psychology*, *46*(3), 319–341.

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology*, *50*, 243–271.

Henderson, J. M., & Hollingworth, A. (2003). Eye movements and visual memory: Detecting changes to saccade targets in scenes. *Perception and Psychophysics*, *65*(1), 58–71.

Hollingworth, A. (2003). Failures of retrieval and comparison constrain change detection in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 388–403.

Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General*, *127*(4), 398–415.

Hollingworth, A., & Henderson, J. M. (1999). Object identification is isolated from scene semantic constraint: Evidence from object type and token discrimination. *Acta Psychologica*, *102*(2–3), 319–343.

Hollingworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(1), 113–136.

Intraub, H. (1997). The representation of visual scenes. *Trends in Cognitive Sciences*, *1*(6), 217–222.

Maguire, E. A., Frith, C. D., & Cipolotti, L. (2001). Distinct neural systems for the encoding and recognition of topography and faces. *Neuroimage*, *13*(4), 743–750.

Marr, D. (1982). *Vision*. San Francisco, CA: Freeman.

Nakamura, K., Kawashima, R., Sato, N., Nakamura, A., Sugiura, M., Kato, T., Hatano, K., Ito, K., Fukuda, H., Schormann, T., & Zilles, K. (2000). Functional delineation of the human occipito-temporal areas related to face and scene processing: A PET study. *Brain*, *123*(9), 1903–1912.

Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, *34*(1), 72–107.

Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology*, *41*(2), 176–210.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

O'Regan, J. K. (1992). Solving the ''real'' mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology*, *46*(3), 461–488.

Potter, M. C. (1975). Meaning in visual search. *Science*, *187*(4180), 965–966.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(5), 509–522.

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology*, *81*(1), 10–15.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition*, *7*(1–3), 17–42.

Rensink, R. A. (2002). Change detection. *Annual Review of Psychology*, *53*, 245–277.

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience*, *3*(Suppl), 1199–1204.

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience*, *5*(7), 629–630.

Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *Journal of Vision*, *3*(6), 440–455.

Rousselet, G. A., Thorpe, S. J., & Fabre-Thorpe, M. (2004). Processing of one, two or four natural scenes in humans: The limits of parallelism. *Vision Research*, *44*(9), 877–894.

Sanocki, T., & Epstein, W. (1997). Priming spatial layout of scenes. *Psychological Science*, *8*(5), 374–378.

Sato, N., Nakamura, K., Nakamura, A., Sugiura, M., Ito, K., Fukuda, H., & Kawashima, R. (1999). Different time course between scene processing and face processing: A MEG study. *Neuroreport*, *10*(17), 3633–3637.

Schyns, P. G. (1998). Diagnostic recognition: Task constraints, object information, and their interactions. In M. J. Tarr & H. H. Bülthoff (Eds.), *Object recognition in man, monkey, and machine* (pp. 147–179). Amsterdam: Elsevier Science Publishers.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time and spatial scale dependant scene recognition. *Psychological Science*, *5*, 195–200.

Schyns, P. G., & Oliva, A. (1997). Flexible, diagnosticity-driven, rather than fixed, perceptually determined scale selection in scene and face recognition. *Perception*, *26*(8), 1027–1038.

Simons, D. J., Chabris, C. F., Schnur, T., & Levin, D. T. (2002). Evidence for preserved representations in change blindness. *Consciousness and Cognition*, *11*(1), 78–97.

Tanaka, J. W., & Presnell, L. M. (1999). Color diagnosticity in object recognition. *Perception and Psychophysics*, *61*(6), 1140–1153.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*, 391–412.

Ullman, S. (1995). Sequence seeking and counter streams: A computational model for bidirectional information flow in the visual cortex. *Cerebral Cortex*, *5*(1), 1–11.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, *5*(7), 682–687.

VanRullen, R., Reddy, L., & Koch, C. (2004). Visual search and dual-tasks reveal two distinct attentional resources. *Journal of Cognitive Neuroscience*, *16*(1), 4–14.

VanRullen, R., & Thorpe, S. J. (2001). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception*, *30*(6), 655–668.

VanRullen, R., & Thorpe, S. J. (2002). Surfing a spike wave down the ventral stream. *Vision Research*, *42*(23), 2593–2615.

Wolfe, J. M. (1999). Inattentional amnesia. In V. Coltheart (Ed.), *Fleeting memories* (pp. 71–94). Cambridge, MA: MIT Press.

Article n°3
2007

Processing scene context: Fast categorization and object interference

---

Vision Research, 47 (26) 3286-3297

Olivier R Joubert, Guillaume A. Rousselet, Denis Fize,
& Michèle Fabre-Thorpe

Available online at www.sciencedirect.com
ScienceDirect
Vision Research 47 (2007) 3286–3297
Vision Research
www.elsevier.com/locate/visres

Processing scene context: Fast categorization and object interference

Olivier R. Joubert [a,b], Guillaume A. Rousselet [c], Denis Fize [a,b], Michèle Fabre-Thorpe [a,b,*]

[a] Université, Toulouse 3, CerCo, UPS, France
[b] CNRS, UMR 5549, Faculté de Médecine de Rangueil, 31062 Toulouse cedex 9, France
[c] Centre for Cognitive Neuroimaging (CCNi), Department of Psychology, University of Glasgow, UK

Received 28 November 2006; received in revised form 26 June 2007

Abstract

The extent to which object identification is influenced by the background of the scene is still controversial. On the one hand, the global context of a scene might be considered as an ultimate representation, suggesting that object processing is performed almost systematically before scene context analysis. Alternatively, the gist of a scene could be extracted sufficiently early to be able to influence object categorization. It is thus essential to assess the processing time of scene context. In the present study, we used a go/no-go rapid visual categorization task in which subjects had to respond as fast as possible when they saw a "man-made environment", or a "natural environment", that was flashed for only 26 ms. "Man-made" and "natural" scenes were categorized with very high accuracy (both around 96%) and very short reaction times (median RT both around 390 ms). Compared with previous results from our group, these data demonstrate that global context categorization is remarkably fast: (1) it is as fast as object categorization [Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience, 13*(2), 171–180]; (2) it is faster than contextual categorization at more detailed levels such as sea, mountain, indoor or urban contexts [Rousselet, G. A., Joubert, O. R., & Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition, 12*(6), 852–877]. Further analysis showed that the efficiency of contextual categorization was impaired by the presence of a salient object in the scene especially when the object was incongruent with the context. Processing of natural scenes might thus involve in parallel the extraction of the global gist of the scene and the concurrent object processing leading to categorization. These data also suggest early interactions between scene and object representations compatible with contextual influences on object categorization in a parallel network.

Keywords: Natural scenes; Fast categorization; Context categorization; Object–context interactions; Parallel processing; Congruency; Object saliency

1. Introduction

Previous studies from our group have demonstrated the very high accuracy and fast speed of the visual system in categorizing different kinds of objects like animals, humans, means of transport or food items. Images flashed for about 20 ms are typically categorized by human observers with high accuracy (94% correct or more), median reaction times around 400 ms, and shortest response latencies around 250 ms (Delorme, Richard, & Fabre-Thorpe, 2000; Fabre-Thorpe et al., 2001; Fabre-Thorpe, Richard, & Thorpe, 1998; Rousselet, Macé, & Fabre-Thorpe, 2003; Thorpe, Fize, & Marlot, 1996; VanRullen & Thorpe, 2001). These short reaction times provide an upper estimate of processing time, as they include the time necessary not only for image processing, but also decisional and motor mechanisms (Bacon-Macé, Macé, Fabre-Thorpe, & Thorpe, 2005; VanRullen & Thorpe, 2001). Despite this limitation, experiments on object categorization in natural scenes have been instrumental in providing temporal constraints on object processing speed.

But typically, these experiments have ignored the relationship between target objects and other elements in the scene. Indeed, in pictures of natural scenes, objects are never isolated; they are seen on a background, surrounded by other objects and various contextual elements.

* Corresponding author.
E-mail addresses: olivier.joubert@cerco.ups-tlse.fr (O.R. Joubert), michele.fabre-Thorpe@cerco.ups-tlse.fr (M. Fabre-Thorpe).

0042-6989/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2007.09.013

Therefore, it is important to determine to what extent scene context might influence object recognition. Information relative to the context of a scene, like semantic consistency (Biederman, Mezzanotte, & Rabinowitz, 1982; Boyce & Pollatsek, 1992; Ganis & Kutas, 2003; Palmer, 1975) or repeated spatial configuration (Chun, 2000), could interact with object information by either facilitating or impairing object visual search and object processing. Although there is strong evidence that the processing of objects is influenced by contextual information, it is still unclear whether context might facilitates object recognition *per se* or might instead facilitate later stages of processing, for instance a decision making stage (Ganis & Kutas, 2003; Henderson, 1992; Henderson & Hollingworth, 1999; Hollingworth & Henderson, 1998). However, in this debate, we lack information about the speed of processing of contextual information, a crucial element needed to determine how early context information might be able to influence object recognition. Mechanisms by which scenes are recognized are still poorly understood, in part because of their complexity. Scenes not only contain objects, but also several non-movable elements with fixed spatial locations such as floor, walls, ceiling, sky, fields, trees, etc. which contribute to the 'gist' of the scene. Different layouts of such fixed elements might rely on different global image features such as spatial envelope properties (openness, naturalness, expansion, symmetry, Oliva & Torralba, 2001, 2006). The fast extraction of such spatial structure of a scene would allow an estimation of the meaning of the scene. Beside this "scene-centered approach", other theories describe scene recognition as the result of the successful identification of some objects in the scene (Friedman, 1979), or the evaluation of spatial links between objects (De Graef, Christiaens & d'Ydewalle, 1990). According to these hypotheses, objects would be systematically processed before scenes (see also Biederman, 1987; Riesenhuber & Poggio, 2000).

A strong argument against these theories is the demonstration that the gist of a scene can be accessed rapidly and accurately even when an image is displayed too briefly to allow an exhaustive processing of the objects in the scene (Biederman, 1972; Biederman et al., 1982; Oliva & Schyns, 1997, 2000; Potter, 1975; Rousselet et al., 2005). The fast processing of briefly presented natural scenes might be explained by the existence of scene specific features that might be used to categorize a scene independently of the objects it contains. To perform scene categorization tasks, subjects could rely on low-level features such as patches of diagnostic colours (Goffaux, Jacques, Mouraux, Oliva, Schyns, & Rossion, 2005; Oliva & Schyns, 2000; Schyns & Oliva, 1994). Alternatively, the spatial structure of the scene might be sufficient on its own to identify scene contexts (Henderson & Hollingworth, 1999; Oliva & Schyns, 2000; Sanocki & Epstein, 1997). Indeed, scene context can still be extracted from filtered scenes containing only low spatial frequencies at which objects cannot be categorized (Schyns & Oliva 1994). Moreover, modelling work suggests that scene classification could rely on specific visual filters that would capture the 'layout of the scene', (Oliva & Torralba, 2001, 2006; Torralba & Oliva, 2003). Such global image signature could be used to determine the general meaning of the scene, or 'gist'. This framework is consistent with the idea that a high-level categorization process does not necessarily depend on high-level representations if representations of lower levels are sufficient to categorize a stimulus in a given task (Schyns, 1998; Ullman, Vidal-Naquet, & Sali, 2002).

Overall, the literature suggests that fast processing of scene context relies to a large extent on visual information that is independent from that used to perform object categorization. However, whether scenes can be categorized as fast or even faster than objects is still a much debated question. Recently, by using a go/no-go paradigm in a 'gist' categorization task, we showed that subjects could discriminate "sea", "mountain", "indoor" and "street" scenes with a very good accuracy (>90%) and short median reaction times (RT) (400–460 ms) (Rousselet et al., 2005). Although such reaction times are relatively fast, object categorization can be faster, with median RT around 400 ms for animal targets (Delorme et al., 2000; Delorme, Rousselet, Macé, & Fabre-Thorpe, 2004; Fabre-Thorpe et al., 1998; Fize, Fabre-Thorpe, Richard, Doyon, & Thorpe, 2005; Rousselet et al., 2003; Thorpe et al., 1996; VanRullen & Thorpe, 2001). However, the reaction time distributions for scenes and objects categorization showed a considerable overlap, arguing against the idea of a systematic processing speed advantage for objects over scenes and leaving open the possibility of large interactions between the two systems in a parallel network.

In the present study, we used broader categories such as natural contexts and man-made contexts. Human subjects might be faster at categorizing scene context at a more general level than the 4 categories (mountain, sea, indoor, and street) used in our previous experiment, allowing more time for interaction between object and context processing. To test this hypothesis, we used the same fast visual categorization task but subjects were asked to categorize the briefly flashed photographs as either "natural" or "man-made" environments. Indeed, compared to the "sea/mountain/indoor/street" experiment, subjects were faster at completing the task. Moreover, when scenes required long processing times to be categorized, a post-hoc analysis revealed a strong interference due to the presence of salient objects.

2. Methods

2.1. Participants

Twelve volunteers (8 men and 4 women, mean age 31, range 23–39, 3 of them left handed) gave their informed written consent. All of them had normal or corrected to normal vision.

2.2. Stimuli

We used photographs of natural scenes from a large commercial CD-ROM library (Corel Stock Photo Libraries). Images (either horizontal or vertical) were in 24-bits jpeg format (16 millions colours), with a size of 768×512 pixels sustaining approximatively a visual angle of $16° \times 11°$. The 1440 images were selected in order to represent equally two categories,

"natural environment" and "man-made environment" (Fig. 1). Within each category, images (50% vertical, 50% horizontal) were as diverse as possible. The "natural environment" category was composed of 720 photographs in which 1/3 were sea scenes, 1/3 were mountain scenes, the last 1/3 contained desert, iceberg, forest or field scenes. The "man-made environment" category also contained 720 photographs in which 1/3 were street scene images (with or without pedestrians), 1/3 were indoor scenes like kitchens, museums, churches, etc., the last 1/3 being composed of aerial views of cities. None of the man-made scenes contained mountains or views of the sea, and none of the natural scenes contained buildings.

2.3. Tasks

Two experimental sessions were run, one with natural scenes as targets, the other one with man-made scenes as targets. The order in which they were performed alternated between subjects. Each session started by a detection task (one series of 96 trials), followed by a categorization training period (48 trials), and the categorization task itself (6 series of 96 trials for each of the two categorization tasks).

During the detection task (go-only task), subjects had to lift their finger as quickly as possible whenever an image appeared, independently of its category (natural and man-made environments were equally likely). The goal of this task was to check the lack of any low-level saliency bias between the two sets of scenes. Stimuli were only seen once in one of the 3 conditions (detection/training/categorization) either as target or distractor.

During the go/no-go categorization tasks, subjects had to lift their index finger as quickly and as accurately as possible (go responses) each time the scene belonged to the target category, and to withhold their responses (no-go responses) when it was a distractor. Each series of categorization tasks contained 50% target trials and 50% distractor trials. In order to avoid possible biases among image sets, conditions were counterbalanced among subjects so that images presented as targets to half of the subjects were presented as distractors to the other subjects.

2.4. Procedure

Subjects sat in a dimly lit room, at 1 m from a computer screen (resolution 1024×768, vertical refresh: 75 Hz) piloted by a PC. Image presentation and measurement of behavioural responses were carried out using the software Presentation (NeuroBehavioral Systems, http://nbs.neurobs.com/).

For both the detection task and the categorization task, each trial started with a fixation cross (0.1° of visual angle) that appeared at the center of a black screen for 300–900 ms randomly, immediately followed by the stimulus presented for two frames (26 ms), also in the middle of the screen. These brief presentations prevented exploratory eye movements.

Motor responses were detected using infrared diodes. For each image (detection task) or for target-images (categorization tasks), subjects had to respond in less than 1000 ms by lifting their finger. Longer reaction times were considered as no-go responses. Following this 1 s period, a black screen was displayed during 300 ms, before the next trial started with the presentation of the fixation cross. A trial lasted between 1600 and 2200 ms.

3. Results

3.1. Detection task

Performance was analyzed separately according to the image category ("man-made" vs. "natural") in order to rule out the existence of low-level saliency biases between

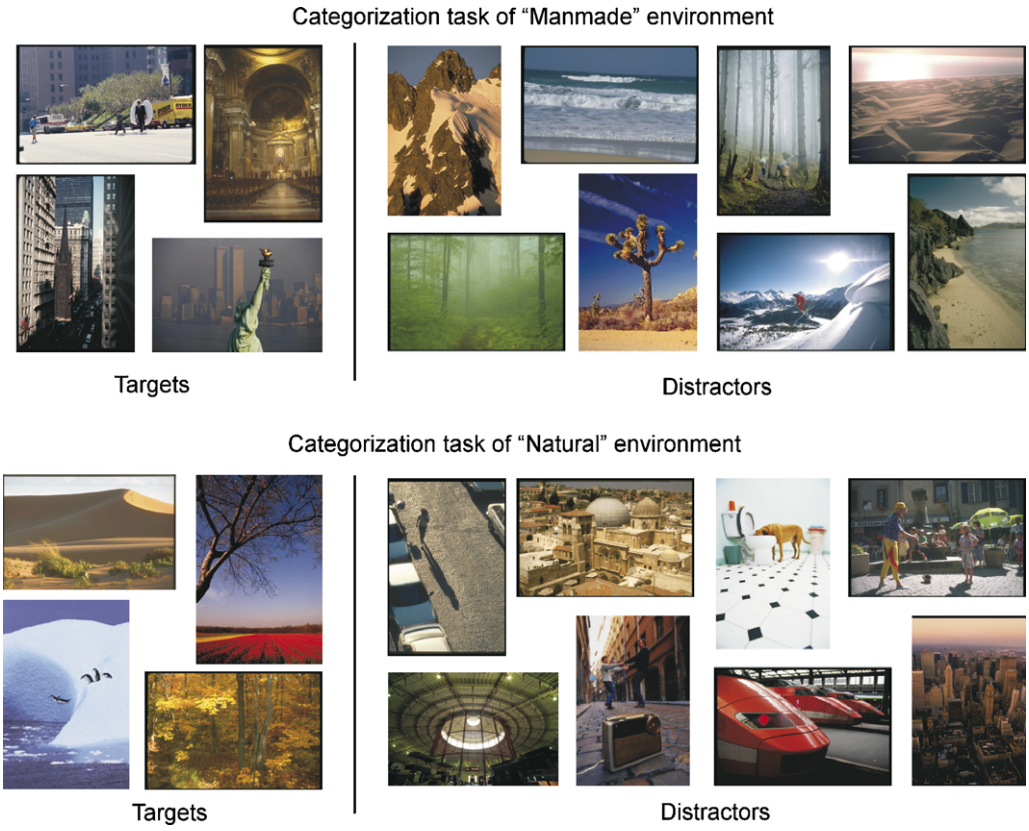

Fig. 1. Tasks and stimuli. In each of the two categorization tasks (man-made or natural environments), the target images were used as distractors in the other task. Note the large variety of stimuli.

the two image groups. Response latencies were very short and extremely similar for the two image categories: the RT distributions for the two image sets were virtually superimposed, and the mean of individual median RT were not statistically different (man-made: 208 ms; natural: 209 ms; paired Wilcoxon test: $Z = -.713$, $p = .476$). Moreover, for this detection task, we computed the minimal RT as the latency from which on the number of correct "go" responses significantly differed from zero. Again, no difference was found (mean of individual minimal RT: 171 ms for both image categories, n.s., paired Wilcoxon tests, $Z = -.863$, $p = .388$). Thus, these results demonstrate that there was no low-level saliency bias between the "man-made" and "natural" sets of selected images, at least in a simple detection task.

3.2. Categorization task

All individual results for the two categorization tasks are summarized in Table 1.

3.2.1. Accuracy

Subjects were very efficient at performing the two categorization tasks, with an accuracy of 96.8% for "natural" pictures and 96.2% for "man-made" pictures (n.s., paired Wilcoxon test: $Z = -1.491$, $p = .136$). Incorrect trials were biased towards false alarms in both tasks. Indeed, subjects were very precise at categorizing target images with only 1% of missed targets in the "man-made" category and 0.9% in the "natural" category. However, they erroneously categorized as targets 6.6% of the distractors in the "man-made" task and 5.4% in the "natural" task. These biases did not differ between the two tasks (paired Wilcoxon test, false alarms: $Z = -1.373$, $p = .17$; missed targets: $Z = 0$, $p = 1$).

3.2.2. Reaction times

Subjects were not only very precise, they were also very fast. Median and mean RT were, respectively, 383 ms and 397 ms in the "man-made" categorization task and 393 ms and 407 ms in the "natural" categorization task (n.s., paired Wilcoxon test, median RT, $Z = -1.134$, $p = .182$; mean RT, $Z = -1.255$, $p = .209$).

As shown in Fig. 2, RT distributions of correct go responses were overall very well superimposed. To confirm this observation, we did a series of permutation tests (one test for each 10 ms time bin) using the number of correct go responses of each subject in each condition (man-made and natural). In a given time bin, all RT values were shuffled which is equivalent to assigning at random the man-made and natural labels at each RT value. The difference between the two "fake" distributions was then computed and stored. This procedure repeated 999 times provided a confidence interval around the mean difference and under the null hypothesis that the two conditions were actually sampled from the same population (Wilcox, 2005). None of the time bins showed a significant difference between the two distributions (1000 permutations, $\alpha = .05$). However, a difference could be seen at the shortest latencies by determining the minimal RT for each task. Here, the minimal RT was defined as the first 10 ms bin from which on the number of correct go responses significantly outnumbered the number of false alarms. The minimal RT computed by pooling together the data from all subjects was 220 ms in the "man-made" task and 280 ms in the "natural" task (Fig. 2). When averaging minimal RTs computed for each subject, the mean minimal RT was at

Table 1
Results of categorization tasks

Subject	Accuracy (%)						Reaction time (ms)					
	Mean		Correct go		Correct no-go		Minimal		Median		Mean	
	Man.	Nat.	Man.	Nat.	Man.	Nat.	Man.	Nat.	Man.	Nat.	Man.	Nat.
OJO	95.3	93.4	99	99.3	91.7	87.5	210	280	322	330	327	339
HKI	95.3	96.2	99	99	91.7	93.4	280	270	346	337	361	351
MFT	97.9	98.1	100	99.7	95.8	96.5	330	320	402	390	420	406
SGA	98.6	99.3	99.7	100	97.6	98.6	270	310	363	394	378	405
NBA	96.2	95.8	100	99.7	92.4	92	270	280	358	354	377	369
CLU	94.6	97.8	99.7	99.2	89.6	96.5	300	290	356	363	367	380
RVR	91.3	93.2	99.7	99.3	83	87.2	230	290	309	330	319	351
SQU	98.6	99.3	99	99.7	98.3	99	320	340	417	415	428	426
FAR	95.1	95.7	97.6	96.5	92.7	94.8	310	300	393	405	407	414
FLE	97.7	97.9	98.3	99.7	97.2	96.2	320	350	423	460	437	465
GFE	99.1	98.3	99	98.6	99.3	97.9	300	370	443	493	462	508
DFI	95	97	97.9	98.6	92	95.5	310	350	461	444	485	466
Mean	96.2	96.8	99	99.1	93.4	94.6	288	313	383	393	397	407
Std	2.3	2	0.8	0.9	4.6	3.9	37.2	33.1	47.9	53.3	51.4	52.7
Min	91.3	93.2	97.6	96.5	83	87.2	210	270	309	330	319	339
Max	99.1	99.3	100	100	99.3	99	330	370	461	493	485	508

The values indicated in the four bottom lines are computed from individual scores.

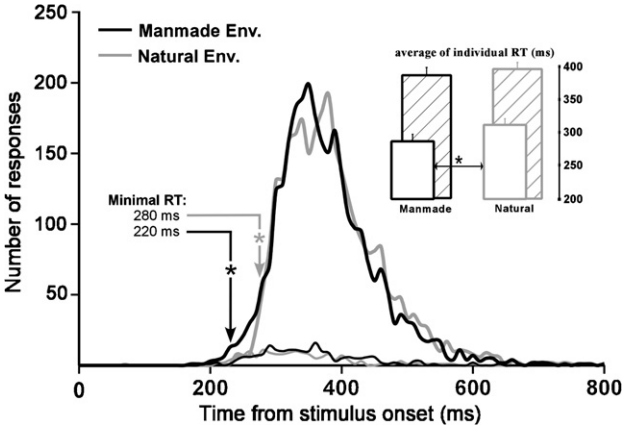

Fig. 2. Speed of responses in context categorization tasks. Reaction time (RT) distributions for correct go-responses (thick curves) and for false alarms (thin curves) are shown for the man-made environment categorization task (black curves) and the natural environment categorization task (grey curves) with the number of responses pooled across all subjects and expressed over time using 10 ms time bins. Minimal RT are determined as the first 10 ms time bin for which correct responses significantly exceed errors. (Top right) Average of individual performance values (see Table 1): minimal RT (empty rectangles) and median RT (hatched rectangles) with their associated standard errors. Asterisks indicate statistically significant differences (Wilcoxon tests, $p < .05$).

288 ms in the ''man-made'' task and 313 ms in the ''natural'' task (Table 1). This 25 ms delay for the natural categorization task compared to the man-made task was significant (paired Wilcoxon test: $Z = -2.136$, $p = .033$) and could be due to a larger set of diagnostic features in ''man-made'' scenes.

The presence of numerous cues within a single picture might be able to speed-up the first wave of visual processing. Diagnostic features for ''natural'' scenes could be more complex and/or of a higher diversity. Indeed, man-made environments contain more straight lines and right angles, as well as more high spatial frequencies than natural environments. Such elements could be diagnostic for models of scene categorization using global features like roughness, expansion, or openness of the scene (Oliva & Schyns, 2000; Oliva & Torralba 2001; Schyns, Jentzsch, Johnson, Schweinberger, & Gosselin, 2003).

The analysis of the reaction time distributions also showed some very long latency responses. Most of the images categorized with long RTs appeared to contain a salient object. A post-hoc analysis aimed at measuring the interference of salient objects in contextual categorization was thus performed in order to check whether the presence of an object could affect context categorization performance.

3.3. Interference of salient objects

Based on the visual inspection of scenes associated with long RTs, we hypothesized that salient objects might interfere with the processing of the background, for instance by capturing attention. Most of the time, salient objects were congruent with the context category; however in some cases salient objects were incongruent with the context, like a man-made object on a natural background or a large biological object (animal, tree, etc.) in a man-made scene (see examples Fig. 3). As the study was not planned for such purpose, incongruencies were more frequent in the case of ''natural'' scenes in our picture sets. In order to evaluate the impact of salient objects on background processing, we designed a protocol to determine as objectively as possible a set of images containing salient (congruent or incongruent) objects. Out of the 12 original subjects, 10 agreed to come again, and were asked to classify all the scenes used in this experiment according to whether or not they contained a salient object, and, if this was the case, whether the object was congruent with the background category of the scene. Subjects had unlimited viewing time. We gave subjects the following definition of saliency: ''which inevitably attracts your attention''. Congruency had to be judged with respect to the ''natural'' and ''man-made'' categories. Because of the frequent presence of humans in urban scenes, humans were only considered as congruent in ''man-made'' scenes.

Out of the 1152 photographs used in the two categorization tasks, 1111 (96.4%) were classified in the same set by 5 subjects or more. In this set of photographs, 948 were considered as presenting no salient object (nSO). In the remaining 163 that contained a salient object (SO), the object was evaluated as congruent (SCO) for 130 photographs and as non-congruent (SnCO) for 33 photographs (see examples, Fig. 3C–E). Based on this labelling, performance in the categorization task was further analyzed.

Subjects were clearly faster at categorizing scene context when no salient object was present (Fig. 3B and F). Indeed, nSO images were categorized with a median reaction time of 393 ms while median reaction time for SO photographs reached 418 ms, a 25 ms RT increase that was significant (paired Wilcoxon test: $Z = 3.062$, $p = .002$). A bootstrap simulation was run to determine if these median RT differences could be obtained by chance. In the simulation, n images (n being the number of images in a given subset (i.e. nSO: $n = 948$, SO: $n = 163$), were randomly sampled, with replacement, from the original pool of N images (i.e. total images: $N = 1111$). Then, the median reaction time for the randomly selected images was computed. This process was repeated 1999 times to compute a 95% confidence interval (CI) for median reaction times. Median RT for SO photographs (418 ms) was significantly longer than expected by chance (95% CI [382–402 ms]), whereas median RT for nSO images (393 ms) fell in the CI for nSO images (95% CI [388–396 ms]). The bootstrap simulation thus demonstrates that the slow down of natural scene processing in the presence of a salient object is very unlikely to be explained by chance.

Within the set of 163 SO images, object congruency also affected median reaction times (SCO images: 409 ms, SnCO images: 451 ms; Fig. 3B and F). The direct comparison of these two conditions did not reach

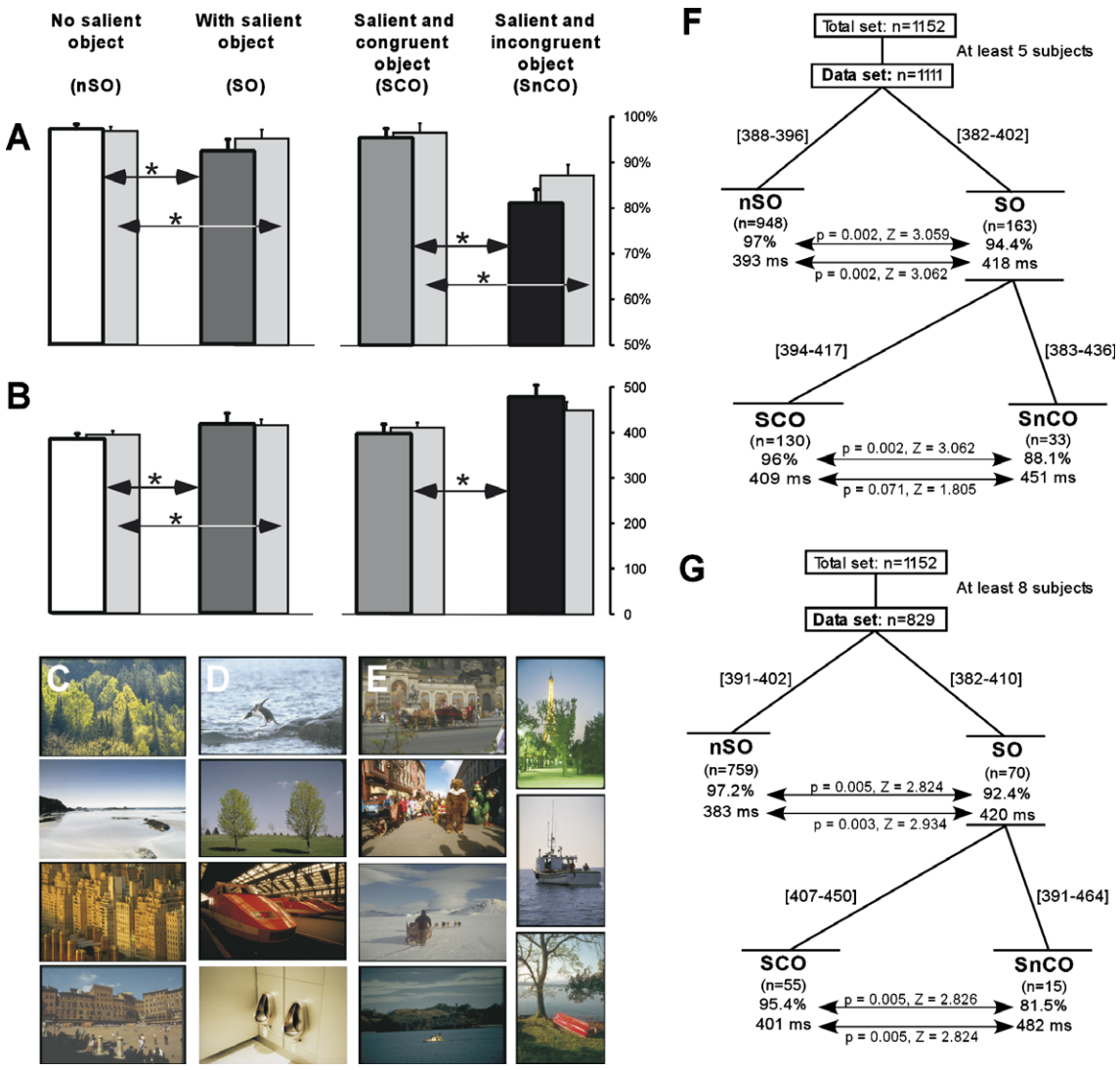

Fig. 3. Interference of salient objects. Performance associated with image sets defined by post-hoc analysis. Results are compared for images with no salient object (NSO) and images with a salient object (SO). Among SO images, performances are compared depending on whether the salient object was congruent (SCO) or incongruent (SnCO) with the environment. (A) Accuracy expressed as the percentage of correct responses with associated standard errors. (B) Median RT for correct go-responses (in ms) with associated standard errors. In (A) and (B), the histograms in front represent values using the image sets delimited by at least 8 of the 10 subjects (G), the histograms behind represent values using the image sets delimited by at least 5 of the 10 subjects. Asterisks indicate statistically significant differences. (C) Examples of stimuli without a salient object. (D) Examples of stimuli with a salient and congruent object. (E) Examples of horizontal and vertical stimuli with a salient but non-congruent object. (F,G) Performance data on the various images subsets (nSO, SO, SCO and SnCO) as defined by at least 5 subjects (F) or 8 subjects (G) see text for details. For each subset the values refer to the number of images (n) considered, the percentage of correct responses and the median RT determined using the responses of all subjects on these stimuli. Arrows show the results of paired Wilcoxon comparisons for accuracy and median RT. The bootstrap confidence intervals computed for RTs as described in the text are indicated in square brackets.

significance probably due to small sets of images (statistical tests in Fig. 3). However, bootstrap simulations showed that the 33 scenes with salient non-congruent objects (SnCO) were associated with median RT of 451 ms, a value that was longer than expected by chance (95% CI [383–436 ms]).

Thus, the saliency and the congruency of objects present in the scene have an influence on the time necessary to perform a context categorization task. This interference due to the presence of a salient object was strengthened by the analysis of performance accuracy.

Accuracy (Fig. 3A and F) was significantly lower for SO images than for nSO images (respectively, 94% and 97% correct, paired Wilcoxon test, $Z = 3.059$, $p = .002$). This accuracy drop appears mainly due to images with salient non-congruent objects (SCO: 96%; SnCO: 88.1%). This 7.9% accuracy drop related to object incongruency was statistically significant (paired Wilcoxon test, $Z = 3.062$, $p = .002$).

The presence of a salient non-congruent object had a large significant deleterious effect on context categorization performance. To appreciate the effect related to object sal-

iency on its own we need to compare performance scores obtained with the 130 SCO images in which the object was salient but congruent to scores obtained with the 948 nSO images that did not contain any salient object. Both permutation and paired Wilcoxon tests showed that the median RT increase and the accuracy drop observed with SCO images were significant (all tests: $p < .021$, all paired Wilcoxon tests: $Z > 2.31$). Second, we run bootstrap simulations by randomly sampling 130 images among the total set of 1111 images and showed that the SCO median RT (409 ms) fell outside the 95% CI [381–403 ms]. The median RT recorded on these 130 SCO images was thus statistically different from the RT on images with no salient object.

To strengthen these results we performed statistical analyses on the more selective set of images in which we only consider the stimuli that were classified in the same subsets by 8 subjects or more (Fig. 3G). Using this more restrictive criterion to define the different subsets, the total number of images kept for analysis dropped to 829 (72% of all the images categorized) among which, 759 were classified as nSO and 70 as SO. Among the 70 photographs with a salient object (SO), 55 images were considered as SCO (congruent object) and 15 as SnCO (non-congruent object).

Similarly to the previous analyses, subjects were more accurate (97.2% vs. 92.4% correct) and faster at categorizing scene context when no salient object was present (median RT 383 vs. 420 ms). The 4.8% accuracy drop and the 37 ms RT increase with SO images were both statistically significant (see all data and statistical results in Fig. 3A, B, and G). A congruency effect was also observed; SnCO images were categorized less accurately (81.5% vs. 95.4% correct) and slower (482 vs. 401 ms) than SCO images. The 13.9% accuracy drop and the 81 ms RT increase were both statistically significant. This accuracy drop off is very large, as we had never obtained such a low accuracy score in any other experiments using the same rapid go/no-go categorization task with intact non-manipulated pictures.

We also evaluated the effect of saliency on its own by comparing performance on SCO images and nSO images. The deleterious effect induced by a congruent object on accuracy was only significant when using a permutation test, the effect on speed was significant using both paired Wilcoxon and permutation tests, but bootstrap simulations were not conclusive (run on the total set of 829 images, they showed that median RT of 401 ms for the 55 SCO images fell within the 95% CI [384–411 ms], indicating that the processing speed observed for those images could be explained by random sampling of the image set). These induced effects on context categorization performance by the presence of a salient object are not as robust as those obtained for incongruent objects. Nevertheless, the interference with context processing by object saliency is shown by an increase in reaction times and an accuracy drop that might need larger set of images to reach robust statistical significance.

In a context categorization task, salient objects present in a scene have a deleterious effect on performance in terms of speed and accuracy. This effect is clear and strong with non-congruent objects. With congruent object the effect is less compelling for accuracy but a tendency is always present that fails to reach significance in some cases. Further research with carefully chosen natural stimuli is needed to evaluate with precision the strength of such interference between object and context processing.

4. Discussion

The main goal of this study was to determine if background categorization can be performed sufficiently rapidly to allow background influences on object categorization. A positive answer to that question would strengthen the idea that scene context can influence object recognition. Indeed, to facilitate object recognition, contextual information has to be extracted rapidly to generate candidate expectations (Bar 2004; Bar & Aminoff, 2003; Bar et al., 2006) or to constrain local analysis (Oliva & Torralba, 2006).

4.1. Context categorization and object categorization

In the two context categorization tasks used, ''man-made'' and ''natural'', subjects reached very good scores. While images were displayed for only 26 ms, they performed the task with high accuracy (96.5% correct) and fast reaction times (median RT of 388 ms on average). An important point is that the rapid categorization task used in the present study had already been used in several studies from our team to assess object processing speed (animals, humans, faces, means of transport, or food objects) and in a first attempt to determine the processing speed of scene context (Rousselet et al., 2005). We can thus directly compare the processing speed of a visual object and of its visual context. A general finding of those experiments is the very high accuracy and the fast speed with which subjects can perform rapid go/no-go object categorization tasks. However, RT can vary substantially between experiments and subjects as shown in Fig. 4. Longer processing times might be needed for some objects, such as food-objects, when compared to animals or means of transport, which is consistent with the idea that decisions about ''edibility'' require additional processing time. Using the same task, the time needed to get at the ''gist of a scene'', as evaluated in Rousselet et al. (2005) with four target categories (sea, mountain, indoor and outdoor scenes), tended to be longer (median RT: 405–463 ms) than observed in most object categorization tasks. In the present experiment, we show that when scene categories are more broadly defined, such as natural environments (that include sea and mountain scenes), or man-made scenes (that include indoor and outdoor scenes), categorization can be done faster (median RT: 383–393 ms). But contextual processing does not appear to be faster than object processing. The two processing streams rather appear to progress in paral-

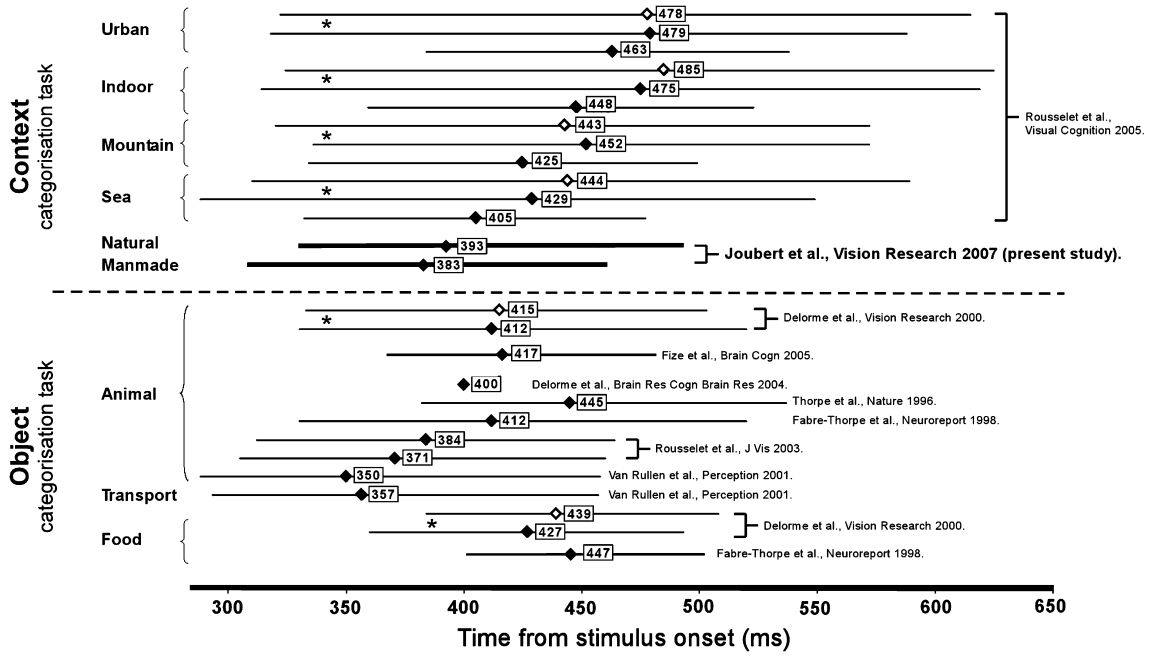

Fig. 4. Object and context categorization tasks. Reaction times obtained in experiments published by our team that used the same go/no-go paradigm with brief central image presentation. For each experiment, the horizontal line extends from minimal to maximal individual median RT; and the across-subject average is indicated by a diamond. Filled diamonds indicates that photographs were presented in colour; empty diamonds indicates that they were presented in greyscale. An asterisk indicates that colour and greyscale stimuli were mixed in the same experimental block. Note the substantial overlap among context categorization tasks reported in this paper and various object categorization tasks.

lel with similar time-course and performance levels which suggests that large interactions have time to take place between object and context processing (Fig. 4).

To better appreciate this overlap, we compared directly the results of the present study with the results obtained in an animal/non-animal categorization task (Rousselet et al., 2003), in which subjects scored an average accuracy of 96.3% and a median RT of 384 ms. In the present context categorization tasks, subjects reached 96.9% and 96.4%, respectively, for natural scenes and man-made environment with corresponding median RT of 393 ms and 383 ms. For similar accuracy scores, not only are median RTs very similar, but the whole RT distributions are well superimposed for the natural, man-made and animal categorization tasks (Fig. 5).

Thus, context categorization can be achieved as fast as object categorization (at least for certain tasks), a result incompatible with theories that describe scene recognition as the result of the successful identification of some objects in the scene. On the contrary, our results suggest that scene and object information might be extracted in parallel with similar temporal dynamics. This in turn opens the possibility of extensive interactions between context and object processing (from context to object categorization, but also from objects to context categorization).

By imposing temporal constraints on the subject's response, briefly flashed photographs and broad categories that could rely on coarse representations, our task is really tackling the early interactions between object and context processing. Such object/background interactions could be strengthened in tasks requiring or simply allowing more time for information processing such as detailed identification tasks, tasks using a larger number of categories or tasks involving explicit verbal responses.

4.2. Coarse vs. detailed contextual information

The comparison of the present study using two broad context categories with our previous study, in which human subjects had to categorize one of four finer target categories (sea, mountain, indoor and outdoor scenes, Rousselet et al., 2005), showed that subjects performed both tasks with a similar accuracy, but with faster RT when categorizing broader categories. On average, they needed 50 ms additional processing time when categorizing one of the four contexts, although sea scenes were categorized faster than the 3 other types of environments. The two studies used similar stimulus sets, thus the main difference concerned the amount of visual analysis needed to perform the task. With the strong temporal constraints of the rapid categorization task, our data might reflect the underlying temporal dynamics of visual processing.

