# Un modèle d'attention visuelle dynamique pour conditions 2D et 3D ; codage de cartes de profondeur et synthèse basée inpainting pour les vidéos multi-vues

Josselin Gautier

N° d'ordre : ANNÉE 2012

**THÈSE / UNIVERSITÉ DE RENNES 1**
*sous le sceau de l'Université Européenne de Bretagne*

pour le grade de

**DOCTEUR DE L'UNIVERSITÉ DE RENNES 1**

*Mention : Informatique*

**Ecole doctorale Matisse**

présentée par


# Josselin Gautier

préparée à l'unité de recherche UMR-6074
IRISA / INRIA
(Composante universitaire : ISTIC)


---

# A Dynamic Visual Attention Model for 2D and 3D conditions ; Depth Coding and Inpainting-based Synthesis for Multiview Videos

**Thèse soutenue à Rennes le 5 décembre 2012**

devant le jury composé de :

**Anne GUÉRIN-DUGUÉ**
Directeur de Recherche, GIPSA Grenoble/ rapporteur

**Frédéric DUFAUX**
Directeur de Recherche, LTCI Paris/ rapporteur

**Fred STENTIFORD**
Professeur, University College de Londres/ examinateur

**Henri NICOLAS**
Professeur, Université Bordeaux 1/ examinateur

**Christine GUILLEMOT**
Directeur de Recherche, INRIA Rennes/ directeur de thèse

**Olivier LE MEUR**
Maître de Conférence, IRISA Rennes / co-directeur de thèse

# Contents

Résumé en français :
De l'usage de la profondeur pour prédire l'attention visuelle en stéréoscopie et pour la chaîne de transmission 3D

# Introduction

Cette thèse porte sur la perception visuelle de contenus 3D et leurs améliorations dans une optique de codage efficace et de synthèse 3D de qualité.

La première partie traite de la perception de la profondeur par le système visuel humain et de l'impact d'un indice de profondeur, la disparité binoculaire, sur l'attention visuelle. Cette première partie est donc un sujet de la communauté vision humaine.

La seconde partie aborde les problèmes de transmission, codage et rendu de scène pour les systèmes 3D émergents tels que la 3DTV et le FTV. Cette partie s'adresse particulièrement à la communauté vision par ordinateur et codage du fait de la recherche de solutions pour des contenus de qualité.

# Partie Vision Humaine : la Perception de la Profondeur

Cette partie traite de la vision spatiale opérée par le Système Visuel Humain (SVH). Des bases théoriques et anatomiques de la vision en général, et du système visuel humain en particulier, sont présentées (annexe 1 de ce résumé). Puis les différents indices de profondeurs stéréoscopiques, oculaires, dynamiques et picturales sont précisément décrits afin d'introduire la principale contribution de cette partie sur l'intégration de la disparité binoculaire dans un modèle d'attention visuelle pour conditions 2D et 3D.

## Chapitre 1 : Les Indices de Profondeur

Le Système Visuel Humain reçoit au travers des rétines deux images bi-dimensionnelles de son environnement visuel. Pourtant celles-ci ne suffisent pas à strictement parler pour estimer la distance des objets à l'observateur. Celui-ci va donc faire appel à différentes sources d'informations ou **indices de profondeur** pour en extraire une organisation plausible des surfaces et objets dans son champ visuel. Il en existe plusieurs types dont la distinction est assez controversée, car purement systémique et non fonctionnelle.

Une première distinction est faite selon que la source d'information de la profondeur concerne l'état des yeux ou de la lumière y entrant. Une seconde selon que la source

d'information implique un seul ou les deux yeux. Une troisième selon que l'information soit extraite d'une image apparaissant sur la rétine de façon statique ou dynamique. Une quatrième selon que l'information donne une distance absolue ou relative aux objets entre eux, et enfin une cinquième séparant les sources apportant une relation de distance numérique de celles donnant une relation d'ordre (du plus proche au plus loin). Nous allons tâcher de décrire ces différents indices selon ces caractéristiques ; on les regroupe au préalable selon le type d'information qu'ils traitent.

### Les indices stéréoscopiques

La distance entre les deux yeux permet de percevoir le monde qui nous entoure de deux points de vues décalés. La perception de la profondeur issue de ce déplacement relatif - la disparité binoculaire - entre deux images rétiniennes est appelée la stéréopsie. La **disparité binoculaire** constitue donc un premier indice fournissant une information quantitative précise mais relative de la distance à des objets statiques. Selon qu'un point est plus proche ou plus loin qu'un point fixé par le regard, il se projettera respectivement vers l'extérieur ou l'intérieur de ce point sur la rétine. On parle alors de disparité "croisée" ou "non-croisée".

Si maintenant, on considère un objet qui ne se déplace pas seulement selon l'axe de profondeur par rapport à l'observateur, mais selon un axe horizontal, ce déplacement va induire une **disparité binoculaire verticale**. Cet objet apparaitra plus gros sur l'oeil droit que sur l'oeil gauche, à la fois selon l'axe horizontal mais aussi vertical. Cette disparité dite verticale entre deux points correspondants du même objets mais de différentes tailles sur chaque rétine est donc une source supplémentaire de profondeur.

Enfin, la **stéréopsie de Da Vinci**, nommée par Nakayama et Shimojo [95] en l'honneur de l'inventeur qui le premier la mentionna, provient des zones visibles depuis un oeil mais apparaissant comme cachées (car occultées) sur l'autre oeil. Le SVH est donc susceptible de déterminer l'ordonnancement de différents plans et surfaces de la scène à partir de cette stéréopsie.

### Les indices oculaires

Liés aux caractéristiques de l'oeil et à son fonctionnement, on en distingue deux principaux : l'accommodation et la convergence.

L'**accommodation** résulte du processus de déformation du cristallin par la contraction-décontraction des muscles ciliaires, ceci afin d'ajuster la mise au point sur les objets plus ou moins proches. Comme la tension à appliquer sur le muscle pour percevoir une image nette sur la rétine est connue par le SVH, cet indice donne une information statique, absolue et quantitative des distances aux objets, mais seulement jusqu'à deux mètres. Au delà, il n'y a plus de déformation du cristallin nécessaire à la mise au point, et donc plus d'information disponible.

Les mouvements de **vergence oculaire** : les yeux se déplacent en sens inverse l'un de l'autre afin de fixer un objet plus ou moins proche, relèvent de la vision binoculaire. En effet, l'angle de convergence varie directement avec la distance à l'objet : un grand angle de convergence pour un objet proche et inversement pour un objet éloigné. Tout comme l'accommodation, cette source d'information est absolue et n'opère qu'à une distance faible (moins d'un mètre).

### Les indices dynamiques

Le déplacement d'un objet sur la rétine au cours du temps fournit une information visuelle d'ordre dynamique au SVH : un flot optique. Sa direction et son amplitude dépendent du mouvement mais aussi de la distance, de la position de l'objet par rapport à l'observateur, ou inversement. Ainsi la **parallaxe de mouvement** est un indice monoculaire basée sur le fait qu'en présence d'un mouvement de l'objet ou de l'observateur les objets proches semblent se déplacer plus rapidement que les objets éloignés. La vitesse apparente d'un objet informe donc de la distance de cet objet, de manière relative, quantitative et monoculaire.

Le mouvement apparent rétinien, qu'il provienne de l'objet ou de l'observateur, implique également l'apparition et la disparition de surface(s) derrière le contour d'un objet occultant en mouvement. Ceci constitue un autre indice lié au mouvement et appelé l'**occultation dynamique** du fait que les zones occultées et découvertes au cours du temps informent de la position ordinale d'une surface par rapport à une autre. L'occultation dynamique est donc un indice relatif et qualitatif. Ainsi l'occultation dynamique est à la stéréopsie de Da Vinci ce que la parallaxe de mouvement est à la disparité binoculaire : une information dynamique issue d'occultation(s).

### Les indices picturaux

Nous faisons l'expérience quotidienne de la position des objets dans l'espace tri-dimensionnel sans même noter qu'ils résultent de processus complexes impliquant différents indices de profondeur, dont ceux cités précédemment. Pourtant, notre expérience ne s'arrête pas en fermant un œil : nous continuons de percevoir un monde tridimensionnel, d'ailleurs même sans bouger ni accommoder. Les indices picturaux sont justement cet ensemble d'indices qui nous permet d'inférer la profondeur à partir d'éléments vues de façon statique et monoculaire.

La **perspective** ou projection centrale, expliquée par la géométrie projective, provient de la transmission et réflexion de la lumière sur des surfaces environnantes avant absorption et traitement par la rétine. Puisque l'œil fait correspondre à chaque point de l'espace visible un point sur la surface rétinienne, la perception visuelle peut-être assimilée à une projection, où la dimension des objets dans l'espace est convertie en dimension angulaire sur l'image.

Ainsi la convergence des lignes parallèles induite par la projection informe de la distance de ces lignes à l'observateur. Une distance constante entre droites parallèles dans l'espace apparait de plus en plus grande à mesure qu'on s'en approche, et de plus en plus petite à mesure qu'on s'en éloigne, jusqu'au point(s) de fuite.

La position relative des objets par rapport à l'horizon informe également de leurs distances. Le SVH s'appuie sur l'hypothèse généralement vraie que les objets reposent sur le sol et qu'ainsi plus ils sont proches de l'horizon plus ceux-ci sont éloignés.

L'indice de **taille relative** est déduit de deux objets supposés de tailles identiques mais projetant des tailles différentes sur la rétine en raison de leurs distances variables. Même si cet indice s'appuie sur une heuristique, il permet d'obtenir une information de profondeur relative mais précise entre deux objets supposés de taille identique.

Contrairement à l'indice précédent, l'indice de **taille familière** fait appel aux propriétés sémantiques d'objets connues de l'observateur dans son champs visuel . En effet, le plus souvent, une table est à 80 cm du sol, un homme mesure autour d'1 m 75 etc.. Si la taille d'un élément est connu du sujet, alors un simple calcul trigonométrique permet d'en retrouver sa distance approximative.

Le **gradient de texture** repose sur la répétition dans l'espace de motifs réguliers dont le changement de taille et de forme nous informe sur leurs distances. La taille des éléments est utilisée par le SVH pour estimer l'orientation et la courbure des surfaces texturées, mais au prix d'une hypothèse: les éléments de texture sont supposés de tailles identiques.

Les **ombres internes** (sur l'objet) et **ombres portées** (sur la surface sur laquelle il repose) permettent de définir la forme des surfaces, leurs courbures dans l'espace et donc le volume des objets. Selon que les ombres internes à l'objet soient situées en bas ou en haut de celui-ci, on le percevra respectivement comme convexe ou concave, car notre système fait l'hypothèse que la source principale d'illumination provient d'en haut. Cette hypothèse implicite est forte mais doit exister pour de bonnes raisons, car écologiquement notre environnement "naturel" est pratiquement toujours illuminé d'en haut. De plus les ombres portées fournissent une information relative et qualitative de la position des objets grâce aux variations lumineuses portées sur les surfaces.

À l'inverse, les différents degrés de luminosité issus de la **perspective aérienne** permettent de déduire la distance relative des objets la composant. La perspective aérienne apparait quand des objets sont vues de très loin ; leur contraste apparait diminué par les particules atmosphériques diffusant la lumière ou bien leur teinte apparait bleutée par le filtrage par l'atmosphère des photons de longueur d'onde courte. Ces différences locales de contrastes et couleurs apportent un indice supplémentaire pour déterminer la profondeur relative, à grande distance cependant.

Enfin, l'information d'**interposition** issue de l'occultation de surfaces les unes par rapport aux autres est primordiale car disponible quel que soit la distance de visualisation. De l'interprétation de contours d'objets en occultant d'autres, une profondeur relative de leur ordonnancement peut être obtenue. En particulier, il a été démontré que le SVH privilégie une séparation forme/fond pour distinguer quel objet est le plus proche de l'autre. Cette propriété perceptuelle opère selon la convexité, les contours fermés ou non, le contraste, la taille, l'orientation etc. des objets entre eux.

Il existe donc un très grand nombre d'indices spécifiques de profondeur. Leur combinaison est l'objet d'études approfondies car elle révèle des fonctions d'identification, de mise en correspondances et de fusion complexes afin d'arriver à une représentation cohérente de la disposition de notre environnement.

## Chapitre 2 : Intégration de la disparité binoculaire dans un modèle d'attention visuelle : de l'importance de l'avant-plan et du temps

La compréhension de la perception de la profondeur est d'importance lorsque l'on présente un indice de profondeur supplémentaire - la disparité binoculaire - au système visuel humain au travers d'écrans dit "3D". Le but de ce chapitre est d'étudier comment la disparité stéréoscopique agit et affecte le déploiement de l'attention visuelle, d'étudier un modèle d'attention visuelle considérant cette stéréoscopie et enfin de proposer un modèle cohérent incluant des caractéristiques visuelles de profondeur.

### Travaux Passés

Différents travaux de recherches récents ont suggéré d'inclure des attributs visuels additionnels à ceux proposés par les modèles d'attention classiques (couleur, luminosité, orientation) afin de les améliorer. On peut citer les modèles incluant la profondeur ou

certaines de ses caractéristiques, ceux incluant le biais de visualisation centré et enfin ceux incluant une multitude de signaux traités à différents niveaux hiérarchiques par le SVH.

Les premiers proposent que la profondeur globale, la disparité stéréo ou des éléments de la vision stéréo soient ajoutés comme caractéristiques supplémentaires à un modèle d'attention existant. Maki et al. [80, 81] proposèrent un modèle basé sur le flot optique et la détection de profondeur et mouvement. La profondeur est alors utilisée pour hiérarchiser la saillance de cibles préalablement sélectionnées. Ouerhani et al. [99] proposèrent d'inclure la profondeur brute comme caractéristique visuelle additionnelle au modèle d'Itti [56] un contraste de profondeur est alors calculé. Plus récemment, Zhang et al. [151] proposèrent un modèle stéréoscopique où la profondeur brute est combinée au mouvement et également aux cartes de saillance statiques de Itti. La fusion de ces trois attributs est réalisée par une combinaison linéaire avec des poids arbitraires. Une des rares tentatives considérant la perception stéréoscopique est le modèle d'attention stéréoscopique de Bruce et Tsotsos [14] traitant la rivalité binoculaire au travers d'une extension d'un modèle à ajustement sélectif.

Suite à de nombreux travaux soulignant le mécanisme initial de recentrage du regard au centre d'un stimulus [127, 10], différents modèles d'attention ont émergé pour prendre en compte de multiples caractéristiques visuelles de bas, moyen et haut niveaux. Judd et al. [62] proposèrent un modèle intégrant au contraste d'intensité, d'orientation et de couleurs d'autres attributs tels que l'a priori de centre (biais centré) et d'horizon, un détecteur de personnes et de visages etc. Une machine à vecteurs de support permet d'entrainer un modèle cumulant tous ces attributs à la fois montants (bottom-up) et descendants (top-down). Les résultats tout comme ceux de [151] soulignent l'importance du biais centré mais également relativise son importance si l'on considère les zones en son sein et en dehors.

Une première étude impliquant les contributions de la profondeur et du biais centré a été réalisée par Vincent et al. [136]. En plus de facteurs haut-niveaux tels que les sources lumineuses et les zones du ciel, les contributions de l'avant-plan et du biais centré sont étudiées quantitativement. Chaque carte caractéristique est binarisée, puis un algorithme d'Espérance-Maximisation (EM) est appliqué à partir des fixations des observateurs pour déterminer les contributions de chaque caractéristique. Les contributions, fixes au cours du temps, indiquent une prédominance de l'avant-plan et du biais central sur d'autres attributs tels que le contraste et les sources lumineuses pour prédire les fixations. Ho-Phuoc et al. [50] suivirent une méthodologie similaire pour étudier le rôle d'attributs bas niveaux du SVH au cours du temps. Pourtant, comme dans [136], le poids de chaque attribut était fixe au cours du temps.

### Expériences

La base de données d'image et de fixations oculométriques nous a été gracieusement fournie par Jansen et al. [58]. Les données acquises sont des photographies stéréoscopiques en niveaux de gris de scènes de forêt conjointes à des mesures de profondeur au travers d'un scanner laser à balayage. Vingt-quatre images sont alors présentées à quatorze participants sur un écran autostéréoscopique, soit en 2D, soit en 3D. En effet cet écran permet d'afficher deux vues distinctes pour chaque œil sans que le port de lunettes soit nécessaire. Un oculomètre Eyelink II enregistre le parcours oculaire de l'œil gauche de chaque observateur. La disparité binoculaire est introduite en remplaçant l'image gauche affichée sur les deux yeux par les images gauche et droite sur l'œil correspondant.

### Étude Comportementale

Jansen et al. ont montré que l'introduction de la disparité altérait les propriétés oculaires basiques tel que le taux de fixation, la longueur des saccades etc. et ce en début de visualisation. Les observateurs tendent à regarder les zones proches d'abord. Au travers d'une étude, nous poursuivons en montrant que la disparité n'a pas d'impact sur la variabilité inter-observateur, et ce quelques soient les intervalles de fixations considérés (dix ou vingt premières ou l'ensemble des fixations). Par contre, nous confirmons par des mesures statistique d'Aire sous la Courbe (AUC) que le degrée de similarité entre les fixations des observateurs en condition 2D et 3D augmente avec le temps : la présence de la stéréoscopie sur image fixe à un effet sur le parcours oculaire.