Why do we need less processing time to access the two broad categories used in the present study? One interpretation considers that low spatial frequency information (processed by the fast magnocellular pathway) is available earlier than high spatial frequency information. Alternatively, scene analysis could proceed from global to local feature analysis (Hughes, Nozawa, & Kitterle, 1996; Navon, 1977). As suggested by Oliva and Torralba

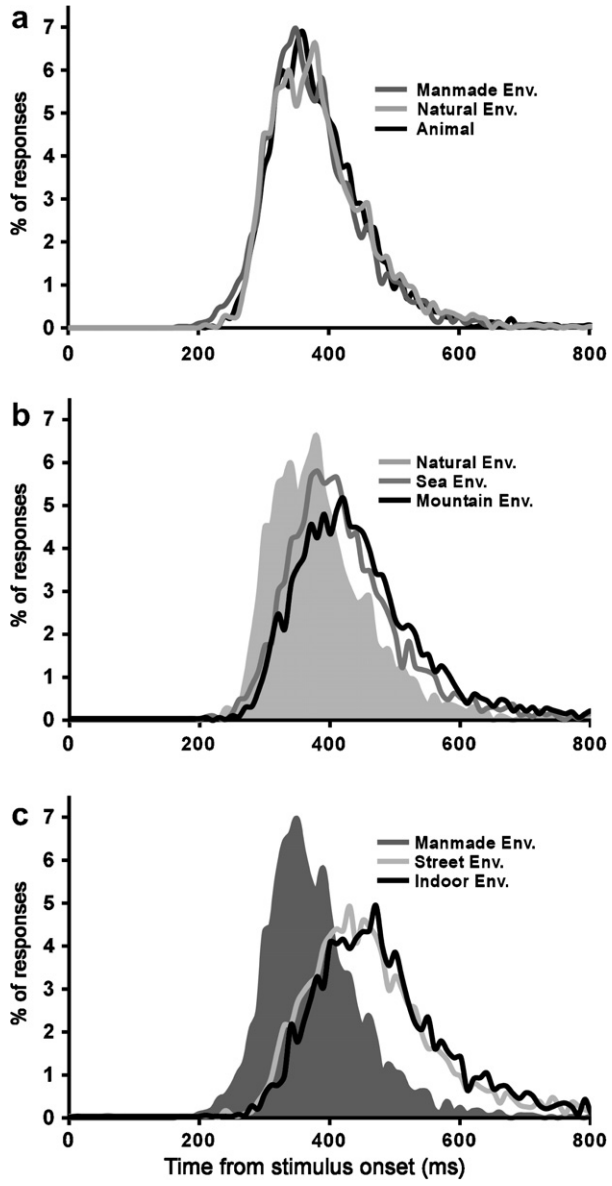

Fig. 5. Object and context categorization tasks: RT distributions. (a) Comparison of correct go-responses RT distributions for 'man-made' and 'natural' environment categorization tasks obtained in the present study and in the 'animal' categorization task reported in Rousselet et al. (2003). (b) Comparison of correct go-responses RT distributions for the 'natural' environment categorization task obtained in the present study and for the 'sea' and 'mountain' environment categorization tasks reported in Rousselet et al. (2005). (c) Comparison of correct go-responses RT distributions for the 'man-made' environment categorization task obtained in the present study and for the 'street' and 'indoor' environment categorization tasks reported in Rousselet et al. (2005). Percentages of responses are expressed over time using 10 ms time bins.

(2006), a low resolution sketch would probably be sufficient to categorize an environment as "natural" or "man-made" at the superordinate level but higher resolution analysis could be needed for basic scene categories such as sea, mountain, indoor and outdoor scenes. These two large context categories can be considered as superordinate classes (Oliva & Torralba, 2006). Indeed, when judging similarity between natural images, human subjects have been shown to spontaneously organize the scenes along an axis running from natural scenes to man-made scenes (Rogowitz, Frese, Smith, Bouman, & Kalin, 1998). These two extreme categories appear thus more distant from each other than finer categories. In object categorization, it is widely accepted that the basic level of categorization is accessed before the superordinate level (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Exploring a taxonomy for environmental scenes with the paradigms used by Rosch et al., Tversky and Hemenway (1983) have also reported an advantage for basic scene categories such as beach, mountains and cities. Our results appear at odds with this classic categorization framework as we report a processing speed advantage for broad categories over more basic scene categories. Taken together, our results suggest that sea or mountain environments might be first recognized as natural contexts. Overall, this categorization hierarchy for visual scenes is in keeping with the "coarse to fine" hypothesis (Hughes et al., 1996; Macé, Joubert, & Fabre Thorpe, 2005; Navon, 1977; Schyns & Oliva, 1994): a coarse view of a scene provides sufficient information to decide whether it belongs to a 'natural' or 'man-made' category, but a finer categorization is likely to require more detailed processing of specific diagnostic features (Fig. 5a and b). Following this line of thought, sea and mountains scenes would be first considered as "natural scenes", indoor or outdoor urban environments as "man-made scenes". Indeed in our previous study (Rousselet et al. 2005), subjects' errors were biased towards distractors that belong to the same "superordinate" category (i.e. mountain in sea scenes categorization).

The preferred access to basic environments reported by Tversky and Hemenway (1983) might be linked to the temporal course of the lexical responses required in their tasks that might have masked an inverted hierarchy in accessing visual categories. More studies are needed to characterize the possible differences between lexical and visual processing time-courses.

4.3. Interaction between processing of context and processing of objects

In this paper, we have shown that the processing of scene context is fast enough to allow for interactions between object processing and context processing. In a fast feed-forward wave of coarse processing, context could be used to select the more likely candidate from the many potential object representations activated (Bar, 2004; Bar & Aminoff, 2003). Moreover, it is important to keep in mind that we investigated context processing in extreme conditions, when new and unrelated pictures are briefly flashed one after another. In daily life, such situations are encountered when zapping from one TV channel to another, or when turning over the pages of a magazine. But context is usually stable over time, and the top–down influence of context on object representations does not

have to fit in with the temporal constraints imposed by the brief image presentation used in our experiment.

On the other hand, although the study was only designed to test the speed at which a coarse categorization of context could be performed, we describe a deleterious effect of object processing on context processing, even though task instructions were clearly oriented towards the analysis of context. In the present study, the presence of a salient object in the scene tends to delay the processing of the background by about 25–37 ms, and induce a 2.6–4.8% accuracy drop. When the object is congruent with the scene context, the processing speed effect appears statistically significant, while the accuracy effect tends to be less robust. On the other hand, processing speed and accuracy are clearly altered when the object is not congruent with its background. Salient and non-congruent objects delay the background with a temporal cost that can reach 81 ms, and induce a large accuracy drop (13.9%). Such results confirm the drop of accuracy observed by Davenport and Potter (2004) in a naming task when manipulating consistency in object and background perception. It clearly shows that early interactions between the processing of objects and backgrounds can take place and need to be further analyzed.

How can we explain such an early effect of object processing on context processing? One way to account for part of this temporal delay when a salient object is present could involve an attentional bottom–up processing bias. However such attentional bias is not supported by several studies indicating that fast object categorization might be done without the need of focused attention. Indeed, in the same rapid animal/non animal categorization task, human subjects can process simultaneously two unrelated scenes with no temporal cost (Rousselet, Fabre-Thorpe, & Thorpe, 2002); they can even do such task in their peripheral visual field when their attention is captured centrally in a dual task paradigm (Li, VanRullen, Koch, & Perona, 2002). Context categorization might also be performed without the need for focused attention, as suggested in the case of the extraction of the spatial layout that represents the 'gist' of the scene (Oliva & Schyns, 2000; Rousselet et al., 2005), a form of processing that might involve cerebral areas independent from object processing areas, such as the parahippocampal cortex (Epstein, Graham, & Downing, 2003; Epstein & Kanwisher, 1998; Goh, Siong, Park, Gutchess, Hebrank, & Chee, 2004; Kanwisher & Wojciulik, 2000). In the present task, the delay due to the presence of a salient object in the scene could be explained by an exogenous capture of attention. Object features could be different enough from context features to capture attention, and lead to the formation of proto-objects (Rensink, 2002) without reaching an explicit level of object representation (Walther & Koch, 2006). On the other hand, performance was even more impaired with incongruent objects, even though the temporal constraints on visual processing are not really compatible with an explicit access to object/context incongruency. How can we explain the early effect of incongruent objects on background processing? Bar and collaborators (Aminoff, Gronau, & Bar, 2007; Bar, 2004; Bar & Aminoff, 2003) have suggested that the parahippocampal cortex (PHC) could mediate the representation of familiar contextual associations. When groups of objects tend to appear together, the populations of neurons selective to these objects would tend to be activated simultaneously. Such familiar contextual associations would be encoded in the PHC. In performing our task and under strict instructions to respond to a given contextual target, top–down preparation of the visual system is presumably maximal and corresponding familiar associations will be activated. For example, if a subject is looking for ''natural environments'', most populations of neurons selective for natural features are probably pre-activated. Through parallel processing, a congruent scene might activate multiple populations of neurons that are usually co-activated. On the other hand, when a salient man-made object appears in a natural environment, it will generate a conflict between populations of neurons that selectively respond to the natural features of the background and the man-made features of the object. With an ''incongruent'' object in a given context, several populations of neurons that do not usually fire together would be active at the same time; the more incongruent features in the scene, the greater the competition between these two populations of neurons and hence the greater the competition between the go and the no-go motor output responses. This possibility is well supported by perceptual decision theories that rely on an ''accumulation of evidence'' (Perrett, Oram, & Ashbridge, 1998): response inhibition of a population of neurons could result in a delay to reach decision threshold. The presence of a salient object will slow the processing of the background but, when the object and context categories are incongruent, decision about the nature of a context would be even slower. It should be noted that object saliency in images was estimated subjectively by the observers. More objective measures of saliency and its systematic control are needed to directly evaluate the temporal cost of object saliency and incongruence in scene processing.

In conclusion, this study shows how fast human subjects are to get at the gist of a scene. The temporal dynamics of background scene processing is clearly compatible with the idea of large interactions with object processing. We have also described a deleterious effect of salient objects on context processing especially for incongruent objects. This is an interesting finding because the question is usually addressed the other way around to evaluate how context might facilitate object recognition and at which level object processing is modulated or constrained by contextual analysis. Object and context could be processed in parallel and may interact extensively all along the first wave of processing. But these interactions should not be thought as unidirectional but rather as reciprocal with both facilitator effects and interferences. Following Davenport and Potter (2004), future experiments should investigate the influence of object/background incongruency on accuracy and pro-

cessing speed in object and in scene categorization/recognition tasks.

Acknowledgments

This work was supported by the CNRS. Financial support was provided to O. Joubert by a Ph.D. grant from the French government. We thank Rufin VanRullen for his valuable comments.

References

Aminoff, E., Gronau, N., & Bar, M. (2007). The parahippocampal cortex mediates spatial and nonspatial associations. *Cerebral Cortex, 17*(7), 1493–1503.

Bacon-Macé, N., Macé, M. J., Fabre-Thorpe, M., & Thorpe, S. J. (2005). The time course of visual processing: Backward masking and natural scene categorisation. *Vision Research, 45*(11), 1459–1469.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience, 5*(8), 617–629.

Bar, M., & Aminoff, E. (2003). Cortical analysis of visual context. *Neuron, 38*(2), 347–358.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmidt, A. M., Dale, A. M., et al. (2006). Top–down facilitation of visual recognition. *Proceedings of The National Academy of Sciences of The United States of America, 103*(2), 449–454.

Biederman, I. (1972). Perceiving real-world scenes. *Science, 177*(4043), 77–80.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review, 94*(2), 115–147.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14*(2), 143–177.

Boyce, S. J., & Pollatsek, A. (1992). Identification of objects in scenes: the role of scene background in object naming. *Journal of Experimental Psychology-Learning Memory and Cognition, 18*(3), 531–543.

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences, 4*(5), 170–178.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science, 15*(8), 559–564.

De Graef, P., Christiaens, D., & d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research, 52*(4), 317–329.

Delorme, A., Richard, G., & Fabre-Thorpe, M. (2000). Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans. *Vision Research, 40*(16), 2187–2200.

Delorme, A., Rousselet, G. A., Macé, M. J., & Fabre-Thorpe, M. (2004). Interaction of top–down and bottom–up processing in the fast visual analysis of natural scenes. *Brain Research. Cognitive Brain Research, 19*(2), 103–113.

Epstein, R., Graham, K. S., & Downing, P. E. (2003). Viewpoint-specific scene representations in human parahippocampal cortex. *Neuron, 37*(5), 865–876.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392*(6676), 598–601.

Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience, 13*(2), 171–180.

Fabre-Thorpe, M., Richard, G., & Thorpe, S. J. (1998). Rapid categorization of natural images by rhesus monkeys. *Neuroreport, 9*(2), 303–308.

Fize, D., Fabre-Thorpe, M., Richard, G., Doyon, B., & Thorpe, S. J. (2005). Rapid categorization of foveal and extrafoveal natural images: Associated ERPs and effects of lateralization. *Brain and Cognition, 59*(2), 145–158.

Friedman, A. (1979). Framing pictures: the role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology-General, 108*(3), 316–355.

Ganis, G., & Kutas, M. (2003). An electrophysiological study of scene effects on object identification. *Brain Research. Cognitive Brain Research, 16*(2), 123–144.

Goffaux, V., Jacques, C., Mouraux, A., Oliva, A., Schyns, P. G., & Rossion, B. (2005). Diagnostic colours contribute to the early stages of scene categorization: Behavioural and neurophysiological evidence. *Visual Cognition, 12*(6), 878–892.

Goh, J. O., Siong, S. C., Park, D., Gutchess, A., Hebrank, A., & Chee, M. W. (2004). Cortical areas involved in object, background, and object-background processing revealed with functional magnetic resonance adaptation. *Journal of Neuroscience, 24*(45), 10223–10228.

Henderson, J. M. (1992). Object identification in context: the visual processing of natural scenes. *Canadian Journal of Psychology, 46*(3), 319–341.

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology, 50*, 243–271.

Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology-General, 127*(4), 398–415.

Hughes, H. C., Nozawa, G., & Kitterle, F. (1996). Global precedence, spatial frequency channels, and the statistics of natural images. *Journal of Cognitive Neuroscience, 8*(3), 197–230.

Kanwisher, N., & Wojciulik, E. (2000). Visual attention: insights from brain imaging. *Nature Reviews Neuroscience, 1*(2), 91–100.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of The National Academy of Sciences of The United States of America, 99*(14), 9596–9601.

Macé, M. J., Joubert, O. R., & Fabre Thorpe, M. (2005). Entry level at the superordinate level in visual categorization. *9th International conference on cognitive and neural systems* (Vol. 52).

Navon, D. (1977). Forest before trees: the precedence of global features in visual perception. *Cognitive Psychology, 9*(3), 353–383.

Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology, 34*(1), 72–107.

Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology, 41*(2), 176–210.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42*(3), 145–175.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. *Progress in Brain Research, 155PB*, 23–36.

Palmer, S. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition, 3*, 519–526.

Perrett, D. I., Oram, M. W., & Ashbridge, E. (1998). Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformations. *Cognition, 67*(1–2), 111–145.

Potter, M. C. (1975). Meaning in visual search. *Science, 187*(4180), 965–966.

Rensink, R. A. (2002). Change detection. *Annual Review of Psychology, 53*, 245–277.

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience, 3*(Suppl), 1199–1204.

Rogowitz, B., Frese, T., Smith, J., Bouman, C., & Kalin, E. (1998). Perceptual image similarity experiments. In *SPIE conference human vision and electronic imaging*. San Jose, California.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 382–439.

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience, 5*(7), 629–630.

Rousselet, G. A., Joubert, O. R., & Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition, 12*(6), 852–877.

Rousselet, G. A., Macé, M. J., & Fabre-Thorpe, M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *Journal of Vision, 3*(6), 440–455.

Sanocki, T., & Epstein, W. (1997). Priming spatial layout of scenes. *Psychological Science, 8*(5), 374–378.

Schyns, P. G. (1998). Diagnostic recognition: task constraints, object information, and their interactions. *Cognition, 67*(1–2), 147–179.

Schyns, P. G., Jentzsch, I., Johnson, M., Schweinberger, S. R., & Gosselin, F. (2003). A principled method for determining the functionality of brain responses. *Neuroreport, 14*(13), 1665–1669.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science, 5*(4), 195–200.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520–522.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network, 14*(3), 391–412.

Tversky, B., & Hemenway, K. (1983). Categories of environmental scenes. *Cognitive Psychology, 15*(1), 121–149.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience, 5*(7), 682–687.

VanRullen, R., & Thorpe, S. J. (2001). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception, 30*(6), 655–668.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks, 19*(9), 1395–1407.

Wilcox, R. R. (2005). *Introduction to robust estimation & hypothesis testing*. New York, NY: Academic Press.

PARTIE III

DES INTERACTIONS OBJET/CONTEXTE PRECOCES ET BIDIRECTIONNELLES

Un hamster de laboratoire dit à son congénère :
"J'ai dressé le savant !
A chaque fois que j'appuie sur le bouton, il m'apporte des graines."
Bernard Werber

1. Des interactions objet/contexte de différentes natures

1.1. Des interactions physiques aux interactions de plus haut-niveau

Comment interagissent les différents aspects globaux et locaux des stimuli. L'un des importants volets de recherche a impliqué les tâches de recherche visuelle qui ont conduit Treisman et Gelade en 1982 à proposer leur « Feature-integration theory » (Treisman, 1982). Lorsque les sujets reçoivent la consigne de détecter une cible définie par une unique dimension diagnostique au milieu d'un ensemble de distracteurs ne partageant pas les mêmes traits, la cible "pop-out" et les temps de réaction ne sont pas influencés par le nombre de distracteurs. La recherche visuelle est donc effectuée en parallèle sur l'ensemble des composants du stimulus. Si au contraire, la cible est définie par deux (ou plus) dimensions physiques (« conjunction search ») dont l'une et/ou l'autre se trouve également présente chez les distracteurs, le temps de réaction des sujets augmentent avec la quantité de distracteurs. Cette interférence des distracteurs peut être difficilement envisagée comme une interférence conceptuelle étant donné que l'effet pop-out est observé, entre autres, avec des stimuli composés de formes non-figuratives. Dans ces tâches de recherche visuelle, la dimension physique du stimulus dans sa globalité, l'ensemble cible et distracteurs, a donc une influence importante sur la détection d'une partie du stimulus.

En outre, la théorie de la Gestalt se base sur l'hypothèse importante que le percept global transcende la somme des percepts locaux. Ce que l'on perçoit n'est donc pas l'addition linéaire de l'ensemble des formes ou traits physiques perçus et traités par les aires visuelles bas-niveau, mais un percept les transcendant comme on peut le constater dans les études portant sur les contours illusoires (Figure 26). Les défenseurs de la Gestalt ont d'ailleurs mis en évidence le fait qu'un stimulus dans sa globalité pouvait être appréhendé avant les différents sous-ensembles qui le constituent. Dans ce sens, la perception d'un sous-ensemble peut être influencée par le percept global du stimulus.

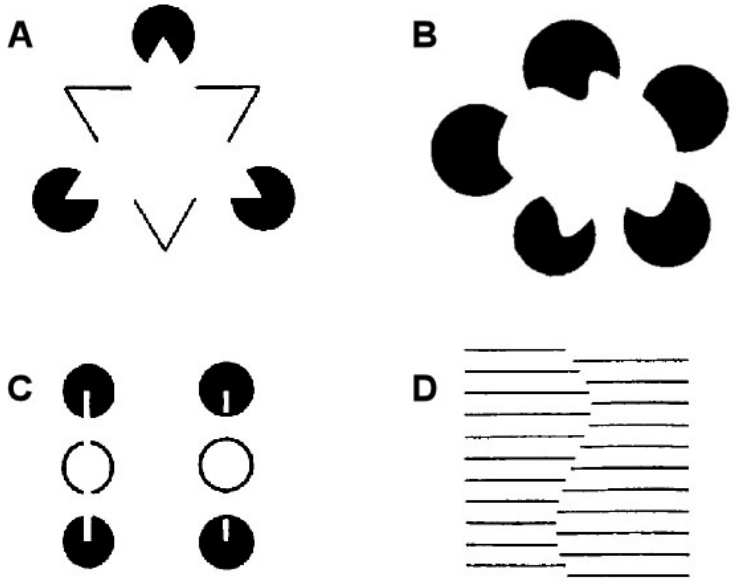

Figure n° 26 : Tirés de Rubin, Nakayama & Shapley, 1996 et de von der Heydt, Peterhans & Baumgartner, 1984. Stimuli illustrant les principes de la «Gestalt Theory ». Certains percepts (A : triangle, B : blob, C : barre, D : courbe) émergent de leur contexte respectif alors qu'aucun contour ne les figure. Ces représentations de complexité moyenne transcendent les caractéristiques bas-niveau des stimuli.

Cependant, des influences non-physiques du contexte ont également pu être observées sur la reconnaissance d'objets locaux. Un des exemples les plus connus a été fourni par l'étude de Navon intitulée « Forest before trees » (Navon, 1977, Navon, 1983) déjà décrite précédemment. Utilisant des stimuli représentant une large lettre constituée par de petites lettres, par exemple un large « S » constitué de petits « H », il démontre une précédence du traitement de l'information globale sur le traitement local. En effet, les sujets (1) réalisent de meilleures performances dans la détection de la lettre globale, (2) sont largement influencés par l'information de la lettre globale alors qu'ils sont impliqués dans une tâche de reconnaissance des petits caractères (aucun effet inverse n'est constaté). Des études ultérieures démontreront que cet avantage des lettres globales serait modulé par la taille des lettres, les fréquences spatiales induites et l'angle visuel auquel les stimuli apparaissent. Les caractères locaux seraient apparentés à une texture de la lettre globale (Kimchi, 1988). Ces résultats suggèrent une fois de plus, que l'influence de l'analyse globale des stimuli sur les traitements locaux est purement physique. De plus, Navon a également montré que (3) les sujets impliqués cette fois-ci dans une tâche de catégorisation auditive de lettre phonologique, atteignaient de moins bonnes performances lorsqu'une lettre visuelle globale présentée simultanément était incongruente, alors qu'aucun effet de l'incongruence des lettres visuelles

locales n'a été mis en évidence. Ce dernier résultat reposant sur des interactions multi-sensorielles ne laisse que peu de place à l'hypothèse d'une interaction purement physique, mais suggère au contraire des interactions objet/contexte dans une autre dimension, probablement conceptuelle.

Cette précédence et influence du contexte sur la perception de l'objet se retrouve également au niveau de la lecture. Il a en effet été démontré que nous présentions de meilleures performances dans la reconnaissance de lettres lorsque ces dernières étaient présentées au sein d'un mot réel plutôt qu'au sein d'un non-mot ou encore isolées (Johnston & McClelland, 1974). De manière similaire, un mot est plus efficacement identifié lorsque intégré dans un paragraphe (Lefton & Fisher, 1976).

Ces résultats peuvent être mis en parallèle avec le phénomène d' « extension boundary effect » décrit par Intraub (Intraub, 1999, Intraub & Richardson, 1989). Lors de la perception d'une scène naturelle, nous construirions un schéma mental de la scène contenant des informations absentes de la scène perçue. En effet, des sujets humains ayant préalablement visionné des scènes naturelles pendant 15 secondes ont reçu pour tâche de se remémorer les scènes et de les dessiner aussi fidèlement que possible (sans pour autant que la qualité esthétique soit mise en cause, Figure 27). Les résultats montrent que les sujets ont tendance à faire apparaître dans les dessins des éléments pertinents qui n'étaient pas présents dans les stimuli, mais qui auraient pu s'y trouver si la caméra avait embrassé la scène de plus loin. Ils ont également tendance à dessiner leurs objets dans leur intégralité, alors qu'ils pouvaient être en partie masqués par d'autres objets dans la scène originale. Il existe donc une représentation conceptuelle globale de la scène transcendant les objets locaux la composant et influençant leur description.

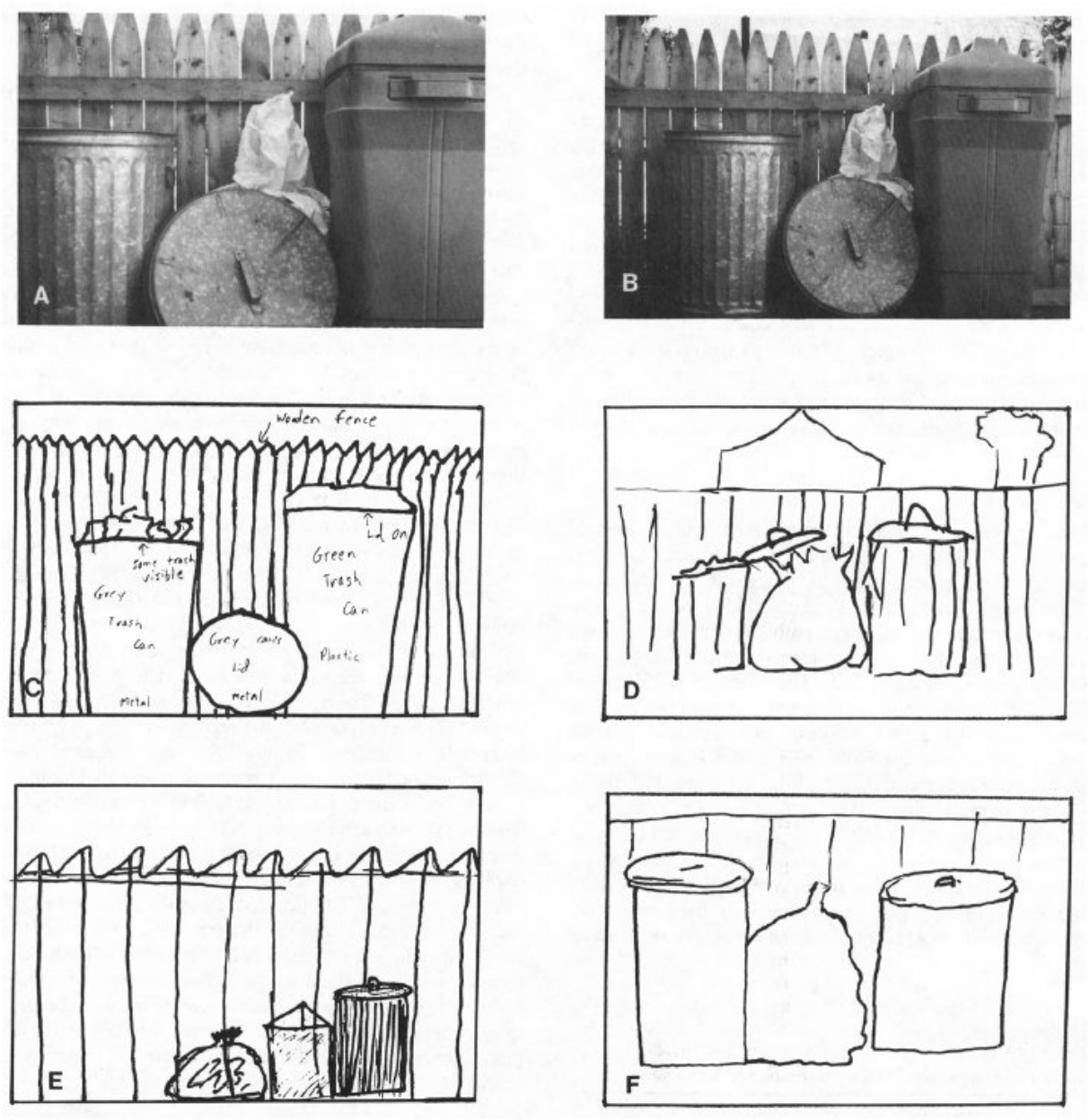

Figure n°27 : Tiré de (Intraub & Richardson, 1989) : Les sujets impliqués dans une tâche de rappel mnésique de photographies (C,D,E,F) ont tendance à représenter des éléments visuellement non présents dans la scène originale (A et B). Les sujets ont de plus tendance à représenter les photographies selon un angle de vue plus profond, et à représenter l'ensemble des contours de l'objet, même ceux masqués par occlusion.

1.2. Le problème de la segmentation et du liage perceptif…

Bien souvent dans notre environnement, les objets ne sont pas complètement apparents mais en partie cachés par d'autres objets. Ce phénomène d'occlusion pose des questions majeures, notamment concernant la compréhension des mécanismes de segmentation et de liage perceptif soit-disant pré-requis à la construction d'une représentation globale de l'objet.

III.1. Des interactions objet/contexte des différentes natures

Comment le système visuel parvient à grouper ensemble les différentes entrées rétiniennes semblant appartenir au même objet tout en ignorant les entrées relatives à d'autres objets ? Quels sont les mécanismes permettant la ségrégation de l'objet et de son contexte ? Si les caractéristiques physiques sur lesquelles reposent les opérations de liage perceptif et de segmentation sont encore inconnues, il semble que synchronie et asynchronie des décharges neuronales puissent en partie expliquer la manière dont le système visuel groupe les informations relatives à un même objet (Usher & Donnelly, 1998). Cette synchronisation se manifeste par exemple au sein d'une composante oscillatoire à hautes fréquences (30 hz) entre 210 et 290 ms lors de la présentation d'un triangle cohérent (vs. incohérent), que celui-ci soit réel ou illusoire (Tallon, Bertrand, Bouchet & Pernier, 1995). De fait, la segmentation ne serait pas basée sur un traitement computationnel local des bordures, mais sur l'intégration des informations simultanément traitées. Néanmoins, le domaine de recherche est large. Si de nombreuses études ont été menées afin de préciser les mécanismes impliqués, il reste encore de nombreuses lacunes à combler. Je ne souhaite pas cependant faire ici une revue de la littérature, mais plutôt mettre en exergue un aspect primordial de la recherche sur la ségrégation objet/contexte, à savoir le décours temporel des traitements sous-jacents. Est-ce que la ségrégation de l'objet/contexte est une étape nécessaire à la catégorisation et/ou à la reconnaissance de l'objet, étape qui devrait nécessairement être accomplie plus précocement au sein du système visuel.

Depuis les premières études sur la question, il est apparu comme évident que la segmentation de l'objet survenait très tôt au niveau du système visuel, et dans tous les cas, avant sa reconnaissance (Driver & Baylis, 1996, Driver, Davis, Russell, Turatto & Freeman, 2001, Marr, 1982 , Nakayama & Shimojo, 1992, Nakayama, Shimojo & Silverman, 1989 , Rock, Nijhawan, Palmer & Tudor, 1992). Ainsi, le système visuel ne pourrait pas reconnaître un objet avant d'avoir assemblé, réuni, les différentes informations visuelles relatives à l'objet. Dans ce courant de pensées, une des questions devenues primordiales fut alors de savoir si les mécanismes de ségrégation nécessitaient ou non l'intervention d'une composante attentionnelle. Les chercheurs suggèrent depuis que les premiers mécanismes de segmentation seraient pré-attentifs et mèneraient à la considération de « proto-objets » tandis que des étapes ultérieures de segmentation nécessitant une composante attentionnelle mèneraient à la construction d'un percept abouti (Driver & Baylis, 1996, Driver et al., 2001, Duncan, 1984, Palmer & Rock, 1994, Treisman, 1986).

D'autres chercheurs vont cependant à l'encontre de ce premier courant de pensées et défendent l'idée d'une reconnaissance de l'objet préalable à sa ségrégation. Par exemple, une étude de Peterson utilisant des stimuli bistables du type Rubin's vase-faces présentés à l'endroit ou à l'envers, montrent que des sujets humains ayant reçu pour consigne de traiter visuellement la figure blanche (ou la figure noire) des stimuli figure 28, considèrent plus longtemps la figure porteuse de sens. Les auteurs en déduisent que la segmentation bas-niveau de l'objet est directement influencée par sa reconnaissance sémantique préalable. (Peterson, 1994, Peterson & Gibson, 1994a, Peterson & Gibson, 1994b).

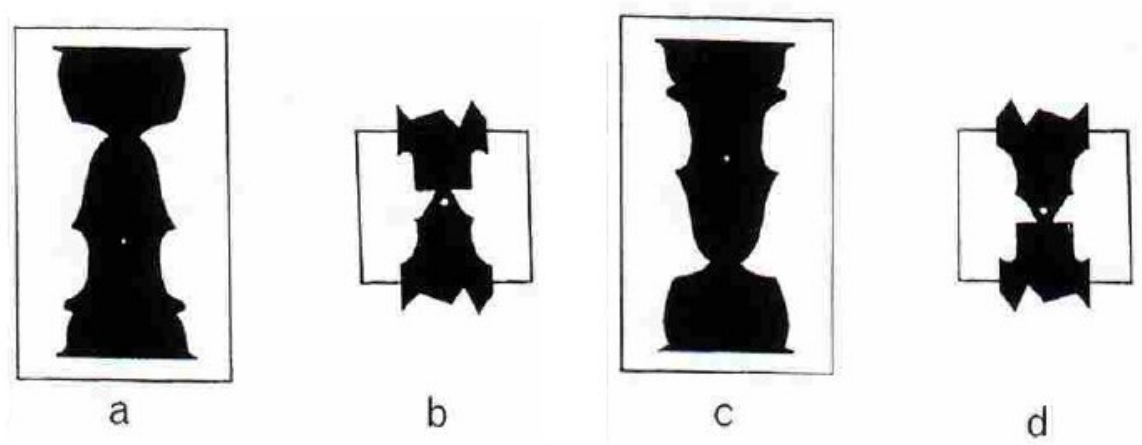

Figure n°28 : Tiré de (Peterson, 1994). Deux exemples de stimuli bi-stables (a) et (b) dans lesquels la figure blanche est porteuse de sens uniquement lorsque présentée à l'endroit (c) et (d) : deux moitié de femmes en robe, et deux visages se faisant face.

L'article n°4 de ce mémoire comparant la catégorisation d'objet sur fond isolé et au sein de d'un contexte apportera de nouvelles réponses à la question du décours temporel des mécanismes de ségrégation.

1.3. Avant l'influence contextuelle, l'apprentissage des régularités

L'influence du contexte sur la reconnaissance des objets n'est envisageable qu'à la condition d'avoir préalablement appris de façon consciente ou non, des règles physiques, spatiales ou sémantiques liant objet et contexte. Un des points majeurs de cet apprentissage réside dans la discrimination des règles d'intérêt. Comment savoir si deux évènements A et B co-occurrents sont une cooccurrence aléatoire ou sous-tendent une structure d'intérêt (Atick,

1992, Barlow, 1989) ? Tout est question de probabilité : si A et B ont une forte probabilité d'apparaître de manière co-occurrente, alors la règle « si A alors B » est grandement avérée. Au contraire, les cooccurrences aléatoires auront une faible probabilité de survenir. Dans notre monde, la probabilité qu'un éléphant soit dans la savane n'est pas certaine mais très forte. Si certaines règles de la vie de tous les jours ont ainsi été apprises en partie consciemment, il s'avère que d'autres règles liant contexte et objet peuvent être apprises de manière totalement inconsciente ainsi que l'ont démontré Chun et Jiang. Leurs sujets étaient impliqués dans une tâche de recherche visuelle, l'objet cible (par exemple T) apparaissait parmi des distracteurs (par exemple L). L'ensemble de ces "L" formaient une structure spatiale qui pouvait être nouvelle ou se répéter sans que le sujet n'en soit averti. Dans le cas où elle se répétait, la cible était toujours à la même place au sein des distracteurs. Les sujets étaient plus rapides à détecter la cible dans ces configurations "familières" mais étaient incapables de reconnaître explicitement ses structures spatiales comme étant "familières". Les sujets avaient donc implicitement appris des associations entre objet et contexte (Chun & Jiang, 1998). Cette expérience répliquée lors d'une étude en magnéto-encéphalographie révèle une activité différentielle entre les conditions de contextes familiers et de contextes nouveaux avant 100 ms. Cette activité mesurée dans les régions occipitales bilatérales reflèterait l'effet précoce d'une mémoire implicite (Chaumon, Drouet & Tallon-Baudry, 2008). De même, les sujets sont capables d'apprendre de manière rapide et automatique des conjonctions de formes qui leur semblent par la suite familières. Les sujets peuvent ainsi encoder en parallèle différentes statistiques d'une structure composée de formes telles que les relations absolues ou relatives entre ces formes, ou encore leurs probabilités de cooccurrence (Fiser & Aslin, 2001). Plus proche de notre cadre de travail puisque les auteurs ont utilisé des scènes naturelles, un apprentissage contextuel de la position d'un objet cible parmi un groupe de distracteurs a été démontré lorsque l'information globale de la scène était répétée (vs. information locale ; (Brockmole, Castelhano & Henderson, 2006). Ainsi, parmi d'autres propriétés de l'objet, sa position probable au sein d'un type de contexte pourrait être encodée en mémoire. Etant donné que ce genre de cooccurrences se répète constamment tout au long de notre développement, nous apprendrions de manière plus ou moins inconsciente certaines règles de régularité qui permettraient à notre cerveau de générer des hypothèses sur notre environnement perceptif.

1.4. Interactions objet/contexte au sein des scènes naturelles

Nous avons vu dans le chapitre I que la représentation d'une scène serait, pour certains auteurs, inférée à partir d'une identification réussie de quelques objets au sein de la scène (Antes et al., 1981, Friedman, 1979), à partir des relations spatiales existant entre les objets (De Graef et al., 1990), ou à partir de caractéristiques physiques globales spécifiques de la scène tels que les géons (Biederman, 1981, Biederman, 1995) ou encore les fréquences spatiales (Oliva & Schyns, 1997, Oliva & Torralba, 2001, Schyns & Oliva, 1994).

Mais une fois la représentation de la scène mise en place, qu'en est-il de l'influence du contexte sur la localisation et la reconnaissance des objets ? Il est souvent considéré qu'une fois la représentation perceptuelle de la scène construite, le système visuel va activer une représentation mnésique de cette catégorie de scènes: "le *schéma"* (Antes et al., 1981, Biederman, 1981, Friedman, 1979, Loftus & Mackworth, 1978), permettant de formuler des attentes concernant les objets susceptibles d'y être présents ainsi que leurs tailles ou leurs positions probables. Ce modèle du schéma n'est pas nécessairement sériel, mais pourrait s'appliquer par interférences et facilitations mutuelles et parallèles entre les représentations de l'objet et de la scène (Metzger & Antes, 1983). Ce phénomène pourrait également résulter d'interactions avec les informations feedback compétitives et coopératives provenant des aires de haut-niveau. Tandis que des traitements précoces et inconscients des traits locaux seraient effectués par les aires de bas-niveau, des influences tardives d'une intégration aboutie de l'ensemble du stimulus pourraient nous aider à préciser nos perceptions locales.

Dans ce sens, la catégorie de la scène doit être précisée précocement au cours de l'analyse de la scène afin de permettre l'émergence de traitements descendants facilitant l'identification des objets. Afin de vérifier cette hypothèse (parmi d'autres) d'un point de vue psychophysique, deux types de paradigme ont largement été utilisés pour l'étude des influences contextuelles sur la reconnaissance des objets : l'enregistrement des performances comportementales dans des tâches de détection/catégorisation d'objets au sein de scènes généralement flashées et l'enregistrement de mouvements oculaires durant l'exploration des scènes.

Les paragraphes suivants ont pour objectif de dresser une revue suffisante mais sans nul doute non exhaustive des études menées sur la réalité des interactions entre objets et contexte au sein des scènes.

L'influence d'un schéma perceptif sur l'identification des objets

L'une des premières études marquantes testant l'influence du contexte sur la reconnaissance des objets fut proposée par Palmer en 1975 (Palmer, 1975). Stephen Palmer présentait à ses sujets une scène dessinée à la main ou un écran blanc pendant 2 secondes avant d'afficher un objet cible 1300 ms plus tard, pendant une courte durée. Les objets cibles pouvaient donc faire suite à un contexte congruent, incongruent, ou neutre (écran blanc). De plus, lorsque le contexte était incongruent, les objets cibles pouvaient avoir (ou non) une forme très similaire à un objet congruent avec la scène. Un exemple de scène et des objets cibles utilisés est présenté figure 29. Notons que l'objet cible ne figurait pas au sein de la scène amorce.

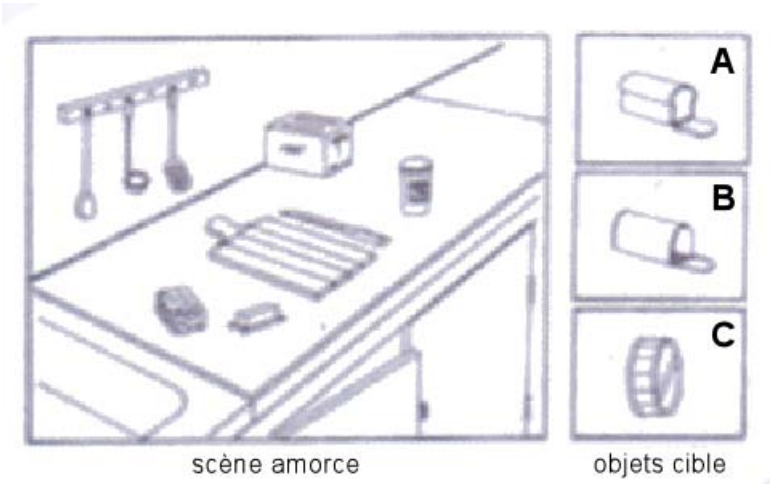

Figure n°29 : Tiré de (Palmer, 1975). Exemple de scène contextuelle utilisée en amorçage. Les objets cibles pouvaient être congruent avec l'amorce (A), non congruent mais de forme similaire à un objet congruent (B), non-congruent et ne ressemblant à aucun objet susceptibles de se trouver dans la scène (C).

Les résultats de cette étude montrent de meilleures performances de la part des sujets lorsque objet cible et contexte étaient congruents. Les performances diminuaient si aucun contexte n'était présenté et atteignaient le score le plus bas lorsque le contexte était incongruent. De plus, dans la condition non-congruente, les performances étaient d'autant plus mauvaises que l'objet cible ressemblait fortement à un objet susceptible de faire partie de la scène. Cette étude fut ainsi la première à suggérer une influence contextuelle.

Par la suite, Biederman a testé la capacité des sujets à reconnaître ou à chercher des objets cible au sein de scène spatialement cohérente ou incohérente (Biederman, 1972,

Biederman, Glass & Stacy, 1973, Biederman et al., 1974). Les scènes incohérentes étaient des photographies découpées en 6 carrés réorganisés par la suite aléatoirement. Une fois de plus, les meilleures performances de reconnaissance, ou les plus courts temps de réaction en recherche visuelle furent obtenus lorsque la congruence du contexte était préservée. Si ces résultats suggèrent une influence avérée du contexte, on peut aussi soupçonner l'existence de certains biais dans le paradigme, telle la présence supplémentaire de contours liés aux découpages dans la condition incongruente qui pourraient bruiter l'information disponible. De plus, aucune évidence n'est apportée sur la nature de l'influence contextuelle qui pourrait être dans ces expériences davantage spatiale que sémantique. Biederman a tenté de répondre à ces questions par une nouvelle étude menée en 1982 (Biederman et al., 1982). La tâche était celle d'une reconnaissance d'un objet cible au sein d'une scène présentée 150 ms. La scène était précédée d'une description verbale d'objet à comparer avec un objet dont la localisation était indiquée par un indice à la fin de la présentation de la scène (Figure 30 D). L'objet pouvait être adapté à la scène ou violait certaines règles d'interactions avec le contexte incluant la probabilité d'apparition, la position, la taille, le support, ou l'interposition.

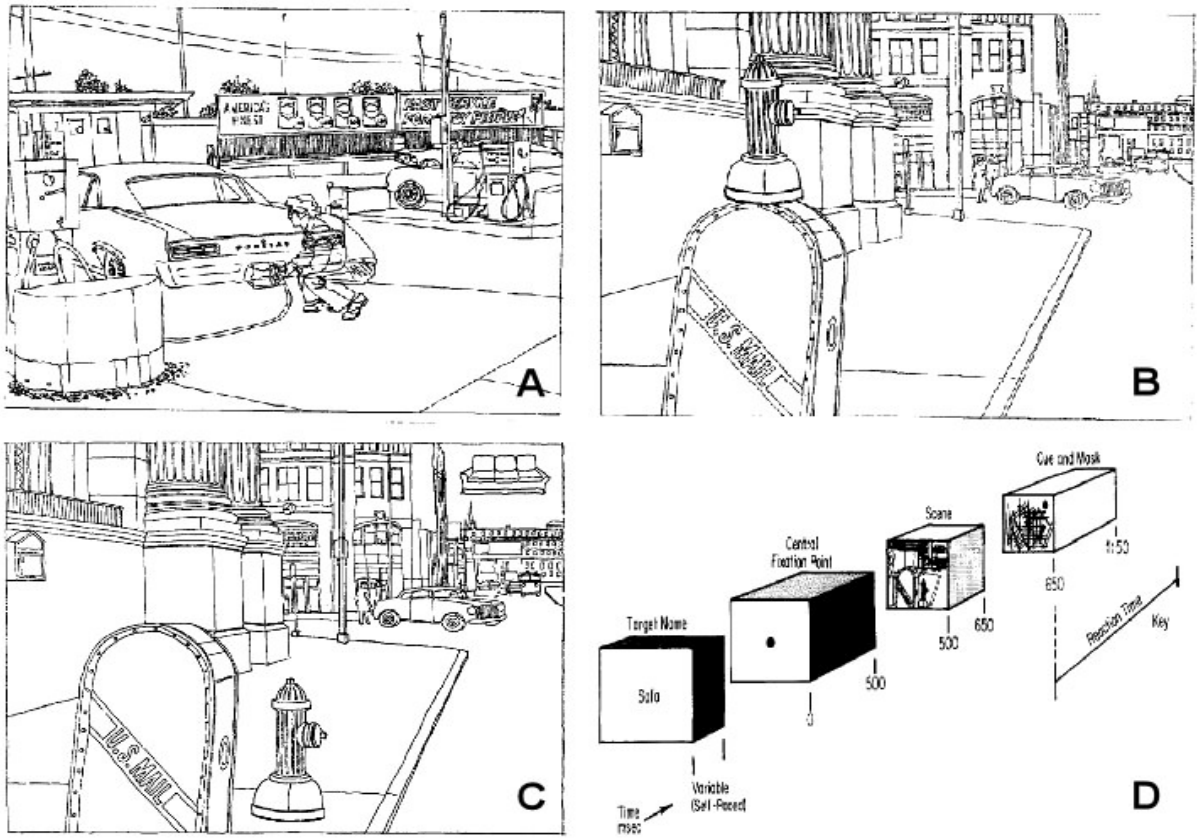

Figure n°30 : Tiré de (Biederman et al., 1982). A. Exemple de violation d'interposition entre l'homme et sa voiture. B. Violation de support entre la boite aux lettres et la borne d'incendie. C. Violation simultanée de support, de taille et de probabilité du sofa dans le ciel de la scène de rue. D. Décours temporel d'un essai.

Les résultats révèlent que les objets violant une relation d'interaction (exceptée l'interposition) sont reconnus moins précisément et plus lentement que les objets cohérents avec leur contexte. Cette interférence est d'autant plus grande que le nombre de violations est important. De plus, aucune différence n'est observée entre les conditions de violation sémantique (probabilité d'apparition, taille, position) et de violations syntactiques (support, interposition). Biederman et collaborateurs en concluent alors que l'influence contextuelle sur l'identification des objets surviendrait lors de l'analyse perceptive des objets, et dans tous les cas avant l'identification complète de l'objet. Cependant, les relations de support et d'interposition peuvent également être considérées comme sémantiques, puisqu'elles sont directement liées à l'identité de l'objet : un oiseau peut voler, un verre est transparent (Henderson, 1992). Cette influence contextuelle serait par contre indépendante de la congruence des autres objets au sein de la scène (Boyce, Pollatsek & Rayner, 1989). En effet, si les scènes présentant plusieurs objets en contexte exercent une influence sur l'identification d'un objet-cible, aucune influence de congruence n'est démontrée lorsque les autres objets de la scène sont présentés hors contexte. Enfin, pour Boyce (Boyce et al., 1989), un contexte incongruent n'interfèrerait pas plus qu'une absence de contexte.

En accord avec ces résultats, un schéma perceptif de la scène serait susceptible d'influencer via des informations descendantes les mécanismes perceptifs de l'identification d'objet.

Une autre alternative : l'influence d'un schéma mnésique de la scène sur la prise de décision

Face à cette série d'études, Hollingworth mettra l'accent sur le fait que détection et identification d'objets sont deux tâches différentes (Hollingworth & Henderson, 1998). Dans les expériences où la description de l'objet cible est présentée avant la scène, il est difficile de défendre l'hypothèse d'une aide contextuelle à l'identification des objets, les réponses pouvant être grandement biaisées par des pré-activations liées au label de l'objet pré-indiqué. Les sujets auraient ainsi plutôt tendance à affirmer la présence d'une vache si la scène présentée est une scène de ferme. Au contraire, ils auront davantage tendance à répondre négativement si une scène de rue suit le même label d'objet « vache », indépendamment de la

présence ou de l'absence de l'objet dans la scène test. Hollingworth et Henderson suggèreront ainsi des résultats biaisés dans les expériences de détection d'objet de Biederman après avoir testé l'influence contextuelle, cette fois-ci dans une réelle tâche d'identification d'objet. Utilisant le même type de stimuli, ils ont testé les sujets sur une réplique modifiée du paradigme de Biederman, dans lequel ils affichaient le label de l'objet testé après la présentation de la scène. Les auteurs ont également remplacé le choix forcé « yes/no » par un choix forcé entre deux labels d'objets susceptibles d'apparaître dans la scène. Selon ce paradigme modifié, aucune facilitation du contexte congruent n'est démontrée. Ce résultat s'accorde avec l'hypothèse d'un schéma mnésique des scènes différent du schéma perceptif préalablement cité.

Allant dans le même sens, il a été montré que des sujets peuvent catégoriser plus rapidement un nom d'objet écrit au sein d'un dessin d'objet congruent (vs. non-congruent). Or de tels résultats peuvent difficilement s'expliquer par une influence contextuelle perceptive. De manière tout à fait intéressante, cette influence contextuelle du dessin de l'objet sur la catégorisation du nom d'objet disparaissait lorsque l'ensemble était intégré au sein d'une scène contextuelle (Mathis, 2002).

Le schéma mnésique ne fournirait pas d'informations descendantes aux mécanismes perceptifs d'identification mais affecterait la disponibilité de l'information au moment de la réponse, l'intégration en mémoire, et le recouvrement des informations mnésiques après identification.

Influence de la scène globale ou des objets voisins ?

Dans les paradigmes d'exploration oculaire, les mouvements des yeux des sujets sont enregistrés lorsqu'ils explorent visuellement une scène pour en effectuer l'encodage mnésique (Antes et al., 1981, Friedman, 1979, Loftus & Mackworth, 1978) ou pour y détecter des non-objets (De Graef et al., 1990). Une grande partie de l'information visuelle étant capturée durant les fixations, la variable mesurée dans ce type de paradigme fut le plus souvent la durée des fixations oculaires sur un objet qui serait d'autant plus longue que les processus d'identification de l'objet en question sont complexes. Il a ainsi été démontré que les sujets dirigeaient rapidement leurs yeux vers les régions informatives de la scène qu'ils exploraient parfois dès la première fixation (Antes, 1974, Loftus & Mackworth, 1978). De plus, les durées de fixations sur les objets sont d'autant plus courtes que la présence de l'objet en question est

probable au sein de la scène (Antes et al., 1981, De Graef et al., 1990, Friedman, 1979, Loftus & Mackworth, 1978) et qu'il s'y trouve dans une position spatiale cohérente (De Graef et al., 1990). Cette précocité de l'effet de la congruence du contexte est cependant remise en question par De Graef qui n'a pu l'observer que sur des saccades plus tardives (De Graef et al., 1990). D'un point de vue mnésique, l'encodage d'un objet non-congruent n'appartenant pas à la frame (schéma) de la scène nécessiterait des traitements locaux longs et approfondis pour une meilleure mémorisation (Friedman, 1979 , Loftus & Mackworth, 1978). De fait, un changement appliqué sur un objet non-congruent serait alors plus facilement détectable (Friedman, 1979). Au contraire, les objets attendus dans une "frame de scène" donnée seraient automatiquement détectés (Friedman, 1979). La durée des fixations étant une mesure « online » ne donnant lieu à aucune prise de décision immédiate, l'hypothèse d'un schéma mnésique influençant les mécanismes post-identification de prise de décision est difficilement envisageable.

Ces résultats tendent à confirmer que l'identification des objets n'est pas basée sur des traitements purement ascendants et pré-conceptuels mais influencée par le schéma de la scène.

Le traitement local et l'amorçage « intra-niveau »

Pourtant, certaines questions restent de mises. La présence d'objets extrêmement diagnostiques d'une scène reste plus probable que certaine dans une scène (un frigo dans une cuisine). (1) Doit-on imaginer que le schéma de la scène ne serait pas activé en l'absence des objets fortement diagnostiques ? (2) De plus, comment envisager que le schéma de scène puisse coder l'ensemble des configurations possibles pour une catégorie de scène donnée ? Sur la base de ces observations, Henderson et De Graef défendent une nouvelle fois l'idée que le contexte global ne peut pas influencer les traitements perceptifs de l'objet mais influencerait uniquement les étapes ultérieures à son identification (De Graef et al., 1990, De Graef et al., 1992, Henderson et al., 1987). Ils proposent l'hypothèse alternative d'une amorce inter-objets. L'influence contextuelle observée dans les études décrites auparavant pourrait être due, non pas à une influence de la scène globale, mais plutôt à l'influence des objets les uns sur les autres (« amorçage intralevel », Henderson et al., 1987). Les représentations des objets qui ont été récemment identifiés en une position donnée amorceraient les représentations des objets sémantiquement reliés ou associés, et spatialement proches (Bar & Ullman, 1996). Selon cette explication, les objets congruents seraient traités

plus rapidement quand ils sont vus dans une scène parce qu'ils sont amorcés par des objets précédemment traités, et non grâce à des informations descendantes provenant du schéma global de la scène. Cette hypothèse inclut une composante attentionnelle nécessaire à la construction d'une représentation aboutie des objets qui circuleraient d'un objet donné à l'objet voisin, et ainsi de suite (voir De Graef, 1998). Dans l'étude de Boyce (1989), l'apparente influence du contexte contrôlée comme indépendante des autres objets présents dans la scène pourrait s'expliquer par le fait qu'en supprimant le contexte, les auteurs ont supprimé dans le même temps une région du contexte pouvant être considéré comme un objet (piscine, frigo…). D'autres études montrent que les influences contextuelles sont conservées lorsqu'un objet contextuel fovéal est inclus dans une scène dépourvue de contexte, (Auckland, Cave & Donnelly, 2007, De Graef et al., 1992, Henderson, 1992, Henderson et al., 1987). De manière intéressante, il a également été démontré de meilleures performances dans l'identification d'un objet cible présenté à côté d'un distracteur sémantiquement relié et fonctionnellement interactif (par exemple, un objet cible « verre » présenté à côté d'un objet distracteur « pichet » dont le bec est dirigé vers le verre) vs. fonctionnellement non-interactif (Green & Hummel, 2006). Les objets fonctionnellement groupés seraient donc également groupés d'un point de vue perceptif. A noter que dans cette étude, aucune composante attentionnelle n'était impliquée.

En résumé, dans le cadre de l'amorçage intra-niveau, la prétendue influence contextuelle résulterait en fait de l'interaction entre les routines d'identification des objets opérant sur des régions spatialement limitées de la scène (Henderson, 1992).

Finalement, il aura fallu attendre 2004 pour qu'une étude évaluant l'impact de la congruence entre objet et contexte sur la reconnaissance du contexte ou de l'objet dans les scènes naturelles soit publiée (Davenport & Potter, 2004). A partir de photos naturelles en couleur, les auteurs créèrent différents types de stimuli : des objets collés sur des contextes congruents, sur des contextes non-congruents, ou sur un fond uniforme, ainsi que des contextes sans objets en premier plan. Ces stimuli étaient présentés 80 ms suivis d'un masque. Les sujets devaient alors selon la consigne dénommer l'objet en premier plan, le contexte, ou les deux. Les auteurs montrent ainsi que (1) les objets sont plus précisément dénommés dans un contexte congruent, (2) les contextes sont mieux dénommés lorsqu'ils contiennent un objet en premier plan congruent, (3) les objets sont mieux dénommés sur fond isolé que dans un

contexte tandis que les contextes sont dénommés aussi précisément avec ou sans objet en premier plan, (4) finalement, l'effet de congruence persiste quand objets et contextes doivent être dénommés en parallèle. Dans une autre étude, Davenport démontre que cet effet de congruence contextuelle est indépendante du nombre d'objets en premier plan, mais également que la congruence d'autres objets en premier plan influence la reconnaissance de l'objet cible (Davenport, 2007).

Au vu de ces résultats, les auteurs suggèrent l'idée d'un modèle où objet et contexte seraient traités de manière interactive, les représentations sémantiques incluant objet et contexte apparié influençant la reconnaissance des objets mais également du contexte.

A noter que dans leurs études, les contextes sont globalement moins bien dénommés que les objets en premier plan suggérant l'intérêt particulier que le système visuel apporte aux objets. Il est de plus important de préciser que ces études ne contrôlent pas les caractéristiques physiques des stimuli, n'évaluent jamais le décours temporel des interactions sous-jacents, et caractérisent des réponses purement sémantiques nécessitant des traitements perceptifs majoritairement aboutis.

Si l'étude présentée dans l'article n°4 de ce chapitre aborde les mêmes questions de congruence entre objet et contexte selon une approche relativement similaire, elle répond en outre à des questions assez différentes tout en contrôlant de nombreux aspects qui auraient pu induire des biais dans les résultats obtenus par Davenport et Potter.

Ces études sur les influences contextuelles et leurs résultats parfois conflictuels ont donné naissance à une série de modèles contradictoires que je vais maintenant m'attacher à décrire.

2. Modèles d'interaction objet/contexte

Dans la reconnaissance de l'objet et de façon évidemment schématique et même simpliste, on peut situer deux phases de traitements : les traitements perceptifs et les traitements conceptuels. Les traitements perceptifs correspondraient à l'ensemble des processus depuis la transcription rétinienne jusqu'à la description structurale aboutie de l'objet. Les traitements conceptuels quant à eux correspondraient à l'ensemble des processus permettant la comparaison et la mise en correspondant du percept avec son concept et associant au percept tous les avoirs relatifs à l'objet.

Une fois de plus, ces définitions sont volontairement larges et ne précisent ni les aires impliquées, ni les latences auxquelles traitements perceptifs et conceptuels s'effectuent. Les étapes de traitements et leurs durées varient sans nul doute en fonction des objets, en fonction de la tâche, et en fonction des conditions visuelles. Enfin, la frontière entre percept et concept est floue, et sûrement dynamique.

2.1. « Perceptual model » et l'architecture triadique

Le modèle perceptif

Supporté par de nombreuses études évaluant l'influence du contexte sur la catégorisation des objets, ce modèle propose que les connaissances dérivées des informations contextuelles de la scène (le schéma, ici perceptif) telles que la relation spatiale entre les objets influenceraient l'analyse perceptive des objets présents dans la scène (Biederman et al., 1982, Boyce et al., 1989, Metzger & Antes, 1983, Palmer, 1975). Le schéma perceptif aurait tendance à faciliter les traitements perceptifs des objets congruents avec la scène, et potentiellement perturberait ceux des objets incongruents. Cette interaction entre schéma et traitements perceptifs pourrait avoir lieu à différents niveaux du système visuel. Ce modèle implique une reconnaissance de la scène au moins aussi rapide que la reconnaissance des objets et est en accord avec des résultats comportementaux caractérisés par de meilleures performances dans la reconnaissance d'objets congruents.

L'architecture triadique de Rensink

Le modèle plus complexe proposé par Rensink est aussi en accord avec l'idée d'un schéma de la scène influençant le traitement perceptif des objets (Rensink, 2000, Rensink, 2002). Cependant, comme nous allons le voir, l'approche est assez différente. Tout d'abord, son modèle se divise en 3 systèmes largement indépendants (Figure 31). Le premier système purement perceptif permet la création de proto-objets, c'est à dire des percepts volatiles d'objets inaccessibles à la conscience et sans concepts associés. Le second système est un système non-attentionnel mais à capacité limitée permettant la mise en place d'un cadre de référence impliquant le gist, la structure spatiale et le schéma de la scène. Au sein de ce deuxième système, le gist permettrait de préciser les objets d'intérêts, la structure spatiale fournirait une localisation probable de ces objets. De l'interaction entre le gist et la structure spatiale émergerait le schéma facilitant globalement la perception des objets.

Enfin, le troisième est un système attentionnel à capacité limitée qui permet la stabilisation d'un seul percept d'objet induisant ainsi l'émergence de son concept. A noter que ces 3 systèmes fonctionneraient en continu, suggérant la possibilité de traitements parallèles des objets et du contexte

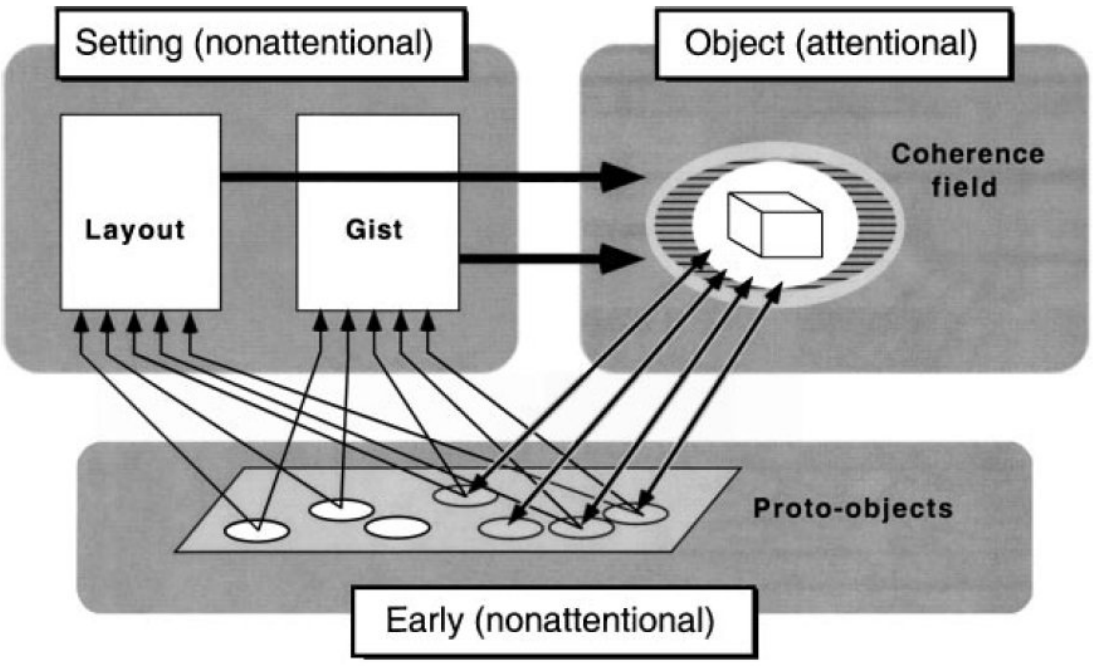

Figure n°31 : Tiré de (Rensink, 2002). Les 3 systèmes de l'architecture triadique.

Lors de la perception d'une nouvelle scène, les traitements perceptifs permettraient la construction d'un proto-objet associé à chaque objet desquels pourraient émerger la structure spatiale, le gist, et par là même le schéma de la scène. Une fois le schéma de la scène évalué, les proto-objets relatifs à des objets attendus (congruents) et donc compris dans le schéma seraient automatiquement détectés et de ce fait, ne nécessiteraient pas de traitements supplémentaires pour être associés à leurs concepts. Au contraire, les proto-objets d'objets incongruents non inclus dans le schéma, nécessiteraient une prise en charge attentionnelle pour la mise en place de traitements conceptuels. Une fois le concept d'un objet incongruent construit, la sémantique associée à ce concept serait alors disponible et le schéma de la scène pourrait être potentiellement mis à jour. Le système attentionnel ayant une capacité limitée, les traitements conceptuels ne pourraient s'effectuer à un instant donné que sur un objet incongruent. La représentation stable de notre environnement ne serait qu'illusion, résultant d'un transfert rapide de l'attention d'un objet vers un autre. Ainsi, ce modèle d'architecture triadique peut se placer dans les modèles perceptuels puisque le schéma de la scène via un système attentionnel influence la prise en charge des proto-objets par les traitements conceptuels.

2.2. « Priming model » Le modèle d'amorçage et le modèle de Bar

Au contraire des modèles perceptifs, les modèles d'amorçage supportent l'idée d'une influence contextuelle haut-niveau lors de la formation du concept de l'objet, et plus précisément, lors de la mise en correspondance entre percept et concept. Le schéma de la scène permettrait de pré-activer l'ensemble des représentations d'objets congruents susceptibles d'apparaître dans la scène modulant ainsi la quantité d'informations perceptives à traiter pour établir une correspondance avec leur concept (Bar & Ullman, 1996, Friedman, 1979, Palmer, 1975). Dans ce sens, moins d'informations perceptives seraient nécessaires pour la formation d'un concept d'objet congruent (vs. incongruent) expliquant les meilleures performances comportementales obtenues dans l'identification d'objets congruents.

Le modèle de Bar

Le modèle proposé par Bar est basé sur les propriétés physiologiques des systèmes magnocellulaire et parvocellulaire, sur notre capacité à reconnaître un objet ou une scène à partir des informations grossières fournies par les fréquences spatiales basses, et sur les résultats comportementaux enregistrés dans les tâches de reconnaissance d'objets sous influence contextuelle (Bar, 2004). Nous verrons également dans un chapitre ultérieur (Part III, chap. 2.2.5) que ce modèle s'inscrit efficacement dans un modèle neuroanatomique détaillé et largement documenté.

D'un point de vue physiologique, une information visuelle grossière et achromatique au sein du système magnocellulaire serait plus rapidement intégrée que l'information visuelle détaillée transitant au sein du système parvocellulaire (dans ce mémoire, Part. I, chap. 3.3.3). Ainsi des représentations grossières de l'objet et de la scène extraites de l'information contenue dans les basses fréquences spatiales pourraient être rapidement construites (respectivement dans le cortex pré-frontal et dans le cortex para-hippocampique, Figure 32). L'interaction entre les représentations grossières de la scène et de l'objet (au sein du cortex inféro-temporal) pourrait alors contraindre suffisamment le répertoire des objets possibles pour en déduire une identité probable de l'objet. Plus tard, la complétion des informations par l'intégration supplémentaire des informations contenues dans les fréquences spatiales hautes permettrait d'affiner notre représentation de l'objet.

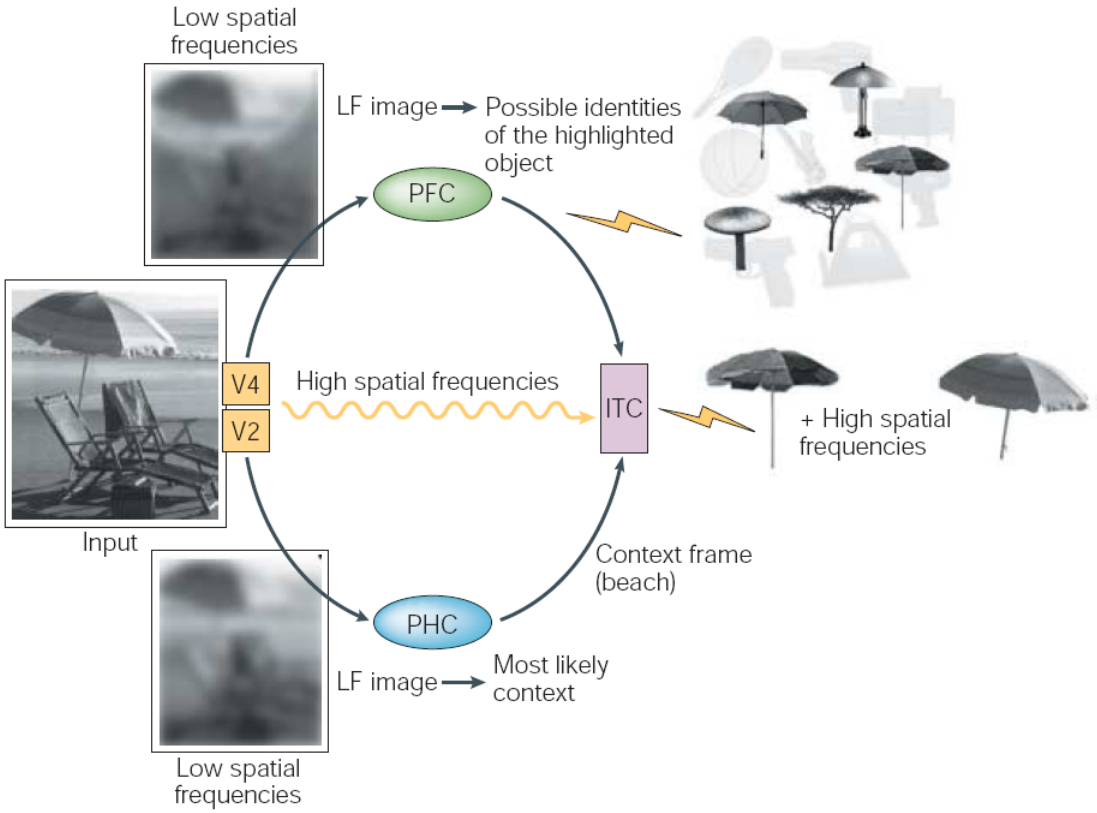

Figure n°32 : Tiré de (Bar, 2004). Modèle proposé pour l'aide contextuelle sur la reconnaissance des objets. Les croisements entre associations contextuelles relatives au contexte grossièrement perçu (plage) et les interprétations possibles de l'objet cible permettent une reconnaissance rapide de l'objet en tant que parasol. L'intégration plus tardive des hautes fréquences spatiales permettra d'identifier le parasol de manière plus spécifique. Seules les principales voies et flux d'informations sont illustrés sur le schéma.

2.3. « Functional isolation model » Le modèle d'isolation fonctionnelle

Le modèle d'isolation fonctionnelle supporte l'idée d'une identification de l'objet isolée des attentes dérivées de la scène (Hollingworth & Henderson, 1998 , Hollingworth & Henderson, 1999). Ce modèle compatible avec certaines théories sur l'identification d'objet (Biederman, 1987, Bulthoff & Edelman, 1992) suggère que les traitements d'analyse ascendants seraient suffisants à reconnaître les objets. Les résultats expérimentaux démontrant une facilitation de l'identification des objets congruents seraient biaisés par les conditions expérimentales, et seraient dus à une influence de la scène lors de la prise de décision (Hollingworth & Henderson, 1998, Hollingworth & Henderson, 1999).

2.4. « Interactive model » Le modèle interactif

Ce dernier type de modèle, récent et très peu détaillé, s'oppose complètement au modèle d'isolation fonctionnelle et suggère que les traitements relatifs aux objets et au contexte opèrent de manière interactive (Davenport, 2007, Davenport & Potter, 2004, Humphreys & Riddoch, 1995). Ce modèle fut proposé par Davenport et Potter après avoir révélé pour la première fois une influence de l'objet sur la dénomination de contexte au sein de scènes naturelles. Puisque le contexte influence la reconnaissance des objets, et qu'à l'inverse, l'objet influence la reconnaissance de son contexte et des autres objets, cela suppose que les traitements relatifs aux objets et à leur contexte se contraignent mutuellement.

A noter cependant qu'aucune précision n'est apportée sur le niveau du système visuel auquel les interactions entre ces traitements ont lieu. Etant donné que dans cette dernière étude, les sujets devaient donner une réponse tardive (tâche de dénomination), les interactions entre objet et contexte pourraient très bien être ultérieures aux traitements perceptifs. A noter également que le modèle proposé par Bar pourrait parfaitement s'intégrer dans ce modèle interactif. Si l'auteur ne discute que de l'influence contextuelle sur l'objet, il semble qu'aucune restriction ne soit de mise concernant une influence de l'objet sur son contexte.

Nous verrons qu'au vu des données décrites dans l'article n°4, ce modèle interactif apparaît des plus séduisants, nos données corroborant grandement l'idée d'interactions bi-directionnelles précoces entre les traitements de l'objet et de son contexte. Des précisions seront de plus apportées sur les niveaux de traitements sujets aux interactions.

2.5 Quelques données cliniques

Tel que rapporté dans la Partie I Chapitre 3.2.1, la principale aire semblant s'activer spécifiquement dans le traitement d'une scène globale est le cortex para-hippocampique (PHC), et plus précisément la PPA, ainsi que le gyrus fusiforme. Cette activation est précoce, dès 130 ms après la présentation des stimuli (Aguirre et al., 1998, Bar & Aminoff, 2003, Epstein et al., 1999, Epstein & Kanwisher, 1998). Une deuxième activation de ces aires est également enregistrée après 230 ms. Ces deux phases d'activation sont interprétées par Bar comme la mise en place d'une première représentation grossière de la scène suivie d'une seconde représentation plus affinée. Le rôle du PHC dans le traitement spatial des scènes et de

l'information topographique a d'ailleurs été largement confirmé (Aguirre, Detre, Alsop & D'Esposito, 1996, Maguire, Burgess, Donnett, Frackowiak, Frith & O'Keefe, 1998, Stern, Corkin, Gonzalez, Guimaraes, Baker, Jennings, Carr, Sugiura, Vedantham & Rosen, 1996), ainsi que son importance dans l'intégration d'informations à grande échelle (Levy et al., 2001). Cependant, d'autres études laissent penser que le PHC n'est pas spécifique d'un traitement de la scène mais s'activerait lors de traitements associatifs et contextuels plus généraux (Aminoff, Gronau & Bar, 2007, Bar & Aminoff, 2003, Burwell, Saddoris, Bucci & Wiig, 2004, Dale, Liu, Fischl, Buckner, Belliveau, Lewine & Halgren, 2000, Duzel, Habib, Rotte, Guderian, Tulving & Heinze, 2003, Halgren, Baudena, Heit, Clarke, Marinkovic, Chauvel & Clarke, 1994, Mendez & Cherrier, 2003 , Sperling, Chua, Cocchiarella, Rand-Giovannetti, Poldrack, Schacter & Albert, 2003) incluant aussi bien les associations entre objet et contexte que les associations entre divers objets.. Les informations de contextes spatiaux seraient plus fortes dans la partie postérieure du PHC et les informations contextuelles non-spatiales plutôt dans la partie antérieure.

Qu'en est-il alors des interactions visuelles entre objets et contexte ? Lors de la comparaison des ERPs obtenus suite à la présentation d'objets en contextes congruents et incongruents, Ganis et Kuta ont eux aussi montré une activité différentielle dans le PHC (Ganis & Kutas, 2003) cependant très tardive (300-500 ms). Dans le même sens, Bar et Aminoff ont présenté à des sujets des photographies d'objets associés de façon plus ou moins forte à des contextes donnés et ont révélé par IRMf une activité dans la PPA modulée par la force de l'association entre objet et contexte (Bar & Aminoff, 2003). Ils ont de plus isolé un second foyer d'activation dans le cortex rétrosplénial (RSC) impliqué dans l'analyse de l'information spatiale (Cooper & Mizumori, 2001). Finalement, les auteurs ont révélé un troisième foyer d'activation au niveau du sulcus occipital supérieur (Bar, 2004, Bar & Aminoff, 2003) dont il propose l'implication dans une mise à jour continue du contexte et une facilitation de la reconnaissance des objets et de la scène via des informations descendantes.

Intégrant ces données neuroanatomiques à son modèle, Bar propose que le PHC serait une sorte de « buffer » organisant et associant des informations (1) multisensorielles en provenance du RSC et du gyrus cingulaire, (2) visuo-spatiales en provenance du cortex parietal postérieur (voie dorsale), (3) auditives depuis le gyrus temporal supérieur, (4) somatosensorielles depuis l'insula, et visuelles depuis TEO et le cortex périrhinal. Dans le cadre de l'influence démontrée du contexte sur la reconnaissance de l'objet, Bar propose que les informations contenues dans les basses fréquences spatiales d'une scène pourraient

rapidement être envoyées dans leur globalité vers le PHC et centrées, via l'intervention de l'attention, sur l'objet d'intérêt dans le PFC. Ces deux types d'informations respectivement intégrées dans le PHC et le PFC seraient après traitement renvoyées vers le cortex inféro-temporal et leurs comparaisons aboutiraient à l'identification du label de l'objet, ce qui expliquerait une identification plus facile dans le cas des objets congruents.

A côté de cet ensemble d'études menées par le groupe de Moshe Bar et synthétisé dans un modèle encore spéculatif sur plusieurs points, rares sont les données neuro-anatomiques caractérisant les associations entre objet et contexte. On peut cependant citer quelques travaux tels ceux d'Altmann qui à l'aide d'un protocole d'adaptation IRMf, démontre que les cellules du LOC codant en partie les informations contextuelles d'un objet sont modulées par la segmentation figure/fond de l'objet en question (Altmann, Deubelius & Kourtzi, 2004). Egalement inscrite dans un paradigme d'adaption IRMf et utilisant des photographies comme stimuli, une étude menée par Goh (Goh, Siong, Park, Gutchess, Hebrank & Chee, 2004) confirme une sensibilité des gyrii fusiformes aux traitements des objets et une certaine sélectivité des régions para-hippocampiques aux contexte des scènes. De plus, des régions para-hippocampiques plus rostrales ainsi que l'hippocampe droit seraient spécifiquement impliqués dans l'association objet/contexte. Ainsi, des patterns IRMf d'adaptation différents dans le LOC, l'hippocampe et les aires temporo-médiales entre sujets contrôles et personnes âgées ou amnésiques témoigneraient de la difficulté de ces sujets tests à encoder de nouvelles associations contextuelles (Chee, Goh, Venkatraman, Tan, Gutchess, Sutton, Hebrank, Leshikar & Park, 2006, Chun, 2000).

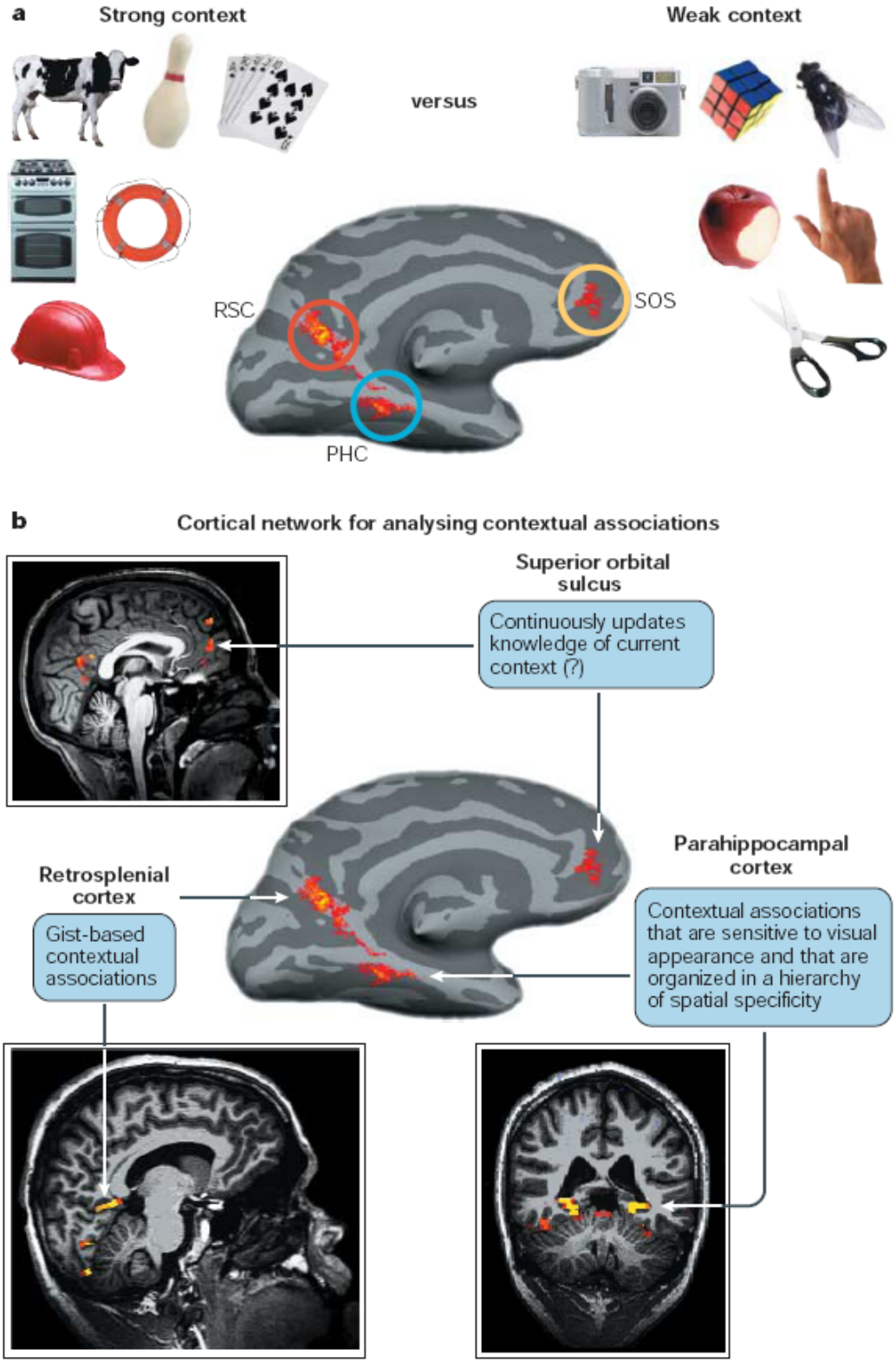

Figure n°33 : Tiré de (Bar, 2004). Aires corticales impliquées dans le traitement contextuel. (a) Zones d'activations obtenues par méthode soustractives en IRMf entre la perception d'objets fortement associés à un contexte spécifique et la perception d'objets non associés à un contexte unique. Le rôle supposé de chaque aire corticale activée est précisé en (b).

3. Retour sur l'article n°3 : Une influence peu soupçonnée de l'objet sur la catégorisation du contexte

3.1. Analyse post-hoc : méthodologie

Comme je l'ai décrit dans le chapitre consacré à l'article n°3, nous avons dans cette étude impliqué nos sujets dans une tâche de catégorisation de contexte à un niveau superordonné « Env. Man. » vs. « Env. Nat. ». Les premières analyses des réponses comportementales go/no-go nous ont permis d'estimer une précision moyenne des sujets autour de 96% pour des temps de réaction médian inférieurs à 400 ms.

Une analyse grossière des performances par image nous a permis de noter que certaines scènes présentées durant l'expérience avaient tendance à être catégorisées plus tardivement. Ces scènes présentaient souvent un objet saillant en premier plan. Nous avons donc réalisé une analyse post-hoc pour définir objectivement l'effet de l'effet d'un objet saillant sur la catégorisation du contexte.

Etant donné que la saillance est une valeur difficilement mesurable, nous avons privilégié l'utilisation d'un questionnaire électronique à partir duquel les sujets ayant passé l'expérience ont pu dire subjectivement si un objet saillant était présent dans chacune des scènes étudiées. S'ils affirmaient la présence d'un objet saillant, ils devaient alors par la suite préciser si cet objet saillant était congruent ou incongruent avec son contexte en termes de catégorie « naturel » vs. « manufacturé ». De fait, un chien dans une maison était considéré comme incongruent avec son contexte, le chien étant un objet naturel dans un contexte manufacturé. Au contraire, un canapé au milieu de la route aurait été classé comme congruent avec son contexte étant donné qu'il représente un objet manufacturé dans un environnement manufacturé.

A noter que plusieurs images contenaient des êtres humains (classés comme congruents dans les milieux manufacturés).

Précision et comportement ont ainsi été ré-analysés selon ces 3 nouvelles conditions : (1) Scènes sans objet saillant, (2) Scènes avec un objet saillant congruent, et (3) Scènes avec un objet saillant incongruent.

3.2. Résultats résumés

La présence d'un objet saillant congruent dans la scène entraîne une baisse de précision moyenne non significative de 2.6 à 4.8 % tandis qu'un retard significatif de la prise de décision de 25 à 37 ms est observé.

Si l'objet saillant est de plus incongruent avec son contexte, les performances des sujets laissent apparaitre une baisse de précision supplémentaire de 14% avec des temps de réactions médians augmentés de 81 ms supplémentaires en moyenne.

3.3. Discussion

Ces résultats démontrent une interférence des objets sur la catégorisation visuelle rapide du contexte global des scènes naturelles. Cette interférence est d'autant plus grande que les objets sont saillants et sémantiquement incongruents. Cet effet peu soupçonné jusque là avait déjà été décrit dans une seule autre étude à ma connaissance (Davenport & Potter, 2004). Cependant, cette étude menée par Davenport utilisait des objets collés au sein des scènes, les rendant de fait potentiellement plus saillants que des photographies présentant naturellement des objets saillants dans leur environnement : l'influence de l'objet sur la reconnaissance du contexte pouvait être alors un biais lié aux manipulations physiques des images. L'analyse post-hoc décrite dans ce chapitre confirme l'interférence des objets sur la catégorisation du contexte indépendamment de biais physiques potentiels. Des implications de cette interférence de l'objet sur la nature des traitements sous-jacents seront avancées dans la discussion du chapitre suivant.

4. Influence immédiate du contexte sur la catégorisation de l'objet à un niveau superordonné

4.1. Objectifs

Nous avons dans l'étude précédente révélé une interférence des objets sur la catégorisation du contexte, suggérant d'ores et déjà des interactions entre les traitements catégoriels de l'objet et du contexte au sein des scènes naturelles.

L'expérience ultime consistait donc à tester l'influence du contexte sur la catégorisation des objets. Cette étude, acceptée dans « Journal of Vision », teste les performances des sujets dans une tâche de catégorisation go/no-go animal/non-animal, les distracteurs étant des objets manufacturés, apparaissant sur des contextes divers. Nous avons donc pour cette expérience extrait un ensemble d'animaux et d'objets de leur contexte original pour les coller par la suite dans de nouveaux contextes. Cependant, cette manipulation obligatoire des stimuli ne devant pas être sans conséquences sur la reconnaissance des objets, nous en avons d'abord étudié les effets.

4.2. Les objets isolés ne sont pas catégorisés plus rapidement que les objets en contexte

Protocoles

Dans une première expérience, nous avons donc évalué la capacité des sujets à catégoriser des objets : (1) dans leurs contextes originaux (photographie intacte), (2) collés sur un fond gris uniforme, ou (3) collés dans un autre contexte congruent. Les contextes congruents des animaux consistaient en des « Env. Nat » tandis que les contextes congruent des objets distracteurs étaient des « Env. Man. ». Les stimuli étaient construits de façon à préserver au mieux la taille absolue des objets, leur position dans l'image, les contrastes locaux, les informations de luminance au travers des 3 conditions. De plus, animaux et objets étaient positionnés de manière à respecter les règles de support et d'interposition définies

comme syntactiques par Biederman, ainsi que leur taille relative (par rapport au contexte). Cette expérience avait deux objectifs principaux : évaluer le traitement de l'objet en contexte par rapport à celui de l'objet isolé et caractériser les biais potentiels liés à la manipulation des images sur les performances de catégorisation.

Résultats résumés

Lorsque les animaux et objets distracteurs sont présentés dans leurs photographies originales, les sujets parviennent à effectuer la tâche de catégorisation animal/non-animal avec 96 % de réussite, et ce en moins de 380 ms. De manière surprenante, la présentation de l'objet isolé de son contexte (la ségrégation de l'objet étant déjà effectuée) n'a que très peu d'effet sur la catégorisation des objets (temps de réaction médian augmenté de 5 ms). Au contraire, les biais physiques liés au collage des stimuli ont pour conséquence une perte de précision de 2% et une augmentation des temps de réaction médians de 11 ms.

Discussion

Alors que depuis de nombreuses années, un nombre important d'études suggèrent une précédence de la ségrégation de l'objet sur les traitements menant à sa reconnaissance (Biederman & Shiffrar, 1987, Kosslyn, 1987, Marr, 1982), les résultats de cette étude démontrent qu'un objet isolé n'est pas catégorisé plus rapidement qu'un objet en contexte. Une explication pourrait venir du fait que les stimuli utilisés dans notre étude étaient plus complexes et écologiques que dans les études précédentes. En effet, comparée aux dessins faits main, l'utilisation de photographies couleurs pourrait améliorer la reconnaissance des objets et des scènes (Oliva & Schyns, 2000, Wurm et al., 1993) Néanmoins, une étude assez similaire menée par Davenport et basée sur la présentation brève de photographies couleurs révèle également de meilleures performances pour les objets isolés dans une tâche de dénomination (Davenport & Potter, 2004). Une hypothèse alternative réside dans la différence entre les tâches à effectuer et les traitements perceptifs sous-jacents. En effet, catégorisation et identification n'impliquent pas l'extraction et l'intégration de la même quantité d'informations. La catégorisation peut se baser sur la détection de quelques traits physiques

diagnostiques de la catégorie cible tandis que l'identification nécessiterait des traitements perceptifs plus aboutis. Afin de tester cette hypothèse, il serait intéressant de comparer les performances de sujets impliqués dans une catégorisation plus fine d'objets (chien/non-chien par exemple) isolés et en contexte. Cette hypothèse va dans le sens d'une première analyse globale de la scène qui serait affinée par la suite par des traitements plus locaux et plus fins pour lesquels la ségrégation serait alors nécessaire. Une dernière explication réside dans le fait que ces performances similaires seraient le reflet d'effets opposés. La condition isolée par exemple bénéficierait d'une facilitation due à une ségrégation déjà effectuée mais sans l'apport potentiel d'une aide contextuelle. A l'inverse, l'objet en contexte ne bénéficierait pas d'une ségrégation déjà effectuée mais de l'aide contextuelle. Des études supplémentaires sont nécessaires pour une meilleure compréhension des données obtenues.

L'autre donnée intéressante obtenue ici concerne l'effet délétère de la manipulation des stimuli. Coller un objet dans un autre contexte même congruent entraîne une perturbation des performances. L'ensemble des précautions prises pour respecter toutes les règles mises en évidence par Biederman (support, échelle, interposition etc…) n'a pas suffit et d'autres éléments sont sans doute à prendre en considération dans les scènes naturelles. Il est donc clair que la manipulation de photographies de scènes naturelles n'est pas triviale et qu'elle introduit des violations dont on ne connaît pas encore l'origine (éclairage de l'objet, ombrages…). Le système visuel, particulièrement performant pour traiter ce type de scènes est également facilement perturbé dès que la "régularité" de ces scènes est altérée par des manipulations comme les insertions d'objets utilisées ici.