Il reste à quantifier plus précisément lequel. En étudiant en début de visualisation la distribution des fixations sur l'écran selon les 2 conditions 2D et 3D, on montre que le biais centré est présent et proche dans les deux cas. Une étude ANOVA 2x3 portant sur la distance au centre de l'écran avec les facteurs 2D-3D et trois intervalles de visualisation montre pourtant qu'il existe une différence statistiquement significative pour l'intervalle médium (de la 11ème à la 20ème fixation) et tardif (de la 21ème à la 30ème fixation). La disparité moyenne des fixations montre également que le facteur stéréoscopique est significatif mais pas le facteur temps. Un t-test de Bonferroni montre que la disparité a une influence en début de visualisation : on regarde effectivement de façon significative les zones les plus proche lors des 10 premières fixations.

Puisque la disparité a une influence non négligeable sur le parcours oculaire au cours du temps, nous avons ensuite cherché à évaluer la robustesse et la capacité de prédiction de 3 modèles d'attention visuelle existants (Itti [56], Bruce [13] et Le Meur [72]) à la nouvelle condition stéréoscopique. La baisse de performance attendue n'est pas si évidente : elle apparait effectivement mais faiblement avec le modèle de Itti, mais ne montre pas de distinctions claires pour les modèles de Bruce et Le Meur.

### Étude Statistique

Nous proposons d'intégrer différents attributs visuels dans un modèle d'attention dit "dynamique" : les contributions de caractéristiques visuelles retenues varient au cours du temps. Une étude statistique selon l'approche Espérance-Maximisation (EM) de Vincent et al. [136]est employée : le mélange additif de différents attributs, chacun associé à un poids donné, et pour un rang de fixation donné, est observé. Cette analyse vise ainsi à séparer au cours du temps les contributions de caractéristiques visuelles "montantes" ou bas-niveau d'autres attributs dits "descendants" ou haut-niveau.

Le biais centré est modélisé par une fonction gaussienne bi-dimensionnelle, en accord avec [151], et dont les propriétés sont fixes au cours du temps.

Le supposé biais de profondeur issu des observations précédentes est modélisé au travers d'une séparation forme/fond ou avant-plan/arrière-plan, selon les observations [114] et propositions [136] de la littérature. Un simple filtrage passe-bas et passe-haut des valeurs à mi-amplitude est appliqué sur les cartes de profondeur pour générer des cartes en niveaux de gris d'attributs d'avant et d'arrière plan.

La combinaison finale inclue donc les cartes de saillance selon le modèle de Itti, la modélisation du biais central, de l'avant-plan, de l'arrière-plan ainsi qu'un attribut dit "uniforme" modélisé par une distribution uniforme.

Les résultats de l'algorithme EM, c'est à dire les contributions de chaque attribut à l'attention au cours du temps, montrent d'importantes différences selon la présence ou

non de la disparité binoculaire. Le biais central est proéminent à la première fixation, mais s'estompe rapidement jusqu'à un niveau de contribution stable mais important dès la 3ème fixation et ce quelques soient les conditions. La disparité promeut l'avant-plan comme attribut participant à la saillance jusqu'à la 17ème fixation en condition 2D mais surtout en 3D. À l'opposé, l'arrière-plan participe très faiblement à la saillance et ce en 3D uniquement. L'attribut uniforme reste faible jusqu'à la période dite tardive de visualisation, il est alors difficile de conclure sur une éventuelle participation de l'arrière plan tardive. Cette analyse temporelle a été réitérée sur les modèles de Bruce et de Le Meur, cela donne des tendances entre attributs très proches.

Ainsi ces premiers résultats suggèrent que l'avant-plan participe à la prédiction des zones saillantes, en 2D mais d'autant plus en 3D : une caractéristique de profondeur telle que l'avant-plan contribue à l'attention même en 2D probablement car la profondeur peut être inférée d'indice de profondeurs monoscopiques : les indices picturaux. Ces résultats montrent également que l'attention est susceptible d'être liée au mécanisme perceptuel de séparation et d'organisation forme-fond.

**Vers des Modèles Dynamiques**

Puisque la combinaison linéaire dynamique de cinq attributs est établie, il est alors possible d'utiliser ces pondérations apprises pour générer des cartes de saillance "adaptées" prenant en compte des caractéristiques visuelles bas-niveau ainsi que les biais de centre et de profondeur.

Ainsi pour chaque fixation, une combinaison différente donne une carte adaptée temporellement. Cette combinaison est apprise sur la moitié des images de tests puis testée sur la moitié restante. Les cartes prédites sont alors comparées à la vérité terrain, c'est à dire aux fixations des observateurs pour un rang de fixation donnée. Par une mesure de similarité entre les fixations et la carte (le critère AUC moyenné sur toutes les images), les résultats montrent que les performances du modèle adaptée sont bien supérieures au performances du même modèle de Itti non-adapté.

Une analyse plus poussée est réalisée cette fois à partir des 3 modèles initiaux et avec l'ajout d'un autre critère de mesure, la saillance du chemin oculaire normalisé ou Normalized Scanpath Saliency (NSS). Une valeur NSS étant donnée pour chaque couple "image x fixation" par participant, les valeurs sont moyennées sur toutes les fixations des participants, images et rangs de fixation. Les résultats confirment que les modèles de Itti, Bruce et Le Meur qui combinent à leur saillance initiale le biais centré, l'avant plan et l'arrière plan améliorent significativement les modèles existants, et ce pour les deux critères AUC et NSS. La méthode proposée a donc amélioré la prédiction de manière significative pour les deux critères, les deux conditions et pour tous les modèles de saillance. Les performances étaient déjà améliorées en combinant les attributs avec des pondérations identiques, mais l'utilisation de pondérations apprises donne des résultats encore meilleurs.

Pourtant, la question se pose concernant la validité de ces résultats sur des images issues d'un contexte et perçues dans des conditions de test différentes. En effet, les poids ont été appris sur la moitié d'une base constituée uniquement de photographies de forêt, puis testés sur l'autre moitié. Le parcours oculaire mais également la perception de la profondeur des scènes a pu influencé le déploiement visuel. De plus, il ne s'agissait que de stimuli en luminance affichés en niveaux de gris, ce qui peut potentiellement influencer les performances des modèles initiaux et adaptés. Pour vérifier ces hypothèses, une test supplémentaire a été mis en place sur des images constituées soit des scènes urbaines ou soit de scènes de forêt. De plus les images étaient affichées en couleur. Les performances du modèle de Itti initial ainsi que de sa version adaptée selon notre approche mais sur des

images couleurs sont très proches. Par contre, le contexte urbain diminue les performances globale de notre méthode et invite à réaliser des améliorations: soit réapprendre les poids selon cet autre contexte ou bien intégrer des attributs supplémentaires. Dans tous les cas, même si le poids à donner à l'avant-plan et au biais centré peut être relativisé, l'intégration de ces caractéristiques dans un modèle de saillance a montré son efficacité.

Il semble acquis également que l'avant-plan est un élément participant activement en début de visualisation à l'exploration de la scène, c'est une caractéristique potentielle participant à un second niveau d'organisation forme/fond de l'attention bottom-up.

Enfin les résultats finaux soulignent l'apport qu'il y a à considérer temporellement - ou fixation par fixation - chaque caractéristique indépendamment. C'est une méthode qui se révèle cohérente au regard des mécanismes de sélection et de compétition mis en jeu par le SVH.

# Partie Vision par Ordinateur : Codage Vidéo 3D et Applications

Cette partie présente les représentations et standards existants de codage pour la 3D, puis introduit deux contributions portant sur le codage de cartes de profondeurs et sur la synthèse de vue par inpainting basé profondeur.

## Chapitre 3 : Introduction à la Représentation et au Codage 3D Multi-Vues

Dans ce chapitre un panorama des principales représentations et méthodes de compressions 3D est dressé. Ces représentations intermédiaires ne peuvent être expliquées sans leur format de données d'entrée, leur complexité ni même sans les compressions et rendus qu'elles permettent. Aussi ceux-ci seront explicités dans la mesure du possible. À la suite, les standards de compression 3D passés, présents et futurs sont décrits.

### Les représentations 3D

**Les Pré-requis** La 3DTV ou télévision en 3D implique qu'un ensemble de vues et donc de vidéos doivent être affichées en temps-réel, pour un parc d'écran varié. La captation, la transmission et l'affichage de multiples vues permettent ainsi d'améliorer le confort visuel ou d'augmenter le nombre de téléspectateurs potentiels devant un écran autostéréoscopique.

Ainsi la qualité et la cohérence du flux vidéo parmi les multiples points de vue doivent être garanties. Il est désormais admis que l'adoption des écrans 3D ne se fera que si la qualité perçue est au moins égale à celle perçue en 2D en haute définition. Enfin, la progressivité et la rétro-compatibilité c'est à dire la capacité d'un couple représentation-codec à rester compatible avec un logiciel ou matériel précédent est également à prendre en compte.

La télévision à point de vue libre (ou FreeViewpoint TV ou FTV en anglais) est une technologie complémentaire à l'usage de multiples points de vue et permettant à l'observateur de contrôler de manière interactive son point de vue dans une scène vidéo. Ici encore, plus le champ de caméra est dense et large et meilleure sera la qualité de reconstruction. Pourtant, un compromis doit être trouvé selon l'application entre le nombre de caméras, la capacité des réseaux et la faible complexité des décodeurs. Le niveau de détail restitué, la compression et la flexibilité des représentations-compressions sont des facteurs primordiaux pour la FTV comme pour la 3DTV.

Il existe différentes représentations qui assurent la mise en forme de données acquises par de multiples caméras, les représentations basées :

- Images : il s'agit de la vidéo **stéréoscopique** conventionnelle, telle que diffusée par les chaines de télévision 3D actuelles, et la vidéo **multi-vues** (ou MVV). Cette dernière est plus flexible car elle permet d'afficher plusieurs paires locales de vues stéréo selon la position de l'observateur à l'écran par exemple. On peut également citer les fonctions plénoptiques qui visent à obtenir une connaissance étendue d'une scène au travers d'une grille (et non plus un alignement) de caméras. Cette technique est actuellement très coûteuse tant du point de vue du studio que du téléspectateur.

- Images et Profondeurs : elles offrent un bon compromis coût/flexibilité par la transmission de la géométrie de la scène au travers de cartes de profondeurs. En effet,

avec la texture et la profondeur d'une vue -"2D+Z"- on peut extrapoler d'autres vues adjacentes par projection. Avec de multiples vues de textures et de profondeurs transmises - on parle de vidéo multi-vues plus profondeur (ou MVD) -, un rendu basé image et profondeur (DIBR) permet de synthétiser de nouvelles vues intermédiaires et donc de réduire le nombre de caméras nécessaires. Pour éviter d'avoir à transmettre le contenu fortement redondant spatialement (entre vues), une alternative consiste à représenter l'information sous forme de couches ou de plans de texture et de profondeur (ou LDI). Une LDI est une matrice 3D composée à la fois de pixels visibles et occultés dans le même point de vue. Les LDI et ses extensions (I-LDI et DES) ont l'inconvénient d'avoir des couches où les pixels sont disposés de façon éparse, ce qui n'est pas évident à compresser à moins d'utiliser une approche basée bloc compatible avec des mécanismes existants de prédictions-compensations.

- Surfaces : les maillages de polygones sont largement utilisés dans la communauté image de synthèse et infographie tridimensionnelle, au point d'être des primitives manipulables par les cartes graphiques d'aujourd'hui à l'échelle de millions. Mais de tels ensembles sont coûteux à transmettre et des techniques proposent la simplification par la progressivité (spatiale ou temporelle) de maillages. La NURBS pour surface B-spline rationnelle non-uniforme est un élément de représentation adéquat pour les logiciels de conception assistée par ordinateur du fait de leur définition géométrique : une surface polynomiale continue par morceaux. Par contre, le raffinement ou la déformation locale de NURBS nécessite des modifications à l'échelle de tout le modèle, ce qui en fait une représentation inadéquate aux contraintes de compacité et de déformations temporelles de la 3DTV et de la FTV.

- Soupe de Polygones : à partir de données MVD, l'idée est de tirer parti des primitives polygonales (compacité, continuité de surface, etc.) pour représenter la géometrie de la scène à la place des cartes de profondeur. Les vidéos de textures ou couleurs sont donc représentées et transmises selon les standards video 2D, mais une soupe de polygones est créée à partir des cartes de profondeur par une décomposition quadtree. Les quads sont récursivement divisés selon leur discontinuité en profondeur ou leur mauvaise approximation de la géométrie de la scène. À l'étape de rendu (ou synthèse de vues), une triangulation permet l'élimination de "cracks" alors que les effets "fantômes" ont été éliminés pendant une étape de réduction du quadtree. Finalement cette représentation permet un rendu de qualité en transférant la complexité à l'étape de construction de la soupe de polygones et non à l'affichage.

- Points : les surfaces de points, particules ou éléments de surfaces (surfels) peuvent également être utilisés en tant que primitives de représentation basées surface. À la place d'un ensemble de points échantillonnés sur les surfaces, les normales aux surfaces et les couleurs de ces points sont enregistrées. Au rendu, un étalement de ces points ou "'splatting" permet de rendre des scènes plus complexes qu'avec l'usage de polygones. En outre, la technique du "splatting" permet de s'affranchir des cracks et des effets de crénelage ("aliasing"). La qualité finale de synthèse fait de cette méthode une bonne candidate pour la 3DTV et FTV.

- Volumes : les représentations volumétriques consistent en une décomposition de la scène en unités de volume : soit par des modèles discrets, soit par des modèles basés primitives. Ces derniers combinent des cylindres, des "supershapes", "hyperquadrics" et autres formes polynomiales sous forme d'opérations agrégées par un graphe. Malgré leur compacité, il s'avère difficile de restituer des formes et objets

naturels au travers de ces primitives. Les modèles discrets décomposent à l'inverse l'espace en unités de volume appelés voxels, dont la taille peut être raffinée jusqu'à une représentation correcte de la scène. Chaque voxel contient les propriétés de segment de surface la composant. Il est alors codé au sein d'une structure de données en arbre appelé "octree". L'avantage de cette représentation en voxel dans un contexte multi-vues est qu'un seul et même modèle est codé, l'accès aux voxels voisins est rapide par la structure en arbre, et le rendu est donc peu complexe. Pourtant, l'unité cube limite la qualité, en particulier lorsque la caméra est trop proche des surfaces basées cube et à la différences des polygones.

Ainsi il existe des façons très variées de représenter l'espace d'une scène, chaque méthode convenant à des usages spécifiques. Il semble à l'heure actuelle que les représentations basées images et images plus profondeurs soient les plus prometteuses du fait entre autre de la rétro-compatibilité qu'elles permettent. Les formats de codage de ces représentations basées images (stéréo et multi-vues) et images plus profondeurs (2D+Z, Multi-vues plus profondeur) vont être détaillés dans la partie suivante.

### Les standards vidéos 3D

Deux types de standards vidéos 3D peuvent être distingués, ceux reposant sur des codecs vidéos existant et les améliorant par une fonctionnalité 3D et ceux "non-restrictifs" ne spécifiant pas d'algorithmes de codage particulier mais plutôt une boite à outils pour le conteneur vidéo. Les codecs basés représentation stéréo vont être d'abord décrits avant de détailler ceux basés image(s) plus profondeur.

**Les codecs stéréo** L'idée principale de ces codecs est d'exploiter la forte redondance spatiale existant entre la vue gauche et droite.

- Le MPEG-2 MVP repose sur le standard MPEG-2 et y ajoute un codage scalable ou la vue supplémentaire est codée au niveau du GOP (Group of Pictures) par une prédiction hybride basée à la fois sur le mouvement et la disparité entre les deux vues. Cette vue peut être décodée ou non et garantie donc la rétrocompatibilité avec le profil "Main" de MPEG-2.
- Le message SEI de H/264/MPEG-4 AVC est également une amélioration de H.264 qui vise à supporter la vidéo stéréo mais pas uniquement (composition alpha, augmentation de la précision de codage etc.). Mais à la différence de MPEG-2 MVP, le signalement SEI distingue les vues gauches et droites au sein du train binaire par le signalement à 1 ou 0 des vues gauches ou droites.

### Les codecs 2D + Profondeur

- Le MPEG-4 MAC définit des composantes auxiliaires multiples (MAC) qui peuvent être employées pour diverses applications. Ces composantes en niveaux de gris peuvent servir à décrire soit la transparence d'un objet vidéo, sa forme, la texture (luminance plus chrominance) qu'il occulte, et donc également la disparité ou la texture d'une vue additionnelle.
- Le MPEG-4 AFX spécifie la manière de représenter un modèle graphique 3D. Les objets synthétiques de haut-niveau sont spécifiés par leur géométrie, texture, et leur

codage. Ce format est compatible avec la réprésentation LDI car il spécifie une structure à base d'images en profondeur elles-mêmes composées de points de texture et de profondeur.

- Le MPEG-C Part 3 décrit un format de représentation et codage basé image plus profondeur. Les cartes de profondeurs sont encodées comme des séquences 2D conventionnelles auxquelles on signale leurs paramètres. Mais le standard de compression des images et profondeurs n'est pas spécifié en tant que tel. Ce standard vise donc à garantir l'interopérabilité des contenus indépendamment des technologies de transmission.

**Un codec 3D multi-vues : MVC** Ce format de représentation et codage de multiples vues a été développé comme une extension de H.264/MPEG-4 AVC. Il exploite les redondances inter-vues par la réutilisation de mécanismes tels que la prédiction de vecteurs de mouvement. Le flux principal de la vue de référence est codé indépendamment pour garantir la rétro-compatibilité.