4.3. Les objets sont plus rapidement catégorisés dans un milieu congruent

Protocole

Dans une seconde expérience, nous avons estimé l'effet de la congruence du contexte sur la catégorisation des objets. Animaux et objets distracteurs pouvaient ainsi apparaître soit dans un « Env. Man » soit dans un « Env. Nat. ». Les mêmes précautions de construction des stimuli que dans la première expérience ont été prises, et pour s'affranchir des effets dus aux

manipulations des stimuli révélés dans l'expérience précédente, tous les stimuli présentaient des objets ou animaux collés dans un contexte que ce contexte soit congruent ou incongruent. Si aucune interaction entre les traitements catégoriels de l'objet et du contexte n'a lieu, les performances de catégorisation dans ces deux conditions devraient alors être similaires. Si des interactions existent, une étude de ces interactions en fonction des temps de réaction du sujet devrait mettre en évidence le délai nécessaire à l'établissement de ces interactions.

Dans une troisième expérience, nous avons finalement testé si les effets observés dans la seconde expérience étaient généralisables. Pour cela, la moitié des animaux cibles et des objets distracteurs ont été collés dans d'autres contextes congruents et incongruents.

Résultats résumés

L'analyse des données révèle que la présentation des objets dans des contextes incongruents induit une baisse de précision de 10% associée à une augmentation des temps de réaction médian de 16 ms. Les distributions des temps de réaction des réponses correctes et les courbes d' révèlent que cet effet de congruence affecte même les réponses les plus précoces, suggérant ainsi une influence immédiate du contexte sur la catégorisation des objets.

L'effet de congruence du contexte est retrouvé dans la troisième expérience en testant de nouveaux sujets sur un nouvel ensemble de stimuli. Cet effet est donc robuste et indépendant des ensembles de stimuli construits.

Enfin l'effet de congruence ne dépend pas de la saillance de l'objet dans l'image. Que l'objet soit très saillant, moyennement ou peu saillant l'effet de congruence se retrouve dans tous les cas.

Discussion

L'effet de la congruence du contexte est quant à lui très clair. Alors que les mêmes biais physiques sont présents dans la condition congruente et dans la condition incongruente, notamment l'effet de collage, les objets restent plus efficacement catégorisés en contexte

congruent. Cette congruence est plutôt à envisager d'un point de vue perceptif que d'un point de vue sémantique. En effet, vu les contraintes temporelles imposées par la tâche, il est difficilement possible d'envisager que la sémantique de la scène puisse influencer la catégorisation des objets lors des réponses les plus précoces. Les 400 ms mesurées par les temps de réaction médians suffisent juste à déterminer la catégorie superordonnée de l'objet ou du contexte. Dans ce temps réduit, le cerveau humain est loin d'avoir le temps d'être confronté aux règles de régularités probabilistes telle que : « un crocodile est habituellement vu dans les marécages ». Il a uniquement le temps d'accéder à des représentations vagues telles qu' « Env. Man. » ou «Env. Nat », catégories finalement plus perceptives que sémantiques. Ainsi l'effet de congruence pourrait résulter d'une accumulation d'informations plus consensuelles lorsque l'animal, objet naturel, est présenté dans un environnement naturel. Cela suppose l'existence de traits physiques diagnostiques communs aux catégories naturelles, que ce soit des catégories d'objet ou de contexte, On peut également imaginer l'existence de populations neuronales sélectives aux caractéristiques physiques animales et de populations neuronales sélectives aux caractéristiques des environnements naturels qui auraient été habituées à décharger de façon simultanée, à l'opposé des populations neuronales codant pour les « Env. Man » moins souvent activées en présence d'animaux. La coactivation répétée de populations neuronales permettrait d'établir des liens fonctionnels facilitateurs entre ces populations neuronales. Cette hypothèse explique aussi bien les interférences de l'objet sur la catégorisation du contexte décrites dans le chapitre précédent que l'influence contextuelle démontrée dans l'étude présente. Dans tous les cas, objets et contextes superordonnés seraient traités en parallèle et de manière purement ascendante selon des décours temporels similaires laissant libre place à des interactions continues entre populations neuronales diagnostiques. Au contraire, les effets d'interférence seraient dus à la coactivation de populations neuronales "conflictuelles" car activées simultanément que de façon très ponctuelle.

On peut imaginer que pour des catégorisations plus fines d'objets (ours/non-ours par exemple) pour lesquelles 40-60 ms de traitements supplémentaires seraient nécessaires, l'influence du contexte s'affinerait également. Les différences congruent/non-congruent pourraient alors apparaitre entre catégories basiques de contexte.

Article n°4

Early interference of context congruence on object processing in rapid visual categorization of natural scenes

---

Journal of Vision
(sous presse)

Olivier R Joubert, Denis Fize, Guillaume A. Rousselet,
& Michèle Fabre-Thorpe

Early interference of context congruence on object processing in rapid visual categorization of natural scenes

Olivier R. Joubert[1,2]
1. Université de Toulouse, CerCo, UPS
2. CNRS, UMR 5549, Faculté de Médecine de Rangueil, Toulouse, France

Denis Fize[1,2]
1. Université de Toulouse, CerCo, UPS
2. CNRS, UMR 5549, Faculté de Médecine de Rangueil, Toulouse, France

Guillaume A. Rousselet[3]
3. Centre for Cognitive Neuroimaging (CCNi), Department of Psychology, University of Glasgow, UK

Michèle Fabre-Thorpe[1,2]
1. Université de Toulouse, CerCo, UPS
2. CNRS, UMR 5549, Faculté de Médecine de Rangueil, Toulouse, France

Whereas most scientists agree that scene context can influence object recognition, the time-course of such object/context interactions is still unknown. To determine the earliest interactions between object and context processing, we used a rapid go/no-go categorization task in which natural scenes were briefly flashed and subjects required to respond as fast as possible to animal-targets. Targets were pasted on congruent (natural) or incongruent (urban) contexts. Experiment 1 showed that pasting a target on another congruent background induced performance impairments, whereas segregation of targets on a blank background had very little effect on behavior. Experiment 2 used animals pasted on congruent or incongruent contexts. Context incongruence induced a 10% drop of correct hits and a 16ms increase in median reaction times, affecting even the earliest behavioral responses. Experiment 3 replicated the congruency effect with other subjects and other stimuli, thus demonstrating its robustness. Object and context must be processed in parallel with continuous interactions possibly through feed-forward co-activation of populations of visual neurons selective to diagnostic features. Facilitation would be induced by the customary co-activation of "congruent" populations of neurons whereas interference would take place when conflictual populations of neurons fire simultaneously.

Keywords: natural scenes, early scene processing, object categorization, object-context interactions, congruence

Introduction

In our everyday rich and complex surrounding world, objects are embedded in visual scenes. Because of repetitive co-occurrence of objects or because of co-occurrence of a given object in a specific contextual frame or schemata, our brains can

generate expectations (Bar & Ullman, 1996, Biederman, Rabinowitz, Glass & Stacy, 1974, Palmer, 1975). These expectations in terms of objects will not be the same when walking in a busy street or along a path in the country, so that object recognition can be facilitated (a car in the street) or perturbed (a telephone box in the country) by such expectations. Our visual system is able to handle statistical regularities and object co-occurrences in our complex visual world even during passive viewing of visual scenes (Fiser & Aslin, 2001). It can learn relevant covariations and use implicit memory representations to guide search behavior (Chun & Jiang, 1999, Jiang & Chun, 2001) and also use size, orientation and location of the object in a scene to make hypotheses about its identity when not given enough information (Oliva & Torralba, 2007).

The idea according to which a consistent context can facilitate object detection, recognition and naming is generally accepted (Biederman, Mezzanotte & Rabinowitz, 1982, Boyce & Pollatsek, 1992, Boyce, Pollatsek & Rayner, 1989, Palmer, 1975). These studies employed line drawing of scenes and objects. In Palmer's study, subjects could analyze a scene for 3 seconds before the presentation of the object they had to identify and all processing stages of the visual pathway had ample time to be influenced by top-down knowledge and expectations. In Biederman's study the name of the object to look for (in a subsequently cued location) was provided before stimulus presentation, again leaving time for expectation to influence behavior and performance was impaired with incongruent objects. Eye-tracking studies confirm this congruence effect by showing that objects inconsistent with a scene tended to be fixated longer than consistent ones (De Graef, Christiaens & d'Ydewalle, 1990, De Graef, De Troy & D'Ydewalle, 1992). However, the hypothesis of a contextual influence on object processing has been challenged by Hollingworth & Henderson (Hollingworth & Henderson, 1998, Hollingworth & Henderson, 1999) who reported that after eliminating guesses and response biases, no advantage was found for the detection of consistent objects (over inconsistent ones). They proposed that object identification processes are isolated from knowledge about the world.

If the extent to which object processing is influenced by its context is still an ongoing debate, the time course of context/object interaction is even more controversial. If context and object processing interfere, such interactions could happen late, after activation of semantic information (Ganis & Kutas, 2003). Alternatively, context could also affect the perceptual processing of object and influence early processing, or set constraints on its possible interpretations. A model proposed by Bar and collaborators suggests fast interference between context and object. According to this model, rapid coarse processing

of a scene, possibly through the dorsal magnocellular pathway, would be used to activate the most likely possible object(s) in a contextual frame (Bar, 2004, Bar, Kassam, Ghuman, Boshyan, Schmid, Dale, Hamalainen, Marinkovic, Schacter, Rosen & Halgren, 2006). Indeed, a coarse "blurred" representation of the contextual frame might be sufficient to guide object processing. The structure of a scene image can be estimated by the basis of global image features that provide a "statistical summary" of its spatial layout properties. Thus, natural image statistics could also be used in scene categorization (Fiser & Aslin, 2001, Torralba & Oliva, 2003), allowing feed-forward processing of scene content and providing early contextual information that can influence object processing (Oliva & Torralba, 2006, Oliva & Torralba, 2007).

Getting at the gist of a scene can be achieved at a glance (Potter & Faulconer, 1975, Potter & Levy, 1969). Using a rapid categorization task frequently used to study the time course of object processing (Fabre-Thorpe, Delorme, Marlot & Thorpe, 2001, Mace, Thorpe & Fabre-Thorpe, 2005, Rousselet, Mace & Fabre-Thorpe, 2003, Thorpe, Fize & Marlot, 1996), we have recently shown that categorizing the global gist of a scene is remarkably fast (Joubert, Rousselet, Fize & Fabre-Thorpe, 2007), with a time-course similar to that of object categorization. In comparison, contextual categorization at more detailed levels such as sea, mountain, urban indoor or outdoor contexts is more time-consuming (Rousselet, Joubert & Fabre-Thorpe, 2005). Contrary to everyday life in which the visual system has ample time to be preset by contextual expectations despite eye and head movements, our rapid categorization task uses natural photographs flashed for only 26 ms or less and thus does not allow time for top-down expectancy influences based on context. In daily life this situation is more likely to happen when one zaps between TV channels or turns over the pages of a magazine. In such situations, contextual information has to be processed in parallel with object features.

Earlier studies have reported contextual influence on object processing with such briefly presented scenes. Davenport and Potter (Davenport & Potter, 2004) used manipulated photographs containing a salient object that was consistent (or not) with its context. The scenes were presented for only 80 ms then masked, and subjects were asked to name the object. They reported that objects were named more accurately in semantically related contexts than in non-consistent contexts. Recent studies (Davenport, 2007; Joubert et al., 2007) have also shown that context and object processing could interact in both directions: context can influence object processing but salient objects can also disturb context processing. In both studies the accuracy of background reports was influenced by the consistency of foreground objects. When subjects were just required to categorize scenes as natural or urban in a rapid categorization task (Joubert et al., 2007), the presence of an in-

congruent salient object in the scene induced an accuracy drop of about 10% correct and delayed correct responses by as much as 80 ms.

The evidence of such interactions in rapid processing suggests that context and object processing must interact early, but the existence of such early interactions has not yet found support from associated brain activity. Analyzing EEG associated with the processing of objects embedded in a congruent or incongruent context, Ganis and Kutas (2003) observed that the earliest signs of activity related to congruity vs. incongruity were recorded in a late 300-500 ms window. This result supports theories postulating a late effect of context on object processing that would depend upon activation of semantic information. Thus, it still remains to be determined when the earliest influences of context on object processing take place. In Davenport's experiments, subjects were asked to provide a verbal response and no reaction times were provided. The manual go/no-go categorization task often used in our group requires a motor response that has to be provided "as quickly and accurately as possible". In such task, median RT are generally around 400 ms or less (Joubert et al., 2007); responses might not even require conscious representations (Thorpe, Gegenfurtner, Fabre-Thorpe & Bulthoff, 2001). Such rapid responses and the precise quantification of reaction times can provide information about the time course of contextual influence on fast object processing.

This was the aim of the present study. Subjects were asked to perform an animal/non-animal fast categorization task on briefly flashed natural scenes using a very large number of stimuli to avoid possible biases. The brief presentation of the stimuli (26 ms) prevented eye movements and scene exploration. Natural scenes were manipulated so that context and object could be either congruent or not. The robustness of the effect was studied by pasting objects in various congruent or non-congruent contexts. In a preliminary experiment, we controlled for the effect of simply manipulating scenes and pasting objects without interfering with context meaning. In the absence of contextual effects, results from our current study could argue for the "functionnal isolation" model (Henderson & Hollingworth 1999). Alternatively, they could support their "priming" model if a delay is needed for scene context to influence object processing. Finally, immediate interactions between context and object processing flows would be compatible with their "perceptual schema" model. Such early effects of context could also support interactions between parallel streams of visual information in a feed-forward wave of processing (Rousselet, Fabre-Thorpe & Thorpe, 2002, VanRullen & Thorpe, 2002; Mace et al., 2005).

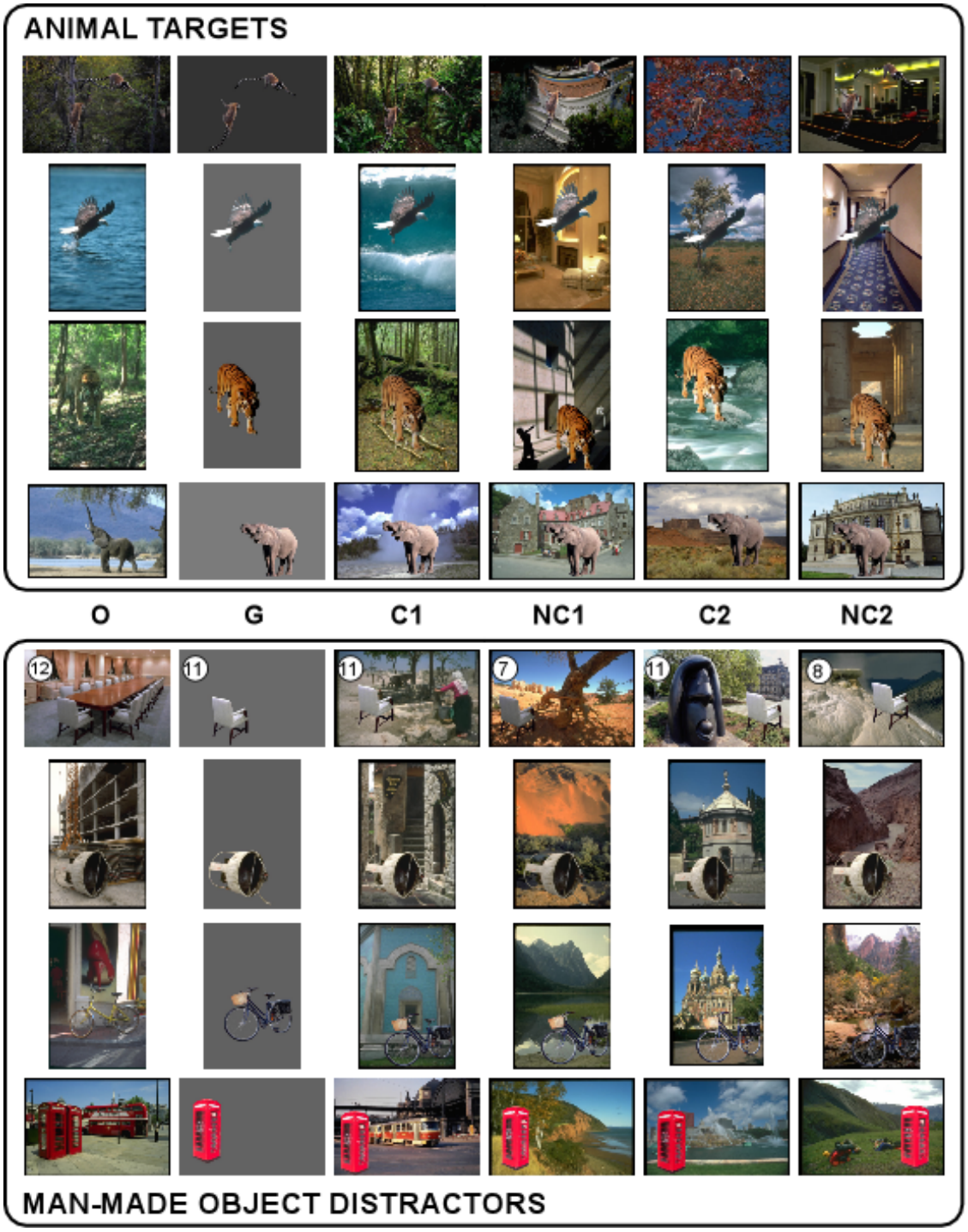

Figure 1: Examples of animal targets and man-made object distractors used in the animal categorization task. Different subsets were used in the 3 experiments: original scenes (O), isolated objects on grey background (G), objects pasted on two different congruent contexts (C1 & C2), and objects pasted on two different non-congruent contexts (NC1 & NC2). Context congruence was considered in terms of natural vs. manmade and scale, position and support relations were respected as much as possible. Examples using objects cropped from original scenes are shown in the two top rows for target and distractors. Examples using similar objects taken from the Hemera library are shown in the two bottom rows for targets and distractors. The first row for distractors illustrates the kind of man-made objects that elicited false alarms when out of its context, especially when pasted on non-congruent (natural) contexts. The number (top left of each image) indicates the number of subjects (out of 12) who correctly withheld their go response when presented with that scene.

Experiment 1: effect of object segregation and stimulus manipulation

The first experiment was designed to evaluate the impact of pasting an object on a neutral or consistent background on fast object categorization performance.

Method

Subjects and Task

Twelve volunteers (8 men, mean age 27, range 23-30, 3 of them left handed) gave their informed written consent. All of them had normal or corrected to normal vision. Subjects performed a rapid visual go/no-go categorization task. They were asked to lift their finger as quickly and as accurately as possible (go responses) each time the picture included an animal, and to withhold their responses (no-go responses) when there was no animal (Thorpe et al., 1996). Each subject performed 12 test series of 96 trials preceded by a training series of 48 trials. In each series, target and distractor trials were equally likely.

Stimuli

All horizontal and vertical scenes (768 x 512 pixels, 8° x 5° of visual angle) included a foreground object that was a man-made object for distractor scenes and an animal for target ones. In order to disentangle context influence and object pasting (whether performance could change simply because of stimulus manipulation), each object was seen by every subject in 3 conditions: (1) *object in the original non altered scene*, (2) *isolated object presented on a grey background*, and (3) *isolated object pasted in another congruent context*. Because contextual processing takes longer at more detailed categorization levels (Joubert et al., 2007, Rousselet et al., 2005), the congruence of an object with its context was defined in terms of "Man-made" vs. "Natural" terms, whereby a sofa is usually found in a man-made environment and a leopard more likely to be found in a natural environment. The order in which these 3 conditions were presented was counterbalanced across subjects. All conditions were randomly interleaved and in equal proportions in each series.

The subset of 384 *original pictures* (see examples in Figure 1.O) were all selected from a large commercial CD-ROM library (Corel Stock Photo Libraries) and included 192 "Man-made" scenes in which there was one or more man-made objects and 192 "Natural" scenes in which there was one or more animals (natural object). Images were in jpeg 24-bit format (16 millions colors). Within each category, images were as diverse as possible.

For half of the stimuli (192), man-made and natural foreground objects were cropped manually from images. To avoid excessively sharp edges we applied Paintshop (version 7.0.0.2, Jasc Software Inc.) progressive transparency on the contour (2 pixels wide). This procedure being very time consuming, the other half of the stimuli were built with objects from the Hemera Photo Objects library. The objects were chosen to be as similar as possible to the object present in the original images and found under the same label in the Hemera library (see Figure 1). Again with these objects, progressive transparency was also used on the contours in order to avoid sharp edges and to allow a good integration of objects in their new background.

The subset of stimuli, *objects on grey background*, were built by pasting the isolated object in the same location as in the original image on a grey background for which luminance was adjusted to the global luminance of the original scene (see examples in Figure 1.G). To build the set of *objects pasted on a new congruent context*, we tried to control as many features as possible (see examples in Figure 1.NC1). For each original scene, we chose one picture of the same context category (man-made or natural) with a roughly similar background in terms of orientation, global luminance and spatial layout. Within this new picture, the location of the object was selected to be as close as possible to the original scene, taking into account orientation and coherence (support, interposition, scale, Biederman et al., 1982). The local luminance was evaluated at this position. Using the YCbCr color system (a way of encoding RGB information in which Y is the luminance component and Cb and Cr are the blue and red chrominance components) that allows the color channel values to be preserved, we adjusted the object luminance relative to the background local luminance in order to keep the same local contrast between object and background. Finally, we pasted the adjusted object onto the selected background.

Procedure

Subjects sat in a dimly lit room, at 1 m from a computer screen (resolution 1024 x 768, vertical refresh: 75 Hz) piloted by a PC computer. Image presentation and behavioral response measurements were carried out using Presentation software (NeuroBehavioral Systems, http://nbs.neuro-bs.com/).

Each trial started with a fixation cross (1° of visual angle) that appeared at the centre of a black screen for 300-900 ms randomly. As soon as the cross disappeared, the stimulus was displayed for two frames (26 ms), also in the middle of the screen. These brief presentations prevented exploratory eye movements. To start stimulus presentation, subjects had to place their fingers on a response pad equipped with infrared diodes that allowed micro-second precision. After the image presentation, a black screen was displayed for 1000 ms, during which period subjects were required to respond by a finger lift if the image was a target. Longer reaction times were considered as no-go responses. Following this 1 sec period, a black screen was displayed during 300 ms before the next trial started. A trial lasted between 1600 and 2200 ms.

Statistics

We used the non-parametric two-way repeated measure Friedman test to evaluate statistical differences across subjects among the three conditions (original, isolated and pasted). When the Friedman test showed a statistical difference, a paired Wilcoxon test, Bonferonni corrected, was used to perform pairwise comparisons (in the text, p-values are corrected).

In order to provide additional information on the robustness of the effect observed on individual data, we computed confidence intervals using Monte-Carlo simulations to test for significant differences between conditions for each subject (Figure 2.B) using the following procedure. For each pairwise comparison, responses on each trial (go/no-go or reaction time) from the two conditions, containing n and m number of images, were pooled together and randomly shuffled. Then n image responses were assigned to a 'fake' subset 1 and the m others to a 'fake' subset 2. Averaged performance was thus computed for the two fake subsets and the difference between them was stored. This procedure was run 2000 times, providing a confidence interval around the null hypothesis that the two conditions were actually sampled from the same population. These confidence intervals are plotted in figures that report individual subjects' differences.

Results

All individual results for the 3 experimental conditions are summarized in Table 1 and in Figure 2.

	Accuracy (%)									Reaction time (ms)		
	Global			Targets			Distractors			Median		
Subject	Original	Isolated	Pasted	Original	Isolated	Pasted	Original	Isolated	Pasted	Original	Isolated	Pasted
MRO	98.2	97.7	96.1	98.4	99.0	94.3	97.9	96.4	97.9	399	396	407
JSN	92.2	92.2	87.8	92.7	97.4	85.9	91.7	87	89.6	324	315	335
RVR	94.3	96.1	85.9	98.4	99.0	91.1	90.1	93.2	80.7	409	407	410
JMO	96.6	95.1	93.8	98.4	99.5	94.8	94.8	90.6	92.7	324	313	334
NGU	94.3	97.4	95.8	98.4	100.0	99.0	90.1	94.8	92.7	371	367	384
IBA	98.4	97.9	96.4	99.5	97.9	94.8	97.4	97.9	97.9	410	401	413
JMA	97.4	99.0	95.1	97.4	99.0	93.2	97.4	99	96.9	377	374	400
LBA	97.4	97.7	94.8	97.4	99.0	92.2	97.4	96.3	97.4	434	430	431
SVI	95.6	95.6	96.1	98.4	99.5	94.8	92.7	91.7	97.4	351	340	371
NBA	97.1	93.8	95.6	99.5	99.5	96.9	94.8	88	94.3	348	342	366
JFO	95.3	98.2	96.4	94.8	99.0	94.8	95.8	97.4	97.9	444	452	457
MMA	97.7	92.7	95.1	99.5	99.0	97.4	95.8	86.5	92.7	328	319	339
Mean	96.2	96.1	94.1	97.7	99.0	94.1	94.7	93.2	94.0	376	371	387
Std.	1.9	2.3	3.5	2.0	0.7	3.4	2.9	4.4	5.0	43	47	39
Min	92.2	92.2	85.9	92.7	97.4	85.9	90.1	86.5	80.7	324	313	334
Max	98.4	99.0	96.4	99.5	100.0	99.0	97.9	99.0	97.9	444	452	457

Table 1: Individual accuracy and reaction time in the 3 conditions of experiment 1: original scenes, isolated objects on grey background and pasted objects on another congruent context (see examples in Figure 1). The four bottom lines indicate mean, standard deviation, minimal and maximal scores computed from individual scores.

Accuracy

Subjects were very efficient at performing the animal categorization task in all three conditions. Global accuracy (correct go and no-go responses) reached 96.2% with original images, 96.1% with isolated objects on a grey background and 94.1% with pasted objects on new natural contexts. Accuracy differences between the 3 conditions were not statistically significant (Friedman test: n.s. $\chi^2_r = 4.696$, df = 2, p = .096, Figure 2.A.). However, the individual performance analysis revealed a significantly lower accuracy in the pasted condition when compared to the original condition in 9 out of 12 subjects (Figure 2.B). Considering separately accuracy on target and distractor trials provided more information (Table 1). The global accuracy decrease was clearly due to target trials (97.7% vs. 94.1%; Friedman test: $\chi^2_r = 17.522$, df = 2, p<.00001; paired Wilcoxon test on go-responses: Z = 2.847, p = .012), whereas accuracy was not statistically different on distractor trials (94.7%

vs. 94%; n.s. Friedman test: $\chi^2_r = 1.644$, df = 2, p = .439). No differences were observed between original scenes and isolated objects.

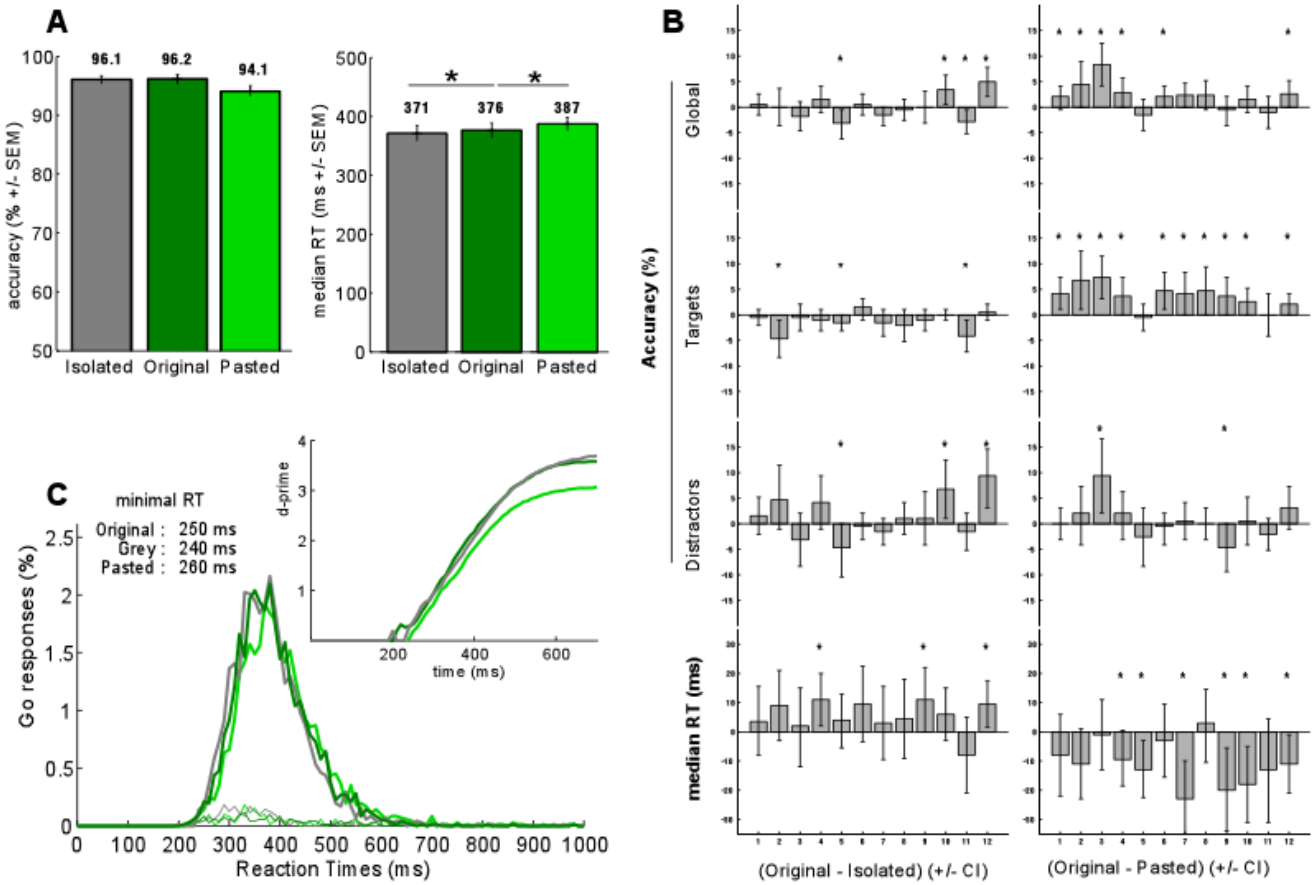

Figure 2: Performance obtained in the 3 conditions of experiment 1: isolated objects, original scenes, and objects pasted on congruent contexts. (**A**) Global accuracy expressed as the percentage of correct responses and median RT for correct-go responses (in ms) are shown with associated standard errors of the mean. Asterisks indicate statistically significant differences between the original and the isolated or pasted conditions. (**B**) Effects on individual accuracy (in percentage of correct responses) and performance speed (Median RT in ms). For each subject, the score obtained in each condition, isolated object (left column) and pasted object (right column), is subtracted from the score obtained in the original condition. Accuracy performance is shown globally and separately on targets and on distractors. Asterisks indicate significant differences (permutation test, $p<.05$, 1000 resamples) between conditions. Note that the accuracy drop in the pasted condition is due for most subjects to a drop in target detection. (**C**) RT distributions for correct go-responses (thick curves) and for false alarms (thin curves) are shown for the original (deep green), isolated (grey) and pasted on congruent context (light green) conditions, with the number of responses pooled across all subjects and expressed over time using 10 ms time bins. Minimal RT, determined as the first 10 ms time bin for which correct responses significantly exceed errors (targets and distractors were equally likely) was observed at 250, 240, and 260 ms for the original, grey and pasted conditions respectively. At the top right, d' curves using signal-detection theory sensitivity measures were plotted as a function of time with 10 ms time bins. Cumulative numbers of hits and false alarms responses were used to calculate $d'=z_{hits} - z_{FA}$ at each time point where z is the inverse of the normal distribution function (Macmillan & Creelman, 2005). d' curves corresponding to the time course of performance give an estimation of the processing dynamics for the entire subject population. d' curve in the "pasted" condition shows a shift towards longer latencies (10-20 ms) from the very beginning and reach a lower plateau.

An additional observation can be made concerning the subjects' response bias. When categorizing original scenes and isolated objects, subjects tended to respond "animal". Incorrect trials were biased towards false alarms (original: 5.3%, isolated: 6.8%), whereas subjects missed very few targets (respectively 2.3% and 1%). This bias towards targets was significant (paired Wilcoxon test, $Z > 2.552$ and $p < .033$ in both conditions). No such bias was observed for the pasted condition (false alarms 6%, missed targets 5.9%; paired Wilcoxon test, $Z = .275$, $p = 1$).

Reaction times

Subjects were also very fast, with median reaction times (RT) around 380 ms. Median RT were computed on correct target trials. Compared to the original condition, the mean median RT was slightly shorter with isolated targets (376 ms vs. 371 ms; paired Wilcoxon test: $Z = 2.512$, $p = .036$), and longer with pasted targets (376 ms vs. 387 ms; paired Wilcoxon test: $Z = 2.867$, $p=.012$). Although RT differences were small, they were also very robust at the individual level (see Table 1 and Figure 2.B). These differences are also illustrated by the RT distributions (Figure 2.C).

To evaluate how accuracy varies with response latency, d' curves were computed for the 3 conditions (Figure 2.C). The d' curves showed that the information processing rate was very similar for the original and isolated conditions. By contrast, the d' curve for the "pasted" condition is shifted towards longer latencies from the very beginning and reaches a lower plateau corresponding to an accuracy drop. Such a clear shift (10-20 ms) from the shortest RT indicates that pasting an object on a new natural context is not a trivial manipulation and is able to slow down object categorization even for the earliest responses.

As described in the methods, half of the manipulated "objects" were totally identical to the object in the original scene while the other half were chosen from the Hemera library to be very similar. We thus checked whether these two image subsets lead to similar categorization performances in all 3 conditions. Behavioral performances on the two image subsets were very similar in each condition. In the original condition, the source images of Corel and Hemera subsets were similarly difficult to categorize (accuracy: 98% vs. 97.4%; median RT: 374 ms vs. 379 ms). In the pasted condition, similar drops of performance were observed for both subsets relatively to the original condition (93.1% vs. 95.1%, 10 ms RT increase in both cases). Finally, global accuracy and median RT for the two grey subsets were also very similar (Corel animals: 99.1% and 371 ms; Hemera animals: 98.8% and 370 ms). Friedman tests computed on individual accuracy and median RT using condi-

tion (3 levels) and image subset (2 levels) as factors confirmed the absence of subset effect after adjusting for condition effects (accuracy : $\chi^2_r = 1.76$, $p = .1851$; median RT: : $\chi^2_r = 0$, $p = 1$). Thus, we can conclude that the subsets of stimuli using identical and Hemera animals were associated with virtually identical performance and were similarly affected by the experimental manipulations.

To sum up, the first experiment showed that segregating an object from its background has –surprisingly- very little effect if any (at least in an interleaved protocol) on rapid categorization performance: compared with original photographs, accuracy is not improved by isolating target objects and median RT is only shortened by 5 ms. This result is strengthened by the finding that pasting an object on another congruent context has a cost both in terms of response speed (10 ms increase in median RT) and in terms of global accuracy (>2%). This accuracy cost is larger when only considering targets trials (>3.5%). Experiment 1 showed that a performance drop can be due to simple stimulus manipulations.

Such manipulations may have affected the saliency of foreground objects. To determine whether objects were as physically salient in original vs. pasted context, we used the saliency toolbox (Walther & Koch, 2006), inspired by the computational model of visual attention from Itti and Koch (Itti & Koch, 2001). The most salient zone was found on (or close to) the animal in 74% of the original images but in only 60% of the pasted stimuli (41% vs. 33% for man-made objects). To prevent interferences from such manipulation with the evaluation of context congruence, all stimuli in the following experiments used pasted objects in contextual scenes.

Experiment 2: Effect of context congruence on object processing

Experiment 2 used only objects pasted on congruent or non-congruent contexts to evaluate the effect of contextual incongruence in rapid visual categorization.

Method

Subjects

Twelve volunteers (8 men, mean age 28, range 22-33, 3 of them left handed) gave their informed written consent. All of them had normal or corrected to normal vision. Three of them participated in the first experiment (see Table 2).

Stimuli

In the experiment 2, 768 stimuli were used: (1) 384 foreground *objects with congruent context* (192 man-made objects and 192 animals), which were those used in the first experiment, and (2) the same 384 foreground *objects pasted on a non-congruent context.* In this experiment, all stimuli contained a pasted object. Pictures of *objects with congruent or non-congruent contexts* were built as reported in experiment 1. Context was defined as congruent if an animal was pasted on a natural context, and as non-congruent if it was pasted on a man-made context, and conversely for a man-made object. Contexts were defined as man-made and natural following Joubert et al. (Joubert et al., 2007): "natural environment" contexts included sea, mountain, desert, iceberg, forest and field scenes; "man-made environment" contexts included street scenes (with or without pedestrians) and indoor scenes like kitchens, museums, and churches. None of the man-made scenes contained mountains or views of the sea, and none of the natural scenes contained buildings. Scenes came from the same photograph gallery as the one used in Joubert et al. (2007) in which the speed of context categorization per se was evaluated. For each of the 4 subsets of stimuli, man-made objects/animals pasted in congruent/incongruent contexts, the mean eccentricities defined as the distance between the fixation point and the center of the object (man-made or animal) were below 48 pixels corresponding to 0.5° of visual angle. Object eccentricity was thus similar for all subsets.

Task and Procedure

Task and Procedure were identical to those in experiment 1. Subjects performed the animal categorization task for 8 series of 96 trials; all series contained, randomly interleaved, 25% animal targets in congruent contexts, 25% animal targets in non-congruent contexts, 25% man-made objects in congruent contexts and 25% man-made objects in non-congruent con-

texts. Each subject saw all target and non-target objects randomly displayed over the experimental series once in a *congruent context*, another time in a *non-congruent context*, but never in the same series.

Results

All global and individual results are summarized in Table 2 and Figure 3.

	Accuracy (%)						Reaction time (ms)	
	Global		Targets		Distractors		Median	
Subject	C	NC	C	NC	C	NC	C	NC
NGU	95.6	85.9	95.8	87.5	95.3	84.4	406	437
SCR	91.9	84.6	92.2	80.7	91.7	88.5	430	436
MLA	94.8	87.8	94.8	85.9	94.8	89.6	427	434
TMA	95.1	88.5	92.7	82.3	97.4	94.8	407	413
RVR	88.5	78.1	86.5	71.4	90.6	84.9	334	354
APA	93.5	83.3	92.7	80.2	94.3	86.5	437	446
NBO	95.6	89.1	96.9	87.5	94.3	90.6	412	428
SBE	88.5	79.2	83.3	70.3	93.7	88	533	547
JMA	94.8	86.7	96.9	88.5	92.7	84.9	322	343
CHE	93.2	86.5	89.1	78.1	97.4	94.8	378	387
LLA	93.2	87.8	92.2	86.5	94.3	89.1	393	415
FRE	90.6	84.6	85.4	76.6	95.8	92.7	387	407
Mean	92.9	85.2	91.5	81.3	94.4	89.1	406	421
std.	2.6	3.5	4.5	6.3	2.0	3.6	54	52
Min	88.5	78.1	83.3	70.3	90.6	84.4	322	343
Max	95.6	89.1	96.9	88.5	97.4	94.8	533	547

Table 2: Individual accuracy and reaction time in the 2 conditions of experiment 2: pasted on congruent context (C) and pasted on non-congruent context (NC, see examples C1 and NC1 in Figure 1). The four bottom lines indicate mean, standard deviation, minimal and maximal scores computed from individual scores. Subjects NGU, RVR and JMA have also performed experiment 1.

Accuracy

There was a clear effect of context congruence on global accuracy. Subjects scored 92.9% correct when animals and distractor objects were presented in a congruent context (respectively natural and man-made), but they only scored 85.2% when both target objects and distractor-objects were presented in a non-congruent context. A paired Wilcoxon test showed that

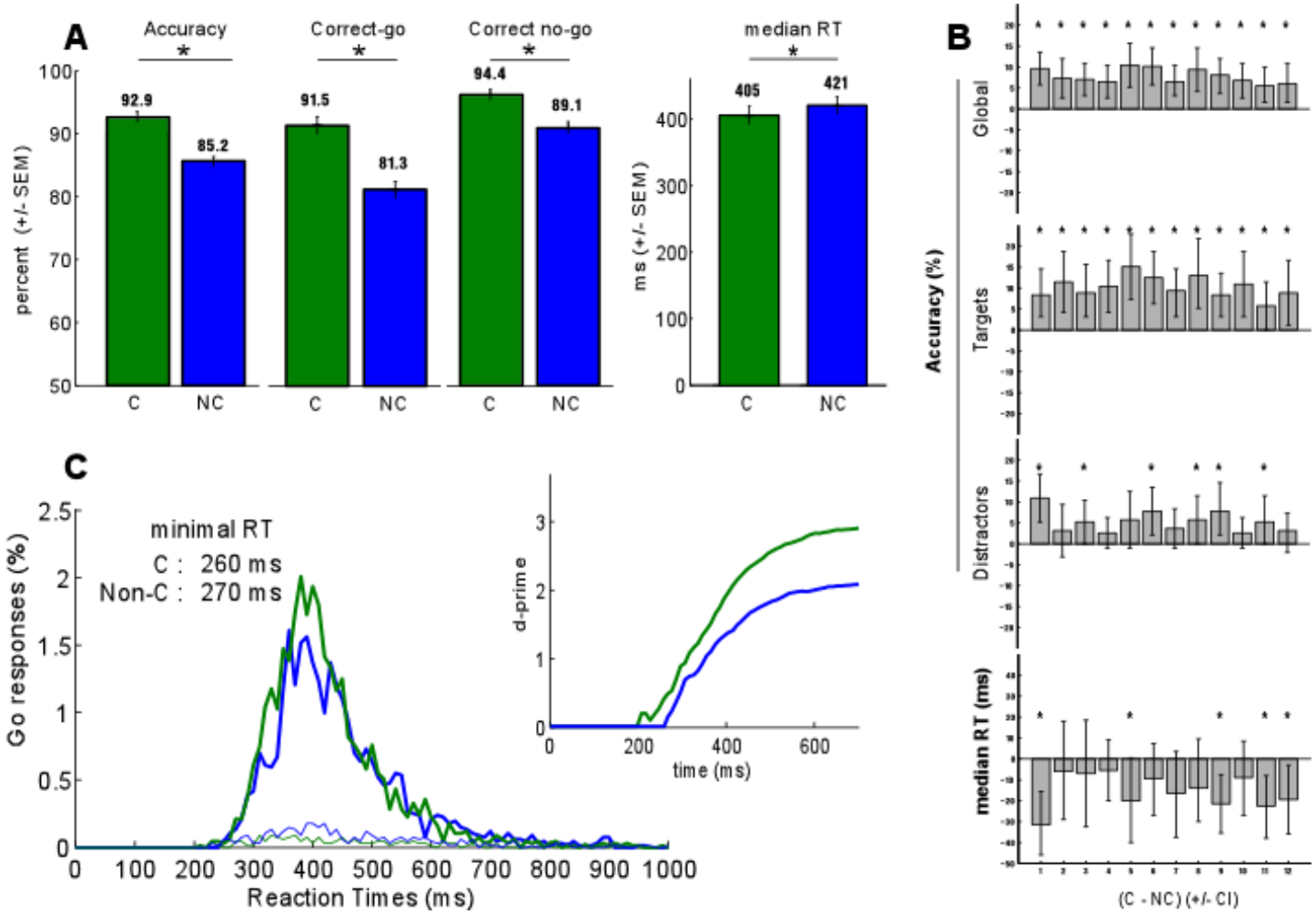

Figure 3: Performance in the 2 conditions of experiment 2: objects pasted in congruent contexts and in non-congruent contexts. **(A)** Accuracy (global accuracy and accuracy on targets and distractors) and median RT for correct-go responses (in ms) are shown with associated standard errors of the mean. Asterisks indicate statistically significant differences. Categorizing objects in congruent contexts was more accurate and faster than objects in non-congruent contexts independently of the image status (target/distractor). **(B)** Congruency effects on individual accuracy and median RT. For each subject, the score obtained in the non congruent condition was subtracted from the score obtained in the congruent condition. Accuracy is shown globally and separately on targets and on distractors. Asterisks indicate significant differences (permutation test, $p<.05$, 1000 resamples) between conditions. The congruent context advantage was present for all subjects. **(C)** RT distributions for correct go-responses (thick curves) and for false alarms (thin curves) are shown for the congruent (green) and non-congruent (blue) conditions with the number of responses pooled across all subjects and expressed over time using 10 ms time bins. Minimal RTs (see caption Figure 2) were observed at 260 and 270 respectively. At the top right, the d' curves computed for each condition show that, in the non congruent condition, the d' curve is shifted (20-30 ms) towards longer latencies and reaches a lower plateau.

this 7.7% (95% percentile bootstrap confidence interval: 6.9-8.7%) decrease in accuracy was statistically significant ($Z = 3.059$, $p = .002$, Figure 3.A) and Monte-Carlo simulations revealed that this decrease was significant for all subjects (Figure 3.B). It is interesting to notice that when analysing separately correct and incorrect go responses, subjects tended to produce more go responses with natural scene contexts than with man-made ones. This was true for correct go responses towards animals (91.5% and 81.3% respectively, paired Wilcoxon test: $Z = 3.062$, $p = .002$, Figure 3.A) and individually significant

for 11 subjects among 12 (Figure 3.B). This was also true for the incorrect go-responses (false alarms) produced towards man-made objects that were embedded in natural (non-congruent) contexts (false alarms: 10.9%). False alarms were considerably reduced when man-made objects were in man-made contexts (5.6%). This difference was statistically significant across subjects (paired Wilcoxon test p = .002, Z = 3.065; Figure 3.A) and very consistent considering individual results (Figure 3.B). The number of false alarms produced towards man-made objects presented in the context of man-made scenes were very similar in experiments 1 and 2 (6% vs. 5.6%), but a decrease in correct go-responses towards targets presented in a natural context was observed (94.1% in exp1 and 91.5% in exp2). This decrease might be due to the group of subjects tested but more likely (as accuracy is usually a robust measure) to the fact that the congruent pictures used in experiment 1 and 2 were intermixed with conflictual stimuli.