La redondance inter-vues est exploitée par l'estimation et la compensation de disparité (ou DCP). Ainsi une image de type B (Bi-directionnelle) d'une vue hors référence peut être prédite à la fois de ses images voisines temporellement mais également des images des vues adjacentes. Les images de type P des vues hors références sont obtenues par compensation de disparité à partir des ancres (de type I) dans la vue de référence.

**Un codec 3D multi-vues plus profondeur : 3DVC** 3DVC est une solution récente qui vise à permettre la compression et le rendu d'un nombre arbitraire de vues plus profondeur. Différentes versions existent, mais les quelques fonctionnalités du codec 3DVC basé sur HEVC vont être décrites :

Le train binaire est arrangé de manière à ce que la vue de référence puisse être facilement extraite. De plus, l'encodeur peut être réglé pour que les vidéos dites de "texture" soient codées indépendamment des vidéos de profondeur.

Concernant le codage de textures, on peut citer :

- la prédiction compensée en (disparité comme MVC),
- la prédiction inter-vues basée sur une synthèse de vues pour déterminer si une unité de codage (ou CU) doit être codée ou non sur les zones découvertes.
- la prédiction de mouvement inter-vues, ceci afin d'exploiter le fait que les vecteurs de mouvements entre des blocs correspondants dans différentes vues sont très similaires
- la prédiction du résidu inter-vues. Il s'agit donc d'une prédiction des erreurs de prédiction...

Concernant le codage de profondeur, on trouve :

- L'usage de 4 modes pour la prédiction intra de cartes : basé wedgelet avec ou sans prédiction en intra à partir de la texture ou de la profondeur, ou basé contour.
- L'héritage des paramètres de mouvement issus de la texture. Puisque un point dans l'espace est associé à un pixel définit par une valeur unique de luminance, chrominance et profondeur, son mouvement sur l'image de couleur sera identique sur la carte de profondeur, on peut donc le déduire des vecteurs de mouvements de l'image.

- Une optimisation débit-distorsion basée synthèse de vue : la fonction de coût lagrangien est adaptée pour inclure les distorsions sur la vue synthétisée dues à différentes quantifications des CU d'une carte de profondeur donnée.

Ce chapitre a donc souligné la multiplicité des représentations et des codages associés. Les représentations basées images plus profondeurs apparaissent comme un bon compromis complexité, qualité de rendu, facilité de mise en œuvre et coût de compression.

## Chapitre 4 : Compression de Cartes de Profondeur basée Contour pour le Codage 3D

### Introduction

Le développement de méthodes de compressions de cartes de profondeurs s'est récemment accéléré par l'émergence de solutions 3D diverses. La transmission de la géométrie de la scène est nécessaire au rendu de nouvelles vues basé profondeurs (DIBR) pour un affichage adapté au type d'écran stéréoscopique, à sa distance de visualisation, à sa technologie, etc.

Il apparait donc comme essentiel de compresser efficacement et de reconstruire précisément les cartes de profondeurs pour générer des points de vues virtuels de qualité : la distance aux objets, leurs surfaces mais surtout leurs contours doivent être préservés. Une inadéquation locale entre des valeurs de pixels de couleurs et de profondeurs induira nécessairement des distorsions à l'étape de rendu, plus ou moins visibles selon la disparité mais aussi selon les différences chromatiques entre objets. Une distorsion à la frontière entre deux objets engendrera des artefacts potentiellement très visibles car impliquant des couleurs d'avant-plan sur l'arrière-plan et inversement.

Dans ce chapitre une méthode alternative au codage de profondeur basé wedgelet est proposée : une compression de carte basée codage des contours sans perte.

En effet les cartes de profondeurs possèdent deux caractéristiques importantes à préserver mais sur lesquelles on peut s'appuyer pour un codage efficace. Les propriétés de régularité voir d'uniformité en profondeurs des surfaces des objets tout d'abord : des larges zones de profondeur peuvent être interpolées au sein d'un même objet. La seconde propriété intéressante est la présence de contours souvent abrupts entre deux objets et donc entre deux plans de profondeurs.

Alors que Merkle et al. [88] proposent que ces régions présentant des gradients soient approximées par des fonctions linéaires par morceaux mais séparées par des segments, nous proposons que ces surfaces puissent être entièrement reconstruites par l'interpolation des valeurs de pixels situés aux contours des surfaces et donc des objets.

Nous revisitons la proposition de Mainberger et al. [79], consistant en un codage de contour d'image de type "cartoon", en apportant des améliorations pour répondre aux contraintes de haute qualité et faible débit du contexte MVD.

Dans la prochaine section, le processus d'encodage est décrit, puis le décodage et l'interpolation sont explicités. Les résultats et performances face à des codecs de l'état de l'art sont donnés selon des méthodes d'évaluation de qualité objectives et subjectives, ceci afin de donner du poids à cette approche que nous estimons répondre à des critères perceptuels.

### Encodage

L'encodage est réalisé en 5 étapes : la détection de contours, l'encodage de la position des pixels, et l'encodage des valeurs en luminance des pixels de contours, des graines et des

bords de la carte.

**La détection de contours** doit répondre à des critères qualitatifs que sont : une détection correcte (l'algorithme doit trouver autant de vrais contours que possible), une bonne localisation (les contours doivent être détectés et positionnés aussi proches que possible des vrais contours) et une bonne robustesse (la détection de contours doit être autant que possible insensible au bruit).

Dans notre contexte, on doit maximiser la qualité de reconstruction par interpolation tout en minimisant le nombre de contours à encoder. Des tests ont montré que la détection de contours selon la méthode de Canny n'était pas assez précise. Elle peut omettre des pixels de contours et donc provoquer lors de l'interpolation une propagation des valeurs de profondeur entre les surfaces. Ceci est dû à l'étape de pré-filtrage dans la méthode de Canny visant à réduire le bruit.

Nous proposons d'utiliser la méthode de Sobel qui garantie une bonne détection et une bonne localisation. Afin d'éviter l'apparition de contours dus à une sur-détection, ceux-ci sont filtrés selon leurs longueurs. La quantité de contours et donc le débit final varie donc selon le seuil de détection de contour choisi.

**L'encodage de la position de contours** est réalisé par le standard JBIG (Joint Bi-level Image experts Group) dédié à l'encodage d'images binaires.

**L'encodage des valeurs de contours** part de l'observation que les valeurs en luminance des pixels de part et d'autre des contours varient peu et donc minimise l'entropie. Nous sauvegardons les valeurs de pixels de contours selon leurs occurrences le long du contour par un chemin à priorités directionnelles fixes. Puisque les valeurs d'un pixel de contour à l'autre (connexe) varient peu, nous sauvegardons les valeurs différentielles ou résidus (DPCM). Ce train binaire est alors encodé avec un codeur arithmétique. À ces contours sont ajoutés deux choses : les valeurs de pixels aux bords de l'écran et des valeurs de graines disposées soit selon une grille régulière, soit selon une décomposition quadtree. Cette dernière à l'avantage de réduire le nombre de graines en les disposant au voisinage des contours. Les deux méthodes sont testées.

### Décodage et interpolation

**Le décodage des positions et valeurs de pixel de contours** est réalisé successivement. Une fois que l'on sait où se trouve les contours, les valeurs originales sont affectées dans le même ordre qu'à l'encodage.

**Reconstruction des valeurs manquantes** : à ce stade la carte n'est constituée que de pixels aux bords, sur les contours, et sur les graines. Nous avions initialement proposé de remplir les surfaces par une diffusion basée sur les équations différentielles partielles (PDE). Au vu des résultats, nous présentons la seconde approche. Les valeurs manquantes sont récupérées par une simple interpolation bilinéaire.

## Résultats Objectifs

Les performances de notre méthode sont comparées à celles de JPEG-2000 et au récent standard 2D vidéo HEVC en cours de normalisation (HM 4.1 Intra). Nous utilisons la séquence MVD Breakdancers de Microsoft.

Tout d'abord visuellement à qualité objective équivalente, les cartes et surtout les vues synthétisées apparaissent bien meilleures et ce avec des graines disposées selon une grille. En étudiant les performances débit-distorsions selon le critère PSNR, nos cartes reconstruites se placent entre HEVC et JPEG-2000 (avec le meilleur placement de graine selon un quadtree).

Par contre, les performances sur la qualité des vues synthétisées (à partir des deux cartes compressées selon l'une des méthodes et à partir de deux textures originales) dépassent à la fois JPEG-2000 et HEVC dans certaines conditions (pour un débit entre 0.12 et 0.22 bpp sur "Breakdancers").

Au vue des améliorations de notre méthode, visibles mais peu observables selon un critère de qualité objectif tel que le PSNR, nous avons décidé de comparer notre approche à d'autres par des tests en qualité subjective.

### Résultats Subjectifs

Les tests ont été réalisés dans des conditions semblables aux tests mis en place par MPEG pour la normalisation de 3DVC. Les mêmes séquences ont été utilisées. Les tests ont eu lieu au laboratoire IVC de l'IRRCyN de Nantes, et impliquaient l'évaluation de différentes méthodes de codage de profondeur.

L'objectif était donc de mesurer et comparer l'influence des méthodes de compression de profondeurs sur la qualité perçue par les observateurs. Les méthodes retenues pour le codage de profondeur sont 3DVC, H.264, HEVC, JPEG-2000 et notre méthode.

De multiples images décalées spatialement mais fixes temporellement ont été synthétisées à partir de deux uniques couples d'images originales et de profondeurs compressées selon différentes méthodes. Cette succession d'image est concaténée en une vidéo et affichée sur écran aux observateurs. Ce type de vidéo laisse le temps aux observateurs d'apprécier les différentes distorsions produites par une méthode mais pour différents points de vues.

**Conditions de tests** Les tests ont été réalisés par 27 sujets en deux sessions. Chaque sujet devait juger de la qualité d'une vidéo en lui attribuant une note de qualité de 1 (mauvaise) à 5 (excellente), selon la méthodologie ACR-HR. Une vidéo de référence (issue de deux cartes de profondeurs originales) était cachée et affichée parmi les autres vidéos. Le score de chaque vidéo était normalisé par rapport au score de cette vidéo de référence en un score d'opinion moyen normalisé (ou DMOS).

Trois versions de vidéos par méthode ont été préalablement sélectionnées selon qu'elles se rapprochaient d'un rendu de qualité bas, moyen ou élevé (trois classes).

**Observations** Les scores d'opinion moyennés sur les 27 observateurs indiquent différentes tendances selon chaque séquence vidéo.

Tout d'abord, ces tendances sont proches selon que l'on effectue un rendu avec l'algorithme VSRS (View Synthesis Rendering Software) avec le paramètre de fusion activé ou non. La méthode de codage de carte 3DVC donne généralement les meilleurs résultats parmi les classes de basse qualité de vidéo mais peut aussi présenter de loin les plus mauvaises qualités (sur les vidéos "Balloons" et "Kendo" non fusionnées). Ceci soulève la question du meilleure compromis à trouver entre la qualité visée et le bitrate alloué à la profondeur.

Il apparait comme difficile de conclure sur la qualité des méthodes pour les classes de vidéos moyennes, car leurs performances se situent généralement dans l'intervalle de confiance de la référence cachée.

Pour les classes de qualité élevée notre méthode ainsi que H.264 et HEVC parviennent à améliorer la qualité visuelle perçue par rapport à l'usage de cartes originales. Ceci peut s'expliquer par les filtrages implémentés dans H.264 et HEVC, et par le codage sans perte des contours mais filtrant les possibles aspérités de surface selon notre méthode.

Enfin notre méthode dépasse les autres en terme de débit/qualité subjective uniquement pour les séquences "Balloons" et "Book Arrival" au sein des classes de qualité moyenne.

#### Conclusion

Une méthode alternative aux approches basées bloc a été présenté dans ce chapitre, qui consiste en un codage sans perte et séparé de la position des pixels de contours de profondeurs et de leurs intensités.

L'usage de graines disposées selon une décomposition quadtree ainsi qu'une reconstruction par interpolation permet d'améliorer les performances de notre méthode et de la rendre compétitive face aux méthodes de compression images et vidéos présentes et à venir.

Les tests subjectifs confirment la pertinence de l'approche mais aussi ses limites à très bas débit. Ainsi le codage de contours globaux de cartes ou son partitionnement pour un codage également basé contours donnent dans les deux cas de bonnes performances. Il est très probable qu'à l'avenir des améliorations notables puissent être obtenues en considérant les propriétés écologiques de la scène, de ses objets et de son contexte.

### Chapitre 5 : Synthèse de Vues basée Inpainting pour la 3DTV

### Introduction

Les techniques de synthèse de vues se situent en dernière position de la chaine 3D avant l'affichage. Celles-ci visent une reconstruction de nouveaux points de vues, dits "virtuels" car non transmis, avec la meilleure qualité possible.

Un nouveau point de vue peut être obtenu soit par interpolation à partir de multiples points de vues (généralement deux), soit par l'extrapolation à partir d'un seul point de vue.

Nous proposons ici une méthode de synthèse de vue par extrapolation basée sur de l'"inpainting" pour le remplissage de régions découvertes. Les problèmes rencontrés par la synthèse suite à une projection d'une vue uniques puis les méthodes d'inpainting dans ce cadre DIBR vont tout d'abord être décrites avant d'introduire notre approche. Les résultats seront évalués selon plusieurs métriques de qualité objective.

#### Les artefacts de projections

La géométrie partielle de la scène, issue de la profondeur d'un seul point de vue, ainsi que sa texture correspondante ne suffisent pas à synthétiser un nouveau point de vue virtuel. Différents artefacts vont apparaitre:

- Les "cracks" ou pixels éparses inconnus dans la vue projetée car issues d'une image et profondeur échantillonnée selon une grille régulière dans la vue d'origine. Ils sont souvent pré-traités par un filtre médian ou moyen.

- Le “ghosting” ou effet fantôme, dû à des valeurs de pixels mélangeant les couleurs de l'avant et de l'arrière-plan qui après projection apparaitront en décalé et donc de façon très visible soit sur l'avant-plan soit sur l'arrière-plan. Le ghosting est prétraité par la suppression du contour de profondeur sur les zones de texture soumises à du potentiel ghosting.

- Les régions découvertes ou désoccultées, apparaissent pour les mêmes raisons qu'expliquées au chapitre 1 avec la stéréopsie de Da Vinci. Ici, les zones occultées par un avant-plan dans la vue originale apparaitront partiellement découvertes dans un nouveau point de vue. Il s'agira donc de retrouver les éléments de surface contigus à l'arrière-plan nouvellement découverts.

Alors que les deux premiers artefacts sont désormais facilement traités, il n'existe pas de méthode optimale pour le remplissage des zones découvertes au sens où elles ne sont jamais connues mais doivent être remplies.

### Techniques d'Inpainting

L'inpainting vise justement à corriger ou modifier une image d'une manière indétectable au sens perceptuel. Ses applications sont variées, allant de la restauration de photographies anciennes aux récentes techniques de suppression d'objets, de codage d'image, ou encore de résilience aux erreurs de transmission.

On distingue les approches basées pixels des approches basées “template” ou motif. Ces dernières sont détaillées car elle permettent la propagation de texture de surface.

L'importance de l'ordonnancement des patchs sélectionnés dans le voisinage connu a été démontré récemment. Les techniques basées exemple (examplar) ou gabarit (template) proviennent de cette idée de copier-coller des éléments de texture du voisinage selon une certaine priorité.

Criminisi et al. [24] proposent de remplir les régions de textures manquantes en priorité là où l'isophote (ou ligne de niveau) de texture atteint le bord de la zone inconnue de manière frontale (ou orthogonale). La priorité est donc calculée comme le produit d'un terme de confiance, illustrant la quantité d'information fiable autour d'un pixel, autrement dit la fiabilité du patch considéré, avec un terme de donnée introduisant de plus grandes priorités aux patchs dont la structure est frontale aux bords découverts. Les structures sont donc propagées en priorité dans les zones à remplir. Les patchs à recopier sont choisis selon un critère de similarité (norme L2) aux éléments connus du patch considéré.

On peut citer deux approches qui reposent sur cette méthode d'inpainting pour synthétiser des nouvelles vues en complétant les zones découvertes suite à une projection. Oh et al. [97] suppriment les frontières d'avant-plan et les remplacent par les frontières d'arrière-plan du côté opposé de la région découverte. Cette méthode repose sur une hypothèse de connexité entre les zones occultantes et occultés, ce qui n'est pas vérifiée pour les projections de vue éloignées.

Daribo et al. [30] ajoute à la méthode de Criminisi un terme de régularisation dans le calcul de priorité. La priorité d'un patch est mise à jour avec ce terme définit comme la variance inverse de la profondeur du patch. Cela augmente la priorité et donc la propagation de patch situés au même niveau de profondeur mais n'empêche pas de débuter la propagation de l'avant-plan.

## Approche proposée

Nous proposons une nouvelle approche qui vise à reconstruire de façon imperceptible les zones découvertes à partir de la texture des surfaces d'arrière plan. Puisque dans le contexte de synthèse les vues seront projetées selon l'axe horizontal, alors les régions vont se découvrir à partir du côté de l'avant-plan similaire à la direction de projection. Ainsi donc si l'on veut s'assurer de la propagation de texture depuis l'arrière-plan, alors il faut donner la priorité aux pixels du bord opposé à l'avant-plan.