Reaction times

Reaction times were also affected by congruence. Subjects categorized animals with an averaged median RT of 406 ms in the congruent context condition and 421 ms in the non-congruent context condition (Figure 3.A). This 15 ms (95% confidence interval: 11–20 ms) RT increase was statistically significant across subjects (paired Wilcoxon test: Z = 3.059, p = .002) and individually for 5 subjects (Figure 3.B).

Although RT distributions for correct go-responses (Figure 3.C) were nearly superimposed in the initial portion of the distribution, the two global distributions clearly differed from 300 ms onwards as attested by χ^2 tests performed between the correct-go distributions in congruent and non-congruent conditions (χ^2 test: p=.0246). Moreover, a higher proportion of correct-go vs. incorrect go responses was observed in the congruent context condition. Consistent with these observations, the d' curves computed for each condition clearly differed. The d' curve for the non-congruent context condition was shifted towards longer latencies and reached a lower plateau. This shift of about 20-30 ms at minimal RT value (Figure 3.C) shows that incongruence between object and context affects the information accumulation rate even for the earliest responses.

To summarize, experiment 2 revealed very early interactions between object and context processing. In a rapid visual go/no-go task in which stimuli are just flashed for 26 ms, context congruence can influence performance. Such performance

is biased towards go responses with a natural context, and towards no-go responses with a man-made context. Non-congruent contextual information delays object processing even for the earliest responses.

Post-hoc analysis on object saliency

Because natural contexts are usually simpler and more uniform than richer man-made contexts, the congruence effect observed in experiment 2 could be partially explained by a higher saliency of animal and man-made objects in natural contexts. Such higher saliency would not necessary lead to better processing as shown by the worse rejection of man-made objects as distractors in non-congruent natural context (Figure 3.A). We used the same saliency toolbox as in experiment 1 to measure object saliency in congruent and non-congruent contexts, in order to evaluate the congruency effect as a function of object saliency. For each stimulus used in experiment 2, we computed a saliency map which allowed us to define the most physically salient area of the scene based on different properties like luminance, orientations and colors. According to this attentional model, this most salient area should be the most probable land-location for the first eye movement if photographs were flashed longer. However, the brief image presentation used in the present study prevented any eye movement exploration. Indeed a bias was found such that animal and man-made objects were both less salient in artificial contexts. The most salient zone was found outside the animal (or man-made) object in 67% (69%) in artificial contexts, but in only 42% (58%) of the case in natural context. We thus analyzed separately the effect of contextual congruence on 3 different subsets of stimuli: most salient area (1) on animals or man-made objects, (2) sat astride animal/object boundary or (3) outside the foreground object. Results are illustrated in Figure 4. This analysis demonstrated that the effect of contextual congruence observed in experiment 2 was independent of object saliency. Regardless of object saliency (animal or man-made), an accuracy drop and a RT increase were observed for objects pasted in non-congruent contexts (vs. congruent). The robustness of the congruence effect after adjusting for saliency effects was confirmed by non-parametric repeated measure Friedman tests with 2 factors (congruence and saliency) on global accuracy ($\chi^2r = 45.19$, $p < .0001$), correct-go ($\chi^2r = 33.78$, $p < .0001$) , correct no-go ($\chi^2r = 13.51$, $p = .0002$) and median RT ($\chi^2r = 4.55$, $p = .0329$). Such results confirm that the performance impairments observed in this experiment must be attributed to a congruence effect between object and context.

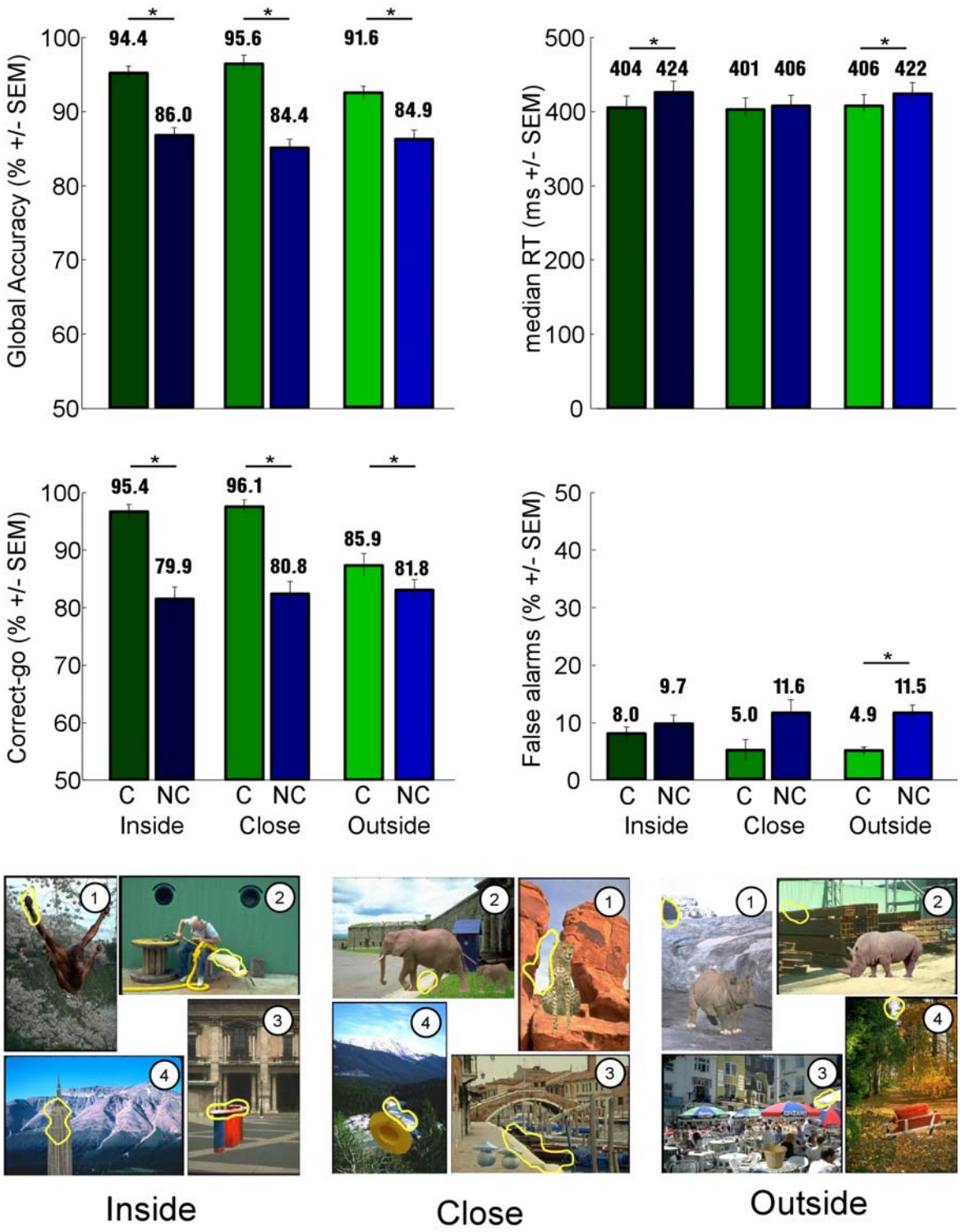

Figure 4: Performance obtained in the congruent vs. non-congruent conditions (C vs. NC) in 3 object saliency conditions: “Inside” when the most salient area of the stimuli was within the pasted object boundaries, “Close” when the most salient area overlapped the pasted object, “Outside” when the most salient area was outside pasted object boundaries. Accuracy and median RT histograms are shown with associated standard errors on the mean. Asterisks indicate statistically significant differences revealed by paired Wilcoxon tests. Results show an advantage for the congruent condition in all conditions of object saliency. For each scene used as stimulus, the most salient area was determined using the Walther & Koch (2006) saliency toolbox (see text). Examples are shown at the bottom for the 3 saliency conditions (inside, close, outside) with the most salient zone outlined in yellow: Animal in congruent (1) and non congruent (2) contexts, man-made object in congruent (3) and non-congruent (4) contexts.

Experiment 3: Robustness of the congruence effect

A third experiment was designed to assess the robustness of the results obtained in experiment 2 by associating the same target and non-target objects used previously with different congruent and non-congruent backgrounds.

Method

Subjects

Twelve volunteers (6 men, mean age 24, range 21-27, 2 of them left handed) gave their informed written consent. All of them had normal or corrected to normal vision. None of them had participated in the first two experiments.

Stimuli

Experiment 3 employed 384 stimuli identical to those in experiment 2, and 384 stimuli in which paired objects switched context (congruent and non-congruent, Figure 1). This second set was built as follows.

Half of the stimuli from experiment 2 were first randomly selected. These 384 stimuli were divided in 4 sets of 96 stimuli (targets in congruent contexts, targets in non-congruent contexts, non-targets in congruent contexts, non-targets in non-congruent contexts). Within each set, stimuli were organized in 48 pairs; care was taken to pair stimuli so that scales and relative positions of the two paired objects were as coherent as possible in both contexts. Then, within each pair, objects were context switched.

Thus, 8 image subsets were considered for analysis: 96 animal objects pasted on two different but congruent natural contexts (C1 and C2), the same 96 animals pasted on two different but non-congruent man-made contexts (NC1 and NC2); 96 man-made objects pasted on two different but congruent man-made contexts (C1 and C2), and the same 96 man-made objects pasted in two different but non-congruent natural contexts (NC1 and NC2). C1 and NC1 were identical stimulus sets to those from experiment 2 whereas C2 and NC2 were new context switched stimuli.

Task and Procedure

Task and Procedure were identical to those in experiment 1. Subjects performed the animal categorization task for 8 series of 96 trials. All series contained 50% animal targets and 50% man-made objects randomly interleaved, with an equal proportion of each context condition. Each subject saw each object pasted on the 4 different backgrounds, but a given object was never seen twice in the same series.

Results

The aim of this third experiment was to check whether the performance impairment induced by contextual incongruence observed in experiment 2 was robust, independently of the specific congruent (C) and non-congruent (NC) contexts used to paste a given object. Here we used non-parametric repeated measure Friedman test with 2 factors: the congruence and the context subset. All global and individual results are summarized in Table 3.

	Accuracy (%)												Reaction time (ms)			
	Global				Targets				Distractors				Median			
Subject	C1	C2	NC1	NC2	C1	C2	NC1	NC2	C1	C2	NC1	NC2	C1	C2	NC1	NC2
OJO	97.4	95.8	92.7	93.2	100	96.9	91.7	89.6	94.8	94.8	93.8	96.9	386	388	398	405
MRU	96.9	95.3	91.1	90.6	95.8	93.8	88.5	86.5	97.9	96.9	93.8	94.8	462	468	486	481
CHA	85.9	87.0	82.3	81.8	92.7	91.7	84.4	84.4	79.2	82.3	80.2	79.2	388	402	397	407
FLA	87.5	87	87.5	87	84.4	82.3	83.3	80.2	90.6	91.7	91.7	93.8	394	416	426	417
CHO	91.7	89.6	87.5	91.2	94.8	93.8	88.5	89.6	88.5	85.4	86.5	92.7	363	362	372	348
MDE	95.3	93.2	92.2	85.4	92.7	88.5	86.5	79.2	97.9	97.9	97.9	91.7	479	457	466	458
EBA	92.2	94.8	89.6	89.1	89.6	92.7	87.5	84.4	94.8	96.9	91.7	93.8	412	418	429	429
CBR	91.1	90.6	82.8	80.7	97.9	99.0	94.8	91.7	84.4	82.3	70.8	69.8	295	297	314	313
EBO	91.7	93.8	85.4	84.4	89.6	91.7	80.2	75	93.8	95.8	90.6	93.8	367	392	389	384
LMO	92.2	91.1	85.9	84.9	86.5	84.4	72.9	70.8	97.9	97.9	99.0	99.0	400	401	416	421
ALA	94.3	94.8	90.6	90.1	91.7	93.8	85.4	85.4	96.9	95.8	95.8	94.8	412	418	424	443
JBL	96.4	91.7	92.2	91.1	96.9	94.8	86.5	87.5	95.8	88.5	97.9	94.8	398	408	426	428
Mean	92.7	92.1	88.3	87.5	92.7	91.9	85.9	83.7	92.7	92.2	90.8	91.2	396	402	412	411
std	3.6	3.1	3.6	4.0	4.7	4.8	5.6	6.3	6.0	6.0	8.2	8.3	47	44	44	46
Min	85.9	87.0	82.3	80.7	84.4	82.3	72.9	70.8	79.2	82.3	70.8	69.8	295	297	314	313
Max	97.4	95.8	92.7	93.2	100.0	99.0	94.8	91.7	97.9	97.9	99.0	99.0	479	468	486	481

Table 3: Individual accuracy and reaction time with the 2 conditions and 4 images subsets in experiment 3: objects pasted on two different congruent contexts (C1 & C2) and pasted on two different non-congruent contexts (NC1 & NC2). (see examples C1, C2, NC1 and NC2 in Figure 1). The four bottom lines indicate mean, standard deviation, minimal and maximal scores computed from individual scores.

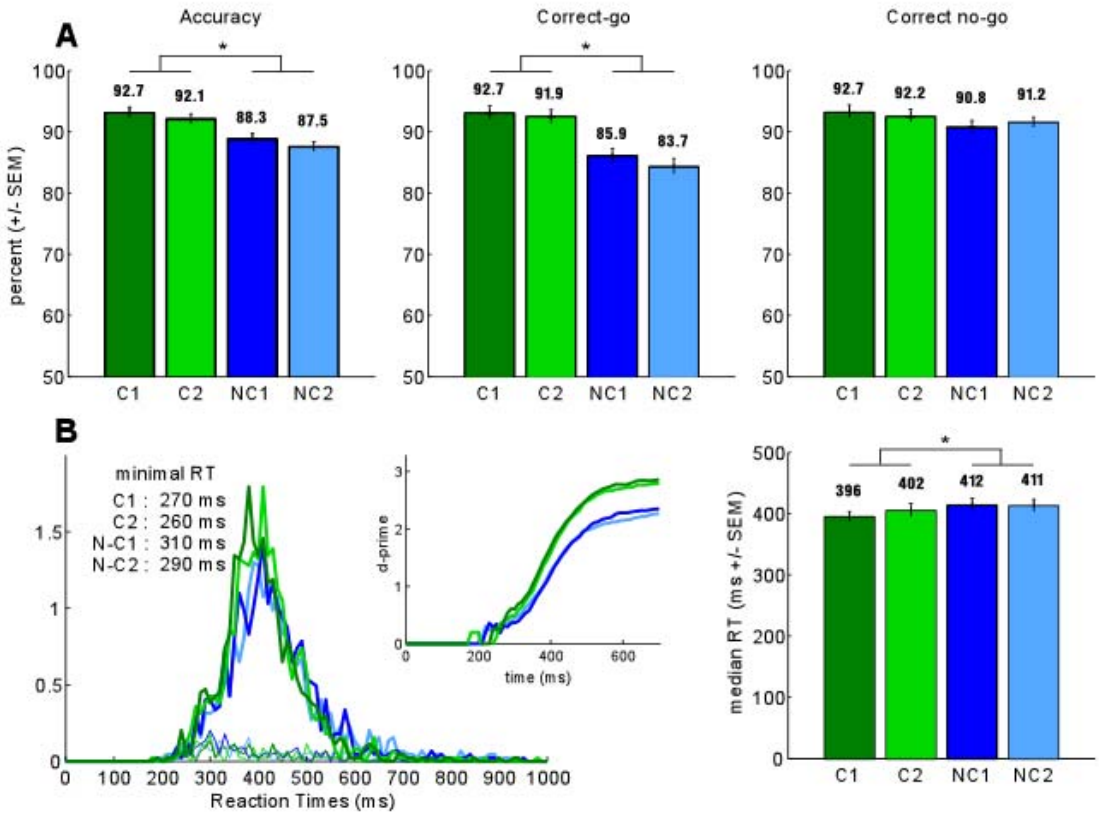

Figure 5: Performance obtained in the 2 conditions and 4 image subsets tested in experiment 3: objects pasted in two different congruent contexts (C1 & C2) and objects pasted in two different non-congruent contexts (NC1 & NC2). **(A)** Accuracy and median RT are shown with associated standard errors of the mean. Asterisks indicate statistically significant differences. The global accuracy drop with images of non-congruent context subsets is mainly due to a drop in target detection. **(B)** RT distributions for correct go-responses (thick curves) and for false alarms (thin curves) are shown for the two congruent image subsets (C1 in deep green & C2 in light green) and the two non-congruent image subsets (NC1 in deep blue & NC2 in light blue), with the number of responses pooled for all subjects and expressed over time using 10 ms time bins. Minimal RT were observed at 270 and 260 ms in the two congruent conditions; at 310 and 290 ms in the two non congruent conditions respectively (see caption Figure 2). At the top right, blue d' curves show a globally later processing dynamic (20-30 ms) when subjects have to categorize objects pasted in non-congruent context (vs. congruent): the interference effect from non-congruent context is robust and immediate, at least for target images.

Accuracy

With objects pasted on congruent contexts, subjects performed the task with high global accuracy for subsets C1 (92.7%) and C2 (92.1%). They showed similar drops of accuracy when objects were shown on non-congruent context, scoring 88.3% and 87.5% correct respectively for subsets NC1 and NC2 (Figure 5.A). Friedman tests revealed a significant effect of congruence after adjusting for possible context subset effects ($\chi^2_r = 12.14$, $p = .0005$) while there was no effect of context subset after adjusting for possible congruence effects (n.s. $\chi^2_r = .35$, $p = .5523$).

It is interesting to notice that, regardless of the subset of stimuli (C1 vs. NC1 or C2 vs. NC2) subjects reached similar accuracies. Moreover, their impairment was similar to the drop of performance displayed by the different group of subjects tested in experiment 2 (92.9% and 85.2% correct respectively on congruent and non-congruent contexts).

Performance was evaluated separately for target and non-target trials, for set 1 and set 2 (Figure 5). We observed no difference ($\chi^2_r = .15$, $p = .6972$), on correct-go responses towards targets between C1 and C2 (92.7% and 91.9% respectively), nor between NC1 and NC2 (85.9% and 83.7). As in experiment 2, subjects tended to respond more to targets when pasted on any congruent context and to withhold their response on any non-congruent context ($\chi^2_r = 21.04$, $p < .0001$). This effect was present for all subjects and replicated the results of experiment 2.

On distractor trials featuring man-made objects, the context influence was not as clear as in experiment 2. When objects were presented in congruent contexts, the percentage of incorrect go-responses reached 7.3% and 7.8% (respectively for C1 and C2); it increased up to 9.2% and 8.8% with non-congruent contexts (respectively NC1 and NC2). This slight (around 1.5%) impairment did not reach significance (Friedman test: n.s. $\chi^2_r = .61$, $p = .4351$) and independent permutation tests confirmed this result (1000 samples, between C1 and NC1: p=.09, C2 and NC2: p=.5523). This result was at odds with those obtained on distractor trials in experiment 2. As experiment 3 used only half of the stimuli of experiment 2, we reanalyzed results from experiment 2 on this restricted subset of stimuli. No difference in accuracy or RT was found on this image subset in experiments 2 and 3.

Reaction Times

Animals in congruent contexts were categorized with median RT of 396 ms and 402 ms, respectively for C1 and C2 images. When animal-targets were embedded in man-made contexts, median RT increased to 412 ms and 411 ms (for NC1 and NC2, respectively). Two way Friedman tests revealed no congruence effect after adjusting for possible context subset bias (n.s. $\chi^2_r = 2.28$, $p = .1308$), and no context subset effect after adjusting for possible congruence bias (n.s. $\chi^2_r = .12$, $p = .7285$). However, more sensitive permutation tests contrasting conditions two by two did show a congruence effect (1000 samples, C1 vs. NC1: p=.0025; C2 vs. NC2: p=.0285), while it confirmed the lack of context subsets bias (1000 samples, C1

vs. C2: p=.1182; NC1 vs. NC2: p=.7784). A small RT increase of about 10-15 ms is thus systematically associated with the decreased accuracy observed when animal targets are presented on non-congruent contexts.

This observation is strengthened by the d' results showing that the information accumulation rate in congruent and non-congruent contextual conditions differed. Although the d' curves were superimposed for identical context conditions, the d' curves associated with either congruent or non-congruent contexts diverged very early (around 280 ms for context subsets 1 and 300 ms for context subsets 2), showing that context incongruence can have in most cases a deleterious effect on object processing from very early on.

General Discussion

The aim of the present study was to evaluate the temporal dynamics of contextual influences on fast object categorization. Experiments 2 and 3 clearly demonstrated a performance impairment in object categorization performance due to incongruent contextual information. This effect is so fast that it affects even on the earliest responses produced by the subjects; it is very robust as it was reproduced regardless of the group of subjects and regardless of the particular contexts in which the objects were pasted.

Experiment 1 provided two additional important results. First, manipulating natural stimuli is not a trivial operation and induces small performance alterations despite all the effort and care involved in making the stimuli. Second, it shows a surprising result, namely that animal categorization is not easier when animal targets are isolated on grey backgrounds rather than embedded in natural scenes. This latter result has to be discussed in relation to the debate about whether figure-ground segregation must precede object recognition.

No performance improvement for isolated targets

Data from experiment 1 replicated the high human performance in ultra-rapid categorization of animals in natural scenes (Fabre-Thorpe et al., 2001, Thorpe et al., 1996). When targets are embedded in their original context, animal categorization is performed with high accuracy and fast RTs (96.2% with 376 ms median RT). Such scores were expected to be improved by the use of isolated animals on grey backgrounds; however, performance showed little if any improvement. With

isolated objects, accuracy was similar (96.1% correct on isolated objects versus 96.2% correct using natural scenes) and the overall median RT showed only a mild decrease of 5ms (371 ms vs. 376 ms). This decrease reached statistical significance, and analysis of individual performances (Figure 2.B) also showed that this tendency was present in most subjects (11 out of 12) but was indeed very small.

Although the use of natural scenes in addressing object recognition processes has recently increased, most of the research has been done so far using isolated objects on uniform backgrounds. Indeed, theories of visual perception and object recognition have often supposed that segregation of objects, or object diagnostic parts, has to precede recognition (Biederman, 1987, Kosslyn, 1987, Marr, 1982). By using isolated objects, the first stage of vision processing, namely "segregation", would be already completed and object recognition could be studied more readily. This idea was reinforced by research in computer vision in which object recognition has generally been assumed to be impossible without segregation. However, the idea that segregation has to precede object recognition has been challenged by results showing that object recognition can influence the initial perception of figure-ground organization in briefly presented stimuli (Peterson, 1994, Peterson & Gibson, 1994a, Peterson & Gibson, 1994b).

Our data support the idea that segregation does not need to precede detection and categorization, although it might have to precede object identification. If segregation was required, targets should be more difficult to categorize in a complex background. An alternative explanation of our results might be to consider that we are faced with two types of facilitation/interference that cancel each other out: even if objects used in the present study show a large diversity of positions, scales and locations, performance with isolated objects on grey backgrounds would benefit from an easier segregation but would also lack contextual facilitation. Following this hypothesis, when objects are embedded in natural scenes, segregation would be more difficult but categorization would benefit from contextual facilitation. However, the visual mechanisms involved in both experimental situations might not be identical. We used a protocol in which all conditions (original, isolated objects, pasted objects) are mixed and equally likely, but objects embedded in scenes were twice as frequent as isolated objects. Subjects might thus favour a strategy that makes use of contextual guidance. Alternatively, subjects could tend to adjust their response speed across trials as a function of the time necessary to respond optimally to stimuli belonging to the hardest condition. According to this latest hypothesis, we thus might observe an advantage for the isolated condition when subjects perform the same task in a blocked design. Further experiments are currently being run using a block design to tackle the

effect of object isolation in a task requiring a more detailed level of categorization (dog/non-dog). Follow-up experiments will also determine whether segregation is needed to categorize objects in the periphery of the visual field.

The "pasting" effect

Experiment 1 also allowed the evaluation of the performance impairment due to stimulus manipulations. Despite the extreme care taken in pasting the isolated objects in new congruent backgrounds, rapid categorization performance was impaired with these manipulated stimuli.

This "pasting effect" was observed on accuracy and response speed. A global decrease of 2.1% of correct responses did not reach significance, but the tendency was observed in most subjects and affected the accuracy on targets more than the accuracy on distractors. This accuracy deficit on target trials was associated with a significant median RT increase of about 10 ms. This effect also was observed even for the earliest responses as illustrated by the shift of the d' curves between original and pasted conditions. This performance impairment could be due to the decreased saliency of the foreground object when pasted on a new context as shown by the saliency analysis. If object saliency could bias early processing, performance should be improved with isolated objects that are of course salient in 100% of the cases. This was not the case although this result might be due to a ceiling effect.

Alternatively, the "pasting" effect could be due to other alterations. We can consider the local physical alteration at the object/context boundary possibly due to techniques used to introduce the object in a new context (see methods). Despite all the careful precautions taken during stimulus manipulation to prevent from reported violation effects (equating spatial layouts, relative scales, supports, object interpositions, Biederman et al., 1982), not all the physical features of realistic photographs could be preserved: object illuminations and shadows might not be coherent, thus violating usual co-occurrence of certain visual features. Pasting effects might result from the violation of such expectations.

Finally, another explanation of the pasting effect might be related to our definition of a "congruent" context. The present experiment considers as equivalent all natural contexts versus all urban contexts, but subjects might have clear expectations about where to find a given animal. For example, giraffes tend to live in dry, open wooded areas in the savannah and might be incongruent in a mountain scene; the mountain scene would be considered as congruent in the present study but

might not be congruent for the giraffe. Although plausible, this explanation does not take into account the fact that recognition of a detailed context such as "sea scenes" or "mountains scenes" takes longer than the recognition of such scenes as "natural context" (Joubert et al., 2007), and that the impairment observed here was significant from the earliest responses.

The "congruence" effect

Experiments 2 and 3 used only manipulated stimuli, thus enabling the evaluation of contextual congruence on its own, preventing interference from other visual regularity violations revealed by Experiment 1. Both experiments clearly showed an effect of context congruence, an effect that was present regardless of the object saliency in the scene. When objects were pasted in a non-congruent context, categorization performance on targets was always worse both in terms of accuracy (10% drop) and response speed (15 ms slower) performance dropped by 10% and about 15 ms on targets. Animals were clearly less easily categorized when presented in an urban context than in a natural context. The results were less robust for the false alarms, induced by manufactured objects presented in a natural (non-congruent) context; further investigations are needed to better understand the pattern of false alarms triggered by the man-made objects.

The contextual effect on target categorization replicated in two experiments provides evidence that context processing strongly interacts with object categorization; this result is strengthened by the fact that subjects had no a priori information about briefly flashed scenes, and were thus free from top-down influences that can be primed when presentations of visual scenes precede the processing of target objects (Palmer, 1975). Obviously, in daily life, our environment is usually stable allowing predictions to be made on the most likely objects to appear. In such cases, context effects on object perception are not time constrained and probably strengthened. On the other hand, the early interactions demonstrated in our experiment might be fully used in circumstances where the surrounding context changes suddenly, as when opening a door and entering a new scene, when making large head movements, when driving a car in a city, zapping from one TV channel to another, or watching family photographs. The second aim of our study was to determine the temporal dynamics of object and context interactions when presented simultaneously. Here the results were very surprising since no minimum delay was necessary to observe a context influence on object categorization. The conflict between objects and their surrounding contexts induced

an additional processing time of about 10-20 ms when considering the minimal input-output processing time (Figure 3 and 5) and increased even more for longer response latencies.

This result clearly argues against the functional isolation model proposed by Henderson and Hollingworth in which objects and contexts are processed independently without interfering (Henderson & Hollingworth, 1999, Hollingworth & Henderson, 1998). Furthermore, these object-context interactions could be bi-directional, since Davenport et al. (2004) and Joubert et al. (2007) provided evidence that salient foreground objects can also influence context processing. Notably, Joubert et al. (2007) have recently shown that the time-course of context and object processing are very similar. This implies that in some cases the context might be categorized faster than the foreground object. In that case, the rapidly processed contexts might interfere with object categorization at a pre-decisional stage. Conversely, for other natural scenes in which salient objects might be categorized faster than context, object processing would also influence context recognition at a pre-decisional stage. In most intermediate cases, one can postulate bidirectional interactions before any decision has been reached on either object or context category. Our results thus support perceptual schema models which propose that the flow of object and context processing can interact early on during perceptual processing.

In a model proposed by Bar and his group (Bar, 2004, Bar, Kassam, Ghuman, Boshyan, Schmidt, Dale, Hamalainen, Marinkovic, Schacter, Rosen & Halgren, 2006), a coarse processing of the context performed through the magnocellular dorsal visual pathway can influence object recognition (Bullier, 2001). This model could well account for interaction of context on object recognition but would have more difficulty for the opposite case. Macé et al. (2005) also emphasized the guidance help that can be provided by the fast magnocellular pathway but, rather than setting this influence through a control exerted by the dorsal visual stream on the ventral visual stream, such interactions were proposed to take place mostly within the ventral visual pathway. Following these views, at each processing stage of the ventral visual stream, the fast magnocellular pathway could feedback information to guide the processing of the slowest parvocellular information in the preceding stages.

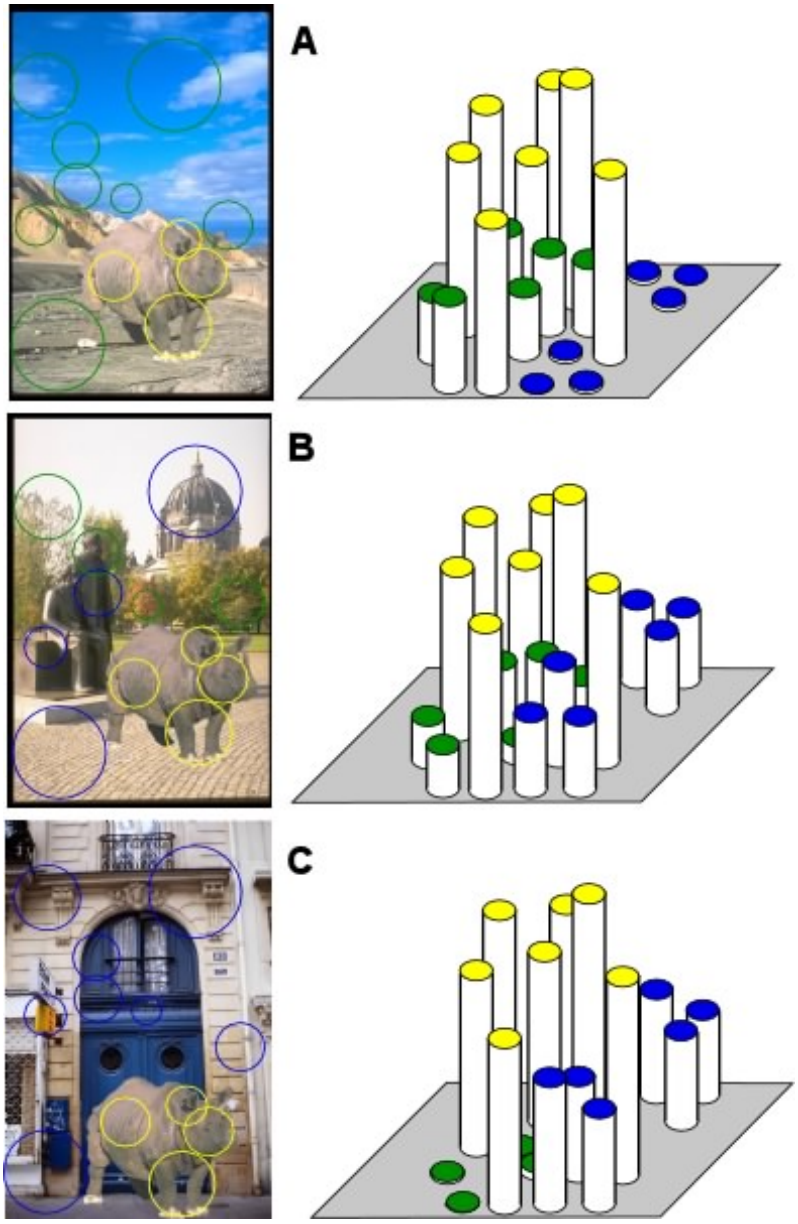

Figure 6: Hypothesised activation of different populations of neurons in the extrastriate areas of the ventral visual stream under different contextual conditions. Neuronal populations specifically activated by animal features (yellow), by natural context features (in green), and by man-made context features (in blue). **(A)** In optimal conditions when an animal appears in a congruent natural context, the co-activation of neuronal populations responding to "natural" and "animal" features that are used through experience to fire simultaneously, facilitates object recognition. **(B)** Intermediate conflictual situations arise when man-made features are presented together with animal and natural features and cause interference. **(C)** Maximal conflict is reached when the contextual information provides only non-congruent information.

Within extrastriate areas in the ventral visual pathway, populations of neurons would tend to strengthen their connections when co-activated by groups of objects that tend to appear simultaneously. In performing our task that requires to respond only to an animal-target, top-down preparation of the visual system is presumably maximal and this preparation

would extend to contextual scenes in which animals are commonly seen. Through parallel processing, an animal in a natural scene would be the expected optimal stimulus and would co-activate multiple populations of neurons (Figure 6.A) that are usually co-activated. On the other hand, when the animal appears in an urban scene, a conflict would arise between populations of neurons that respond to elements of the scene that are not usually co-activated (animal and urban man-made features). Such conflict might range from moderate (when some expected natural features are present in an urban background, e.g. Figure 6.B) to extreme (Figure 6.C). The more incongruent features in the background of the scene, the greater the competition between the neuronal responses to the background and the neuronal responses to the animal target. Facilitation would arise between populations of neurons that have reinforced mutual connections because they tend to fire simultaneously (Hebb, 1949); such learning of visual co-variations has been shown to be implicit (Chun & Jiang, 1999, Jiang & Chun, 2001). Otherwise, interference would take place. Hence, with strong interference, the conflict between go and no-go responses would take longer to resolve or might lead to an incorrect motor decision at the level of the prefrontal cortex (Rousselet, Fabre-Thorpe & Thorpe, 2002, Rousselet, Thorpe & Fabre-Thorpe, 2004). It has recently been postulated that the parahippocampal cortex (PHC) could mediate the representation of familiar object associations (Aminoff, Gronau & Bar, 2007, Bar, 2004, Bar & Aminoff, 2003). Thus, the conflict might be present all along the visual stream; it might be maximal in the PHC that receives information directly from the ventral visual stream (Suzuki, 1996, Suzuki & Amaral, 1994) and would encode recurrent regularities or associations in our surrounding world. The conflict could thus arise in the first feed-forward sweep of the earliest available visual information and might explain why the interaction between object and background can be observed even on the earliest behavioral responses that have been suggested to depend mostly on upon feed-forward processing (Fabre-Thorpe et al., 2001, Mace et al., 2005, VanRullen & Thorpe, 2002).

Implications and conclusion

In the light of these new results, it appears that objects and scene context are processed in parallel and engaged in bidirectional interactions. Among the three possible models proposed by Henderson and Hollingworth (1999), our results rule out the *functional isolation model* in which object identification is not influenced by expectations from scene context, and the

priming model in which contextual influences occur during decisional stage "when a structural description of an object token is matched against long-term memory representation". On the other hand they support the *perceptual schema model*, which describes object and context processing interacting at perceptual stages. The immediate effect of context congruence on object categorization using briefly flashed scenes is also compatible with a feed-forward wave of processing in which facilitation and interference between neuronal populations depends on whether they are usually co-activated or not.

Acknowledgments

This work was supported by the CNRS (Centre National de la Recherche Scientifique), by the French government (Ministère de la recherche et de l'Enseignement supérieur) and by the Fondation pour la Recherche Médicale.

References

Aminoff, E., Gronau, N., & Bar, M. (2007). The parahippocampal cortex mediates spatial and nonspatial associations. *Cereb Cortex, 17* (7), 1493-1503. [Article]

Bar, M. (2004). Visual objects in context. *Nat Rev Neurosci, 5* (8), 617-629. [PubMed]

Bar, M., & Aminoff, E. (2003). Cortical analysis of visual context. *Neuron, 38* (2), 347-358. [PubMed]

Bar, M., Kassam, K.S., Ghuman, A.S., Boshyan, J., Schmidt, A.M., Dale, A.M., Hamalainen, M.S., Marinkovic, K., Schacter, D.L., Rosen, B.R., & Halgren, E. (2006). Top-down facilitation of visual recognition. *Proc Natl Acad Sci U S A, 103* (2), 449-454. [Article]

Bar, M., & Ullman, S. (1996). Spatial context in recognition. *Perception, 25*, 343-352. [PubMed]

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol Rev, 94* (2), 115-147. [Abstract]

Biederman, I., Mezzanotte, R.J., & Rabinowitz, J.C. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cognit Psychol, 14* (2), 143-177. [Abstract]

Biederman, I., Rabinowitz, J.C., Glass, A.L., & Stacy, E.W., Jr. (1974). On the information extracted from a glance at a scene. *J Exp Psychol, 103* (3), 597-600. [PubMed]

Boyce, S.J., & Pollatsek, A. (1992). Identification of objects in scenes: the role of scene background in object naming. *J Exp Psychol Learn Mem Cogn, 18* (3), 531-543. [PubMed]

Boyce, S.J., Pollatsek, A., & Rayner, K. (1989). Effect of background information on object identification. *J Exp Psychol Hum Percept Perform, 15* (3), 556-566. [PubMed]

Bullier, J. (2001). Integrated model of visual processing. *Brain Res Brain Res Rev, 36* (2-3), 96-107. [PubMed]

Chun, M.M., & Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. *Psychological Science, 10* (4), 360-365.

Davenport, J.L., & Potter, M.C. (2004). Scene consistency in object and background perception. *Psychol Sci, 15* (8), 559-564. [PubMed]

Davenport, J.L. (2007). Consistency effects between objects in scenes. Mem Cognit, 35 (3), 393-401. [PubMed]

De Graef, P., Christiaens, D., & d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. Psychol Res, 52 (4), 317-329. [PubMed]

De Graef, P., De Troy, A., & D'Ydewalle, G. (1992). Local and global contextual constraints on the identification of objects in scenes. Can J Psychol, 46 (3), 489-508. [PubMed]

Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *J Cogn Neurosci, 13* (2), 171-180. [PubMed]

Fiser, J., & Aslin, R.N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychol Sci, 12* (6), 499-504. [PubMed]

Ganis, G., & Kutas, M. (2003). An electrophysiological study of scene effects on object identification. *Brain Res Cogn Brain Res, 16* (2), 123-144. [PubMed]

Hebb, D.O. (1949). The Organization of Behavior. A Neuropsychological Theory. (New-York: Wiley).

Henderson, J.M., & Hollingworth, A. (1999). High-level scene perception. *Annu Rev Psychol, 50*, 243-271. [PubMed]

Hollingworth, A., & Henderson, J.M. (1998). Does consistent scene context facilitate object perception? *J Exp Psychol Gen, 127* (4), 398-415. [PubMed]

Hollingworth, A., & Henderson, J.M. (1999). Object identification is isolated from scene semantic constraint: evidence from object type and token discrimination. *Acta Psychol (Amst), 102* (2-3), 319-343. [PubMed]

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. Nat Rev Neurosci, 2 (3), 194-203. [PubMed]

Jiang, Y., & Chun, M.M. (2001). Selective attention modulates implicit learning. *Q J Exp Psychol A, 54* (4), 1105-1124. [PubMed]

Joubert, O.R., Rousselet, G.A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Res, 47* (26), 3286-3297. [PubMed]

Kosslyn, S.M. (1987). Seeing and imagining in the cerebral hemispheres: a computational approach. *Psychol Rev, 94* (2), 148-175. [PubMed]

Macé, M.J., Thorpe, S.J., & Fabre-Thorpe, M. (2005). Rapid categorization of achromatic natural scenes: how robust at very low contrasts? *Eur J Neurosci, 21* (7), 2007-2018. [PubMed]

Macmillan, N.A., & Creelman, C.D. (2005). Detection Theory A user's Guide. (Routledge). [PubMed]

Marr, D. (1982). Vision. (San Francisco, CA: Freeman. [PubMed]

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. *Prog Brain Res, 155*, 23-36. [PubMed]

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends Cogn Sci, 11* (12), 520-527. [PubMed]

Palmer, S.E. (1975). The effects of contextual scenes on the identification of objects. *Memory and Cognition, 3* (5), 519-526.

Peterson, M.A. (1994). Object recognition processes can and do operate before figure-ground organization. *Curr Direct Psychol Sci, 3*, 105-111.

Peterson, M.A., & Gibson, B.S. (1994a). Must figure-ground organization precede object recognition? *Psychol Sci, 5* (5), 253-259.

Peterson, M.A., & Gibson, B.S. (1994b). Object recognition contributions to figure-ground organization: operations on outlines and subjective contours. *Percept Psychophys, 56* (5), 551-564. [PubMed]

Potter, M.C., & Faulconer, B.A. (1975). Time to understand pictures and words. *Nature, 253* (5491), 437-438. [PubMed]

Potter, M.C., & Levy, E.I. (1969). Recognition memory for a rapid sequence of pictures. *J Exp Psychol, 81* (1), 10-15. [PubMed]

Rousselet, G.A., Fabre-Thorpe, M., & Thorpe, S.J. (2002). Parallel processing in high-level categorization of natural images. *Nat Neurosci, 5* (7), 629-630. [PubMed]

Rousselet, G.A., Joubert, O.R., & Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cogn, 12* (6), 852-877. [Abstract]

Rousselet, G.A., Mace, M.J., & Fabre-Thorpe, M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *J Vis, 3* (6), 440-455. [Article]

Rousselet, G.A., Thorpe, S.J., & Fabre-Thorpe, M. (2004). How parallel is visual processing in the ventral pathway? *Trends Cogn Sci, 8* (8), 363-370. [PubMed]

Suzuki, W.A. (1996). Neuroanatomy of the monkey entorhinal, perirhinal and parahippocampal cortices: Organization of cortical inputs and interconnections with amygdala and striatum. *Semin Neurosci, 8* (1), 3-12.

Suzuki, W.A., & Amaral, D.G. (1994). Perirhinal and parahippocampal cortices of the macaque monkey: cortical afferents. *J Comp Neurol, 350* (4), 497-533. [PubMed]

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381* (6582), 520-522. [PubMed]

Thorpe, S.J., Gegenfurtner, K.R., Fabre-Thorpe, M., & Bulthoff, H.H. (2001). Detection of animals in natural images using far peripheral vision. *Eur J Neurosci, 14* (5), 869-876. [PubMed]

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network, 14* (3), 391-412. [PubMed]

VanRullen, R., & Thorpe, S.J. (2002). Surfing a spike wave down the ventral stream. *Vision Res, 42* (23), 2593-2615. [PubMed]

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. Neural Netw, 19 (9), 1395-1407. [PubMed]

PARTIE IV

SYNTHESE, MODELE, ET PERSPECTIVES

Si le monde tel que tu le vois ne te plait pas... Ferme les yeux…
Ouvre les yeux… Un nouveau monde s'offre à toi…

Dans les parties expérimentales de ce mémoire ont été présentés 4 articles dont 3 publiés ou acceptés et 1 en révision, permettant de mieux cerner les décours temporels des traitements perceptifs et catégoriels impliqués dans la catégorisation de scènes naturelles au sein d'un paradigme go/no-go. Ces 4 études font partie intégrante d'un projet de recherche plus vaste visant à caractériser la catégorisation de l'objet, du contexte, et leurs interactions à différents niveaux de traitements et sous différentes contraintes cognitives et temporelles. En effet, j'ai eu l'occasion durant mon doctorat de mettre en place ou de participer à d'autres études impliquant des analyses EEG ou des mouvements oculaires respectivement lors de tâches de catégorisation ou de discrimination catégorielle plus ou moins complexes de scènes naturelles. Ces études ont déjà apporté nombre de données complémentaires; les articles en résultant sont en préparation et devraient donner lieu à publication dans les années à venir. De plus, les expériences menées au cours de ce doctorat ont mis à jour d'autres interrogations qui ouvrent de nouvelles voies de recherche et déboucheront sans nul doute sur de nouvelles études, au sein du CerCo mais également durant mon projet post-doctoral sous la direction de Aude Oliva dans le département Brain and Cognitive Science au MIT.

Face à un tel bilan, j'ai souhaité organiser la partie IV de ce mémoire en 5 sous-parties. Dans un premier temps, je dresserai un bilan des résultats déjà présentés et discuterai leurs implications par rapport à la littérature scientifique déjà publiée sur la catégorisation visuelle. De manière plus spécifique, la dynamique observée des interactions objet/contexte sera discutée par rapport aux modèles relatifs aux interactions objet/contexte décrits dans la partie III (Chap. 2). A l'issue de ces deux premières parties, je proposerai un modèle prenant en compte l'ensemble de ces résultats expérimentaux. Finalement, je présenterai brièvement les études complémentaires que j'ai menées et qui contraignent d'autant plus les hypothèses relatives aux traitements catégoriels de scènes naturelles, et je suggèrerai de nouvelles expériences qui me semblent primordiales à une meilleure compréhension des traitements perceptifs et catégoriels sous-jacents, notamment dans des conditions plus écologiques.

1. Bilan des résultats et implications pour le modèle

Un traitement de la scène global, parallèle, pondéré, donnant lieu à des interactions entre objet et contexte

Dans les quatre articles présentés dans ce mémoire, les expériences décrites reposaient sur l'utilisation d'un paradigme go/no-go dans lequel les stimuli cibles et distracteurs étaient présentés seulement 26 ms. Ainsi aucune exploration oculaire n'était possible. Il est important cependant de préciser qu'aucun masque n'apparaissait après les stimuli ce qui contraint le temps de la prise d'information même si l'on doit considérer la persistance rétinienne de l'image communément estimée aux alentours de 50 ms. Malgré ces contraintes temporelles, les sujets impliqués dans les tâches de catégorisation de contexte ou d'objet à différents niveaux ne font que très peu d'erreurs et atteignent dans la majorité des conditions des précisions supérieures à 94 % de réussite. Cela n'est pas vide de sens. Cela démontre que le contexte, souvent considéré comme d'intérêt mineur et davantage distribué spatialement, peut être appréhendé aussi précisément que des objets d'intérêts plus locaux. Des scores de précision inférieurs n'ont été enregistrés que lorsque les informations disponibles dans les stimuli étaient conflictuelles ou insuffisantes: (1) dans la tâche de catégorisation de contexte à un niveau superordonné lorsqu'un objet incongruent était présent (81.5%), (2) dans la tâche de catégorisation animal/non-animal lorsque le contexte était incongruent (85%), ou encore (3) dans la tâche de catégorisation de contexte lorsque les informations de phase étaient altérées (de plus de 50% : niveau chance). Nous démontrons ainsi que la présence d'une information incongruente a priori non pertinente pour la tâche à effectuer interfère sur les niveaux de précision atteints lors de la catégorisation d'objets ou de contextes (au moins à un niveau superordonné) au sein des scènes naturelles : un objet incongruent interfère sur la catégorisation du contexte tandis qu'un contexte incongruent interfère sur la catégorisation de l'objet. Ces interactions entre les traitements de l'objet et du contexte sont donc bidirectionnelles.

Ces résultats forcent à penser que l'ensemble de la scène est traitée de manière parallèle, même si seule une partie de la scène contient l'information cible. D'autres études vont d'ailleurs dans le même sens. Il a en effet été démontré que le cerveau humain était capable de traiter en parallèle deux scènes naturelles afin de déterminer la présence ou non

d'un animal au sein de l'une d'elles (Rousselet, Fabre-Thorpe & Thorpe, 2002). Les performances de catégorisation sur une scène naturelle apparaissant autour du point de fixation ne sont pas affectées lorsque les scènes naturelles peuvent apparaître de façon aléatoire en 9 excentricités horizontales différentes (jusqu'à 70.5°, Thorpe et al., 2001) montrant la capacité du cerveau à traiter l'ensemble du champs visuel horizontal. Enfin, le cerveau est capable d'effectuer en parallèle une tâche de détection de lettres en région fovéale tout en catégorisant une voire deux scènes présentées simultanément en régions périphériques sans exigence attentionnelle supplémentaire (Fei-Fei, VanRullen, Koch & Perona, 2005, Li, VanRullen, Koch & Perona, 2002 , VanRullen, Reddy & Koch, 2004).

Pourtant, dans les tâches de catégorisation d'objet ou de contexte, les sujets sont capables de mettre l'accent respectivement sur les informations relatives à l'objet ou au contexte même si la zone d'intérêt couvre une moins grande surface spatiale. Toutes les régions de la scène ne sont donc pas traitées avec la même importance ce qui suggère une modulation des traitements ascendants des zones spécifiques de l'objet et du contexte. Pour autant, il est difficilement imaginable qu'une étape de ségrégation précoce de l'objet et de son contexte ait lieu étant donné que des performances similaires sont enregistrées lors de la catégorisation d'objets qu'ils soient isolés ou en contextes congruents. Nos résultats pourraient s'expliquer par des interactions entre traitements catégoriels de l'objet et du contexte à différents niveaux. L'effet de congruence du contexte que nous avons observé semble reposer sur des interactions plutôt haut-niveau (perceptives ou décisionnelles) que bas-niveau (physique). En effet, même s'il semble que les traitements catégoriels des informations cibles soient modulés par la surface qu'elles couvrent dans la scène (Delorme, 2000), Delorme et Fabre-Thorpe, en préparation), cet aspect physique ne suffit pas pour autant à expliquer l'effet de congruence comme le montre l'article 4 dans lequel les mêmes objets étaient utilisés dans les conditions de « contexte congruent » et de « contexte incongruent ». De plus, nous avons également contrôlé que l'effet de congruence était indépendant de la saillance de l'objet d'intérêt (article 4 également), saillance mesurée en prenant en compte les caractéristiques physiques de couleurs, de luminance et d'orientation, et donc par là même de contraste. Enfin, l'effet de congruence révélé dans nos études est probablement sous-estimé en comparaison des conditions où le contexte apparaît de manière continue et stable comme dans la vie de tous les jours où influences descendantes prédictrices s'ajoutent aux interactions au sein des traitements ascendants. Comme nous l'avons évoqué dans la discussion de l'article 4, l'influence facilitatrice du contexte congruent sur la catégorisation des objets pourrait

s’expliquer par des connexions fonctionnelles excitatrices qui s'établiraient au cours du développement entre populations neuronales sélectives à des catégories d’objets et de contextes habituées à décharger simultanément. Au contraire, lors de la présentation d’un objet dans un contexte incongruent, le conflit entre populations neuronales sélectives à l’objet et au contexte serait maximal, ralentissant ainsi l’intégration de l’information visuelle et/ou la prise de décision.

Outre les interactions haut-niveau pouvant sous-tendre l’effet de congruence du contexte, d’autres interactions cette fois-ci plutôt physiques pourraient être envisagées. En effet, nous avons pu observer une baisse de performances lors de la catégorisation d’objet après manipulation des stimuli pour insérer l'objet dans un contexte congruent (de même catégorie superordonnée) différent du contexte original. Cette légère baisse de performance pourrait être attribuée à l’insertion non maîtrisée d’incohérences physiques bas-niveau (non respect de l’ombrage, défauts dans les contours…). Cependant, d'autres hypothèses alternatives peuvent être invoquées telles que l’insertion également non maitrisée d’incongruences contextuelles plus haut-niveau.

Dans tous les cas, l’incohérence n’existe que parce que le contexte qui entoure l'objet doit imposer certaines règles contraignantes sur le traitement de l'objet, il y a donc une interaction significative entre l’objet et son contexte. Ces résultats ne sont pas sans rappeler ceux de Biederman démontrant avec des dessins au trait une moins bonne reconnaissance des objets lorsque ces derniers violaient certaines règles syntactiques et sémantiques par rapport au contexte (Biederman et al., 1982).

Le modèle proposé pour la catégorisation visuelle rapide de scènes naturelles devra donc rendre compte d’un traitement global, ascendant, et parallèle de la scène au sein duquel l’intégration des informations relatives à l’objet ou au contexte devra pouvoir être pondérée par de la tâche à accomplir, potentiellement par une composante descendante pré activant le système à l'exécution de la tâche. Enfin, le modèle devra rendre compte d’interactions entre objet et contexte potentiellement à différents niveaux de traitements : perceptivo-décisionnels (effet de congruence) et peut-être physique (effet lié à la manipulation des stimuli).

Des décours temporels supportant l'hypothèse coarse-to-fine et la notion de distance perceptive entre catégories

Se limiter à l'analyse des scores de précision serait cependant limitatif pour à la compréhension des traitements visuels sous-jacents. En effet, la précision atteinte dans une tâche donnée ne fournit qu'une information partielle sur les capacités de notre cerveau à accomplir la tâche. L'analyse complémentaire des temps de réaction permet de définir en plus la dynamique de la prise d'informations diagnostiques tout en s'affranchissant des différences de stratégies mises en place. De ces analyses temporelles sont ressorties plusieurs évidences : (1) les catégories basiques d'objets et de contextes sont visuellement appréhendées aussi rapidement, mais plus lentement que leurs catégories superordonnées respectives, (2) selon la tâche, certaines informations de la scène semblent très diagnostiques et d'autres simplement facilitatrices, et surtout (3) les interactions entre objet et contexte surviennent très précocement dans le traitement des scènes.

En effet, tandis que les sujets peuvent catégoriser visuellement le contexte d'une scène à un niveau superordonné (Env. Man vs. Env. Nat.) en moins de 400 ms, les informations captées pendant le flash du stimulus doivent encore être traitées pendant 50 ms supplémentaires en moyenne pour définir la catégorie basique de ce même contexte (Mer, Montagne, Intérieurs urbains, scènes de rues, 400-460 ms, Figure 34).

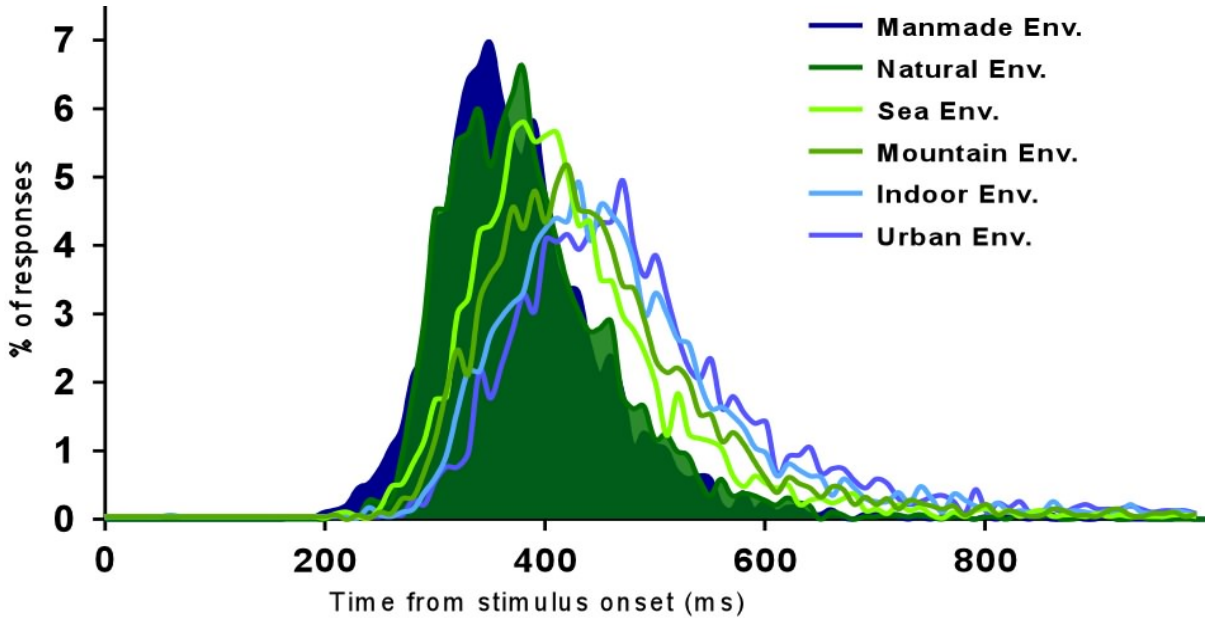

Figure n°34 : Distribution des réponses go correctes au cours du temps lors de tâche de catégorisation de contexte à des niveaux superordonnés et à des niveau basiques.

Ce résultat n'est pas sans rappeler les résultats similaires obtenus dans des tâches de catégorisation visuelle d'objets. Tandis que les sujets sont capables d'effectuer une catégorisation superordonnée animal/ non-animal en moins de 400 ms (Delorme et al., 2000, Fabre-Thorpe et al., 1998, Fize, Fabre-Thorpe, Richard, Doyon & Thorpe, 2005, Macé et al., , Macé et al., 2006, Rousselet et al., 2003, Thorpe et al., 1996, VanRullen & Thorpe, 2001a), l'accès à la catégorie basique d'un animal comme « chien » ou "oiseau" nécessite également 50 ms de traitements additionnels environ (Macé et al., 2006). Ces résultats peuvent non seulement être mis en relation avec la hiérarchie des niveaux de catégorisation définie par Rosch pour les objets et par Tversky pour les contextes (Rosch et al., 1976, Tversky & Hemenway, 1983), mais également être mis en relation avec les hypothèses « coarse-to-fine » des traitements visuels proposés par Schyns et Oliva (Oliva & Schyns, 1997, Schyns & Oliva, 1994). En effet, Rosch définit les catégories superordonnées comme plus générales et plus abstraites que les catégories basiques. Les catégories basiques sont au contraire des catégories plus spécifiques partageant de nombreux attributs communs, tout en possédant d'autres caractéristiques exclusives. Si on considère qu'un premier traitement grossier de la scène a lieu avant que des informations plus fines soient disponibles (hypothèse coarse-to-fine), il apparaît donc logique que les catégories superordonnées soient visuellement déterminées avant les catégories basiques. Dans la tâche de catégorisation de contexte à un niveau superordonné, la brièveté des temps de réaction minimaux enregistrés (260-270 ms) laisse supposer que des représentations visuelles grossières seraient élaborées dès l'aboutissement des traitements rapides de l'information magnocellulaire (d'autant plus que des analyses EEG non publiées sur cette même tâche ont révélé une activité différentielle liée au statut de l'image dès 170 ms après la présentation du stimulus). Les informations visuelles plus fines de la scène ne seraient que plus tardivement disponibles, suite à l'intégration plus lente des informations parvocellulaires. A cela pourrait s'ajouter l'influence de boucles de rétro-action itératives supposant des traitements descendants dépendant de l'intégration des informations aux niveaux visuels supérieurs permettant par exemple de pondérer l'importance diagnostique des informations ascendantes au sein de la voie ventrale.

Ainsi, la préférence pour le niveau basique dans les études menées par Rosch et par Tversky pourrait être spécifique des réponses verbales qui nécessitent un accès au lexique Or on communique beaucoup plus fréquemment avec la terminologie des catégories de base (chiens, fruits, oiseaux, meubles, voiture etc…). Cette fréquence d'utilisation entraînerait une plus grande rapidité d'accès à ces termes privilégiés et la lenteur d'accès aux termes de

catégories superordonnées serait telle qu'elle masquerait l'avantage que nous voyons ici au niveau visuel, les étapes de traitements lexicaux n'étant pas nécessaires dans nos tâches de catégorisation visuelle. Le paradigme go/no-go utilisé dans nos études implique une réponse motrice et s'affranchit donc d'un accès nécessaire au lexique à la présentation de chaque stimulus. L'accès au lexique pourrait être seulement effectué lors de l'intégration de la consigne par le sujet au début d'un bloc d'images (rappelons qu'un bloc comprend en général 100 essais). A partir des consignes données, la préactivation du système visuel à la tâche serait effectuée pour privilégier le traitement des informations les plus diagnostiques de la catégorie cible. De ce fait, la prise de décision ne nécessiterait pas une représentation aboutie de la scène mais seulement l'intégration des informations les plus diagnostiques de la catégorie cible, modulées en partie par les informations spécifiques à la catégorie distractrice. Plus la distance perceptive entre cibles et distracteurs serait importante, plus la nature de ces informations diagnostiques de la catégorie cible pourrait être grossières, et de fait une moins grande quantité d'informations à intégrer suffirait à rendre possible la prise de décision. Nos résultats corroborent d'ailleurs grandement cette notion de distance perceptive, puisque comme l'a suggéré Rosch, cette distance perceptive serait plus grande entre catégories superordonnées qu'entre catégories basiques, les catégories superordonnées ayant moins d'attributs en commun que les catégories basiques. Plus la distance perceptive entre cibles et distracteurs est faible, plus la probabilité que des informations considérées comme diagnostiques de la catégorie cible se trouvent présentes dans les images distractrices est importante. Dans ce sens, dans nos tâches demandant une catégorisation de contexte au niveau de base, il paraît également logique d'observer un plus grand nombre d'erreurs commis sur les contextes appartenant à la même catégorie superordonnée.

Le modèle de catégorisation visuelle rapide de scènes naturelles devra donc considérer la précocité avec laquelle peuvent être extraites les informations diagnostiques de la scène cible pour permettre une rapidité du système visuel plus grande dans des tâches de catégorisation superordonnée que dans des tâches de catégorisation basique. La notion de distance perceptive entre informations diagnostiques de la catégorie cible et des catégories distractrice est importante puisque la hiérarchie "temporelle" serait sous-tendue par une accumulation progressive de l'information diagnostique.

Nature des informations diagnostiques essentielles et des informations secondaires

La rapidité des sujets constatée dans des tâches de catégorisation ayant la même catégorie cible mais limitant les informations disponibles dans les stimuli proposés nous permet de préciser l'influence de certaines caractéristiques visuelles de la scène sur les traitements catégoriels.

Le premier article a ainsi démontré que les sujets humains étaient encore capables de catégoriser le contexte de scènes achromatiques à un niveau superordonné lorsque seulement 55 % de l'information de phase originale était encore disponible et que la diagnosticité de l'information d'amplitude avait été supprimée. Il semble donc que le spectre de phase des images véhicule une certaine quantité d'information diagnostique essentielle à une bonne catégorisation . Or le spectre de phase code la phase des fréquences spatiales en fonction de leur orientation, c'est-à-dire la position relative de chaque sinusoïde spatiale au sein de l'image. C'est grâce aux informations comprises dans le spectre de phase que nous percevons des contours organisés et localisés, contours bien plus prépondérants dans les scènes « Env. Man » que « Env. Nat.» (McCotter, Gosselin, Sowden & Schyns, 2005). Il est d'ailleurs intéressant de rappeler que les informations du spectre de phase prises dans leur globalité suffisent à la localisation des informations de luminance de la scène, sans qu'aucun traitement spatialement localisé ne soit nécessaire. De part leur redondance le long d'orientations spécifiques, les contours diagnostiques des « Env. Man » seraient moins altérés par l'insertion de bruit dans la phase, ce qui expliquerait la stratégie des sujets à baser leurs décisions préférentiellement sur leur détection pour effectuer la tâche.

De fait, les caractéristiques absentes du spectre de phase telles que la couleur et l'amplitude apparaissent dès lors comme secondaires.

En comparant les temps de réaction obtenus lors de la présentation de scènes originales avec ceux obtenus lors de la présentation de scènes sans informations d'amplitude diagnostiques, il apparaît que les informations d'amplitude pourrait avoir un rôle facilitateur. Cet effet facilitateur est particulièrement rapide -au moins pour certaines scènes- puisqu'il est présent dès les réponses les plus précoces, comme le montrent les distributions des réponses correctes go et les courbes d' décalées vers les latences plus longues dans le cas de la condition « amplitude égalisée » (vs. scènes originales).

IV.1 Bilan des résultats et implications pour le modèle

L'article n°2 apporte des informations complémentaires sur l'influence cette fois ci de la couleur dans la catégorisation du contexte à nu niveau basique. Dans cette étude, nous avons en effet comparé les temps de réaction obtenus dans la même tâche alors que les stimuli proposés étaient soit les scènes originales soit les mêmes scènes mais achromatiques. Ces analyses temporelles démontrent que la couleur a une influence positive mais tardive sur la catégorisation. Dans ce sens, la couleur pourrait constituer un complément d'informations pour la catégorisation des scènes relativement difficiles à catégoriser sur la base des caractéristiques achromatiques. Cette facilitation de la couleur a d'ailleurs déjà été décrite dans d'autres études antérieures. Dans une première étude menée par Gegenfurtner, une scène naturelle cible chromatique ou achromatique était présentée aux sujets pendant un temps variable suivie d'un masque coloré, puis deux scènes naturelles test pouvant également être chromatiques ou achromatiques étaient présentées aux sujets qui devaient alors désigner celle qui correspondait à la scène cible (Gegenfurtner & Rieger, 2000). Les auteurs démontrent ainsi l'influence positive de la couleur aussi bien pour la scène cible que pour les scènes tests et suggèrent alors un rôle facilitateur de la couleur aussi bien au niveau de l'encodage de l'information qu'en termes de rappel mnésique.

En 1996, dans une tâche de dénomination de contexte brièvement présentés, Oliva et Schyns démontrent également la diagnosticité des informations chromatiques pour des scènes d' « Env. Nat », mais l'importance des informations de couleur n'est pas reproduit pour les « Env. Man ». En 2000, ils confirmeront l'influence de la couleur uniquement sur la reconnaissance des scènes pour lesquelles la couleur est diagnostique de la catégorie par l'utilisation de scènes achromatiques, aux couleurs normales et aux couleurs inconsistantes (Figure 35, Oliva & Schyns, 2000).

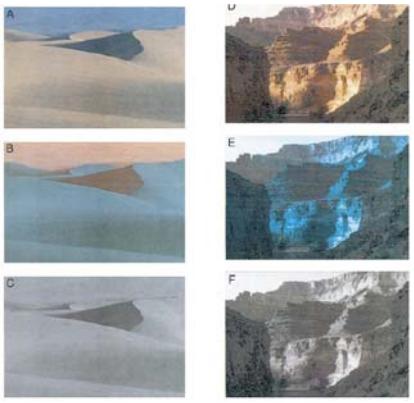

Figure n°35 : Tiré de (Oliva & Schyns, 2000). Exemples d'environnements naturels et manufacturés avec de haut en bas, des couleurs diagnostiques, non diagnostiques et en niveaux de gris

Ces résultats comportementaux ont été confirmés et complétés par la suite par des analyses EEG au cours d'une expérience similaire (Goffaux, Jacques, Mouraux, Oliva, Schyns & Rossion, 2005). Les enregistrements EEG ont alors révélé une composante frontale dont l'amplitude variait en fonction de la diagnosticité des couleurs au sein de la scène. Cette composante survenant 150 ms après l'apparition des scènes composées de leurs couleurs originales, était retardée lors de la présentation de scènes achromatiques, et encore plus lors de la présentation de scènes avec des couleurs anormales (vs. couleurs normales). Cette aide mitigée de la couleur n'est cependant pas spécifique de la reconnaissance des scènes naturelles dans leur globalité mais a également été reportée lors de tâche d'identification et de catégorisation d'objets (Part I. Chap 1).

Ainsi, un modèle de catégorisation visuelle rapide des scènes naturelles aussi complet que possible devrait prendre en compte la prépondérance des informations visuelles de phase, et par là même des contours organisés, sur des informations de couleurs ou d'amplitude moins cruciales.

Dynamique et nature des interactions entre les traitements de l'objet et du contexte

L'un des objectifs clé de ce mémoire était la caractérisation des décours temporels des interactions entre l'objet et son contexte lors de la catégorisation rapide de scènes naturelles. En effet, si de nombreuses études ont déjà mis en évidence une influence du contexte sur la reconnaissance des objets et si une seule étude avait révélé jusqu'alors une influence de l'objet sur la reconnaissance de contexte, aucune à ma connaissance n'a évalué aussi rigoureusement la dynamique de ces interactions au sein de scènes naturelles. Les contraintes temporelles de notre tâche et l'enregistrement des temps de réaction de la réponse manuelle nous ont permis de dresser les courbes de distributions des réponses go pour chaque condition expérimentale ainsi que les courbes d' dérivées s'affranchissant de biais stratégiques et caractérisant l'accumulation de l'information diagnostique. Ainsi, dans la tâche de catégorisation animal/non-animal en contexte congruent et incongruent, la comparaison des décours temporels entre les conditions « contexte congruent » et « contexte incongruent » nous a permis de démontrer que les interactions entre objets et contextes survenaient dès les réponses les plus précoces aux alentours de 260 ms. Ces divergences entre conditions congruente et incongruente sont également observables pour des réponses plus tardives. Ces résultats

suggèrent des interactions entre traitements de l'objet et traitements du contexte non pas ponctuelles mais continues à diverses étapes du traitement ascendant de la scène. Ces interactions pourraient de fait comprendre aussi bien des interactions bas-niveau, comme peut le suggérer la baisse de performance liée aux manipulations des stimuli dans l'article 3, que des interactions de plus haut-niveau tel que l'effet de congruence indépendant des caractéristiques physiques sous-tendant la saillance de l'objet. Pour autant, il faut garder en tête que les 400 ms nécessaires en moyenne à une prise de décision dans notre tâche ne suffisent pas à intégrer une représentation aboutie du contexte ou de l'objet, ce qui rend peu probables des interactions de très haut-niveau basées sur des concepts complexe. De telles interactions pourraient cependant exister et être mises en évidences au cours de tâches imposant des traitements additionnels plus couteux temporellement, telles que des tâches de catégorisation plus fine.

Malheureusement, la dynamique de l'influence de l'objet sur la catégorisation du contexte n'a pu être plus précisément caractérisée étant donné que la mise en évidence de cet effet a résulté d'une analyse post-hoc, et que de fait, nous ne disposions pas d'assez d'essais contenant un objet saillant incongruent pour tracer des distributions convenables. Cependant, les décours temporels du traitement de l'objet et du contexte étant très similaires, il est fort probable que des interférences continues de l'objet sur le contexte seraient également observées. Néanmoins, il serait intéressant de mettre en place une nouvelle expérience de catégorisation de contexte en utilisant par exemple les stimuli de l'article 4 pour confirmer cette hypothèse.

Peut-on déterminer la latence minimale des premières interactions objet/contexte? Complément d'informations avec la discrimination catégorielle rapide de scènes naturelles en choix forcé saccadique

Au cours des études présentées dans l'article n°3 et n°4, nous avons notamment démontré la possibilité de reconnaître la catégorie superordonnée d'une scène en moins de 400 ms, l'influence des objets saillants de la scène sur la catégorisation superordonnée du contexte, ainsi que l'influence du contexte sur la catégorisation superordonnée de l'objet. Cet ensemble de résultats fut obtenu à partir d'expériences basées sur un protocole go/no-go impliquant la production d'une réponse manuelle. Même si cette réponse a été utilisée pour

permettre d'obtenir les temps de réaction les plus courts possibles (une seule action, un seul effecteur, relever plutôt que d'appuyer le doigt), les temps de réaction manuels sont beaucoup plus longs que les temps de réaction oculaires. Nous avons donc voulu utiliser la tâche développée par Kirchner et Thorpe pour évaluer l'analyse du contexte et les interactions objet/contexte dans le cadre de contraintes temporelles plus importantes (Kirchner & Thorpe, 2006). Cette tâche correspond à une tâche de discrimination catégorielle (animal/non-animal) au sein de laquelle chaque essai est constitué de l'apparition d'une croix de fixation, suivie de la présentation simultanée de deux scènes naturelles, l'une apparaissant à gauche et l'autre à droite de l'écran. Tandis que l'une des images appartient à la catégorie cible (contient un animal) préalablement définie lors de la consigne donnée aux sujets, l'autre image est distractrice. L'image cible peut aléatoirement apparaître à droite ou à gauche de l'écran, et les sujets doivent porter leur regard vers la cible le plus précisément et le plus rapidement possible. Enfin, les mouvements oculaires sont enregistrés avec un oculomètre dont on dérive la latence d'initiation de la première saccade et la qualité des réponses.

Les résultats de cette étude montre que les sujets sont capables de diriger leur regard vers la cible contenant un animal dans 90% des essais. De plus, les auteurs observent des temps de réaction minimaux de 120 ms en moyenne. Cette latence très courte impose de sévères contraintes à prendre en compte pour comprendre comment le système visuel dans son ensemble et les traitements impliqués dans la discrimination catégorielle parviennent à intégrer assez d'informations pour réaliser correctement la tâche et déclencher la saccade après un temps de traitement aussi court. En se basant sur de nombreuses données électro-physiologiques et neuro-anatomiques, les auteurs proposent que dans ce type de tâche, les traitements visuels ascendants des informations diagnostiques de chaque hémisphère pourraient court-circuiter les aires inférotemporales. Les informations accumulées jusqu'en V4 seraient ainsi suffisantes pour autoriser une prise de décision déclenchée par l'activation des colliculi supérieurs comme illustré dans le modèle présenté par la figure 36. Il est à noter que biologiquement, une réponse oculaire rapide peut être cruciale et n'a pour seule conséquence -si elle est erronée- que de devoir être corrigée par une autre saccade. La "punition" pour le déclenchement d'une réponse manuelle erronée peut être beaucoup plus coûteuse pour le sujet.

Qu'en est-il de la discrimination catégorielle du contexte et surtout des interactions objet/contexte dans ce nouveau paradigme qui permet d'explorer comportementalement une fenêtre temporelle plus précoce qu'avec la réponse manuelle ? Nous avons utilisé ce protocole

de catégorisation avec réponse oculaire et choix forcé pour tester (1) la catégorisation de contexte, (2) l'influence des objets saillants sur la discrimination catégorielle des contextes, et (3) l'effet de congruence du contexte sur la catégorisation de l'objet. Dans les tâches de discrimination de contexte, les Env. Man et Env Nat pouvaient être dépourvus d'objets saillants ou pouvaient au contraire contenir un animal ou un véhicule. Dans les tâches de discrimination d'objet, nous avons réutilisé ces Env. Man et Env. Nat contenant un animal ou un véhicule. Contrairement à l'étude de l'article n°4, toutes les photos ont été récupérées sur internet et sont donc sans retouches afin d'éviter les biais liés à la manipulation des stimuli.

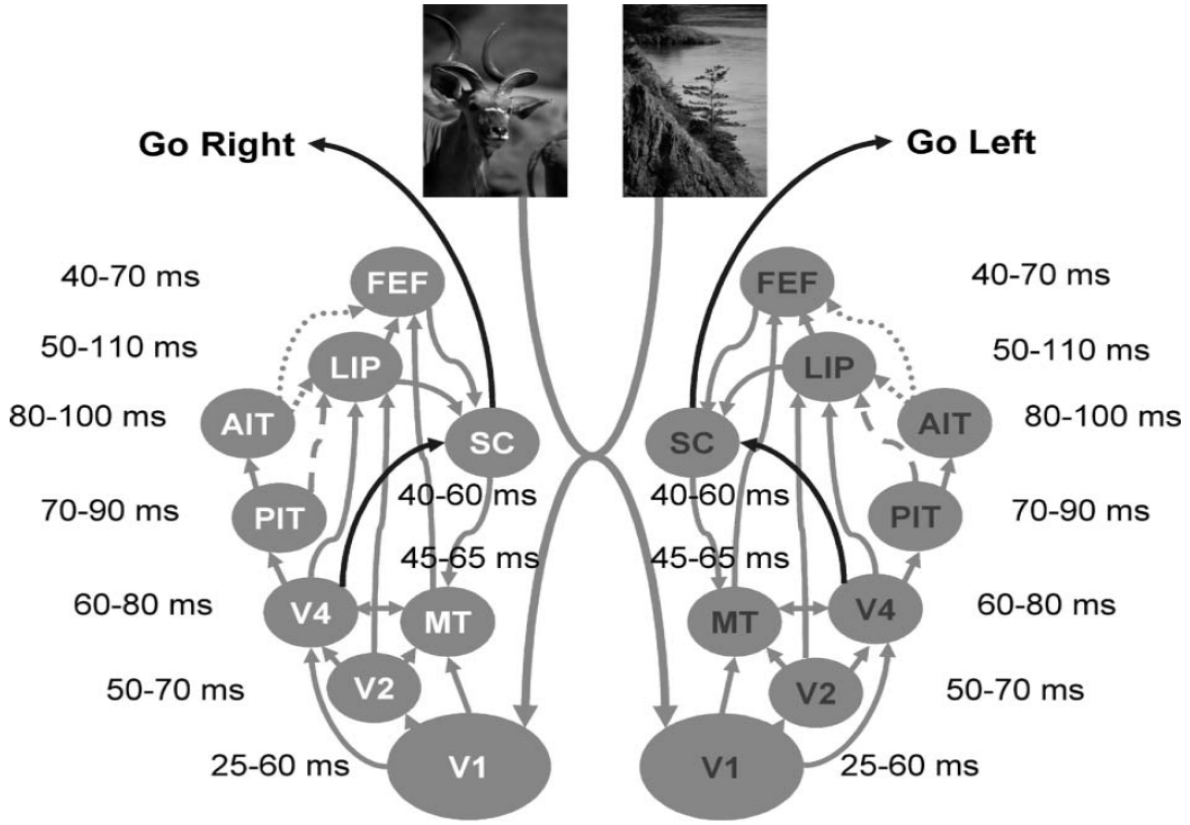

Figure n°36 : Tiré de Kirchner & Thorpe, 2006. Organigramme des connexions anatomiques entre les aires visuelles de la voie ventrale impliquées dans la reconnaissance d'objets et les aires de programmation oculaire impliquées dans le contrôle des mouvements oculaires. L'estimation des latences indiquées sont basées sur des données enregistrées chez le singe, et sont bien sûr supposées plus longues chez l'humain. Le premier nombre correspond à la latence minimale des neurones répondant à un stimulus flashé, le second à la latence moyenne des réponses (Nowak & Bullier, 1997, Thorpe & Fabre-Thorpe, 2001). Les connexions partant du cortex inféro-temporal antérieur et postérieur sont considérées comme insignifiantes voire inexistantes (Bullier, Schall & Morel, 1996, Schall, Morel, King & Bullier, 1995). De par les contraintes temporelles imposées par la réponse en choix forcé saccadique, les auteurs suggèrent que l'information visuelle ascendante pourrait court-circuiter les traitements visuels effectués au niveau de IT.

IV.1 Bilan des résultats et implications pour le modèle

Les tout premiers résultats montrent que, dans un protocole de choix forcé, l'être humain est capable d'effectuer une discrimination catégorielle de contextes efficacement avec un taux de précision moyen d'environ 73%. Cette précision est couplée à des temps de réaction médians extrêmement rapides aux alentours de 200 ms sans qu'aucune des catégories superordonnées de contexte ne soit privilégiée. Si il a déjà été suggéré que les premiers traitements d'encodage sensoriel seraient communs aux tâches de catégorisation et de discrimination catégorielle (Bacon-Mace, Kirchner, Fabre-Thorpe & Thorpe, 2007), il est fort probable que les traitements ultérieurs impliqués dans la tâche de discrimination catégorielle diffèrent de ceux sous-tendant la catégorisation rapide dans le paradigme go/no-go, la présence simultanée de deux images permettant leur comparaison. Ces résultats apportent de nouvelles informations sur les décours temporels des traitements des scènes naturelles qu'il faudra à l'avenir prendre en compte.

Il apparaît qu'une fois de plus, la présence d'objets saillants au sein de la scène rend la tâche de discrimination catégorielle de contexte bien plus difficile à effectuer (61% environ pour les deux types de catégories d'Env.) sans pour autant la ralentir (200 ms environ).
Cependant, de manière étonnante, aucun effet de la congruence du contexte sur la discrimination catégorielle d'objets n'a pu être observé. Comme signalé plus haut, la nature de la tâche est différente. Il ne s'agit pas d'extraire une quantité d'informations diagnostiques suffisante pour répondre (comme on peut supposer que c'est le cas dans le protocole en go/ no-go) mais plutôt d'effectuer une comparaison de la quantité d'informations diagnostiques contenue dans chaque image pour privilégier une des deux scènes comme cible. D'un autre côté, le déclenchement rapide de la saccade oculaire pourrait ne pas laisser suffisamment de temps pour que s'installent les interactions objet/contexte. Ce protocole aurait ainsi permis de déterminer la latence minimale à laquelle traitement de l'objet et traitement de contexte interfère. D'un point de vue neuro-anatomique, si le modèle proposé par Kirchner et Thorpe s'avère correct, l'absence d'effet "congruence contextuelle" dans ce nouveau protocole pourrait suggérer que les interférences objet/ contexte s'effectuent dans des aires visuelles de haut-niveau telles que IT ou peut-être éventuellement V4. L'analyse à venir du point d'impact de la "saccade-réponse" dans l'image corrélée au facteur taille et saillance des objets pourrait nous apporter de précieux indices supplémentaires sur les mécanismes sous-jacents.

Le modèle devra de fait envisager les interactions de congruence contextuelle davantage au niveau d'aires visuelles haut-niveau probablement en IT (éventuellement V4).

2. Discussion des résultats et des modèles déjà proposés

Face à ces hypothèses sur la dynamique et la nature des interactions, il est primordial de revenir sur les modèles et études déjà proposés dans la littérature afin de discuter des divergences entre résultats.

Cependant, avant d'entrer dans la discussion des modèles décrits Partie III Chap. 2, il me semble important de rappeler quelques spécificités de la tâche de catégorisation et des stimuli utilisés lors de nos études sur les interactions objet/contexte.

Outre les contraintes temporelles d'affichage des stimuli, il existe également des contraintes temporelles relatives à la prise de décision. En effet, les sujets, même s'ils disposent dans chaque expérience d'1 seconde pour répondre, sont grandement encouragés à répondre le plus rapidement possible, ce qui limite l'affinement des représentations construites et non abouties sur lesquelles sont basées leurs décisions. D'autant plus qu'une partie du temps de réaction est consacrée à la réalisation de la réponse motrice.

Cette hypothèse m'amène à un deuxième point. Notre tâche de catégorisation ne peut en aucun cas être assimilée à une tâche d'identification ou de reconnaissance d'objet et/ou de contexte. Les identifications complètes de l'objet et du contexte ne sont pas nécessaires pour effectuer la tâche, pas plus qu'un accès au lexique ultérieur à la présentation des stimuli. A mon sens, la catégorisation ne reflète que les premières étapes de traitements de l'identification d'un objet, une catégorisation qui s'affine progressivement jusqu'à préciser l'identité de l'objet. Le seul accès au lexique se fait au moment de l'intégration de la consigne permettant la pré-activation du système pour traiter les informations visuelles caractéristiques à extraire de l'image.

Peut-on alors considérer que la catégorisation est une sorte de détection ? Mon avis sur la question est cette fois ci bien plus mesuré, et toujours d'un point de vue personnel, je ne pourrais y répondre de façon certaine. Pour moi, si l'extraction simple d'une ou plusieurs informations diagnostiques suffit à définir la catégorie cible comme différente de la catégorie distractrice, alors les sujets sont bien impliqués dans une tâche de détection, et le nom donné à notre tâche serait à revoir. Par contre, si la définition de la catégorie cible nécessite une intégration complexe et une organisation spécifique de plusieurs informations diagnostiques, alors la tâche ne peut être assimilée à une simple tâche de détection. Face à la diversité des stimuli utilisés, il est difficile de concevoir que l'extraction de plusieurs traits visuels est suffisante pour effectuer la tâche. Par exemple, parmi les images cibles contenant des

animaux, il était possible de trouver des photographies de mammifères, d'oiseaux, de poissons, de reptiles isolés ou en groupe. Ces animaux diffèrent aussi bien par leurs caractéristiques physiques que par leurs tailles relatives, leurs couleurs, leurs textures ou leurs positions relatives dans l'image. De même les contextes étaient choisis pour être aussi divers que possible. Pour exemple, au sein des « Env. Nat.», on pouvait trouver des scènes de mer, de montagne, de désert, de banquise, de falaise, ou encore des fonds marins, susceptibles qui plus est de contenir des objets manufacturés. La catégorisation de fait reposerait plutôt sur le recoupement d'informations variées que sur leurs simples détections.

Cela précisé, il est alors possible de discuter des modèles déjà proposés dans la littérature. Les résultats issus de nos dernières études permettent de rejeter complètement certains des modèles et mettent en évidence la nécessité d'en compléter certains. Au final, seul le modèle interactif semble prendre en compte l'ensemble de nos résultats.

Les premières études menées par Biederman évaluant l'influence du contexte sur la reconnaissance d'objet avait poussé l'auteur à suggérer un "modèle perceptif" supposant l'existence d'un schéma perceptif de la scène influençant précocement l'analyse perceptive des objets en son sein (Biederman & Ju, 1988, Biederman et al., 1982, Palmer, 1975). A raison, Hollingworth et Henderson préciseront que puisque dans ses expériences, le label de la cible était présenté avant l'apparition de l'image, on ne pouvait considérer que les sujets effectuaient une tâche d'identification, mais plutôt une tâche de détection ou de reconnaissance (Hollingworth & Henderson, 1998). Dans une réelle tâche d'identification au cours de laquelle les sujets devaient tester la correspondance entre un objet indicé au sein de la scène et des descriptions verbales de deux objets présentés après la scène, Hollingworth et al. n'ont pu mettre en évidence une influence du contexte sur l'identification de l'objet et en ont conclu à l'absence d'interactions entre objet et contexte. De ce constat a émergé l'idée du "modèle d'isolation fonctionnelle". Par ce changement de protocole, ces chercheurs se sont effectivement affranchis des pré-activations spécifiques à une tâche de détection ou de reconnaissance. Mais ils ont en contrepartie grandement alourdi les étapes cognitives nécessaires à la prise de décision. En effet, la présentation des descriptions verbales suite au stimulus test implique de comprendre la signification des mots, puis d'accéder au lexique afin de récupérer les représentations visuelles respectives de chaque objet décrit. De fait, ces traitements cognitifs supplémentaires sont grandement susceptibles de « noyer » des effets de contextes moins apparents. Au vu de nos résultats, on ne peut exclure l'hypothèse d'une

influence précoce du contexte. Si on suppose que les traitements catégoriels sont impliqués dans l'identification de l'objet, le modèle d'isolation fonctionnelle semble alors difficilement envisageable. Cependant, on ne peut considérer pour autant que le modèle perceptif soit complet, voir même complètement valide en ce qui concerne la catégorisation. Tout d'abord, il ne prend pas en compte l'influence de l'objet incongruent sur la catégorisation du contexte. De plus, il implique un accès au schéma perceptif de la scène qui influencera par la suite l'analyse perceptive des objets. Si ces étapes supposent l'existence de traitements descendants, elles pourraient bien ne pas avoir le temps de se manifester dans nos expériences où les contraintes temporelles sont conséquentes. Le modèle perceptif ne peut donc être appliqué à nos résultats que dans l'unique condition où le schéma perceptif serait construit de manière ascendante en parallèle avec l'intégration des informations propres aux objets. L'architecture triadique proposée par Rensink (Rensink, 2000) rencontre de fait le même problème : un accès au schéma de la scène préalable à l'identification des objets.

Le modèle d'amorçage quant à lui suppose l'influence d'un schéma mnésique de la scène sur la mise en correspondance entre le percept de l'objet et son concept. La première question que l'on est en droit de se poser dans notre tâche de catégorisation concerne la nécessité de faire appel à un concept. De plus, les interactions physiques précoces révélées dans nos expériences révèlent dans tous les cas des lacunes au sein de ce modèle d'amorçage. Il est possible que des interactions haut-niveau aient lieu, mais elles sont forcément ultérieures aux interactions perceptives de plus bas-niveau.

Le modèle de Bar est communément considéré comme un modèle d'amorçage. Effectivement, dans son modèle, Bar décrit une représentation grossière du contexte intégrée de manière progressive et en parallèle de l'intégration des informations visuelles relatives à l'objet (Bar, 2004). Une fois les représentations de l'objet et du contexte assez abouties, il y aurait interaction entre ces représentations dans le but d'affiner l'identité de l'objet. Nos résultats expérimentaux pourraient facilement rendre compte de ce modèle si on y ajoutait des interactions plus précoces entre les traitements de l'objet et du contexte. Néanmoins, l'hypothèse de représentations grossières de l'objet et du contexte rapidement construites sur la base des informations contenues dans les basses fréquences spatiales, via le système magnocellulaire rapide est en complète adéquation avec les idées que nous souhaitons défendre. Finalement, le modèle interactif proposé par Davenport est le plus fidèle à nos résultats. Ce modèle suppose des traitements relatifs aux objets et au contexte opérant continuellement de manière interactive et se contraignant mutuellement.