En se basant sur cette observation, nous proposons une méthode de remplissage de régions découvertes pour la synthèse de vue basée sur une propagation de structure robuste et directionnelle.

**Tenseur :** Nous remplaçons le terme de donnée de Criminisi basé sur le gradient de couleur par un calcul de tenseur plus robuste. Ce tenseur est également filtré par un noyau Gaussien afin d'être plus robuste au bruit. Ainsi, plutôt que de calculer le produit vectoriel de l'isophote avec la normale au contour, nous visons à propager les plus fortes structures en calculant à partir du tenseur son orientation principale par le vecteur propre $v_2$ et son intensité par son vecteur propre $\lambda_2$. En se basant sur la norme de cohérence définie par Weickert [142], nous introduisons la différence des valeurs propres $\lambda_2$ et $\lambda_1$ ainsi que la norme de $v_2$ dans une nouvelle fonction de calcul du terme de donnée.

**Tenseur 3D :** Le tenseur 3D permet la diffusion de structure non seulement présentes sur la texture couleur mais aussi sur la profondeur. Au tenseur de structure précédent est ajouté la composante profondeur.

**Directionnalité** Comme énoncé, nous visons à propager la structure à partir du bord connexe à l'arrière-plan, soit le côté du bord identique à la direction de projection. Le carte de priorité est simplement mise à zéro sur les côtés du bord haut, bas et opposés à la direction de projection.

Ces trois propositions amènent effectivement à initier la propagation d'une structure robuste depuis l'arrière-plan.

**Recopie de patchs** Une fois que l'on sait effectivement où débuter la propagation, il est alors important de trouver les meilleurs patchs candidats et ce dans l'arrière-plan. Nous restreignons la recherche de patchs candidats au même niveau de profondeur du patch considéré en introduisant dans la métrique de similarité la composante de profondeur. La minimisation de la somme des erreurs quadratiques est donc réalisée à la fois sur les composantes Rouge Vert Bleu et sur la composante profondeur pondérée par trois pour donner autant de poids à la similarité en texture qu'à la similarité en profondeur.

## Résultats

Les résultats de notre méthode sont d'abord évalués et comparés visuellement à d'autres méthodes de l'état de l'art. Puis la capacité des méthodes d'évaluation de qualité objective 2D à juger des dégradations de reconstruction pour différents paramètres de notre méthode est étudiée.

**Comparaison subjective et qualitative** Notre méthode est évaluée sur la séquence MVD "Ballet" de Microsoft, qui à l'avantage de présenter un fort écart entre des caméras non rectifiées d'une part et qui possède différentes orientations de structures dont la propagation s'avère a priori complexe. La vue originale 5 est projetée dans la vue 4 et dans le vue 2 de manière à tester la capacité des méthodes à reconstituer des régions découvertes de plus en plus vastes.

Le rendu avec la méthode de Criminisi n'est pas correct et introduit de sévères artefacts dans les régions remplies. La méthode de Daribo, qui elle repose sur la profondeur, donne de bon résultats lorsque l'on synthétise une vue proche mais échoue à remplir de manière correcte les zones découvertes sur une vue plus éloignée.

Notre méthode donne visuellement les meilleurs résultats, les artefacts sont imperceptibles lors de la synthèse d'une vue proche, mais deviennent visibles pour une vue plus éloignée. À cette distance de la caméra originale, non seulement des distorsions apparaissent et un quart de la surface de l'image est à remplir. Les surface d'arrière-plan sont correctement remplies mais les "coutures" ou liaisons entre régions découvertes et régions connues deviennent visibles. Une amélioration pourrait consister à flouter ces liaisons sans affecter les structures remplies.

**Etude de la pertinence des métriques de qualité 2D** Nous cherchons ici à vérifier que les métriques de qualité 2D peuvent indiquer les dégradations engendrées par différents paramètres de notre méthode: la taille du patch et la taille de la fenêtre de recherche des meilleurs patchs candidats. Ceci a été effectué sur la séquence rectifiée "Balloons" du corpus de vidéos tests de MPEG avec la métrique "PSNR", l'index de similarité de structure "SSIM", l'index de fidélité de l'information "VIF" et enfin le rapport signal-bruit visuel ou "VSNR".

La taille de patch a été variée de $2 * n + 1$ avec $n$ variant de 2 à 14 pixels, et la taille de fenêtre de $2 * N + 1$ avec $N$ variant de 30 à 60 pixels.

Les résultats subjectifs et objectifs semblent être concordant : le meilleure rendu visuel est obtenue avec la taille de fenêtre la plus petite et la taille de recherche la plus grande. Ceci est vrai dans les limites des effets des tailles observées. De plus les distorsions apportées par les cartes n'ont pas été considérées. Des tests ont été réalisés sur les autres séquences du corpus MPEG et donnent des tendances semblables. La pente de qualité visuelle chute avec l'usage de patchs plus grands.

Il apparait donc possible d'évaluer, mais dans un intervalle mesuré très faible, la qualité de synthèse de vue d'une méthode pour ses différents paramètres avec les métriques de qualité objective de l'état de l'art. Cependant le développement d'une métrique de qualité de synthèse "3D" focalisée sur une reconstruction perceptuellement correcte des régions découvertes permettrait de mieux cibler et améliorer la qualité.

## Conclusion

Une méthode de synthèse de vue par extrapolation mono-vue a été proposée. Celle-ci se base sur une méthode d'inpainting par propagation des structures et textures d'arrière-plan. Les résultats visuels confirment la validité de l'approche par rapport aux méthodes de l'état de l'art.

Enfin une étude supplémentaire montre que l'on peut potentiellement utiliser des métriques de qualités objectives 2D pour estimer automatiquement le meilleur paramétrage d'une méthode de synthèse par inpainting. Une amélioration à notre méthode consisterait

à combiner des patchs de différentes tailles et résolution pour une reconstruction encore plus fidèles des zones découvertes.

## Conclusion

Ce mémoire propose diverses contributions à la compréhension de la vision spatiale et ses applications aux systèmes 3D.

La première partie propose d'étendre la capacité de prédiction des modèles d'attention visuels actuels en y intégrant des attributs liés aux biais de visualisation et à la perception de la profondeur dans la scène ; ceci afin de pouvoir s'adapter à des conditions de visualisation de contenus 2D ou 3D.

La seconde partie apporte des contributions à la transmission et au rendu de vidéos multi-vues plus profondeur. Une méthode de compression efficace de codage de cartes de profondeur basée sur le codage de contours sans perte donne des résultats subjectivement prometteurs. Enfin une nouvelle méthode de synthèse de vues basée inpainting directionnel solutionne les limites de l'extrapolation de vues en permettant une reconstruction de qualité.

# Preamble

This thesis is dedicated to the understanding of the human spatial vision and its applicability to the current trend on viewer experience improvement by offering stereoscopic visual content of quality.

The first part focuses on the description of what we know of the dedicated spatial functions of the human visual system i.e. the depth perception and the impact of depth cues on the visual attention. To this end this first part can be regarded as a human vision community subject, in the sense of vision psychologists: how well the algorithm models the human performance in controlled experiments.

The second part of this work addresses the coding and rendering issues of the emerging 3D television. To this goal, several proposals on coding chain standards and rendering issues have been offered. This image coding part clearly targets the computer vision and image coding community in the sense of determining how well the algorithm works in producing useful and quality perception.

I emphasized the traditional concerns of coding community, but clear concerns and links exist with the first part: a definition of qualitative, visually and perceptibly good image rendering. Key perceptual aspects to represent to the end-user will be described and solutions proposed.

However the reader might unfortunately not find a direct application of visual attention to the traditional balance of minimal-cost image transmission at best quality. Nevertheless this work will try to blaze a trail on these still limited but highly-promising possibilities.

What guides this thesis is the comprehension of vision as a biological perceptual process of knowledge acquisition, on an information processing basis. Then these two parts address the vision issue as it is indeed: an interdisciplinary fascinating domain that covers many endeavours.

# General Introduction

In the course of history, human vision has always drawn attention as a perception so very powerful that it was believed to be obvious and veridical as a "window onto reality". Initially described as an inner fire that gave rise to rays emanated from the eye towards a perceived object (Plato, 4th century BC), it took a while to "see" the eye as a pinhole camera where the light is reflected from object surfaces and enters into the eye (with the arabic philoopher Alhazen, 10th century AD).

Thanks to the development of geometric optic in the 17th century, it was understood that the light - after processes such as reflections - enters into the eyes through lenses. However the process of vision is much more than that. The modern vision theorists precisely struggled on the possible mechanisms underlying vision (How?) and its purpose (What for?). More than a sensor, the human visual system gives a veridical but constructed information about the world around us. The fundamental question then became: how a knowledge from the environment arises from this light? The early stages of vision must somehow solve an inverse problem. The light reflected from the 3D world produces 2D images at the back of the eye. This raises an additional question: how does the vision system get from optical images of a scene to knowledge of the objects that gave rise to them?

This thesis is precisely guided by this comprehension of the mechanisms employed by the human visual system to construct a conscious experience of vision.

This subject can be addressed by the study of the binocular disparity mechanisms that strongly participate in a 3D constructed perception of the visual world. Indeed the complex mechanisms that allow the brain to fuse and construct a single image from two slightly translated retinal images are believed to be fundamental because they require considerable cognitive resources for stereo matching, identification etc.

We propose in this thesis to address this issue in the light of recent knowledge and developments in the visual attention domain. The attention mechanisms are thought to form a glue between various processes occurring along the human visual pathways. These are studied through recent technologies of eye tracking and stereoscopic displays.

Specifically, these screens are supposed to re-create to the viewer a new depth sensation. These displays have emerged recently through the development of digital video technologies. Referred to as "3D", these visual systems aim to simulate the binocular disparity as experienced in our 3D world to improve the depth sensation.

Even if this technology dates back to 1838 with the invention of the first stereoscope

by Sir Charles Wheatstone [145, 146], many attempts have failed to install the 3D solutions on a long-term basis for a public audience. The visual discomfort experienced by the public in cinemas during the 1950s durably tarnished the reputation of the 3D. It is now accepted that the successful re-introduction of 3DTV could be realized at the main condition of quality content and viewing comfort at least equal to conventional two-dimensional television. However, an end-to-end 3DTV framework implies reconsidering the way the movies are recorded in studios, how they are edited, transmitted and finally correctly displayed and perceived by end-users. The second part of this thesis is in the scope of this consideration: how to efficiently encode the geometry of the scene while considering the perceived visual quality at receiver side and how to synthesis new virtual views for 3DTV and FTV applications in a plausible way?

The 3D perception is the overall guiding theme for the issues addressed in this thesis. They will be described in the next section before that the main objectives are stated. Then the scope of this thesis will be described.

## Challenges and Objectives

### Visual Attention

The Visual Attention domain emerged in the early 2000s with the understanding and computer modelling of the first stage of visual attentional processes. The seminal work of Itti and Koch [56] on the prediction of the gaze pattern through the modelling of the early cortical stages by separated color, intensity and orientation features opened the way to many researches and subsequent models. Most of these **saliency** models were first bottom-up, but then introduced top-down attributes (related to subject condition, experience, knowledge) to improve their capabilities to predict where the observer's gaze will deploy. Today these biologically plausible hierarchical models compete with statistical models based purely on information processing.

However, few of these past models consider what could be the impact of the depth perception on visual attention, the impact of modifying the natural perception in 3D environment to a 2D flat screen viewing and the impact of re-introducing a simulated stereoscopy or binocular perception on visual attention.

Past researches however underline that an entire visual pathway might be dedicated to the disparity processing [76, 103] within the primary visual cortex also known to participate to bottom-up visual attention.

The first objective of this part is to study the implication of stereoscopy on the visual deployment. Behavioural studies will be realized to precisely study over time the potential effects and biases appearing in monoscopic and stereoscopic conditions.

The second objective is to propose a new model of visual attention adapted to both 2D and 3D conditions which carefully integrate distinct features related to center bias and depth perception dynamically.

### 3D Coding and Synthesis

The two functionalities targeted by the recent emergence of 3D digital technologies are the 3DTV and FTV.

**3DTV** Tri-Dimensional Television or 3DTV is the technology believed to revolutionise the way we will experience videos. The added binocular disparity should improve the depth sensation, the immersion, the visual experience and the overall quality of experience perceived by viewers. Many research projects have been devoted to the design of a complete 3D platform, such as ATTEST [111], the 3DTV network of excellence [2], the 3D4YOU project [1] etc. The main requirement of the 3DTV is its flexibility to support various types of 2D and 3D display technologies. The capture, transmission, synthesis and display from MultiView Videos (MVV) is supposed to cope with this issue but raises the questions of bandwidth and video quality.

**FTV** The Free Viewpoint Television or FTV can be seen as another usage of the Multi-View Videos (MVV) initially available for 3D. The FTV allows the viewer to move his viewpoint freely into a given scene. The viewer can select a preferred viewpoint and then interact with the video content being displayed. However, unlike 3DTV, a larger displacement between cameras (or baseline) is necessary. This implies either synthesizing new content beyond the initial baseline of camera -or extrapolate-, or rendering a virtual view between and from two existing viewpoints. Also, this rendering should be performed in real time to display smooth transitions between views.

The 3D processing scheme that supports the 3DTV and FTV is composed of four main stages: the acquisition, the transmission, the rendering and the display. The objectives of this second part are to propose some solutions to improve both the bitrate of MultiView-plus-Depth video sequences at the transmission stage and the synthesis quality of virtual view at the rendering stage. Thus a new depth compression algorithm and a view synthesis method for extrapolation are presented.

## Document organization

The layout of this thesis is divided into two main parts. The first part addresses the way we perceive the depth in monoscopic and stereoscopic conditions and how it can be employed to improve the predictability of visual attention. The second part proposes two contributions for efficient compression of MVD videos and qualitative view synthesis based on an inpainting technique. Both contributions rely on perceptual mechanisms of spatial vision (presented in appendix A) that consequently aim to render video content of quality.

Part I: Perceiving the Depth

**Chapter 1.** introduces the depth cues known to be selectively integrated by the human visual system to infer the depth of a scene from 2D retinal images. These stereoscopic, ocular, dynamic or pictorial sources of information are described before their potential integrations and combinations are explained.

**Chapter 2.** first introduces a state-of-the-art of the past saliency models aiming to integrate different visual features into a coherent model. Then, a behavioural study on the implication of the depth on the visual attention is realized. Thus, a dynamic visual attention model is proposed that combines relevant saliency features over time.

Part II: 3D Video Coding and Applications to View Synthesis

**Chapter 3.** presents two stages of the 3D framework addressed by the coding and computer vision community: the representation and coding of 3D contents. The requirements of quality, consistency, flexibility for 3DTV and FTV will be detailed before

presenting the main known representations. The subsequent 3D video coding standards will then be over-viewed with a particular focus on the future 3DVC coding and its functionalities.

**Chapter 4.** presents a contribution on the efficient coding of depth maps for results of pertinent visual quality. Recent proposals on depth coding are first presented before a lossless edge based depth map coding method is detailed. The principles of edge detection and coding at the encoder side, but also edge interpolation at decoder side are introduced. Objective and subjective results are then given. They underline the potential but also the limits of the approach.

**Chapter 5.** is devoted to the view synthesis application for 3DTV and FTV. A new method of view rendering based on background structure propagation allows an efficient filling of disoccluded areas and gives qualitative visual results. Then a study is conducted on the appropriateness of objective quality metrics to efficiently measure the visually annoying reconstruction artifacts.

**Appendix A** This Appendix presents a whole state-of-the-art on the spatial vision and its underlying mechanisms. The notions of visual perception, constructive act of vision and the inverse problem solved by the human visual system are first introduced. Then the anatomy and functions of the visual systems are described and linked to the computational approaches to vision. The visual pathways are finally presented with a particular focus on the stereo pathway as an essential component of bottom-up depth perception.

The overall issues addressed by this thesis are illustrated in Figure 1 within a complete 3D framework.

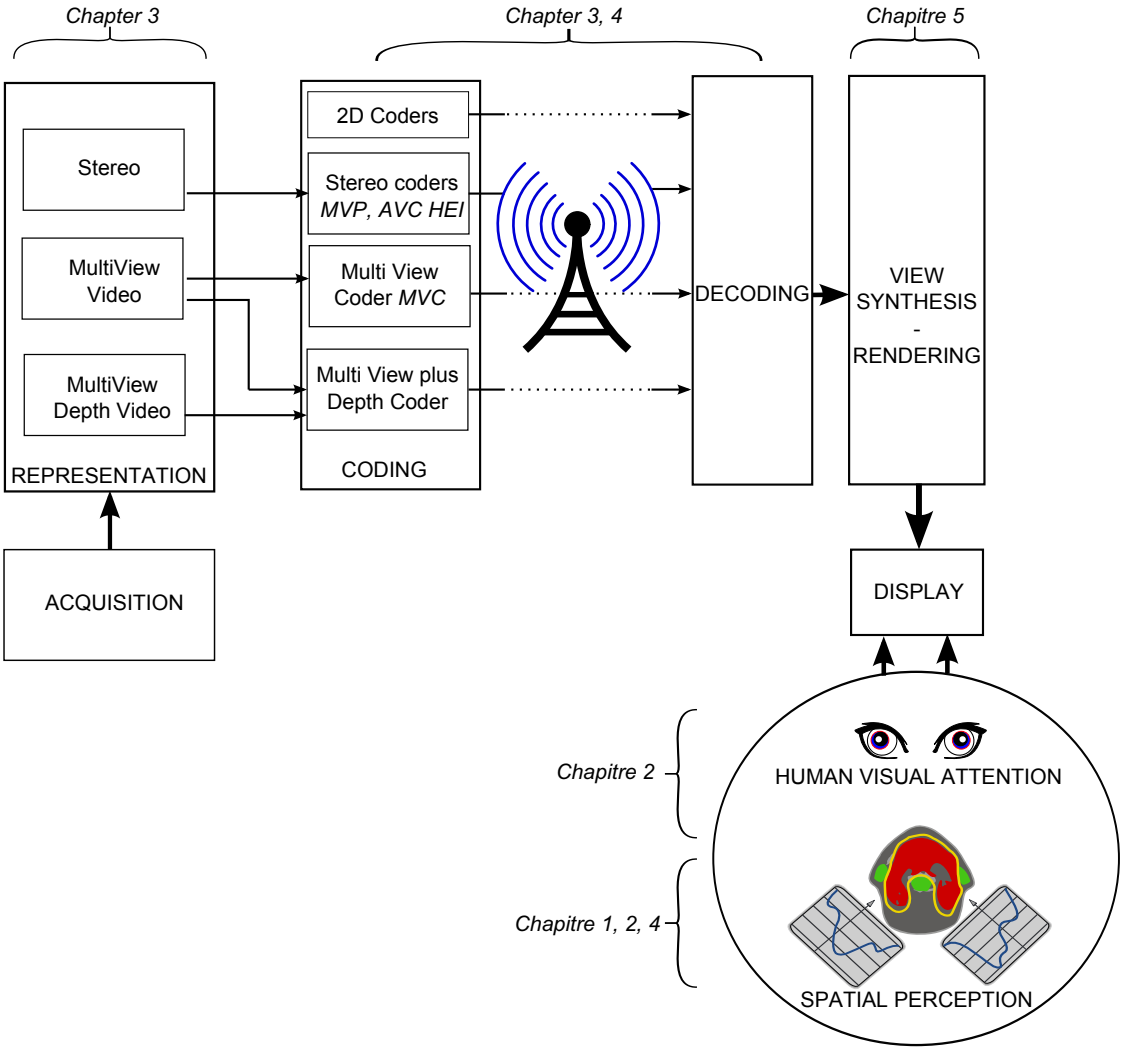



Figure 1: Overview of the 3D framework from acquisition to end-user vision: attention and perception.

# Part I

# Perceiving the Depth

CHAPTER 1

# The Depth Cues

## Introduction

The two human retinas receive the explicit two-dimensional images from their visual environment. The visual system then acquires the two-times-two-dimensional information of the three dimensional world, but this does not suffice to gain an exact knowledge of the third dimension, the distance of the surface of objects from the observer: **the depth**. However, this third depth dimension is lost by the optical projection from points, lines and surfaces of a 3D-world on the 2D retinas. We have seen it is lost because the inverse projection can find theoretically an infinite number of solutions, an infinite number of depth where these points, lines and surfaces could be re-projected. However people are very good in practice to perceive their 3D environment, how is it regained?

The perception of spatial arrangement of objects - their surfaces - in the visual environment with respect to the observer is then a heuristic process aiming to find a solution to an inverse problem. It is heuristic because inference will be made from assumptions about the most plausible solution. Helmoltz[49] concludes that it is the interpretation of the most likely state of affairs in the external world that could have caused the retinal stimulation in accordance with the likelihood principle.

In fact, two interconnected problems must be solved to perceive the spatial arrangement of surfaces from the observer: find the depth i.e. the distance of these surfaces, but also extract their orientation. Perceiving the surface orientation implies retrieving the slant and tilt of the surfaces with respect to the viewer's line of sight. Depth and surface orientation perception are interdependent because orientation of surfaces gives a distance information of its various parts from the observer and reciprocally the distance of its various part inform of its orientation. J.J. Gibson highlighted the first importance of surface perception in vision, what he called the surface layout. This concept was later computationally modelled by Marr and Nishihara in 1978 by a surface-based representation: the 2.5D sketch. Be that as it may, recent vision theorists have devoted much of their works to understand and conceptualize the perception of oriented surfaces in depth, because it is a complex, inferential process whose understanding might lead to the comprehension of higher-level phenomena.

To understand the physiological process of functional areas, one needs to understand the format of this information, what it holds and what it implies. The position of com-

putational theorists is that there must exist distinct processing modules that compute the depth information from separate sources, as proposed Marr with his 2.5D sketch illustrated in Figure A.23. Different inter-dependent modules process different kinds of information and then participate in a global depth perception by constraining the possible depth interpretation at different stages (at the surface stage represented by Figure A.23).

This point of view then supposes that we can classify different sources of depth information integrated by the HVS. There exist various ways and criteria to classify them, highly distinct but also controversial.

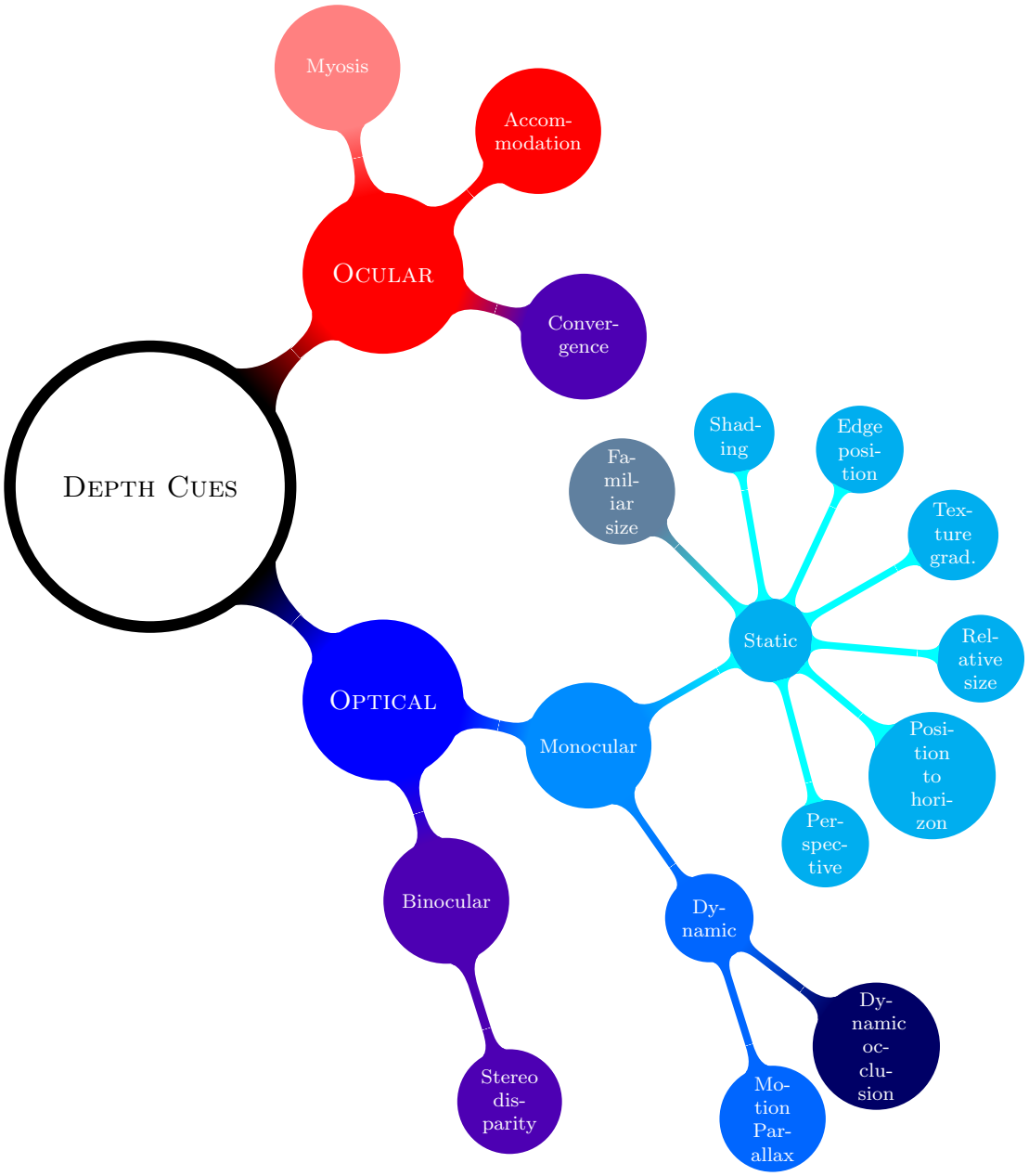



Figure 1.1: Tree graph of the depth sources of information. Three important characteristics are illustrated by the hierarchy: optical vs ocular, monocular vs binocular (blue vs purple), and static vs dynamic. Ocular depth cues are static, but only convergence is binocular (purple). Two other distinctions are illustrated with local colors: the familiar size is absolute while others optical cues are relative (but ocular cues are absolute), the Dynamic occlusion cue is qualitative while all other cues are quantitative.

A first distinction can be made if it concerns the state of the eyes (ocular information) or the light entering the eyes (optical information). A second depending on if the information involves one (monocular) or the both eyes (binocular). A third depending on the source being extracted from still image (static) or from moving objects or observer (dynamic). A fourth according as the information gives an actual distance to objects (absolute) or specifies their position relative to each other (relative), while the fifth separates the numerical distance relation (quantitative) source from the ordinal relation (qualitative). A tree illustrates these distinctions in Figure 1.1. Lots of monocular cues come from the visual world, most of them are called pictorial cues. The Myosis is also called a Depth Of Field (DOF) depth cue, but its contribution is limited if not hypothetical. The rest of the depth cues will be described in the next section. The functional segmentation and separation between these depth cues is limited however. First because they interact, second because there might be some hierarchies or at least some common functional roots. As explained in Appendix **??**, the edge interpretation might comes early in V1, and then be used by different areas, for surface slant analysis (for the perspective and texture gradient cue for example) as for surface separation in depth.

## 1.1 Stereoscopic information

The baseline between the human eyes enables the perception of the world from two slightly translated viewpoints. The perception of the depth issued from the relative displacement -the binocular **disparity**- between the two retinal images is the **stereopsis**. This perception is possible because the two visual fields of each eye overlap in the central region, whose 3D world points project to displaced locations in the two retinas relatively to their distance from the fixation point. This relative local displacement is precisely the binocular disparity. The binocular visual field with its overlapping is illustrated in Figure 1.2.

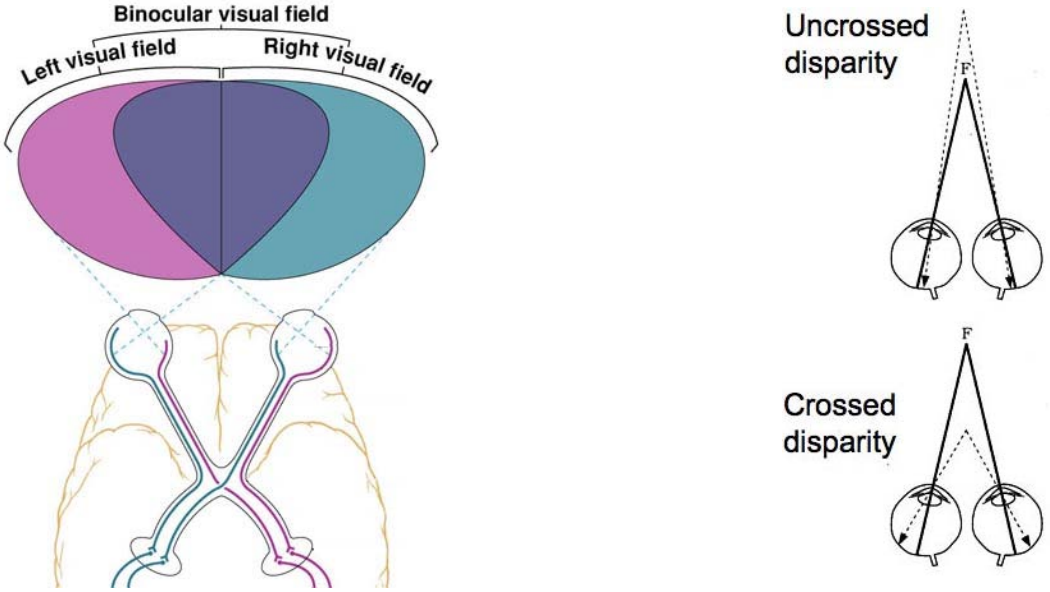



Figure 1.2: left: Illustration of the binocular visual field. Right: Illustration of the uncross disparity versus crossed disparity

### 1.1.1 Binocular disparity

The binocular disparity or stereo disparity is one of the most powerful depth cues, in the sense that it provides an accurate, large, quantitative information on the distance to static objects, at the price of a complex registration and matching of displaced points on the retinas (see section A.3.1, on page 181) .

The *direction* of disparity shows which points are closer and farther than the fixated point, the *magnitude* provides the quantitative information about how much closer and farther they are. As binocular disparity occurs by the lateral displacement of 3D world point position on the left and right retinas, a closer-in-depth point than the fixated point will fall in outward direction on both foveae, this is the **crossed disparity**. At the opposite, a farther point will project in the inward direction, this is the **uncrossed disparity**, see Figure 1.2 (right). Thus the amount of displacement of the points relative to the fixation point (in outward or inward direction) indicates how much farther or closer in depth they are to the fixation point.

**The Horopter** The optical machinery which is the eye implies some limitations. As we have seen, the density of cones and rods is not uniform along the retina, a peak of cone density is actually found in the fovea. The left and right retina images will appear displaced except at the fixation point located in the fovea. Other environmental points could able a stereo matching on both sides of the fixation point, but with a certain limit, mainly from the non-uniform distribution -and then resolution- of the rods and cones along the eye. The horopter then corresponds to the environmental points that project to corresponding points on the two retinae to give a stereo perception.

Geometrically, the theoretical horopter is defined by projecting pairs of retinal points outward through the nodal point (or optical center) of the eye. The set of all environment points then corresponds to a circle, called the Vieth-Muller circle, passing through the fixation points and nodal corresponding points of the eyes. Thus the binocular disparity of the points on this circle is null.

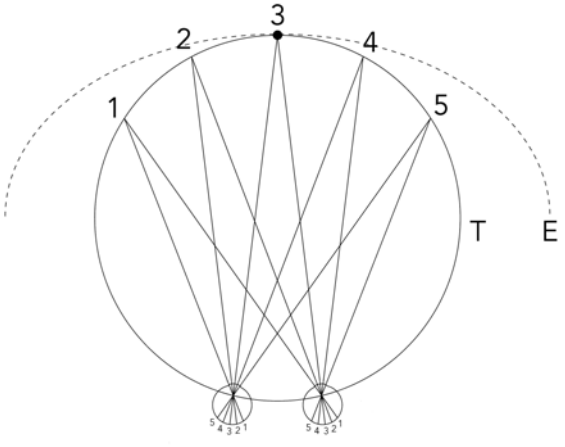



Figure 1.3: The horopter in the horizontal plane of the eyes. The theoretical horopter (T) in the horizontal plane is a circle passing through the nodal points of both eyes and fixation point (3). The Empirical horopter (E) is ellipsoidal and includes the theoretical Vieth-Muller circle.

Thus, the binocular disparity, its direction and amount, define at which depth a given

point is relative to the horopter. Because we normally don't perceive a double image, the points on or close to the horopter are fused into a single experienced image. This area of perceptual fusion on both sides of the horopter is called the Panum's fusional area. A contrario, the points lying outside of the Panum's area allow the perception of disparity as depth.

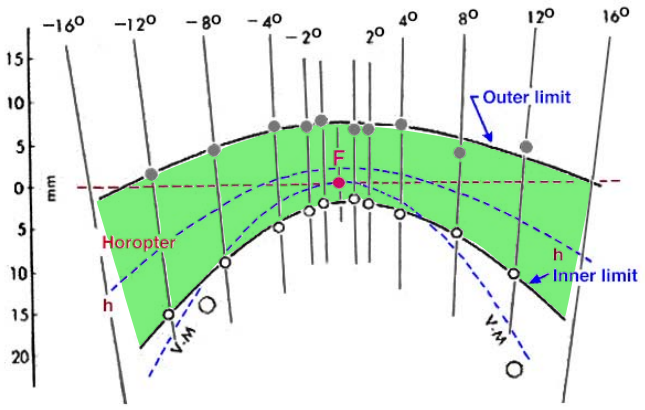



Figure 1.4: The Panum's fusional area (in green) where points are fused into a single image. F is the central fixation point. Points closer or farther from the Panum's area produce double image of crossed and uncrossed respectively binocular disparity. Modified from Ogle K.N., 1964.

**The random dot stereogram and the correspondence problem** A stereogram consists of a pair of images whose elements differ in their relative lateral displacement, such that when viewed **stereoscopically** (one image stimulates one eye, and the other image the other eye), produce an illusion of depth. They were invented by Sir Charles Wheatstone in 1838 to analyse the geometry of binocular disparity.

A random dot stereogram is also composed of a pair of images, but consisting of numerous randomly placed dots. These dots have lateral displacement and can produce convincing perception of depth when viewed stereoscopically. This can be experienced by staring at Figure 1.5.

Because the random-dots precisely do not possess other higher level depth features (forms/shape, shading) than the binocular depth cue, they prove that the binocular/stereo disparity can be perceived without monocular shape information. It does not prove that there is no prior shape analysis before stereopsis. But the stereo perception appears to be independent of the monocular shape cues. This supports the notion of a stereo pathway in the visual system as hypothesized by Livingstone and Hubel.