3. Modèle de catégorisation rapide des scènes naturelles

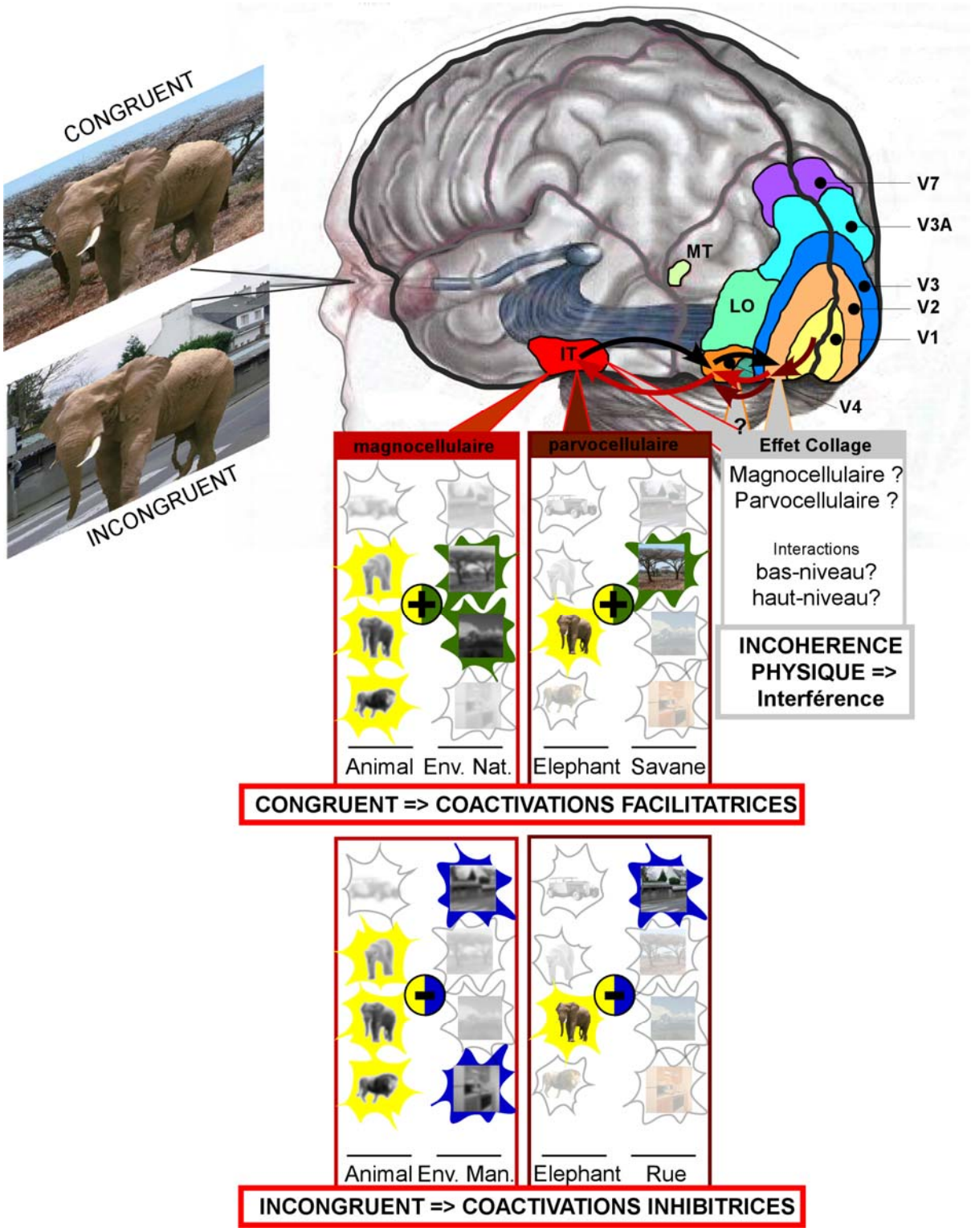

Figure n°37 : Modèle de catégorisation rapide des scènes naturelles : interactions bidirectionnelles entre l'objet et son contexte lors de catégorisations basiques et superordonnées d'objets et de contextes (voir texte)

Dans un système visuel préactivé en fonction de la tâche à accomplir et de la catégorie cible (dans l'illustration, animal), l'information visuelle purement ascendante extraite des stimuli va transiter le long des aires visuelles de la voie ventrale impliquées dans la reconnaissance d'objet (V1-V2-V4-IT…). L'ensemble de l'information ne va cependant pas être traitée à la même vitesse. L'information magnocellulaire (en rouge foncé) va être plus rapidement traitée que l'information parvocellulaire (marron foncé), mais ne suffira à construire qu'une représentation grossière de l'objet et de la scène dans son ensemble (vs. plus fine).

Lorsqu'une scène naturelle représentant un éléphant dans la savane va être présentée au sujet, la première vague de traitements des informations (magnocellulaires) va se propager vers l'avant jusqu'en IT et activer les populations neuronales sélectives aux représentations grossières d'éléments de la scène. On peut imaginer dans ce sens qu'une grande partie des populations sélectives aux différentes catégories d'Env. Nat. ainsi qu'aux différentes catégories d'animaux va être activée. De par les régularités de notre environnement, la probabilité de co-activation de ces populations est importante (les animaux se trouvent souvent dans la nature) et va entrainer une modulation de la force de leurs connexions fonctionnelles (par modulation des poids synaptiques par exemple). La mise en place de telles interactions facilitatrices résulterait en une prise de décision relativement rapide que ce soit dans une tâche de catégorisation animal/non-animal ou dans une tâche Env. Nat/non-Env. Nat. Les populations neuronales sélectives aux catégories d'Env. Man. et d'objets manufacturés seront quant à elle restées inactives. Toujours sur la base d'une vague d'information magnocellulaire, la présentation d'une scène cette fois-ci incongruente (un éléphant dans la rue) va mener à l'activation des populations neuronales sélectives aux catégories d'Env. Man. et d'animaux peu habituées à décharger ensemble, menant ainsi à des interférences conflictuelles. La prise de décision sur une tâche de catégorisation animal/non-animal (mais aussi Env.Man./non-Env. Man.) s'en trouve retardée.

A des latences plus tardives, les informations parvocellulaires sous-tendant une représentation plus fine de l'image vont parvenir en IT. Dès lors, les informations diagnostiques de l'image étant plus affinées, les populations neuronales activées vont être plus sélectives. En fait, la majorité des populations neuronales activées coderont pour la catégorie basique de l'animal (ici éléphant) ou de la scène (savane ou scène de rue). Les populations codant pour les autres catégories basiques d'animaux (autres qu'éléphant) ou les autres catégories basiques d'Env. seront peu ou pas du tout activées. Selon la congruence de la scène

présentée, il pourra de nouveau y avoir coactivation facilitatrice (entre populations neuronales sélectives aux éléphants et à la savane) ou coactivation inhibitrice (entre éléphant et scène de rue). En fait, dans une tâche de catégorisation basique, nous devrions pouvoir observer un retard (peut être plus grand encore) dans la prise de décision lorsque l'objet est incongruent avec son contexte.

A ce modèle s'ajoute des interactions de nature plus physique liées à la manipulation des stimuli. Etant donné le peu d'indices que nous avons sur les décours temporels liées à ces interactions, les aires visuelles impliquées dans ce phénomène restent relativement mal définies. On peut envisager aussi bien un ralentissement de l'encodage sensoriel (en V2-V4) qu'un ralentissement des activations neuronales à plus haut-niveau (en IT).

Enfin, il est important de tenir compte des boucles de rétro-action itératives à chaque étape de traitement et fonctionnant de manière continue au sein de la voie ventrale. Ces traitements en feed-back seraient cependant bien plus importants dans des tâches temporellement moins limitatives que notre tâche de catégorisation visuelle rapide.

4. Etudes en cours et perspectives

Bien évidemment, ce modèle suggérant diverses hypothèses sur les traitements sous-jacents à la catégorisation nécessite d'être davantage argumenté et appelle à un grand nombre d'études. Trois études en cours viennent apporter de nouvelles réponses (et poser de nouvelles questions) à propos de la catégorisation rapide des scènes naturelles.

Etudes en cours

Catégorisations en parallèle : la préactivation du système visuel peut-elle être effective sur deux catégories différentes d'objets?

La capacité du cerveau humain à catégoriser en parallèle deux scènes naturelles différentes (Rousselet et al., 2002), ou encore à effectuer en parallèle deux tâches dont une sur des scènes naturelles en périphérie ne fait désormais plus aucun doute (VanRullen et al., 2004). Nous avons de fait voulu tester la capacité de sujets humains non seulement à pré-activer des populations de neurones sélectives à des traits diagnostiques appartenant à deux catégories différentes, mais également leur capacité à répondre uniquement lorsque deux objets de catégories différentes étaient présents au sein d'une même image. Présenté autrement notre but était de déterminer si des sujets humains sont capables d'effectuer des catégorisations impliquant des opérations de type "OU" (non exclusif) et "ET" entre diverses catégories d'objets. Pour cela, nous avons utilisé 4 types de stimuli différents récupérés sur internet: des scènes naturelles chromatiques et non manipulées contenant soit un animal, soit un véhicule, soit les deux, soit aucun des deux. Les stimuli étaient comme toujours brièvement présentés (26 ms) et les sujets devaient répondre aussi rapidement et précisément que possible selon les consignes de 4 tâches successives différentes : répondre quand vous apercevez un animal dans l'image, quand vous apercevez un véhicule, quand vous apercevez l'un OU l'autre (non exclusif), ou uniquement quand vous apercevez un animal ET un véhicule au sein de la même image. En plus des performances comportementales (réponses correctes et temps de réaction), nous avons pratiqué des enregistrements EEG malheureusement encore non analysés à ce jour.

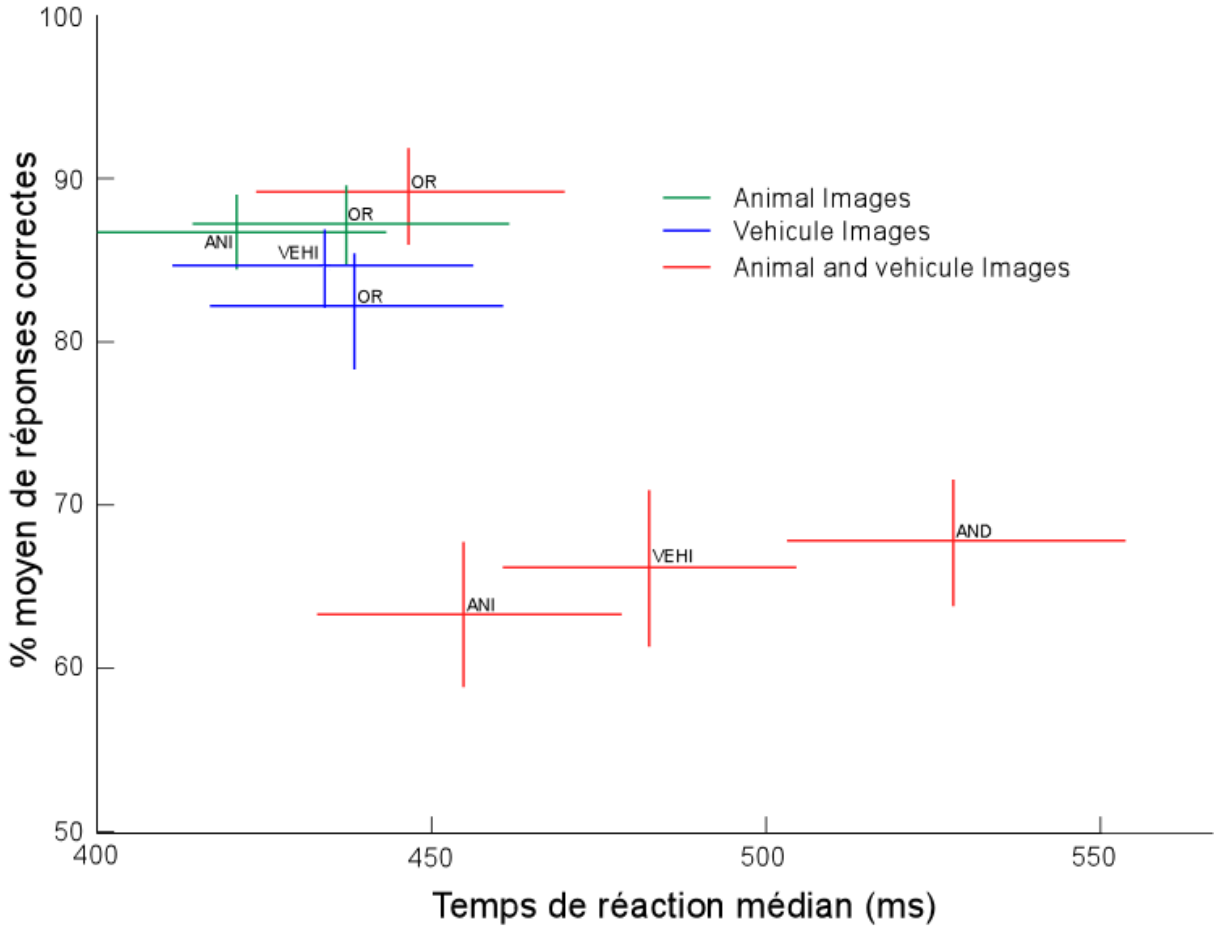

Figure n°38 : Réponses comportementales dans les tâches de catégorisation animal, véhicule, animal OU véhicule (non-exclusif), animal ET véhicule. La précision moyenne dans chaque tâche et pour chaque type d'images est indiquée en fonction du temps de réaction moyen enregistré dans les mêmes conditions. Les axes horizontaux et verticaux correspondent à l'intervalle de confiance (95%) « bootstrappé » sur 1000 tirages. Tandis que nous sommes capables de pré-activer deux catégories d'objets en parallèle, la prise de décision sur la présence simultanée de deux objets dans l'image nécessite des traitements cognitifs supplémentaires.

Les premiers résultats comportementaux sont assez prometteurs :

Tout d'abord, les sujets impliqués dans la tâche OU réalisent des performances globalement équivalentes aux performances obtenues dans une tâche de catégorisation simple animal/non-animal ou véhicule/non-véhicule, aussi bien en termes de précision (environ 85 %) qu'en termes de rapidité (moins de 450 ms). Nous mettons en évidence par ces résultats le fait que des pré-activations simultanées de catégories superordonnées différentes d'objets sont tout à fait possibles. De fait, dans une tâche de catégorisation simple animal/non-animal, il est fort probable que les informations diagnostiques recherchées ne soient pas qu'un type d'informations commun à toutes les catégories basiques d'animaux, mais un ensemble de conjonctions d'informations diagnostiques organisée de manière spécifique et se chevauchant entre catégorie (mammifères, oiseaux, poissons, reptiles, insectes…).

IV.4 Etudes en cours et perspectives

Nous avons également confirmé un effet de la taille des objets à catégoriser. En effet, parmi les images contenant un animal ET un véhicule, il s'avère qu'une partie des scènes naturelles contenait un animal de grande surface et un véhicule de petite surface, ou inversement. Les images ET contenant un animal de petite surface étaient moins bien catégorisées que celles contenant un animal de grande surface, même dans la tâche de catégorisation simple animal/non animal. Il en était de même pour les véhicules. Cet effet de taille semble directement impliqué dans la baisse de performances sur les images ET quelle que soit la tâche (et donc également dans la tâche de catégorisation double), une partie des cibles étant plus difficile à catégoriser.

Lorsque l'on considère uniquement ces scènes contenant un animal et un véhicule, ce biais de taille explique sans soute l'altération de la précision des sujets humains dans une tâche simple de catégorisation "animal" ou de catégorisation "véhicule". En revanche il est à noter que les sujets humains sont capable d'atteindre la même précision en catégorisation ET mais qu'ils prennent leurs décisions plus lentement. En d'autres termes, il faut plus de temps pour s'assurer de la présence de deux objets de catégories différentes au sein d'une même image que d'un seul objet d'une des deux catégories. Etant donné que l'ensemble de la scène est traitée de manière parallèle, nous pouvons supposer que le délai observé dans la tâche ET est lié à une prise de décision soumise à davantage de contraintes cognitives et impliquant des traitements de haut-niveau supplémentaires. Nous espérons extraire des évidences supplémentaires sur cette hypothèse lors d'analyses comportementales plus poussées et de l'analyse des enregistrements EEG.

La ségrégation de l'objet devient-elle cruciale lorsque la décision est basée sur un traitement plus détaillé de l'objet?

Dans l'étude comportementale décrite dans l'article n°4, nous avons démontré qu'il n'était pas plus facile ni réellement plus rapide de catégoriser des animaux isolés de leur contexte que des animaux en contexte remettant ainsi en question l'intérêt d'une ségrégation précoce de l'objet. Cependant, il est envisageable que la catégorisation superordonnée d'objets puisse se baser sur des informations diagnostiques indépendantes des contours de l'objet. Qu'en serait-il dans le cadre d'une catégorisation basique d'objets utilisant des distracteurs appartenant à la même catégorie superordonnée ? Une étude récemment initiée

vise donc à comparer les performances des sujets dans une tâche de catégorisation chien/non-chien isolés et en contexte dans laquelle les distracteurs appartiendraient à la catégorie superordonnée des animaux. Des performances similaires dans les deux conditions mettront à mal les théories qui considèrent comme cruciale une étape de ségrégation de l'objet préalable à leur catégorisation fine. Au contraire, de meilleures performances lorsque les chiens seront isolés suggèreront une importance croissante de la ségrégation avec l'affinement des traitements menant à son identification.

A noter qu'il serait tout aussi intéressant de tester selon un paradigme similaire si l'effet de congruence du contexte croit au fur et à mesure que des catégorisations d'objets plus fines sont demandées.

Perspectives de recherche

Catégorisation du contexte périphérique

Il est intéressant de garder en tête les conditions écologiques de la perception. Dans la vie de tous les jours, notre regard a tendance à se porter plutôt sur les objets d'intérêts que sur leur contexte. De fait, les objets pertinents ont plutôt tendance à se retrouver traités dans la région fovéale suite à une saccade oculaire d'orientation tandis que les contextes seraient majoritairement traités par la rétine périphérique. Cette régularité de la perception semble avoir des conséquences sur les traitements perceptifs, au moins de haut-niveau. En effet, pour rappel, « la représentation complexe d'objets nécessitant une prise d'information détaillée en région fovéale (tel que les visages) serait plutôt effectué par des cellules du LOC fortement associée aux traitements de l'information centrale tandis que la représentation des objets considérés comme plutôt périphériques (bâtiments, objets contextuels) serait fortement associée aux traitements de l'information périphérique (Levy et al., 2001), Part I, Chap 1.2.5). Il serait ainsi intéressant de tester les capacités humaines dans une tâche de catégorisation de contexte et d'objet en périphérie. Puisque notre système visuel est plus habitué à traiter les informations contextuelles en vision périphérique, les performances comportementales enregistrées dans des tâches de catégorisation de contexte périphérique devraient être supérieures à celles enregistrées par Thorpe et al. (Thorpe et al., 2001) lors d'une tâche de

catégorisation d'objets en périphérie. Mais surtout l'influence du contexte sur le traitement de l'objet devrait être plus importante pour des scènes présentées en périphérie.

Interférence de l'objet saillant sur la catégorisation de contexte

S'il semble clair qu'il existe une influence des objets saillants (d'autant plus qu'ils sont incongruents) sur les traitements sous-jacents à la catégorisation du contexte, la nature de ces interactions jusque là peu étudiée reste encore assez incertaine. Une part de cette interférence pourrait s'expliquer par la surface occupée par l'objet dans le contexte. En effet, puisque l'objet masque le contexte, la présence d'un objet induirait une baisse de la quantité d'informations disponibles en provenance du contexte, baisse qui serait fonction de la surface couverte par l'objet. Afin de tester cette hypothèse, il pourrait être intéressant d'évaluer les performances de sujets humains dans une tâche de catégorisation de contexte en utilisant des scènes naturelles au sein desquelles se trouverait un objet congruent, un objet incongruent, ou encore une simple zone noire de forme aléatoire, mais de position et de taille identiques aux objets congruents et incongruents.

Etude de la perception visuelle en condition plus écologique : les scènes dynamiques

Les expériences mentionnées dans la revue des interactions objet/contexte, ainsi que mes résultats présentés ci-dessus, ont été obtenus en manipulant des scènes statiques (images ou photographies) affichées brièvement sur un écran. Aucune donnée expérimentale n'est disponible en situation plus écologique de scènes dynamiques, dans lesquelles objets ou contextes apparaîtraient comme dans le champ d'une caméra effectuant une prise de vue continue sur plusieurs secondes.

Dans le cas de scènes dynamiques, on peut s'attendre à la mise en jeu d'autres mécanismes neuropsychologiques. La catégorie du contexte de la scène étant alors stable, les représentations des objets congruents ou compatibles seraient pré-activés ou « attendus » par le système visuel tandis que les incongruences (détectées par les mécanismes ci-dessus) feraient l'objet de capture attentionnelle ou au contraire de phénomènes d'aveuglement (type *change blindness*). Ce sont ces phénomènes considérés encore comme marginaux et les mécanismes neuro-anatomiques sous-jacents que je voudrais explorer maintenant dans des

environnements dynamiques, étant convaincu qu'ils prennent tout leur sens dans des situations plus naturelles. Dans ce but, deux outils me semblent essentiels : l'IRMf exploité via le protocole d'adaptation (Grill-Spector & Malach, 2001) et l'environnement 3D virtuel modulable.

Le paradigme d'adaptation en IRMf déjà utilisé avec des images statiques, nous permettrait de caractériser les aires cérébrales impliquées dans les traitements cognitifs d'environnements dynamiques. Dans un premier temps, on enregistrerait le signal BOLD généré par des populations neuronales adaptées (dont la sensibilité s'est réduite) par la présentation répétée de contextes de même catégorie (« Manmade » ou « Natural »). Dans un second temps, la catégorie des scènes varierait et la récupération de l'adaptation serait évaluée. Si le signal restait adapté, les populations neuronales en question ne seraient pas sélectives à la catégorie. En calculant le contraste de signal entre ces deux conditions, on localiserait les aires cérébrales sélectives aux deux catégories de contexte (1) tandis que le sujet navigue dans un environnement continu ou (2) lorsqu'on lui présente des scènes statiques tirées de l'environnement 3D. Par comparaison des cartographies cérébrales, nous pourrions préciser parmi les aires sélectives aux contextes « Manmade » et « Natural » celles qui sont spécifiquement sensibles à la nature dynamique du contexte.

Dans une étude ultérieure, je pourrai m'appuyer sur la flexibilité du logiciel d'environnement 3D proposé dans le laboratoire d'Aude Oliva permettant l'apparition d'objets lors de la navigation du sujet pour préciser les interactions objet/contexte en environnement dynamique. Rensink a mis en évidence le phénomène de « change blindness » (Rensink, 2002) révélant ainsi notre incapacité à détecter l'apparition d'objet au sein d'images statiques. Mais qu'en est-il dans un monde dynamique ? Dans quelles mesures les propriétés de cohérence sémantique et de congruence physique influent-elles sur notre capacité à détecter le changement en contexte dynamique ? De part la stabilité du monde 3D utilisé, les représentations d'objets **sémantiquement** cohérents avec l'environnement devraient davantage être pré-activées, et de ce fait capturer l'attention dans une moindre mesure que les

objets non-cohérents. Selon le modèle computationnel proposé par Oliva et Torralba (Oliva & Torralba, 2001), les objets **physiquement** non-congruents devraient davantage perturber notre système perceptif que les objets congruents. Je souhaiterais de ce fait étudier la modulation de l'activité des aires cérébrales sélectives au contexte et la capacité des sujets à détecter les nouveaux objets en fonction de la présence d'objets, saillants ou non, congruents ou non, consciemment perçus ou non.

La performance des sujets à détecter de nouveaux objets apparaissant soudainement dans l'environnement 3D couplée à l'enregistrement de l'activité cérébrale en IRMf et des mouvements oculaires devraient permettre préciser comment la sémantique et les traits physiques d'un objet dans la scène sous-tendent une capture attentionnelle.

Après obtention de mon doctorat, ce projet post-doctoral sera financé par la Fondation Fyssen (http://www.fondation-fyssen.org/) et devrait se dérouler sous la direction d'Aude Oliva dans le département « Brain and Cognitive Science » du MIT. L'expertise de son équipe dans le domaine de la modélisation est en effet un atout majeur pour la réussite du projet. De plus, l'accès à un scanner IRMf à haut champ avec oculomètre intégré proposé par son laboratoire est tout aussi essentiel pour caractériser les aires cérébrales impliquées dans le traitement du contexte dynamique tout en contrôlant rigoureusement les mouvements des yeux.

Ce projet post-doctoral mené dans un laboratoire parfaitement adapté visera donc une meilleure compréhension de la voie visuelle perceptivo-décisionnelle dans sa globalité et d'une composante attentionnelle encore bien méconnue en condition plus écologique.

PARTIE V

BIBLIOGRAPHIE

Adelson, E.H., & Bergen, J.R. (1991). the Plenoptic function and the elements of early vision. In: M. Landy, & J.A. Movshon (Eds.), *Computational Models of Visual Processing* (pp. 3-20). Cambridge: MIT Press.

Afraz, S.R., Kiani, R., & Esteky, H. (2006). Microstimulation of inferotemporal cortex influences face categorization. *Nature, 442* (7103), 692-695.

Aguirre, G.K., Detre, J.A., Alsop, D.C., & D'Esposito, M. (1996). The parahippocampus subserves topographical learning in man. *Cereb Cortex, 6* (6), 823-829.

Aguirre, G.K., Zarahn, E., & D'Esposito, M. (1998). An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. *Neuron, 21* (2), 373-383.

Albright, T.D., Desimone, R., & Gross, C.G. (1984). Columnar organization of directionally selective cells in visual area MT of the macaque. *J Neurophysiol, 51* (1), 16-31.

Allison, J.D., Melzer, P., Ding, Y., Bonds, A.B., & Casagrande, V.A. (2000). Differential contributions of magnocellular and parvocellular pathways to the contrast response of neurons in bush baby primary visual cortex (V1). *Vis Neurosci, 17* (1), 71-76.

Allman, J., Miezin, F., & McGuinness, E. (1985). Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. *Annu Rev Neurosci, 8*, 407-430.

Altmann, C.F., Deubelius, A., & Kourtzi, Z. (2004). Shape saliency modulates contextual processing in the human lateral occipital complex. *J Cogn Neurosci, 16* (5), 794-804.

Aminoff, E., Gronau, N., & Bar, M. (2007). The parahippocampal cortex mediates spatial and nonspatial associations. *Cereb Cortex, 17* (7), 1493-1503.

Andersen, R.A., Bracewell, R.M., Barash, S., Gnadt, J.W., & Fogassi, L. (1990). Eye position effects on visual, memory, and saccade-related activity in areas LIP and 7a of macaque. *J Neurosci, 10* (4), 1176-1196.

Antes, J.R. (1974). The time course of picture viewing. *J Exp Psychol, 103* (1), 62-70.

Antes, J.R. (1977). Recognizing and localizing features in brief picture presentations. *Memory and Cognition, 5*, 155-161.

Antes, J.R., Penland, J.G., & Metzger, R.L. (1981). Processing global information in briefly presented pictures. *Psychol Res, 43* (3), 277-292.

Ariely, D. (2001). Seeing sets: representation by statistical properties. *Psychol Sci, 12* (2), 157-162.

Atick, J.J. (1992). Could information theory provide an ecological theory of sensory processing? *Network: Computation in neural systems, 3* (2), 213-251.

Attneave, F. (1954). Some informational aspects of visual perception. Psychol Rev, 61 (3), 183-193.

Auckland, M.E., Cave, K.R., & Donnelly, N. (2007). Nontarget objects can influence perceptual processes during object recognition. *Psychon Bull Rev, 14* (2), 332-337.

Avidan, G., Harel, M., Hendler, T., Ben-Bashat, D., Zohary, E., & Malach, R. (2002). Contrast sensitivity in human visual areas and its relationship to object recognition. *J Neurophysiol, 87* (6), 3102-3116.

Bacon-Mace, N., Kirchner, H., Fabre-Thorpe, M., & Thorpe, S.J. (2007). Effects of task requirements on rapid natural scene processing: From common sensory encoding to distinct decisional mechanisms. *J Exp Psychol Hum Percept Perform, 33* (5), 1013-1026.

Baddeley, R.J. (1997). The correlational structure of natural images and the calibration of spatial representations. *Cogn Sci, 21* (3), 351-372.

Baker, C.I., Behrmann, M., & Olson, C.R. (2002). Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nat Neurosci, 5* (11), 1210-1216.

Bar, M. (2004). Visual objects in context. *Nat Rev Neurosci, 5* (8), 617-629.

Bar, M. (2005). Top-Down Facilitation of Visual Object Recognition. In: L. Itti, G. Rees, & J.K. Tsotsos (Eds.), *Neurobiology of Attention* (pp. 140-145): Elsevier.

Bar, M., & Aminoff, E. (2003). Cortical analysis of visual context. *Neuron, 38* (2), 347-358.

Bar, M., Aminoff, E., & Ishai, A. (2008). Famous faces activate contextual associations in the parahippocampal cortex. *Cereb Cortex, 18* (6), 1233-1238.

Bar, M., & Ullman, S. (1996). Spatial context in recognition. *Perception, 25*, 343-352.

Barlow, H.B. (1953). Summation and inhibition in the frog's retina. *J Physiol, 119* (1), 69-88.

Barlow, H.B. (1961). Possible principles underlying the transformations of sensory messages. In: W.A. Rosenblith (Ed.) *Sensory Communication* (pp. 217-236): John Wiley & Sons.

Barlow, H.B. (1989). Unsupervised learning. *Neural Comput, 1*, 295-311.

Barrow, H.G., & Tenenbaum, J.M. (1981). Computational vision. *Proceedings of the IEEE, 69* (5), 572 -595.

Bartels, A., & Zeki, S. (2000). The architecture of the colour centre in the human visual brain: new results and a review. *Eur J Neurosci, 12* (1), 172-193.

Bear, M.F., Connors, B.W., & Paradiso, M.A. (2002). Neurosciences : A la découverte du cerveau (

Beauchamp, M.S., Yasar, N.E., Kishan, N., & Ro, T. (2007). Human MST but not MT responds to tactile stimulation. *J Neurosci, 27* (31), 8261-8267.

Bell, A.J., & Sejnowski, T.J. (1997). The "independent components" of natural scenes are edge filters. *Vision Res, 37* (23), 3327-3338.

Biederman, I. (1972). Perceiving real-world scenes. *Science, 177* (4043), 77-80.

Biederman, I. (1981). On the semantics of a glance at a scene. In: M.K.J.R. Pomerantz (Ed.) *Perceptual Organization* (pp. 213 -254). Hillsdale: Lawrence Erlbaum.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol Rev, 94* (2), 115-147.

Biederman, I. (1995). Visual object recognition. In: S.F.K.D.N. Osherson (Ed.) *An invitation to cognitive science* (pp. 121-165): MIT Press.

Biederman, I., & Bar, M. (1999). One-shot viewpoint invariance in matching novel objects. *Vision Res, 39* (17), 2885-2899.

Biederman, I., Blickle, T.W., Teitelbaum, R.C., & Klatsky, G.J. (1988). Object search in nonscene displays. *J Exp Psychol Learn Mem Cogn, 14* (3), 456-467.

Biederman, I., Glass, A.L., & Stacy, E.W., Jr. (1973). Searching for objects in real-world sciences. *J Exp Psychol, 97* (1), 22-27.

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognit Psychol, 20* (1), 38-64.

Biederman, I., Mezzanotte, R.J., & Rabinowitz, J.C. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cognit Psychol, 14* (2), 143-177.

Biederman, I., Rabinowitz, J.C., Glass, A.L., & Stacy, E.W., Jr. (1974). On the information extracted from a glance at a scene. *J Exp Psychol, 103* (3), 597-600.

Biederman, I., & Shiffrar, M.M. (1987). Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task. *J Exp Psychol: Learning, Memory and Cognition, 13* (4), 640-645.

Blakemore, C., & Campbell, F.W. (1969). On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *J Physiol, 203* (1), 237-260.

Booth, M.C., & Rolls, E.T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cereb Cortex, 8* (6), 510-523.

Boussaoud, D., Desimone, R., & Ungerleider, L.G. (1991). Visual topography of area TEO in the macaque. *J Comp Neurol, 306* (4), 554-575.

Bowers, J.S., & Jones, K.W. (2008). Detecting objects is easier than categorizing them. *Q J Exp Psychol (Colchester), 61* (4), 552-557.

Boyce, S.J., Pollatsek, A., & Rayner, K. (1989). Effect of background information on object identification. *J Exp Psychol Hum Percept Perform, 15* (3), 556-566.

Bremmer, F., Distler, C., & Hoffmann, K.P. (1997). Eye position effects in monkey cortex. II. Pursuit- and fixation- related activity in posterior parietal areas LIP and 7A. *J Neurophysiol, 77* (2), 962-977.

Bringuier, V., Chavane, F., Glaeser, L., & Fregnac, Y. (1999). Horizontal propagation of visual activity in the synaptic integration field of area 17 neurons. *Science, 283* (5402), 695-699.

Britten, K.H., Newsome, W.T., Shadlen, M.N., Celebrini, S., & Movshon, J.A. (1996). A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Vis Neurosci, 13* (1), 87-100.

Brockmole, J.R., Castelhano, M.S., & Henderson, J.M. (2006). Contextual cueing in naturalistic scenes: Global and local contexts. *J Exp Psychol Learn Mem Cogn, 32* (4), 699-706.

Bruce, C., Desimone, R., & Gross, C.G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *J Neurophysiol, 46* (2), 369-384.

Bullier, J. (2001a). Feedback connections and conscious vision. *Trends Cogn Sci, 5* (9), 369-370.

Bullier, J. (2001b). Integrated model of visual processing. *Brain Res Brain Res Rev, 36* (2-3), 96-107.

Bullier, J., & Nowak, L.G. (1995). Parallel versus serial processing: new vistas on the distributed organization of the visual system. *Curr Opin Neurobiol, 5* (4), 497-503.

Bullier, J., Schall, J.D., & Morel, A. (1996). Functional streams in occipito-frontal connections in the monkey. *Behav Brain Res, 76* (1-2), 89-97.

Bulthoff, H.H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc Natl Acad Sci U S A, 89* (1), 60-64.

Burwell, R.D., Saddoris, M.P., Bucci, D.J., & Wiig, K.A. (2004). Corticohippocampal contributions to spatial and contextual learning. *J Neurosci, 24* (15), 3826-3836.

Bussey, T.J., & Saksida, L.M. (2002). The organization of visual object representations: a connectionist model of effects of lesions in perirhinal cortex. *Eur J Neurosci, 15* (2), 355-364.

Campbell, F.W., Nachmias, J., & Jukes, J. (1970). Spatial-frequency discrimination in human vision. *J Opt Soc Am, 60* (4), 555-559.

Campbell, F.W., & Robson, J.G. (1968). Application of Fourier analysis to the visibility of gratings. *J Physiol, 197* (3), 551-566.

Chaumon, M., Drouet, V., & Tallon-Baudry, C. (2008). Unconscious associative memory affects visual processing before 100 ms. *J Vis, 8* (3), 10 11-10.

Chee, M.W., Goh, J.O., Venkatraman, V., Tan, J.C., Gutchess, A., Sutton, B., Hebrank, A., Leshikar, E., & Park, D. (2006). Age-related changes in object processing and contextual binding revealed using fMR adaptation. *J Cogn Neurosci, 18* (4), 495-507.

Chun, M.M. (2000). Contextual cueing of visual attention. *Trends Cogn Sci, 4* (5), 170-178.

Chun, M.M., & Jiang, Y. (1998). Contextual cueing: implicit learning and memory of visual context guides spatial attention. *Cognit Psychol, 36* (1), 28-71.

Cooper, B.G., & Mizumori, S.J. (2001). Temporary inactivation of the retrosplenial cortex causes a transient reorganization of spatial coding in the hippocampus. *J Neurosci, 21* (11), 3986-4001.

Cragg, B.G. (1969). The topography of the afferent projections in the circumstriate visual cortex of the monkey studied by the Nauta method. *Vision Res, 9* (7), 733-747.

Crick, F. (1984). Function of the thalamic reticular complex: the searchlight hypothesis. *Proc Nat Acad Sci, 81* (14), 4586-4590.

Dacey, D.M., & Petersen, M.R. (1992). Dendritic field size and morphology of midget and parasol ganglion cells of the human retina. *Proc Natl Acad Sci U S A, 89* (20), 9666-9670.

Dale, A.M., Liu, A.K., Fischl, B.R., Buckner, R.L., Belliveau, J.W., Lewine, J.D., & Halgren, E. (2000). Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron, 26* (1), 55-67.

Daugman, J.G. (1984). Spatial visual channels in the Fourier plane. *Vision Res, 24* (9), 891-910.

Daugman, J.G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J Opt Soc Am A, 2* (7), 1160-1169.

Davenport, J.L. (2007). Consistency effects between objects in scenes. *Mem Cognit, 35* (3), 393-401.

Davenport, J.L., & Potter, M.C. (2004). Scene consistency in object and background perception. *Psychol Sci, 15* (8), 559-564.

Davidoff, J.B., & Ostergaard, A.L. (1988). The role of colour in categorial judgements. *Q J Exp Psychol A, 40* (3), 533-544.

De Graef, P. (1998). Prefixational object perception in scenes: Objects popping out of schemas. In: G. Underwood (Ed.) *Eye guidance in reading and scene perception* (pp. 313-336). Oxford, UK: Elsevier.

De Graef, P., Christiaens, D., & d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychol Res, 52* (4), 317-329.

De Graef, P., De Troy, A., & D'Ydewalle, G. (1992). Local and global contextual constraints on the identification of objects in scenes. *Can J Psychol, 46* (3), 489-508.

De Valois, R.L., & De Valois, K.K. (1988). Spatial Vision. (Oxford: Oxford University Press.

Deco, G., & Rolls, E.T. (2004). A Neurodynamical cortical model of visual attention and invariant object recognition. *Vision Res, 44* (6), 621-642.

Delorme, A. (2000). Traitement visuel rapide des scènes naturelles chez le singe, l'homme et la machine. (p. 293): UPS.

Delorme, A., Richard, G., & Fabre-Thorpe, M. (2000). Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans. *Vision Res, 40* (16), 2187-2200.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu Rev Neurosci, 18*, 193-222.

Desimone, R., Fleming, J., & Gross, C.G. (1980). Prestriate afferents to inferior temporal cortex: an HRP study. *Brain Res, 184* (1), 41-55.

Desimone, R., & Gross, C.G. (1979). Visual areas in the temporal cortex of the macaque. *Brain Res, 178* (2-3), 363-380.

Desimone, R., Schein, S.J., Moran, J., & Ungerleider, L.G. (1985). Contour, color and shape analysis beyond the striate cortex. *Vision Res, 25* (3), 441-452.

DiCarlo, J.J., & Maunsell, J.H. (2003). Anterior Inferotemporal Neurons of Monkeys Engaged in Object Recognition Can be Highly Sensitive to Object Retinal Position. *J Neurophysiol, 89* (6), 3264-3278.

Driver, J., & Baylis, G.C. (1996). Edge-assignment and figure-ground segmentation in short-term visual matching. *Cognit Psychol, 31* (3), 248-306.

Driver, J., Davis, G., Russell, C., Turatto, M., & Freeman, E. (2001). Segmentation, attention and phenomenal visual objects. *Cognition, 80* (1-2), 61-95.

Dubner, R., & Zeki, S.M. (1971). Response properties and receptive fields of cells in an anatomically defined region of the superior temporal sulcus in the monkey. *Brain Res, 35* (2), 528-532.

Duffy, C.J. (1998). MST neurons respond to optic flow and translational movement. *J Neurophysiol, 80* (4), 1816-1827.

Duncan, J. (1984). Selective attention and the organization of visual information. *J Exp Psychol Gen, 113* (4), 501-517.

Duzel, E., Habib, R., Rotte, M., Guderian, S., Tulving, E., & Heinze, H.J. (2003). Human hippocampal and parahippocampal activity during visual associative recognition memory for spatial and nonspatial stimulus configurations. *J Neurosci, 23* (28), 9439-9444.

Edelman, S. (1997). Computational theories of object recognition. *Trends Cogn Sci, 1* (8), 296-304.

Edelman, S., & Bulthoff, H.H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Res, 32* (12), 2385-2400.

Edelman, S., & Duvdevani-Bar, S. (1997). A model of visual recognition and categorization. *Philos Trans R Soc Lond B Biol Sci, 352* (1358), 1191-1202.

Edelman, S., & Intrator, N. (2003). Towards structural systematicity in distributed, statically bound visual representations. *Cogn Sci, 27* (1), 73-109.

Enroth-Cugell, C., & Robson, J.G. (1966). The contrast sensitivity of retinal ganglion cells of the cat. *J Physiol, 187* (3), 517-552.

Epstein, R., Graham, K.S., & Downing, P.E. (2003). Viewpoint-specific scene representations in human parahippocampal cortex. *Neuron, 37* (5), 865-876.

Epstein, R., Harris, A., Stanley, D., & Kanwisher, N. (1999). The parahippocampal place area: recognition, navigation, or encoding? *Neuron, 23* (1), 115-125.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392* (6676), 598-601.

Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *J Cogn Neurosci, 13* (2), 171-180.

Fabre-Thorpe, M., Richard, G., & Thorpe, S.J. (1998). Rapid categorization of natural images by rhesus monkeys. *Neuroreport, 9* (2), 303-308.

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *J Vis, 7* (1), 10.

Fei-Fei, L., VanRullen, R., Koch, C., & Perona, P. (2005). Why does natural scene categorization require little attention? Exploring attentional requirements for natural and synthetic stimuli. *Visual Cogn, 12* (6), 893-924.

Felleman, D.J., & Van Essen, D.C. (1987). Receptive field properties of neurons in area V3 of macaque monkey extrastriate cortex. *J Neurophysiol, 57* (4), 889-920.

Ferrera, V.P., Nealey, T.A., & Maunsell, J.H. (1992). Mixed parvocellular and magnocellular geniculate signals in visual area V4. *Nature, 358* (6389), 756-761.

Field, D.J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *J Opt Soc Am A, 4* (12), 2379-2394.

Field, D.J. (1994). What is the goal of sensory coding? *Neural Comput, 6* (4), 559-601.

Fine, I., MacLeod, D.I., & Boynton, G.M. (2003). Surface segmentation based on the luminance and color statistics of natural scenes. *J Opt Soc Am A Opt Image Sci Vis, 20* (7), 1283-1291.

Fiser, J., & Aslin, R.N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychol Sci, 12* (6), 499-504.

Fize, D., Fabre-Thorpe, M., Richard, G., Doyon, B., & Thorpe, S.J. (2005). Rapid categorization of foveal and extrafoveal natural images: Associated ERPs and effects of lateralization. *Brain Cogn, 59* (2), 145–158.

Frazor, R.A., & Geisler, W.S. (2006). Local luminance and contrast in natural images. *Vision Res, 46* (10), 1585-1598.

Freedman, D.J., Riesenhuber, M., Poggio, T., & Miller, E.K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science, 291* (5502), 312-316.

Friedman, A. (1979). Framing pictures: the role of knowledge in automatized encoding and memory for gist. *J Exp Psychol Gen, 108* (3), 316-355.

Fukushima, K., & Miyake, S. (1982). Neocognitron : A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recogn, 15*, 455-469.

Gallant, J.L., Connor, C.E., & Van Essen, D.C. (1998). Neural activity in areas V1, V2 and V4 during free viewing of natural scenes compared to controlled viewing. *Neuroreport, 9* (9), 1673-1678.

Gandhi, S.P., Heeger, D.J., & Boynton, G.M. (1999). Spatial attention affects brain activity in human primary visual cortex. *Proc Natl Acad Sci U S A, 96* (6), 3314-3319.

Ganis, G., & Kutas, M. (2003). An electrophysiological study of scene effects on object identification. *Brain Res Cogn Brain Res, 16* (2), 123-144.

Gautrais, J., & Thorpe, S. (1998). Rate coding versus temporal order coding: a theoretical approach. *Biosystems, 48* (1-3), 57-65.

Gegenfurtner, K.R., Kiper, D.C., & Levitt, J.B. (1997). Functional properties of neurons in macaque area V3. *J Neurophysiol, 77* (4), 1906-1923.

Gegenfurtner, K.R., & Rieger, J. (2000). Sensory and cognitive contributions of color to the recognition of natural scenes. *Curr Biol, 10* (13), 805-808.

Ghose, G.M., & Ts'o, D.Y. (1997). Form processing modules in primate area V4. *J Neurophysiol, 77* (4), 2191-2196.

Gibson, J.J. (1979). The ecological approach to visual perception. (Boston: Houghton Mifflin).

Gilbert, C.D., Sigman, M., & Crist, R.E. (2001). The neural basis of perceptual learning. *Neuron, 31* (5), 681-697.

Goffaux, V., Jacques, C., Mouraux, A., Oliva, A., Schyns, P.G., & Rossion, B. (2005). Diagnostic colours contribute to the early stages of scene categorization: Behavioural and neurophysiological evidence. *Visual Cogn, 12* (6), 878-892.

Goh, J.O., Siong, S.C., Park, D., Gutchess, A., Hebrank, A., & Chee, M.W. (2004). Cortical areas involved in object, background, and object-background processing revealed with functional magnetic resonance adaptation. *J Neurosci, 24* (45), 10223-10228.

Goldstone, R. (1994). Influences of categorization on perceptual discrimination. *J Exp Psychol Gen, 123* (2), 178-200.

Goodale, M.A., & Humphrey, G.K. (1998). The objects of action and perception. *Cognition, 67* (1-2), 181-207.

Goodale, M.A., & Milner, A.D. (1992). Separate visual pathways for perception and action. *Trends Neurosci, 15* (1), 20-25.

Goodale, M.A., Milner, A.D., Jakobson, L.S., & Carey, D.P. (1991). A neurological dissociation between perceiving objects and grasping them. *Nature, 349*, 154-156.

Green, C., & Hummel, J.E. (2006). Familiar interacting object pairs are perceptually grouped. *J Exp Psychol Hum Percept Perform, 32* (5), 1107-1119.