The question is now to understand how the visual system manages to find corresponding features in the random dot stereogram without any geometry or shape information. Indeed, by considering only the pairs of points lying along the horizontal axis of binocular displacement - 100 points along a given horizontal line in one image - the logically possible pairings is 100!, that is a huge number of possibilities. Computational models to solve this correspondence problem has been developed at a point, line, edge, and local region scale [82].

**Depth resolution** The depth resolution, or depth **sensitivity** of the binocular disparity is very high, just a few seconds of angle (1/3600 of a degree), but its range of operation

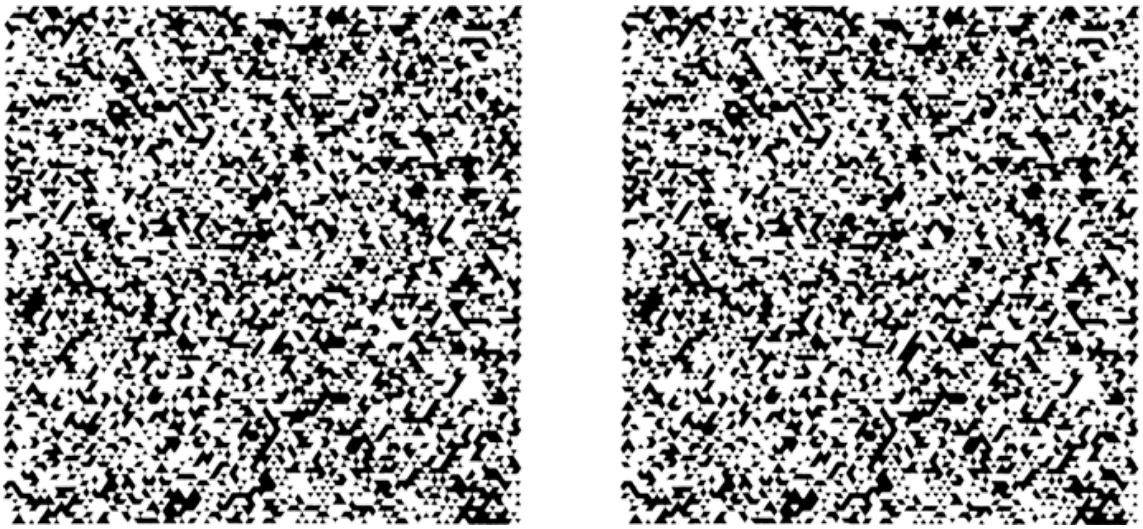


Figure 1.5: A random dot stereogram. These images are obtained from a single array of randomly placed triangles (as dots), a subregion of them being laterally displaced. When viewed with crossed disparity (by crossing the eyes) so the right eye of the left image appears surimposed with the left eye's view of the right image, a square should be perceived floating above the page. From Julesz [63].

.

is limited to 30 meters. Beyond, the ratio disparity/precision becomes too small. We will discuss the sensitivity threshold of different depth cues versus their range of effectiveness at the end of this chapter (see Section 1.5).

### 1.1.2 Vertical disparity

As we have seen, the binocular disparity consists of viewing the same object from two viewpoints. The closer or farther points from this given object will introduce a binocular disparity. But if this object is moving not only along a depth axis, but along a horizontal line from the eyes viewpoint, it will introduce a vertical binocular disparity. Let's consider the upper illustration of Figure 1.6, the object seen by the right eye is bigger in the right retina than in the left one. In fact the object is bigger along the horizontal axis in both directions and also along the vertical axis. This results in vertical disparity between corresponding points of the same object but projected to different sizes on each retina.

### 1.1.3 Da Vinci stereopsis

When we experience binocular vision, we can see different surfaces at different depths, the closer surfaces occluding the farther surfaces. A portion of the farther surface is seen by just one eye, because it is occluded by the closer surface -and then hidden- to the other eye. Then the portion of the farther surface to the right of the closer surface is only seen by the right eye but not the left, and conversely for the left eye. This is illustrated in Figure 1.7 (a).

Nakayama and Shimojo (1990) [95] rediscovered this other form of stereopsis and named it in honor of Leonardo da Vinci who was the first to describe it. The depth information comes from the geometry adjacent to the occluding depth edges where each monocularly viewed region always belong to the **farther surface**. This is precisely this property that

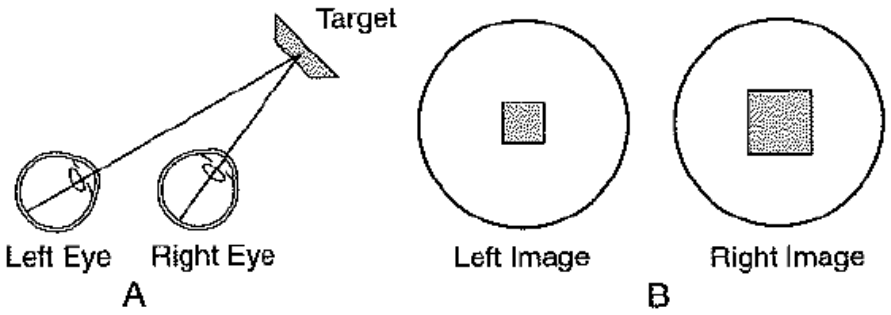



Figure 1.6: Illustration of the vertical disparity. Here a surface of an object (A) is closer to one eye than the other, differences in size on each retina (B) leads to horizontal and vertical disparity.

will be used in the proposed algorithm of the automatic occlusion filling method described in Chapter 5.

Recently, Assee and Qian [4] proposed a computational model of da vinci stereopsis based on depth-edge-selective V2 cells. The model relies on a coarse-to-fine disparity energy computation in V1 followed by disparity-boundary selective units in V2. They demonstrate with random-dot stereogram that the model V2 stage can find the location and the eye-of-origin of monocularly occluded regions while improving disparity map computation.

## 1.2 Ocular information

The ocular information cues are related to characteristics of the eyes and its internal and external components. The focus of the lens is called **accommodation** and the angle between the two eyes' optical axis is the **vergence**.

### 1.2.1 Accommodation

For each eye, the optical focus of the lens is controlled by the ciliary muscles around it. By applying different tensions on the lens, the shape of the lens varies temporarily: thin to focus on light from faraway objects, and thick for nearby ones (see Figure 1.8).

So, if the HVS has information about the tension to apply on the muscles that control the lens shape, then it has **absolute** information about the distance of the object to focus on (considering the visual system is properly "calibrated").

The visual system interprets the proper focus to apply on the retina by blurriness/sharpness of edges in high spatial frequencies. This is out of scope and will not be further described.

This accommodation cue is monocular because the focus is applied by the ciliary muscles on each eye separately. Also, it is absolute and quantitative because the visual system has information about the tension to applied to the muscles, from which it can retrieve information about the distance to the focused object.

Different studies have shown accommodation is a weak but useful source of depth information at close distance, not especially to make direct judgement about distance but rather to evaluate the size of objects (Wallach and Floor, 1971). But beyond 2 meters, accommodation provides hardly no depth information, because no lens deformation needs

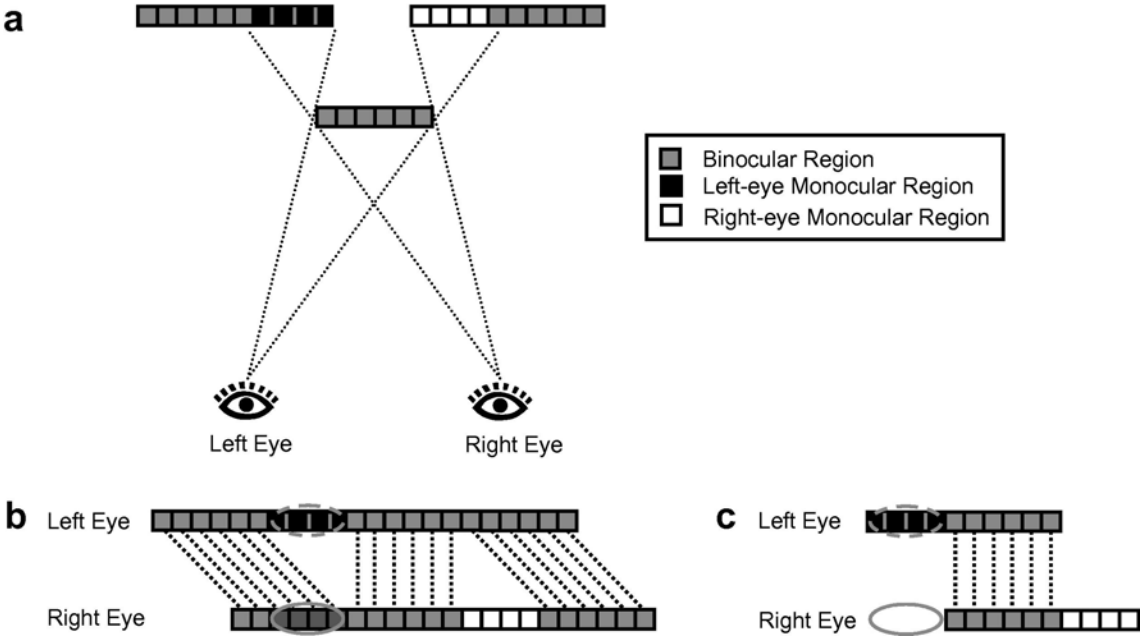



Figure 1.7: Occlusion geometry and the role of monocular vs binocular cells in solving da Vinci stereopsis. (a) Schematic diagram of a scene where a near surface occludes a background. The dotted lines indicate the extent to which the near surface occludes the far surface from each eye. (b) Images seen by the left and right eyes for the scene in (a) when fixation is at the near surface. (c) A special case of (b) when the binocular background is assumed to be featureless. For all panels, gray squares indicate binocular regions, and black and white squares represent left- and right-eye-only monocular regions, respectively. In (b) and (c), the dotted lines indicate correspondence between two eyes' images. Ovals indicate the RFs of monocular cells, with dashed and solid lines representing left- and right-eye-only RFs, respectively. From [4]

to be applied at this distance, muscles controlling its shape being at their most relaxed state.

### 1.2.2 Vergence

The eyes also have the capacity to turn inward into their orbit. The angle between the two lines of sight can be varied to fixate an object and make the incoming light fall on both foveae. This is the convergence process which can also inform -partially- the distance the eyes fixate to. The angle of vergence varies directly with the distance to the fixated object. A close object will be fixated with a large convergence angle while a distant object will be fixated with a small convergence angle (see Figure1.9).

This source of depth information is then binocular unlike the accommodation, but provides also an absolute information about the distance. The convergence can be expressed as indicated in Figure 1.10.

From this asymptotic behavior we can see that the angle of convergence decreases rapidly up to one meter, but very little after as it becomes asymptotic. That means that the vergence control information is a reliable and accurate information up to two meters -as for the accommodation case- but provides hardly any accurate information beyond this distance.

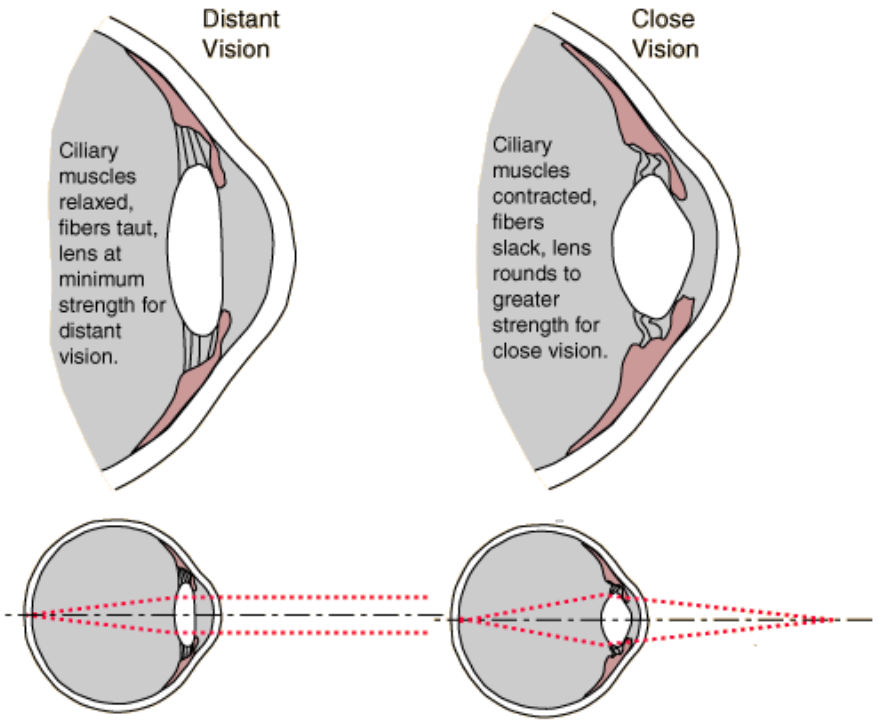



Figure 1.8: Upper right: The eye accommodates for close vision by tightening the ciliary muscles which make the crystalline lens more rounded. Lower right: The light rays from close objects diverge and then need more refraction for focusing, realized by the thicken lens. Left: the opposite process occurs for distant vision, with relaxed ciliary muscles that make the lens thinner to diminish refraction necessary for distant focus.

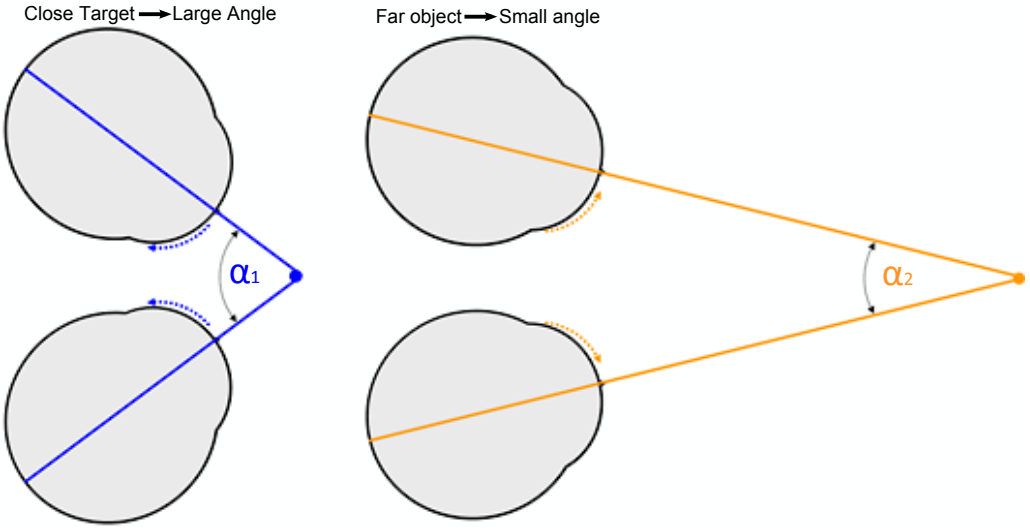



Figure 1.9: The depth information issued from convergence. The angle of convergence between the two eyes depends on the distance to the object. Large angles ($\alpha_1$) for close objects and small angles ($\alpha_2$) for far objects.

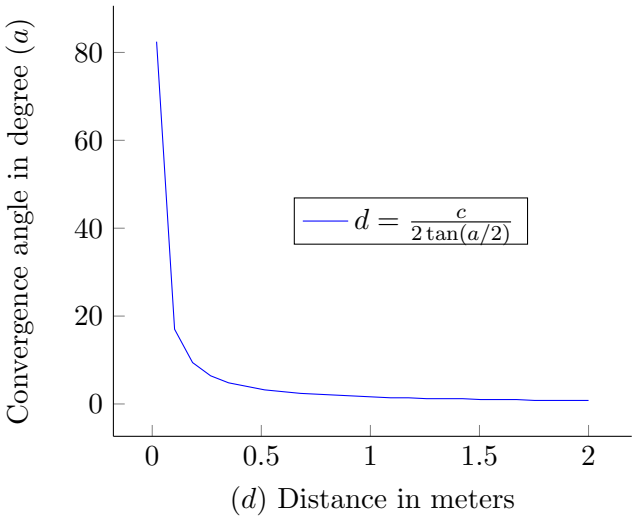



Figure 1.10: Vergence plotted as a function of distance. The vergence comes to the asymptote beyond 1 meter.

It can also be observed that accommodation and vergence are interdependent and co-vary: change in the distance to an object will imply related change in both accommodation and vergence. Accommodation and vergence are then dependent contributions to depth perception and are among the few cues that give absolute distance to a fixation point nevertheless at close distance.

## 1.3 Dynamic cues

Displacement of object on the retinas over time provides dynamic visual information to the human visual system: a retinal image motion, or "optic flow". If an observer moves, a fixated object moves in an environment or the whole environment is moving around an observer, a distance can be perceived from this movement.

The direction and rate at which objects are retinally displaced depend's on the **motion**, but also on the **distance** and the **position** from the object to the observer. Thus the depth-from-motion occurs from a differential position over time (motion) of paired points due to their different depths relative to the fixation point: motion parallax.

### 1.3.1 Motion parallax

The perception that two points at different distances from the observer move at different retinal velocities as the observer viewpoint is changing is an illustration of motion parallax. The differential motion of pairs of points due to their different depths relative to the fixation point provides motion and relative depth information. An analogy can be done with the binocular disparity (see section 1.1). Binocular disparity is related to the difference of position between a pair of displaced points taken at the same time, while motion parallax involves the difference of position between a pair of displaced points over time. As was the case for binocular disparity, the nature of retinal motion parallax is related to the distance to objects in the environment but also to the observer's fixation point: how much closer or farther is an object from the fixated object. It is then a relative depth cue.

It should be remarked that in a natural situation, we don't experience the motion parallax in itself: environmentally static objects are perceived as static, even if their retinal

images move with the body or eye movement. We don't experience a retinal motion or displacement, but instead the actual position of objects, as it was the case with binocular disparity. This is called position constancy.

**Optic flow: case of a moving observer** The optic flow is the generalization of the theoretical motion parallax between two points to the multiple points in space. Gibson (1996) advanced that when an observer is moving, the image motion is structured and depends of the structure of the 3D environment, its oriented surface, and the observer's motion. He introduced the concept of motion gradients to describe the motion of regions, their quantity (speed) and direction.

An observer moving leftward while fixating a point in the middle of a line along depth (by example a straight road directed to the horizon) will lead to a relative motion leftward for the closer points on the line (and the closer they are, the faster they move), rightward for the farthest one, as illustrated in Figure 1.11

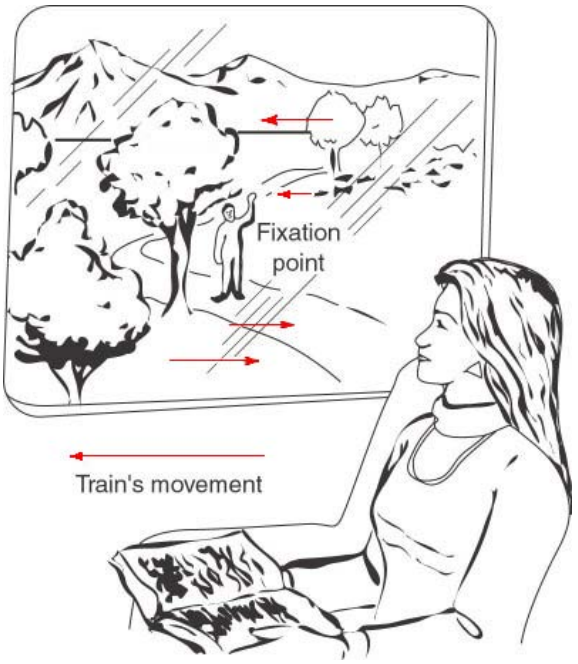



Figure 1.11: Motion gradients (red arrows) produced by a moving observer and the resulting optic flow.

### 1.3.2 Optic flow: case of moving objects

An object moving in the visual field with respect to the observer also enables the depth perception.

Wallach and O'Connell [137] described this phenomenon in 1953 as the kinetic depth effect (KDE). They cast the shadow of a 3D bent-wire figure. When the figure is static, the wire is stationary, no depth is perceived. Then, when the wire figure is rotated, it pops into a 3D shape. See illustration in Figure 1.12.

Recovering depth information from object rotation is nevertheless ambiguous, the 2D retinal motion could likely be perceived as a figure that deforms over the time. But people perceived instead a rigid object consistent with the moving image : an object in rotation. The visual system seems to be selectively tuned to perceive -to infer depth from- a rigid

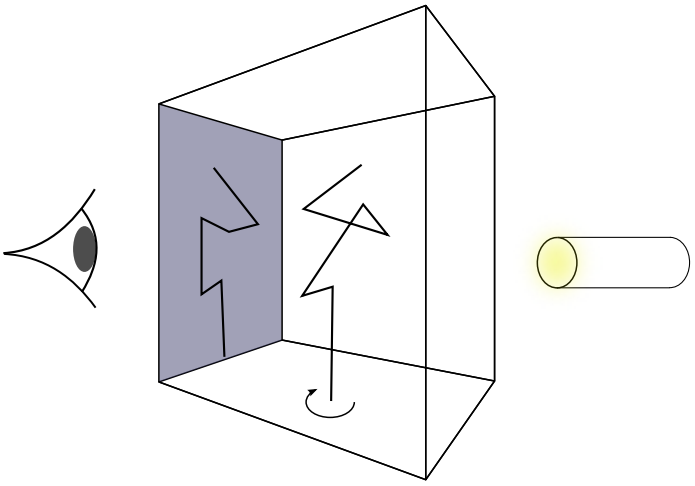


Figure 1.12: The kinetic depth effect. A bent-wire occupying a volume is illuminated from behind, and its show is projected onto a sceen viewed by an observer.

motion rather than a plastic motion. The constructivists could answer that rigid motions are more highly probable to occur in natural conditions than plastic ones. Gestalt theorists instead could argue that the rigid interpretation is the simplest and is then preferred by the visual system. This illustrates how different theories with different principles such as likelihood and Prägnanz can well predict the same outcome.

### 1.3.3 Dynamic Occlusion

The motion appearing to the retina wherever it comes from objects or observer, also can imply the appearance and disappearance of surfaces behind a moving edge [42]. This is called Dynamic Occlusion precisely because the occlusion occurring over time informs the ordinal positions of surfaces between each other: this is a relative and qualitative depth cue.

Because the edge always belongs to the closer surface (foreground) while the appearing/disappearing surface to the farther surface (background), it must be detectable as texture instead of uniform surface, it is also called Accretion/Deletion of Texture cue. As a parallel can be drawn between motion parallax and binocular disparity, accretion/deletion of texture revealed over the time is revealed across the views of left and right eyes in da Vinci stereopsis : occlusion information.

## 1.4 Pictorial information

We experience object position perception in the three-dimensional environment without even noticing they are the process of different sources of depth information, stereoscopy and motion. But by closing one eye without moving, the world continues to look three-dimensional. In the same way, paintings and photographs manage to give compelling impression of depth. How ?

The pictorial information cues are the set of cues responsible for depth inferred from static, monocularly viewed pictures. They are very efficient to interpret visual scenes even from 2D pictures where both stereo and motion are inoperative. They can even overcome

the binocular information if it is reversed by an optical system switching the images going to the left and right eyes.

### 1.4.1 Perspective projection

Light is transmitted in straight lines and is reflected by environmental surfaces onto the retina. The **perspective projection** renders this phenomenon by mapping the real-world objects and scenes to their optical images.

As we know, this projection induces a dimensionality reduction: the 3D world objects become 2D images and the depth dimension is lost. Among the pictorial cues that help to recover - partially - this lost dimension, the perspective perception is one of the more powerful and has been known for a long time. Renaissance artists and painters depicted depth realistically on flat canvas. Alberti in 1436 discovered the first this property and proposed a drawing method respecting perspectives, as illustrated by the engraving from Vignola in 1611 (see Figure 1.13). Then the word *perspective* originates from "fenestra aperta", the "window to the outside".

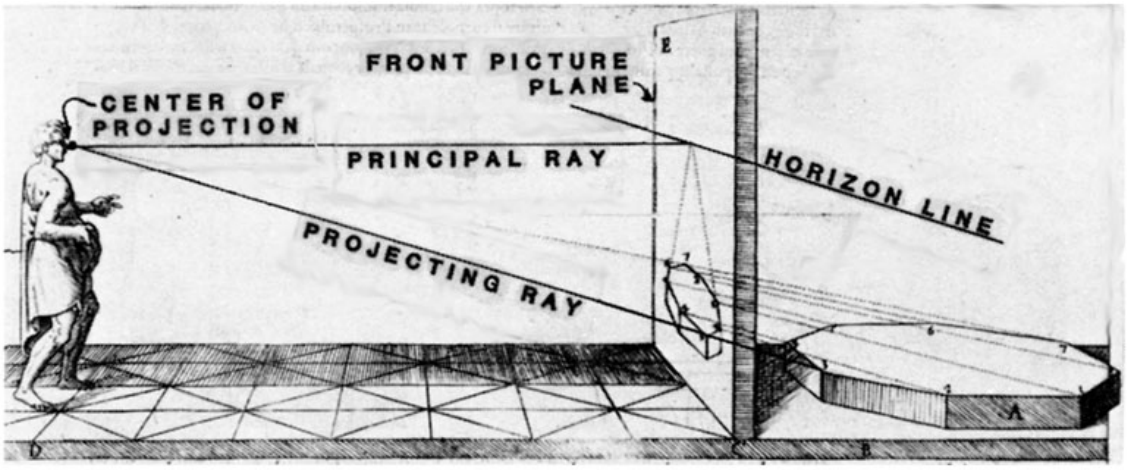



Figure 1.13: Perspective depth information from a real world scene can be traced on a 2D surface by the contours of object(s). Representation of Alberti's window (perspective drawn using a front picture plane). Engraving (modified) from G. B. Vignola, La due regole della prospettiva practica, 1611.

**The convergence of parallel lines** are induced by the perspective projection. The parallel lines in the three-dimensional world appear to project in 2D image as lines converging toward a vanishing point on the horizon line. Constant distance between parallel lines appears increasingly large at close distance, and increasingly small at far distance, up to a vanishing point. Depending on the set of parallel lines, notice that there can be an infinite number of vanishing points along the horizon line of the ground plane, as there can be other horizon lines for other oriented planes in the 3D space.

**Position relative to the Horizon** The heights of objects in the 2D image plane relative to the horizon inform as well of their distance. This is related to the perspective projection but does not involve converging lines or vanishing points. When you experience in your visual field a landscape of mountains and trees, the trees toward the bottom appear closer than the ones farther up, while at the opposite the clouds toward the top appear closer

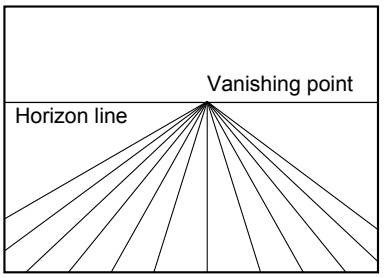



Figure 1.14: Convergence of parallel line in a 2D image projection.

than the clouds farther down. This sub-part of the perspective depth cue involves the spatial layout of object in the visual field. The objects that are closer to the horizon -even not visible- appear at a farther distance. The visual system indeed hypothesizes that the objects rely on a level plane, so that the closer to the horizon they are, the farther they lie on this plane. Quantitative information can even be deduced from a visible or deductible horizon line [119]. Let's consider the Figure 1.15. The visual system can retrieve by geometrical relations the distance $d = h * cotan(\alpha)$ , that is the distance $d$ to objects derived from the horizon angle $\alpha$ and the height from observer $h$. This is called the egocentric distance to the distance of object onto plane, because it relies on the use of our own height above the ground plane.

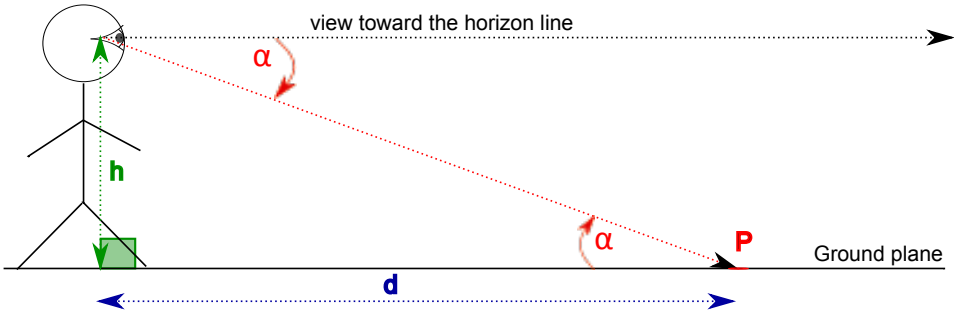



Figure 1.15: Distance as function of the horizon angle to a point on a surface. The distance $d$ to a point on a surface is the product of the perpendicular height to the surface $h$ times the cotangent of the angel $\alpha$. From [101].

Recently, Ozkan and Braunstein [100] experimented and confirmed that higher objects equal in projected size in a 3D scene were judged larger and farther when it was at or below the implied horizon. The reverse effect appeared when the higher object was above the horizon, where it was judged smaller and farther back than the lower object. Thus they confirmed that the relation between judged size and judged distance of objects depends on the positions of the objects relative to the horizon.

### 1.4.2 Relative size

The relative size cue comes from two supposed identical size objects projecting to different sizes onto the retina due to their different distances. Then, more distant objects project to a smaller image onto the retina, while closer objects project a bigger image. The retinal size of the image of an object - measured in degrees of visual angle - then depends on its distance, and can be computed easily by trigonometry. The height of an object $H$ seen at a visual angle $\alpha$ inserts into a right triangle. The distance to the object $d$ is then the ratio of the height $h$ over the tangent of the angle, as illustrated by the Figure 1.16:

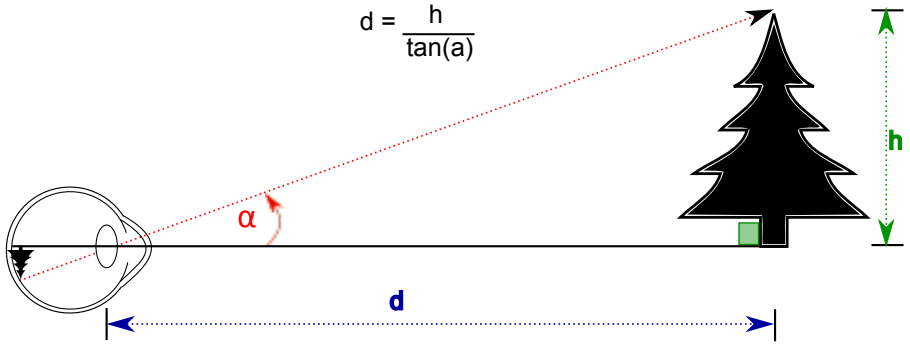



Figure 1.16: The size-distance relation. The projection illustrates that the distance to an object can be retrieved from its size $h$ and the tangent of its visual angle $\alpha$. From [101].

In order to solve the equation, the height of the object $h$ must be known, and this may not always be the case, except if you consider familiar objects as we will do in next section. Consider a single line or a single plane as illustrated in Figure A.3. You can't tell from such form whether you are looking at a small form nearby or a larger one farther away, because there is an infinite number of possible sizes, as there is an infinite number of solutions to inverse projection. But when you see two very similar objects very probably also similar in size, you can situate them in depth just by comparing their relative image sizes. The visual system then uses a heuristic, it assumes that two visually identical objects have the same objective size to express their relative distance from their relative image sizes. The depth can then be computed by adding a heuristic assumption to an undetermined equation.

### 1.4.3 Familiar size

Contrary to the relative size cue, the familiar size cue relies on the semantics of the two identical objects involved in the visual field. Indeed most of objects in nature have a range of size that can be clearly perceived and approximated through the viewer's **visual experience**. Tables are about 80 cm off the floor, cars about 1.5 m high, adult men 1.75 m and so on. As stated in previous section, if the size of an object is known to the observer, then the size-distance equation can be solved and the distance retrieved. This phenomenon has also largely been used by artists and can be illustrated in Figure 1.17. Recent findings based on fMRI recordings suggest that the ventral stream regions of the visual cortex showing category specificity are modulated by the perceived size and distance of visual stimuli [18]. The familiar size depth cue is then a "high-level" in the sense that it necessitates high-level cognitive resources such as memories, object identification and so on.

Figure 1.17: Familiar size as a depth cue. The figure to the left is perceived as closer than the figure to the right.

### 1.4.4 Texture gradients

We experience in our structured modern visual world many patterns of textures: regular elements whose pattern is repeating over space. Examples include the wooden floor and tiles of a house, but also natural blads of grass in a lawn or pebbles on the beach. The systematic change of size and shape of these repeating texture elements on an environmental 3D surface is called the **texture gradients**.

It is actually an important depth cue because it informs the different depths of points along the surfaces but especially their orientations and curvatures. The element size and shape appear to provide independent sources of information about surface orientation [125]. The element size is indeed used by the visual system to estimate the orientation of the textured surface, at the price of an assumption: the texture elements are supposed to be identical in size. It illustrates -as in the case of relative and familiar size- how a heuristic assumption can resolve an undetermined solution. The Figure 1.18 illustrates how our visual system is fooled by this assumption. The texture gradients are indeed illusory and lead to perceived depth in this picture.

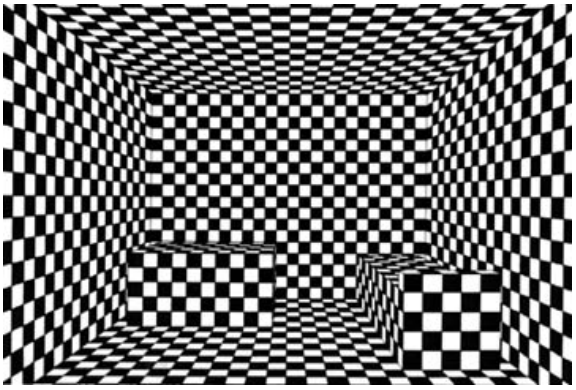


Figure 1.18: Artifical texture gradients.

We will see in chapter 5 the importance of respecting local texture gradients -their orientation and size- in order to fill in plausibly some missing surface regions.

### 1.4.5 Shading and shadow information

The differences in the amount of light reflected from a given surface come from the different orientations of this surface relative to the light source. These differences are called the **shading**. The shapes of surfaces curved in depth and their volumes can be efficiently recovered from their shading by the HVS. Consider three spheres as illustrated in Figure 1.19 (a): first a theoretical case without shading. (b) a more practical case with the shading added: the sphere made of matte material diffuses the light uniformly in all directions. It is illuminated by a single light source: variations in the amount of reflected light appear, the brightest part being at the surface normal pointing back to the light source.