Grieve, K.L., Acuna, C., & Cudeiro, J. (2000). The primate pulvinar nuclei: vision and action. *Trends Neurosci, 23* (1), 35-39.

Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: as soon as you know it is there, you know what it is. *Psychol Sci, 16* (2), 152-160.

Grill-Spector, K., Knouf, N., & Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. *Nat Neurosci, 7* (5), 555-562.

Grill-Spector, K., Kushnir, T., Hendler, T., Edelman, S., Itzchak, Y., & Malach, R. (1998). A sequence of object-processing stages revealed by fMRI in the human occipital lobe. *Hum Brain Mapp, 6* (4), 316-328.

Grill-Spector, K., & Malach, R. (2001). fMR-adaptation: a tool for studying the functional properties of human cortical neurons. *Acta Psychol (Amst), 107* (1-3), 293-321.

Gross, C.G. (1994). How inferior temporal cortex became a visual area. *Cereb Cortex, 4* (5), 455-469.

Gross, C.G. (2008). Single neuron studies of inferior temporal cortex. *Neuropsychologia, 46* (3), 841-852.

Gross, C.G., Bruce, C.J., Desimone, R., Fleming, J., & Gattass, R. (1981). Cortical visual areas of the temporal lobe: three areas in the macaque. In: C.N. Woolsey (Ed.) *Cortical Sensory Organization - Multiple Visual Areas*, 2 (pp. 187-216). New York: Humana Press.

Gross, C.G., Rocha-Miranda, C.E., & Bender, D.B. (1972). Visual properties of neurons in inferotemporal cortex of the Macaque. *J Neurophysiol, 35* (1), 96-111.

Guerin-Dugue, A., & Oliva, A. (2000). Classification of scene photographs from local orientations features. *Pattern Recogn Lett, 21* (13-14), 1135-1140.

Guyader, N., Chauvin, A., Peyrin, C., Herault, J., & Marendaz, C. (2004). Image phase or amplitude? Rapid scene categorization is an amplitude-based process. *C R Biol, 327* (4), 313-318.

Hadjikhani, N., Liu, A.K., Dale, A.M., Cavanagh, P., & Tootell, R.B. (1998). Retinotopy and color sensitivity in human visual cortical area V8 [see comments]. *Nat Neurosci, 1* (3), 235-241.

Halgren, E., Baudena, P., Heit, G., Clarke, J.M., Marinkovic, K., Chauvel, P., & Clarke, M. (1994). Spatio-temporal stages in face and word processing. 2. Depth-recorded potentials in the human frontal and Rolandic cortices. *J Physiol Paris, 88* (1), 51-80.

Hanazawa, A., & Komatsu, H. (2001). Influence of the direction of elemental luminance gradients on the responses of V4 cells to textured surfaces. *J Neurosci, 21* (12), 4490-4497.

Haxby, J.V., Grady, C.L., Horwitz, B., Ungerleider, L.G., Mishkin, M., Carson, R.E., Herscovitch, P., Schapiro, M.B., & Rapoport, S.I. (1991). Dissociation of object and spatial visual processing pathways in human extrastriate cortex. *Proc Natl Acad Sci U S A, 88* (5), 1621-1625.

Hayward, W.G., & Williams, P. (2000). Viewpoint dependence and object discriminability. *Psychol Sci, 11* (1), 7-12.

Henderson, J.M. (1992). Object identification in context: the visual processing of natural scenes. *Can J Psychol, 46* (3), 319-341.

Henderson, J.M., Brockmole, J.R., & Gajewski, D.A. (2008). Differential detection of global luminance and contrast changes across saccades and flickers during active scene perception. *Vision Res, 48* (1), 16-29.

Henderson, J.M., & Hollingworth, A. (1999). High-level scene perception. *Annu Rev Psychol, 50*, 243-271.

Henderson, J.M., Pollatsek, A., & Rayner, K. (1987). Effects of foveal priming and extrafoveal preview on object identification. *Journal of Experimental Psychology : Human Perception and Performance, 13*, 449-463.

Herault, J., Oliva, A., & Guerin-Dugue, A. (1997). Scene categorisation by curvilinear component analysis of low frequency spectra. *ESANN'97* (pp. 91-96). Brugge.

Hillyard, S.A., Vogel, E.K., & Luck, S.J. (1998). Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence. *Philos Trans R Soc Lond B Biol Sci, 353* (1373), 1257-1270.

Hochstein, S., & Ahissar, M. (2002). View from the top. Hierarchies and reverse hierarchies in the visual system. *Neuron, 36* (5), 791-804.

Hollingworth, A., & Henderson, J.M. (1998). Does consistent scene context facilitate object perception? *J Exp Psychol Gen, 127* (4), 398-415.

Hollingworth, A., & Henderson, J.M. (1999). Object identification is isolated from scene semantic constraint: evidence from object type and token discrimination. *Acta Psychol (Amst), 102* (2-3), 319-343.

Hubel, D.H., & Wiesel, T.N. (1959). Receptive fields of single neurones in the cat's striate cortex. *J Physiol, 148* (3), 574-591.

Hubel, D.H., & Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol, 160* (1), 106-154.

Hubel, D.H., & Wiesel, T.N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J Physiol, 195* (1), 215-243.

Humphreys, G.W., & Riddoch, M.J. (1994). Attention to within-object and between-object spatial representations: multiple sites for visual selection. *Cogn Neuropsychol, 11* (2), 207-241.

Humphreys, G.W., & Riddoch, M.J. (1995). Separate coding of space within and between perceptual objects: Evidence from unilateral visual neglect. *Cogn Neuropsychol, 12* (3), 283-311.

Hupe, J.M., James, A.C., Payne, B.R., Lomber, S.G., Girard, P., & Bullier, J. (1998). Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature, 394* (6695), 784-787.

Intraub, H. (1979). The role of explicit naming in pictorial encoding. *journal of Experimental Psychology : Human Learning and Memory, 6*, 604-610.

Intraub, H. (1997). The representation of visual scenes. *Trends in Cognitive Sciences, 1* (6), 217-222.

Intraub, H. (1999). Understanding and remembering briefly glimpsed pictures: Implications for visual scanning and memory. In: V. Coltheart (Ed.) *Fleeting Memories: Cognition of brief visual stimuli* (pp. 47-70). Cambridge, Mass.: MIT Press.

Intraub, H., & Richardson, M. (1989). Wide-angle memories of close-up scenes. *J Exp Psychol Learn Mem Cogn, 15* (2), 179-187.

Irvin, G.E., Norton, T.T., Sesma, M.A., & Casagrande, V.A. (1986). W-like response properties of interlaminar zone cells in the lateral geniculate nucleus of a primate (Galago crassicaudatus). *Brain Res, 362* (2), 254-270.

Ito, M., Tamura, H., Fujita, I., & Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J Neurophysiol, 73* (1), 218-226.

Johnston, J.C., & McClelland, J.L. (1974). Perception of letters in words: seek not and ye shall find. *Science, 184* (142), 1192-1194.

Jolicoeur, P., Gluck, M.A., & Kosslyn, S.M. (1984). Pictures and names: making the connection. *Cognit Psychol, 16* (2), 243-275.

Joubert, O.R., Rousselet, G.A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Res, 47* (26), 3286-3297.

Kanwisher, N. (2000). Domain specificity in face perception. *Nat Neurosci, 3* (8), 759-763.

Kanwisher, N. (2001). Faces and places: of central (and peripheral) interest. *Nat Neurosci, 4* (5), 455-456.

Kanwisher, N., McDermott, J., & Chun, M.M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci, 17* (11), 4302-4311.

Kaping, D., Tzvetanov, T., & Treue, S. (2007). Adaptation to statistical properties of visual scenes biases rapid categorization. *Visual Cogn, 15* (1), 12-19.

Kimchi, R. (1988). Selective attention to global and local levels in the comparison of hierarchical patterns. *Percept Psychophys, 43* (2), 189-198.

Kirchner, H., & Thorpe, S.J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Res, 46* (11), 1762-1776.

Knierim, J.J., & Van Essen, D.C. (1992). Visual cortex: cartography, connectivity, and concurrent processing. *Curr Opin Neurobiol, 2* (2), 150-155.

Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J Neurophysiol, 71* (3), 856-867.

Koenderink, J.J. (1984). Geometrical structures determined by the functional order in nervous nets. *Biol Cybern, 50* (1), 43-50.

Kosslyn, S.M. (1987). Seeing and imagining in the cerebral hemispheres: a computational approach. *Psychol Rev, 94* (2), 148-175.

Kourtzi, Z., & Kanwisher, N. (2000). Cortical regions involved in perceiving object shape. *J Neurosci, 20* (9), 3310-3318.

Kuffler, S.W. (1953). Discharge patterns and functional organization of mammalian retina. *J Neurophysiol, 16* (1), 37-68.

Lamme, V.A. (1995). The neurophysiology of figure-ground segregation in primary visual cortex. *J Neurosci, 15* (2), 1605-1615.

Lamme, V.A., & Roelfsema, P.R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci, 23* (11), 571-579.

Lamme, V.A., Super, H., & Spekreijse, H. (1998). Feedforward, horizontal, and feedback processing in the visual cortex. *Curr Opin Neurobiol, 8* (4), 529-535.

Lefton, L.L., & Fisher, D.F. (1976). Information extraction during visual search: a developmental progression. *J Exp Child Psychol, 22* (2), 346-361.

Lerner, Y., Hendler, T., Ben-Bashat, D., Harel, M., & Malach, R. (2001). A hierarchical axis of object processing stages in the human visual cortex. *Cereb Cortex, 11* (4), 287-297.

Leventhal, A.G., Thompson, K.G., Liu, D., Zhou, Y., & Ault, S.J. (1995). Concomitant sensitivity to orientation, direction, and color of cells in layers 2, 3, and 4 of monkey striate cortex. *J Neurosci, 15* (3 Pt 1), 1808-1818.

Levy, I., Hasson, U., Avidan, G., Hendler, T., & Malach, R. (2001). Center-periphery organization of human object areas. *Nat Neurosci, 4* (5), 533-539.

Lewis, M.B., & Edmonds, A.J. (2003). Face detection: mapping human performance. *Perception, 32* (8), 903-920.

Li, F.F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proc Natl Acad Sci U S A, 99* (14), 9596-9601.

Loftus, G.R., & Mackworth, N.H. (1978). Cognitive determinants of fixation location during picture viewing. *J Exp Psychol Hum Percept Perform, 4* (4), 565-572.

Loftus, G.R., Nelson, W.W., & Kallman, H.J. (1983). Differential acquisition rates for different types of information from pictures. *Q J Exp Psychol A, 35* (Pt 1), 187-198.

Logothetis, N.K., & Pauls, J. (1995). Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cereb Cortex, 5* (3), 270-288.

Logothetis, N.K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Curr Biol, 5* (5), 552-563.

Logothetis, N.K., & Sheinberg, D.L. (1996). Visual Object Recognition. *Annu Rev Neurosci, 19*, 577-621.

Loschky, L.C., & Larson, A.M. (2008). Localized information is necessary for scene categorization, including the Natural/Man-made distinction. *J Vis, 8* (1), 1-9.

Loschky, L.C., Sethi, A., Simons, D.J., Pydimarri, T.N., Ochs, D., & Corbeille, J.L. (2007). The importance of information localization in scene gist recognition. *J Exp Psychol Hum Percept Perform, 33* (6), 1431-1450.

Lund, J.S., Henry, G.H., MacQueen, C.L., & Harvey, A.R. (1979). Anatomical organization of the primary visual cortex (area 17) of the cat. A comparison with area 17 of the macaque monkey. *J Comp Neurol, 184* (4), 599-618.

Lund, J.S., Lund, R.D., Hendrickson, A.E., Bunt, A.H., & Fuchs, A.F. (1975). The origin of efferent pathways from the primary visual cortex, area 17, of the macaque monkey as shown by retrograde transport of horseradish peroxidase. *J Comp Neurol, 164* (3), 287-303.

Macé, M. (2006). Représentations visuelles précoces dans la catégorisation rapide de scènes naturelles chez l'homme et le singe. (Toulouse: Paul Sabatier.

Macé, M., Fabre Thorpe, M., & Joubert, O.R. Dog or animal ? What comes first in vision. *Perception, 34, suppl* (8)

Macé, M., Joubert, O.R., & Fabre Thorpe, M. (2006). Entry level at the superordinate level in visual categorization. *Proceedings of the 9th International Congerence on Cognitive and Neural Systems*, 52 (

Mace, M.J., Thorpe, S.J., & Fabre-Thorpe, M. (2005). Rapid categorization of achromatic natural scenes: how robust at very low contrasts? *Eur J Neurosci, 21* (7), 2007-2018.

Mack, M.L., Gauthier, I., Sadr, J., & Palmeri, T.J. (2008). Object detection and basic-level categorization: sometimes you know it is there before you know what it is. *Psychon Bull Rev, 15* (1), 28-35.

MacKay, W.A., & Riehle, A. (1991). Correlates of preparation of arm reach parameters in parietal area 7A of the cerebral cortex. In: J.R.G.E. Stelmach (Ed.) *Tutorials in motor neuroscience* (pp. 347-356): Kluwer.

MacKay, W.A., & Riehle, A. (1992). Planning a reach: spatial analysis by area 7a neurons. In: G.E.S.J. Requin (Ed.) *Tutorials in motor behavior II* (pp. 501-514): Elsevier.

Maguire, E.A., Burgess, N., Donnett, J.G., Frackowiak, R.S., Frith, C.D., & O'Keefe, J. (1998). Knowing where and getting there: a human navigation network [see comments]. *Science, 280* (5365), 921-924.

Majaj, N.J., Carandini, M., & Movshon, J.A. (2007). Motion integration by neurons in macaque MT is local, not global. *J Neurosci, 27* (2), 366-370.

Mallat, S. G. (1987). A theory of multiresolution signal decomposition: the wavelet representation. GRASP Lab Technical Memo. University of Pennsylvania,Department of Computer and Information Science, MSCIS-87–22.

Malpeli, J.G., & Baker, F.H. (1975). The representation of the visual field in the lateral geniculate nucleus of Macaca mulatta. *J Comp Neurol, 161* (4), 569-594.

Mante, V., Frazor, R.A., Bonin, V., Geisler, W.S., & Carandini, M. (2005). Independence of luminance and contrast in natural scenes and in the early visual system. *Nat Neurosci, 8* (12), 1690-1697.

Marr, D. (1982). Vision. (San Francisco, CA: Freeman.

Marr, D., & Hildreth, E. (1980). Theory of edge detection. *Philos Trans R Soc Lond B Biol Sci, 207* (1167), 187-217.

Marr, D., & Nishihara, H.K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proc R Soc Lond B Biol Sci, 200* (1140), 269-294.

Martinez, A., Anllo-Vento, L., Sereno, M.I., Frank, L.R., Buxton, R.B., Dubowitz, D.J., Wong, E.C., Hinrichs, H., Heinze, H.J., & Hillyard, S.A. (1999). Involvement of striate and extrastriate visual cortical areas in spatial attention. *Nat Neurosci, 2* (4), 364-369.

Mathis, K.M. (2002). Semantic interference from objects both in and out of a scene context. *J Exp Psychol Learn Mem Cogn, 28* (1), 171-182.

Maunsell, J.H. (1992). Functional visual streams. *Curr Opin Neurobiol, 2* (4), 506-510.

Maunsell, J.H., & van Essen, D.C. (1983a). The connections of the middle temporal visual area (MT) and their relationship to a cortical hierarchy in the macaque monkey. *J Neurosci, 3* (12), 2563-2586.

Maunsell, J.H., & Van Essen, D.C. (1983b). Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation. *J Neurophysiol, 49* (5), 1127-1147.

McCotter, M., Gosselin, F., Sowden, P., & Schyns, P.G. (2005). The use of visual information in natural scenes. *Visual Cogn, 12* (6), 938-953.

McIntyre, J., Zago, M., Berthoz, A., & Lacquaniti, F. (2001). Does the brain model Newton's laws? Nat Neurosci, 4 (7), 693-694.

Mendez, M.F., & Cherrier, M.M. (2003). Agnosia for scenes in topographagnosia. *Neuropsychologia, 41* (10), 1387-1395.

Merchant, H., Battaglia-Mayer, A., & Georgopoulos, A.P. (2001). Effects of optic flow in motor cortex and area 7a. *J Neurophysiol, 86* (4), 1937-1954.

Merigan, W.H., & Pham, H.A. (1998). V4 lesions in macaques affect both single- and multiple-viewpoint shape discriminations. *Vis Neurosci, 15* (2), 359-367.

Metzger, R.L., & Antes, J.A. (1983). The nature of processing early in picture perception. *Psychol Res, 45* (3), 267-274.

Milner, A.D. (1995). Cerebral correlates of visual awareness. *Neuropsychologia, 33* (9), 1117-1130.

Mirabella, G., Bertini, G., Samengo, I., Kilavik, B.E., Frilli, D., Della Libera, C., & Chelazzi, L. (2007). Neurons in Area V4 of the Macaque Translate Attended Visual Features into Behaviorally Relevant Categories. *Neuron, 54* (2), 303-318.

Mishkin, M., & Ungerleider, L.G. (1982). Contribution of striate inputs to the visuospatial functions of parieto- preoccipital cortex in monkeys. *Behav Brain Res, 6* (1), 57-77.

Mishkin, M., Ungerleider, L.G., & Macko, K.A. (1983). Object vision ans spatial vision: two cortical pathways. *Trends Neurosci, oct.*, 414-417.

Moore, C., & Engel, S.A. (2001). Neural response to perception of volume in the lateral occipital complex. *Neuron, 29* (1), 277-286.

Morgan, M.J., Ross, J., & Hayes, A. (1991). The relative importance of local phase and local amplitude in patchwise image reconstruction. *Biological Cybernetics, 65*, 113-119.

Motter, B.C., & Mountcastle, V.B. (1981). The functional properties of the light-sensitive neurons of the posterior parietal cortex studied in waking monkeys: foveal sparing and opponent vector organization. *J Neurosci, 1* (1), 3-26.

Munk, M.H., Nowak, L.G., Girard, P., Chounlamountri, N., & Bullier, J. (1995). Visual latencies in cytochrome oxidase bands of macaque area V2. *Proc Natl Acad Sci U S A, 92* (4), 988-992.

Murphy, G.L. (1991). Parts in object concepts: experiments with artificial categories. *Mem Cognit, 19* (5), 423-438.

Murphy, G.L., & Smith, E.E. (1982). Basic level superiority in picture categorization. *J Verbal Learning & Verbal Behavior, 21*, 1-20.

Murphy, G.L., & Wisniewski, E.J. (1989). Categorizing objects in isolation and in scenes: what a superordinate is good for. *J Exp Psychol Learn Mem Cogn, 15* (4), 572-586.

Murray, E.A., & Bussey, T.J. (1999). Perceptual-mnemonic functions of the perirhinal cortex. *Trends Cogn Sci, 3* (4), 142-151.

Murray, E.A., & Bussey, T.J. (2001). Consolidation and the medial temporal lobe revisited: methodological considerations. *Hippocampus, 11* (1), 1-7.

Nakayama, K., & Shimojo, S. (1992). Experiencing and perceiving visual surfaces. *Science, 257* (5075), 1357-1363.

Nakayama, K., Shimojo, S., & Silverman, G.H. (1989). Stereoscopic depth: its relation to image segmentation, grouping, and the recognition of occluded objects. *Perception, 18* (1), 55-68.

Navon, D. (1977). Forest before trees: the precedence of global features in visual perception. *Cognit Psychol, 9* (3), 353-383.

Navon, D. (1983). How many trees does it take to make a forest? *Perception, 12* (3), 239-254.

Nealey, T.A., & Maunsell, J.H. (1994). Magnocellular and parvocellular contributions to the responses of neurons in macaque striate cortex. *J Neurosci, 14* (4), 2069-2079.

New, J., Cosmides, L., & Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proc Natl Acad Sci U S A, 104* (42), 16598-16603.

Nobre, A.C., Allison, T., & McCarthy, G. (1998). Modulation of human extrastriate visual processing by selective attention to colours and words. *Brain, 121* (Pt 7)(2-3), 1357-1368.

Norman, G.R., Brooks, L.R., Coblentz, C.L., & Babcook, C.J. (1992). The correlation of feature identification and category judgments in diagnostic radiology. *Mem Cognit, 20* (4), 344-355.

Nowak, L.G., & Bullier, J. (1997). The timing of information transfer in the visual system. In: K.S. Rockland, J.H. Kaas, & A. Peters (Eds.), *Extrastriate visual cortex in primates,* 12 (pp. 205-241). New York: Plenum Press.

O'Craven, K.M., Downing, P.E., & Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature, 401* (6753), 584-587.

Oliva, A. (2005). Gist of the Scene. In: L. Itti, G. Rees, & J.K. Tsotsos (Eds.), *Neurobiology of Attention* (pp. 251-256): Elsevier.

Oliva, A., & Schyns, P.G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognit Psychol, 34* (1), 72-107.

Oliva, A., & Schyns, P.G. (2000). Diagnostic colors mediate scene recognition. *Cognit Psychol, 41* (2), 176-210.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42* (3), 145-175.

Olshausen, B. (2003). Sensory coding in the natural environment. *Network Comput Neural Syst, 14* (3), 369-370.

Olshausen, B.A., & Field, D.J. (1996a). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381* (6583), 607-609.

Olshausen, B.A., & Field, D.J. (1996b). Natural image statistics and efficient coding. *Network Comput Neural Syst, 7*, 333-339.

Op De Beeck, H., & Vogels, R. (2000). Spatial sensitivity of macaque inferior temporal neurons. *J Comp Neurol, 426* (4), 505-518.

Op de Beeck, H., Wagemans, J., & Vogels, R. (2001a). Can neuroimaging really tell us what the human brain is doing? The relevance of indirect measures of population activity. *Acta Psychol (Amst), 107* (1-3), 323-351.

Op de Beeck, H., Wagemans, J., & Vogels, R. (2001b). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nat Neurosci, 4* (12), 1244-1252.

Op de Beeck, H., Wagemans, J., & Vogels, R. (2003). The effect of category learning on the representation of shape: dimensions can be biased but not differentiated. *J Exp Psychol Gen, 132* (4), 491-511.

Oram, M.W., & Perrett, D.I. (1992). Time course of neural responses discriminating different views of the face and head. *J Neurophysiol, 68* (1), 70-84.

Ostergaard, A.L., & Davidoff, J.B. (1985). Some effects of color on naming and recognition of objects. *J Exp Psychol Learn Mem Cogn, 11* (3), 579-587.

Palmer, S., & Rock, I. (1994). Rethinking perceptual organization: The role of uniform connectedness. *Psychon Bull Rev, 1* (1), 29-55.

Palmer, S.E. (1975). The effects of contextual scenes on the identification of objects. *Mem Cognit, 3*, 519-526.

Pasupathy, A., & Connor, C.E. (2001). Shape representation in area V4: position-specific tuning for boundary conformation. *J Neurophysiol, 86* (5), 2505-2519.

Pasupathy, A., & Connor, C.E. (2002). Population coding of shape in area V4. *Nat Neurosci, 5* (12), 1332-1338.

Perrett, D.I., Rolls, E.T., & Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Exp Brain Res, 47* (3), 329-342.

Peterson, M.A. (1994). Object recognition processes can and do operate before figure-ground organization. *Curr Direct Psychol Sci, 3*, 105-111.

Peterson, M.A., & Gibson, B.S. (1994a). Must figure-ground organization precede object recognition? *Psychol Sci, 5* (5), 253-259.

Peterson, M.A., & Gibson, B.S. (1994b). Object recognition contributions to figure-ground organization: operations on outlines and subjective contours. *Percept Psychophys, 56* (5), 551-564.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects [see comments]. *Nature, 343* (6255), 263-266.

Potter, M.C. (1975). Meaning in visual search. *Science, 187* (4180), 965-966.

Potter, M.C. (1976). Short-term conceptual memory for pictures. *J Exp Psychol Hum Learn Mem, 2* (5), 509-522.

Price, C.J., & Humphreys, G.W. (1989). The effects of surface detail on object categorization and naming. *Q J Exp Psychol A, 41* (4), 797-827.

Quiroga, R.Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature, 435*, 1102-1107.

Quraishi, S., Heider, B., & Siegel, R.M. (2007). Attentional modulation of receptive field structure in area 7a of the behaving monkey. *Cereb Cortex, 17* (8), 1841-1857.

Rensink, R.A. (2000). The dynamic representation of scenes. *Visual Cogn, 7* (1), 17-42.

Rensink, R.A. (2002). Change detection. *Annu Rev Psychol, 53*, 245-277.

Rensink, R.A., O'Regan, J.K., & Clark, J.J. (1997). TO SEE OR NOT TO SEE: the need for attention to perceive changes in scenes. *Psychol Sci, 8* (5), 368-373.

Reynolds, J.H., Chelazzi, L., & Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *J Neurosci, 19* (5), 1736-1753.

Richmond, B.J., Wurtz, R.H., & Sato, T. (1983). Visual responses of inferior temporal neurons in awake rhesus monkey. *J Neurophysiol, 50* (6), 1415-1432.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat Neurosci, 2* (11), 1019-1025.

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nat Neurosci, 3 Suppl*, 1199-1204.

Robinson, D.L., Goldberg, M.E., & Stanton, G.B. (1978). Parietal association cortex in the primate: sensory mechanisms and behavioral modulations. *J Neurophysiol, 41* (4), 910-932.

Rock, I., Nijhawan, R., Palmer, S., & Tudor, L. (1992). Grouping based on phenomenal similarity of achromatic color. *Perception, 21* (6), 779-789.

Rockland, K.S., & Pandya, D.N. (1979). Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. *Brain Res, 179* (1), 3-20.

Roelfsema, P.R., Lamme, V.A., & Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature, 395* (6700), 376-381.

Rolls, E.T., Sanghera, M.K., & Roper-Hall, A. (1979). The latency of activation of neurones in the lateral hypothalamus and substantia innominata during feeding in the monkey. *Brain Res, 164*, 121-135.

Rolls, E.T., Stringer, S.M., & Trappenberg, T.P. (2002). A unified model of spatial and episodic memory. *Proc R Soc Lond B Biol Sci, 269* (1496), 1087-1093.

Rolls, E.T., Treves, A., & Tovee, M.J. (1997). The representational capacity of the distributed encoding of information provided by populations of neurons in primate temporal visual cortex. *Exp Brain Res, 114* (1), 149-162.

Rolls, E.T., Treves, A., Tovee, M.J., & Panzeri, S. (1997). Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. *J Comput Neurosci, 4* (4), 309-333.

Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognit Psychol, 8*, 382-439.

Rossi, A.F., Desimone, R., & Ungerleider, L.G. (2001). Contextual modulation in primary visual cortex of macaques. *J Neurosci, 21* (5), 1698-1709.

Rousselet, G.A., Fabre-Thorpe, M., & Thorpe, S.J. (2002). Parallel processing in high-level categorization of natural images. *Nat Neurosci, 5* (7), 629-630.

Rousselet, G.A., Joubert, O.R., & Fabre-Thorpe, M. (2005). How long to get to the “gist” of real-world natural scenes? *Visual Cogn, 12* (6), 852-877.

Rousselet, G.A., Mace, M.J., & Fabre-Thorpe, M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *J Vis, 3* (6), 440-455.

Rousselet, G.A., Thorpe, S.J., & Fabre-Thorpe, M. (2004). How parallel is visual processing in the ventral pathway? *Trends Cogn Sci, 8* (8), 363-370.

Rubin, N., Nakayama, K., & Shapley, R. (1996). Enhanced perception of illusory contours in the lower versus upper visual hemifields. *Science, 271* (5249), 651-653.

Ryan, T.A., & Schwartz, C.B. (1956). Speed of perception as a function of mode of representation. *Am J Psychol, 69* (1), 60-69.

Sanocki, T., & Epstein, W. (1997). Priming spatial layout of scenes. *Psychol Sci, 8* (5), 374-378.

Schall, J.D., Morel, A., King, D.J., & Bullier, J. (1995). Topography of visual cortex connections with frontal eye field in macaque: convergence and segregation of processing streams. *J Neurosci, 15* (6), 4464-4487.

Schein, S.J., Marrocco, R.T., & de Monasterio, F.M. (1982). Is there a high concentration of color-selective cells in area V4 of monkey visual cortex? *J Neurophysiol, 47* (2), 193-213.

Schiller, P.H., & Malpeli, J.G. (1977). Properties and tectal projections of monkey retinal ganglion cells. *J Neurophysiol, 40* (2), 428-445.

Schiller, P.H., Malpeli, J.G., & Schein, S.J. (1979). Composition of geniculostriate input ot superior colliculus of the rhesus monkey. *J Neurophysiol, 42* (4), 1124-1133.

Schyns, P.G. (1997). Categories and percepts: A bi-directional framework for categorization. *Trends Cogn Sci, 1*, 183-189.

Schyns, P.G. (1998). Diagnostic recognition: task constraints, object information, and their interactions. *Cognition, 67* (1-2), 147-179.

Schyns, P.G., & Murphy, G.L. (1994). The ontogeny of part representation in object concepts. In: *The Psychology of Learning and Motivation*, 31 (pp. 305-349): Academic Press.

Schyns, P.G., & Oliva, A. (1994). From blobs to boundary edges: evidence for time- and spatial-scale-dependent scene recognition. *Psychol Sci, 5* (4), 195-200.

Schyns, P.G., & Rodet, L. (1997). Categorization creates functional features. *J Exp Psychol Learn Mem Cogn, 23* (3), 681-696.

Sengpiel, F., & Hubener, M. (1999). Visual attention: spotlight on the primary visual cortex. *Curr Biol, 9* (9), R318-321.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proc Natl Acad Sci U S A, 104* (15), 6424-6429.

Shapley, R., & Perry, V.H. (1986). Cat and monkey retinal ganglion cells and their visual functional roles. *Trends Neurosci, May*, 229-235.

Sherman, S.M., Wilson, J.R., Kaas, J.H., & Webb, S.V. (1976). X- and Y-cells in the dorsal lateral geniculate nucleus of the owl monkey (Aotus trivirgatus). *Science, 192* (4238), 475-477.

Shulman, G.L., Sullivan, M.A., Gish, K., & Sakoda, W.J. (1986). The role of spatial-frequency channels in the perception of local and global structure. *Perception, 15* (3), 259-273.

Shulman, G.L., & Wilson, J. (1987). Spatial frequency and selective attention to local and global information. *Perception, 16* (1), 89-101.

Sigala, N. (2004). Visual categorization and the inferior temporal cortex. *Behav Brain Res, 149* (1), 1-7.

Sigala, N., Gabbiani, F., & Logothetis, N.K. (2002). Visual categorization and object representation in monkeys and humans. *J Cogn Neurosci, 14* (2), 187-198.

Sigala, N., & Logothetis, N.K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature, 415* (6869), 318-320.

Sigman, M., Cecchi, G.A., Gilbert, C.D., & Magnasco, M.O. (2001). On a common circle: natural scenes and Gestalt rules. *Proc Natl Acad Sci U S A, 98* (4), 1935-1940.

Sillito, A.M., Grieve, K.L., Jones, H.E., Cudeiro, J., & Davis, J. (1995). Visual cortical mechanisms detecting focal orientation discontinuities. *Nature, 378* (6556), 492-496.

Simoncelli, E.P., & Olshausen, B.A. (2001). Natural image statistics and neural representation. *Annu Rev Neurosci, 24*, 1193-1216.

Smith, C., Carey, S., & Wiser, M. (1985). On differentiation: a case study of the development of the concepts of size, weight, and density. *Cognition, 21* (3), 177-237.

Sperling, R., Chua, E., Cocchiarella, A., Rand-Giovannetti, E., Poldrack, R., Schacter, D.L., & Albert, M. (2003). Putting names to faces: successful encoding of associative memories activates the anterior hippocampal formation. *Neuroimage, 20* (2), 1400-1410.

Stern, C.E., Corkin, S., Gonzalez, R.G., Guimaraes, A.R., Baker, J.R., Jennings, P.J., Carr, C.A., Sugiura, R.M., Vedantham, V., & Rosen, B.R. (1996). The Hippocampal Formation Participates in Novel Picture Encoding Evidence from Functional Magnetic Resonance Imaging. *Proc Natl Acad Sci U S A, 93*, 8660-8665.

Suzuki, W.A., Miller, E.K., & Desimone, R. (1997). Object and place memory in the macaque entorhinal cortex. *J Neurophysiol, 78* (2), 1062-1081.

Tajfel, H., & Wilkes, A.L. (1963). Classification and quantitative judgment. *British Journal of Psychology, 54*, 101-114.

Tallon, C., Bertrand, O., Bouchet, P., & Pernier, J. (1995). Gamma-range activity evoked by coherent visual stimuli in humans. *Eur J Neurosci, 7* (6), 1285-1291.

Tanaka, J.W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognit Psychol, 23*, 457-482.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu Rev Neurosci, 19*, 109-139.

Tarr, M.J., & Bulthoff, H.H. (1998). Image-based object recognition in man, monkey and machine. *Cognition, 67* (1-2), 1-20.

Tarr, M.J., Williams, P., Hayward, W.G., & Gauthier, I. (1998). Three-dimensional object recognition is viewpoint dependent. *Nat Neurosci, 1* (4), 275-277.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381* (6582), 520-522.

Thorpe, S., & Imbert, M. (1989). Biological constraints on connectionist modelling. In: R. Pfeifer, Z. Schreter, F. Fogelman-Soulié, & L. Steels (Eds.), *Connectionism in perspective* (pp. 63-92). Amsterdam: Elsevier.

Thorpe, S.J., & Fabre-Thorpe, M. (2001). Neuroscience. Seeking categories in the brain. *Science, 291* (5502), 260-263.

Thorpe, S.J., & Gautrais, J. (1997). How can the visual system process a natural scene in under 150 ms? On the role of asynchronous spike propagation. *ESANN1997 - 5th European Symposium on Artificial Neural Networks* (pp. 79-84). Bruges, Belgium: D Facto.

Thorpe, S.J., Gegenfurtner, K.R., Fabre-Thorpe, M., & Bulthoff, H.H. (2001). Detection of animals in natural images using far peripheral vision. *Eur J Neurosci, 14* (5), 869-876.

Thorpe, S.J., Rolls, E.T., & Maddison, S. (1983). The orbitofrontal cortex: neuronal activity in the behaving monkey. *Exp Brain Res, 49* (1), 93-115.

Tomita, H., Ohbayashi, M., Nakahara, K., Hasegawa, I., & Miyashita, Y. (1999). Top-down signal from prefrontal cortex in executive control of memory retrieval [see comments]. *Nature, 401* (6754), 699-703.

Tootell, R.B., Mendola, J.D., Hadjikhani, N.K., Liu, A.K., & Dale, A.M. (1998). The representation of the ipsilateral visual field in human cerebral cortex. *Proc Natl Acad Sci U S A, 95* (3), 818-824.

Torralba, A., & Oliva, A. (2002). Depth estimation from image structure. *IEEE Trans Pattern Analysis & Machine Intelligence, 24* (9), 1226-1238.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network, 14* (3), 391-412.

Treisman, A. (1982). Perceptual grouping and attention in visual search for features and for objects. *J Exp Psychol Hum Percept Perform, 8* (2), 194-214.

Treisman, A. (1986). Features and objects in visual processing. *Sci Am, 255* (5), 114-125.

Treisman, A.M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognit Psychol, 12* (1), 97-136.

Treue, S. (2001). Neural correlates of attention in primate visual cortex. *Trends Neurosci, 24* (5), 295-300.

Treue, S., & Maunsell, J.H. (1996). Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature, 382* (6591), 539-541.

Trotter, Y., & Celebrini, S. (1999). Gaze direction controls response gain in primary visual-cortex neurons. *Nature, 398* (6724), 239-242.

Turatto, M., & Galfano, G. (2000). Color, form and luminance capture attention in visual search. *Vision Res, 40* (13), 1639-1643.

Tversky, B., & Hemenway, K. (1983). Categories of environmental scenes. *Cogn Psychol, 15* (1), 121-149.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *J Exp Psychol Gen, 113* (2), 169-197.

Tyler, L.K., Stamatakis, E.A., Bright, P., Acres, K., Abdallah, S., Rodd, J.M., & Moss, H.E. (2004). Processing objects at different levels of specificity. *J Cogn Neurosci, 16* (3), 351-362.

Ullman, S. (1995). Sequence seeking and counter streams: a computational model for bidirectional information flow in the visual cortex. *Cereb Cortex, 5* (1), 1-11.

Ungerleider, L.G., & Haxby, J.V. (1994). 'What' and 'where' in the human brain. *Curr Opin Neurobiol, 4* (2), 157-165.

Usher, M., & Donnelly, N. (1998). Visual synchrony affects binding and segmentation in perception. *Nature, 394* (6689), 179-182.

Vailaya, A., Jain, A., & Zhang, H.J. (1998). On image classification: city images vs. landscapes. *Pattern Recogn, 31* (12), 1921-1935.

Van Essen, D.C., Maunsell, J.H., & Bixby, J.L. (1981). The middle temporal visual area in the macaque: myeloarchitecture, connections, functional properties and topographic organization. *J Comp Neurol, 199* (3), 293-326.

Van Essen, D.C., & Zeki, S.M. (1978). The topographic organization of rhesus monkey prestriate cortex. *J Physiol, 277*, 193-226.

van Hateren, J.H., & van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc R Soc Lond B Biol Sci, 265* (1394), 359-366.

Vanduffel, W., Tootell, R.B., & Orban, G.A. (2000). Attention-dependent suppression of metabolic activity in the early stages of the macaque visual system [In Process Citation]. *Cereb Cortex, 10* (2), 109-126.

VanRullen, R., Reddy, L., & Koch, C. (2004). Visual search and dual tasks reveal two distinct attentional resources. *J Cogn Neurosci, 16* (1), 4-14.

VanRullen, R., & Thorpe, S.J. (2001a). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception, 30* (6), 655-668.

VanRullen, R., & Thorpe, S.J. (2001b). The time course of visual processing: from early perception to decision- making. *J Cogn Neurosci, 13* (4), 454-461.

Vogel, J., & Schiele, B. (2004). A Semantic Typicality Measure for Natural Scene Categorization. *26th DAGM Symposium*, 3175 (pp. 195-203). Tübingen, Germany: Springer-Verlag.

Vogels, R. (1999a). Categorization of complex visual images by rhesus monkeys. Part 1: behavioural study. *Eur J Neurosci, 11* (4), 1223-1238.

Vogels, R. (1999b). Categorization of complex visual images by rhesus monkeys. Part 2: single-cell study. *Eur J Neurosci, 11* (4), 1239-1255.

Vogels, R., Biederman, I., Bar, M., & Lorincz, A. (2001). Inferior Temporal Neurons Show Greater Sensitivity to Nonaccidental than to Metric Shape Differences. *J Cogn Neurosci, 13* (4), 444-453.

Vogels, R., & Orban, G.A. (1996). Coding of stimulus invariances by inferior temporal neurons. *Prog Brain Res, 112*, 195-211.

von der Heydt, R., Peterhans, E., & Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science, 224* (4654), 1260-1262.

Wallis, G., & Rolls, E.T. (1997). Invariant face and object recognition in the visual system. *Prog Neurobiol, 51* (2), 167-194.

Watson, J.D., Myers, R., Frackowiak, R.S., Hajnal, J.V., Woods, R.P., Mazziotta, J.C., Shipp, S., & Zeki, S. (1993). Area V5 of the human brain: evidence from a combined study using positron emission tomography and magnetic resonance imaging. *Cereb Cortex, 3* (2), 79-94.

Westheimer, G. (2001). The Fourier theory of vision. *Perception, 30* (5), 531-541.

Wichmann, F.A., Braun, D.I., & Gegenfurtner, K.R. (2006). Phase noise and the classification of natural images. *Vision Res, 46* (8-9), 1520-1529.

Wiesel, T.N., & Hubel, D.H. (1966). Spatial and chromatic interactions in the lateral geniculate body of the rhesus monkey. *J Neurophysiol, 29* (6), 1115-1156.

Wilson, P.D., Rowe, M.H., & Stone, J. (1976). Properties of relay cells in cat's lateral geniculate nucleus: a comparison of W-cells with X- and Y-cells. *J Neurophysiol, 39* (6), 1193-1209.

Wolfe, J.M., Friedman-Hill, S.R., Stewart, M.I., & O'Connell, K.M. (1992). The role of categorization in visual search for orientation. *J Exp Psychol Hum Percept Perform, 18* (1), 34-49.

Wurm, L.H., Legge, G.E., Isenberg, L.M., & Luebker, A. (1993). Color improves object recognition in normal and low vision. *J Exp Psychol Hum Percept Perform, 19* (4), 899-911.

Zeki, S. (1980). The representation of colours in the cerebral cortex. *Nature, 284* (5755), 412-418.

Zeki, S. (1983). The relationship between wavelength and color studied in single cells of monkey striate cortex. *Prog Brain Res, 58*, 219-227.

Zeki, S.M. (1969). Representation of central visual fields in prestriate cortex of monkey. *Brain Res, 14* (2), 271-291.

Zeki, S.M. (1971). Cortical projections from two prestriate areas in the monkey. *Brain Res, 34* (1), 19-35.

Zeki, S.M. (1974). Functional organization of a visual area in the posterior bank of the superior temporal sulcus of the rhesus monkey. *J Physiol, 236* (3), 549-573.

Zeki, S.M. (1978). The cortical projections of foveal striate cortex in the rhesus monkey. *J Physiol, 277*, 227-244.

AUTEUR : **Olivier Joubert** DIRECTRICE DE THESE : **Dr. Michèle Fabre-Thorpe**

RAPID CATEGORIZATION OF NATURAL SCENES : THE OBJECT, THE CONTEXT, AND THEIR INTERACTIONS

SUMMARY

In a world governed by physical laws, our brain is able to extract some invariants and to generate expectations to precise our visual percept. However, as soon as we turn on the light of a dark room, as soon as we zap to a new tv channel, we understand nearly instantaneously the content of these new complex scenes. How rapidly are we able to extract the gist of natural scenes? Which influence has the context on the recognition of relevant object ? What are the necessary visual informations and the underlying visual processing ? The two first papers of this thesis show our capacity to recognize context superordinate categories of natural scenes in less than 400 ms while the access to basic categories need about 50 ms of additional processing. They reveal also that the presence of salient objects interfere with rapid context categorization. These short processing times that are very similar to those obtained in object categorization tasks suggest a global and parallel processing of the whole scene. The third paper show that the context congruency with the object influe immediately on object categorization processing and the last one analyse the necessary visual information in context processing. Based on such data, I support the idea of early crossed interactions between parallel bottom-up visual processings of object and context. A final study aimed at determining the minimal latency for such interactions by using ocular response rather than manual response. The implications of these fundamental results are discussed in the perspective of future researchs.

KEY-WORDS
visual perception, categorization, go/no-go, natural scenes, interactions between objet and context, amplitude and phase spectra.

AUTEUR : **Olivier Joubert** DIRECTRICE DE THESE : **Dr. Michèle Fabre-Thorpe**

CATEGORISATION RAPIDE DES SCENES NATURELLES : L'OBJET, LE CONTEXTE, ET LEURS INTERACTIONS

RESUME

Dans un monde régi par les lois physiques, notre cerveau est capable d'extraire des invariants et de générer des attentes pour préciser notre percept visuel. Pourtant, en éclairant une pièce, ou en naviguant à travers les chaînes de télévision, nous comprenons quasi-instantanément l'essence de ces nouvelles scènes naturelles. A quelle vitesse peut-on extraire une représentation sémantique globale des scènes? Quelle est l'influence du contexte sur la reconnaissance de l'objet d'intérêt? Quelles sont les informations visuelles nécessaires et la nature des traitements sous-jacents? Les deux premiers articles démontrent notre capacité à reconnaître la catégorie superordonnée du contexte d'une scène en moins de 400 ms tandis que l'accès au niveau basique nécessite 50 ms de traitement additionnel. Ils démontrent aussi que la présence d'objets saillants interfère sur la catégorisation rapide du contexte. Ces temps de traitements très similaires à ceux enregistrés dans la catégorisation rapide d'objets suggèrent un traitement global et parallèle de l'ensemble de la scène. Le troisième article montre que la congruence (incongruence) du contexte avec l'objet influence immédiatement le traitement de l'objet et le dernier précise les informations visuelles à la base de l'analyse du contexte. Je défends ainsi l'idée d'interactions bidirectionnelles précoces entre les traitements visuels ascendants et parallèles de l'objet et du contexte et recherche dans une dernière étude la latence minimale de ces interactions en remplaçant réponse manuelle par réponse oculaire. L'implication de ces résultats fondamentaux est discutée dans la perspective des recherches à venir.

MOTS-CLES
perception visuelle, catégorisation, paradigme go/no-go, scènes naturelles, interactions objet/contexte, spectres d'amplitude et de phase.

Doctorat en Sciences Cognitives

Soutenu le **30 Septembre 2008**
Salle des Thèses – Université Toulouse III – Paul Sabatier

Réalisé au **Centre de Recherche Cerveau et Cognition**
UMR 5549 (CNRS-Université Paul Sabatier Toulouse 3)
Faculté de Médecine de Rangueil. 31062 Toulouse CEDEX9