As with other depth and surface orientation cues, the perception of depth also relies on heuristic assumptions. For example our visual system assumes that the main illumination has its source from above. This is illustrated in Figure 1.19(d) which appears to be convex bumps that "pop-out" to the viewer while the Figure 1.19(e) is perceived as concave dents. This perception is veridical until you turn the page upside down: the reverse direction of illumination thus reverses the perceived convex and concave forms. This hidden assumption is quite strong but must be have some reasons to exist as it is. It makes sense indeed, because our visual environment is almost always illuminated from above.

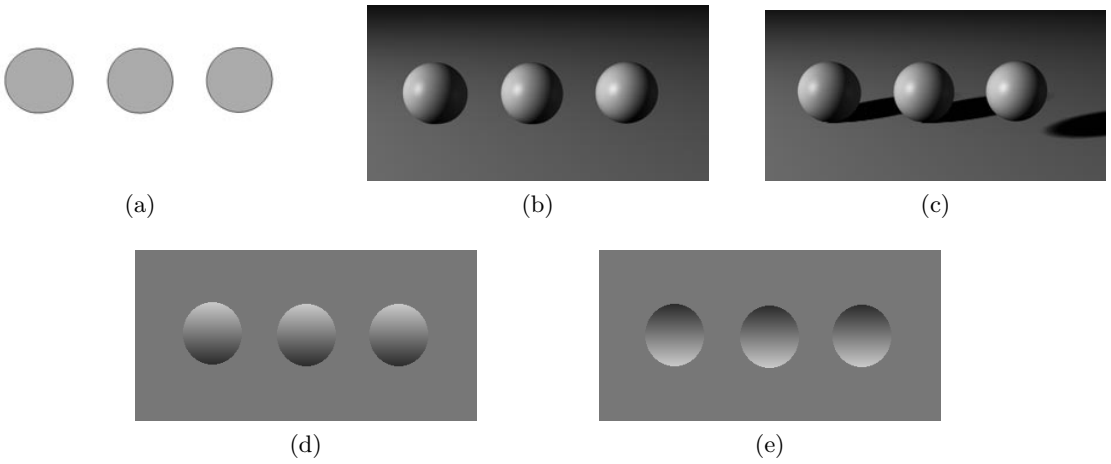



Figure 1.19: (a) An image of three spheres without shading nor shadows.(b) The shading is added. The distances of the different parts to the viewer can be perceived.(c) The shadows are added. The perceived distance of the spheres changes with the different positions of their cast shadows. Lower row: direction of illumination and perceived convexity: (d) the spheres appear to be convex, the opposite of the spheres in (e) that look like concave dents.

Concerning the hypothetical localization of the shading and shadow process along the visual pathway, some studies [106] showed their plausible analysis in V1. The orientation of surfaces with shading can be recovered using the receptive fields present in the primary visual cortex.

The shadows of objects falling on the surface of neighbouring objects are another source of depth information. The Figure 1.19(c) shows three spheres. The sphere on the right appears to be closer to the viewer than the other spheres. Its shadow is distant from the bottom of the ball and seems to indicate it is floating above the surface. The shadings and cast shadows then provide relative and qualitative depth information on the position and

distance to objects. They are independent components visible by variations of light onto the object surface - shading - and onto its neighbouring surface(s) - the cast shadows.

### 1.4.6 Aerial perspective

Figure 1.20: The aerial perspective locates the far objects between each other, as here with the mountains.

The aerial perspective occurs when objects are viewed from very long distances. Their contrast can be diminished by the atmospheric particles that scatter the light. Also in the case of mountain landscapes, the far mountains appear bluish. The additional atmosphere through which a far mountain can be seen scatter longer wavelengths while letting the short wavelengths of blue through.

The resulting local differences in contrast and color are not sufficient in themselves to indicate an absolutely accurate distance, but they give an extra sense of relative depth when used with other depth cues at far distances.

### 1.4.7 Edge information

The interposition and occlusion of objects between each other means one can differentiate and order them along their depth (see illustration in Figure 1.21. This information comes from the interpretation of the object edges and the supposed occultation of light from an object by an opaque one closer to the viewer.

The edge information depth cue is thus a relative depth cue, specifying an ordinal depth relation: the given information is qualitative. It can only inform about the occlusion and interposition of objects between each other: a cat is closer than a door etc. But it works only within the practical limit of the human eye resolution: the edge information can give consistently accurate ordinal depth information at very far distances.

**Computational theories** Different computational theories have emerged [44] to model the edge interpretation based on clever heuristics. They place emphasis on the intersections of edges (or vertices or junctions) and classify them in different categories. In particular, the T-junctions refer to the occlusion situation where the top of the T corresponds to the occluding edge while the stem to the occluded edge. Later, Huffman et al.[54] specified all types of vertex when viewing trihedral angles and demonstrated that the edge interpretation could be simplified by local constraints at each vertex.

Once edge have been extracted, four types of edges can be distinguished: the orientation, depth, illumination and reflectance edges.(see Figure 1.22). Orientation edges

Figure 1.21: Illustration of the overlapping or interposition of people place in front of each other (left) and the possible interpretation of their distance relative to the camera (light gray surface for the closer, darker for the farther).

appear on discontinuities in surface orientation: two surfaces orientations meet along an edge (also called albedo boundary). Depth edges appear at spatial discontinuities between surfaces. One surface occludes another that extends with space behind it. Illumination edges appear due to illumination differences marked locally, such as at the edges of a shadow. Reflectance edges are formed where the light-reflecting properties of the surface is modified, like painting on a uniform surface.

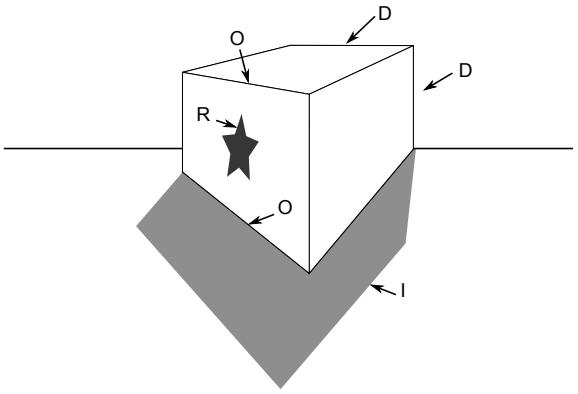



Figure 1.22: The four kinds of luminance edges: orientation edges (O) due to surface orientation changes at edges of objects, depth edges (D) due to spatial discontinuity in depth between surfaces, illumination edges (I) due to shadows and reflectance edges (R) due to change in surface pigments or material. Modified from [101].

**Figure/ground organization** The visual system organizes the visual world in a coherent spatial layout of object surfaces. In this way the interposition of objects with their edges clearly identifies which one is closer to the observer than the other. This interpretation appears to be binary and exclusively paired when we consider the Figure 1.23.

This perceptual organization of virtual figures that pop-out from the background is known as figure/ground organization, where a thing-like region is the figure, and its background-like region is the ground. The psychologist Edgar Rubin discovered this effect

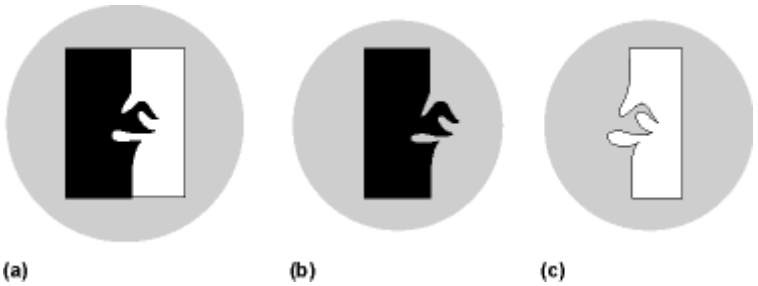



Figure 1.23: Figure-ground organization. The pattern in (a) usually organized as a black figure on a white ground or a white figure on a black ground. The shapes of these two figures are very different despite the fact that the central contour is the same in the two cases, as is clearly indicated in (b) and (c). (From Rock, 1975.)

in 1921 with phenomenological subjective experiments of figure and ground separately. They are opposed on different properties:

- **Figure** appears thing-like, closer to observer, bounded by contours that define its shape
- **Ground** appears not thing-like, farther from observer, extended behind contours that doesn't help to define its shape.

Then, the figural region is interpreted to be closer to the observer, while its shape defines the contour: the contour belongs to the figure. Interestingly, Rock extends this study in a recognition memory experience. Subjects were presented similar series of ambiguous figure/ground stimuli (see Figure 1.23). Half of them had to remember one of the figure, either white or black, and were shown afterwards a series of figure or grounds. Subjects remembered the figural tests shapes while they didn't for ground shapes. Importantly, they didn't perceive the ground as having its own shape.

Thus the principle of figure-ground organization is an element participating in the edge interpretation cue, it is therefore a Gestalt cue that follows their perceptual grouping principles.

**The ecological importance of figure/ground cues** Distinct factors have been isolated by psychophysiologists as favoring the discrimination of figure and ground regions. They have been defined by isolating one factor from the others, so in the case of multiple conflicting factors, the outcome is practically hard to define.

- Attention: the element that tend to draw the first the attention is the figure.
- Surroundedness: a region surrounded by another tend to be perceived as figural.
- Size: smaller region is perceived as figural.
- Orientation: vertical and horizontal regions tend to be perceived more often as figure.
- Position: upper regions tend to be perceived as figure, below the ground regions.
- Contrast: regions with higher contrast are often perceived as figural.
- Parallelism: regions whose contours are parallel are also taken as figural.

- Convexity: all else being equal, convex regions have the tendency to be perceived as figure, while concave ones as ground [89].

It is known that three fundamental processes of perceptual organization are region segmentation, grouping and then probably parsing. It is very probable that the figure/ground organization operates before region grouping because the input elements must already have been differentiated from ground before being grouped. Thus figure/ground organization must operate early in the perception before higher-level region organization could occur. This is precisely what neurophysiologists Qui et al. confirmed electrophysiologically. They found neural correlates of figure-ground assignment in V2 within the 10-25 ms of the onset of response activity.

Recent computer scientists use a totally different approach to understand vision and depth interpretation by figure/ground organization in particular. Instead of trying to match the human behaviour by programs resolving similar vision problem -such as edge interpretations- or using phenomenological or electrophysiological experiments, they propose to analyse the statistics of regions along the edge of natural images.

Fowlkeset al.[38] then quantify the extent to which figural region tends to be smaller, more convex and above ground regions on a large collection of natural images. Results confirmed that these Gestalt cues are valid in the ecological sense: they found statistically more convex, above ground and smaller regions that indeed belong to the figure region. Burge et al. [15] recently confirmed the ecological validity of the figure-ground cue by both statistical and psychophysical analysis: they verified quantitatively whether the observers internalize those statistics. Regarding its usefulness they proposed that convexity should be reclassified as a metric, quantitative depth cue.

## 1.5 The depth cues integration and combination

We have seen that a huge number of depth information sources can be perceived and used by the human visual system. Indeed, this number is greater than any other property of perception in any modality of perception. However, the comprehension on how perceivers integrate and combine these cues is relatively low. Does the HVS integrate them separately, at different stages: how?

Cutting et Vishton [27] proposed replacing them in their potential contexts of usage. They extend the depth resolution or the "just discriminable depth thresholds" functions to nine sources of information originated by Nagata in 1987 [94] and illustrated in Figure 1.24. They assume that more potent sources of information are associated with smaller depth-discrimination thresholds: these threshold functions then reflect suprathreshold utility. These functions can in turn be classified in three distinct types of space around the observer: the personal, action and vista space. They suggest that within each subspace a subset of sources act in consort with different relative strenghts to offer to the perceiver a "big picture" of depth information.

All these sources inform somehow on the surfaces oriented in depth, but the HVS manages to combine them into a coherent depth representation: how? Scientists have proposed three different combinations: one source dominates one or multiple others, a compromise is achieved between conflicting sources or interacting sources reach a single coherent solution.

Landy et al. (1995) [68] proposed a hybrid solution based on compromise and interaction among depth sources. The weak fusion originally assumed no interaction between

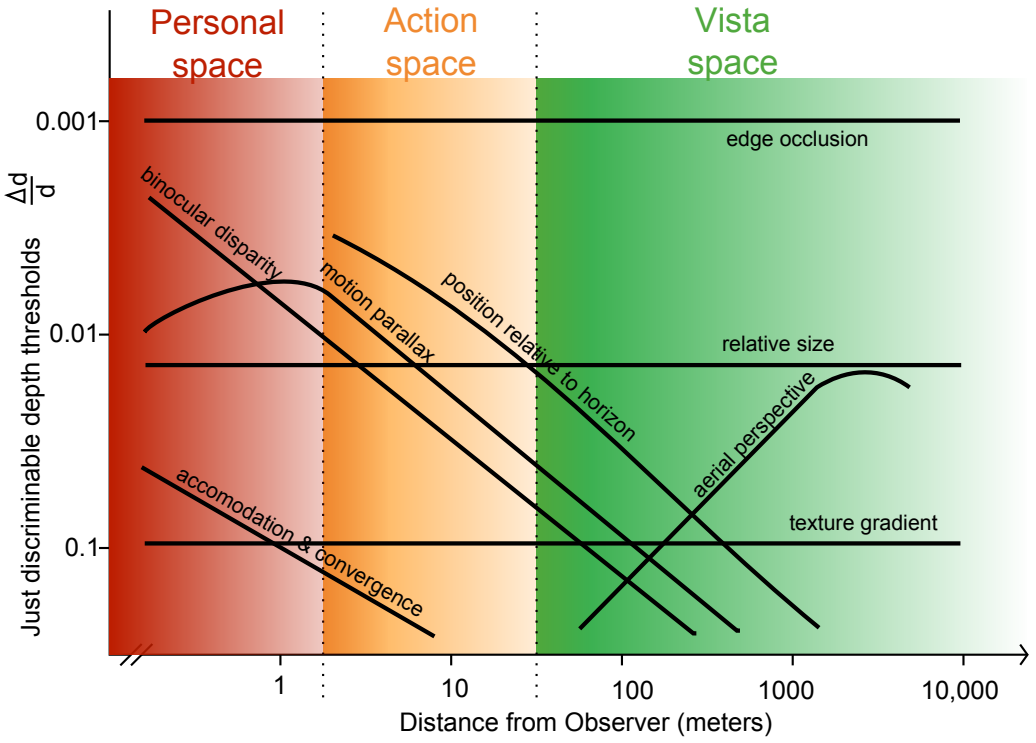



Figure 1.24: Just-discriminable depth thresholds versus the log-distance from the observer, from 0.5 to 5000 meters, for nine different sources of information about layout. Originated by Nagata (1981), extended and reproduced from Cutting and Vishton (1995). The array of function is idealized and can vary with the environmental conditions.

sources, each depth estimate being computed independently and in parallel. These estimates could then be integrated by different rules of combination such as average, addition and multiplication.

Because interactions among cues appear to be relevant for the human visual system, Landy et al. [68] later proposed the "modified weak fusion" that allows for certain interactions among depth sources. For example, the binocular disparity is a relative depth source, but coupled with an absolute source like accommodation or convergence, it could upgrade the original disparity to a level of absolute depth for the whole visual field, what they called the **promotion** to a "depth map".

Up to now, relatively few researches have been devoted to the big question of the visual combination of depth sources into a coherent representation of visible surfaces in a 3D visual world layout. The perceptual evidences are often conflicting and unsystematic; the computational research -helped by recent imaging technologies- on this subject is only beginning.

## Conclusion

There exist a large number of specific sources of depth information. They accomplish one of the most complex goal of vision: resolving the indeterminacy of spatial vision issued by the projection of the 3D world light onto the 2D retinae. The spatial vision acts to resolve the inverse problem of projection and gives effortlessly a coherent 3D perception of a scene to the viewer.

Even if the depth dimension is lost, the vision system relies on multiple cues (see the Figure 1.25) supposed to be integrated through different stages of image and surface processing to prove a coherent representation of the layout of the scene.

Figure 1.25: Painting of Gustave Caillebotte. Paris Street, Rainy Day, 1877. Art Institute of Chicago. This painting illustrates different pictorial depth cues.

Different classifications are distinguished according to anatomical (binocular or monocular), physical (optical or ocular, static or dynamic) or metrical (absolute or relative, quantitative or qualitative) characteristics. Beyond this distinction, it is important to see that some cues can interact to provide a veridical and upgraded representation of the depth.

The range of operation of the depth sources can help to give new insights on the plausibility of different depth cue interaction according to their depth range and depth sensitivity threshold.

In the path to a better comprehension of human vision and brain, the study of depth cue combination is an open and promising research area that without doubt will receive further attention.

CHAPTER 2

# Integrating the Binocular Disparity in Visual Attention: of the Importance of Foreground and Time

## 2.1 Introduction

As seen in previous chapters, the perception of the depth is a process of integration, combination and fusion of different depth cues.

The comprehension of the depth perception is actually a hot topic for video applications. With the digital era of communication, an emerging trend of named "3D" displays promote a new user-experience based on a new depth dimension. Actually, this new dimension consists of the display and projection of a pair of images onto our eyes that simulates one additional source of depth information, the binocular perception experienced in natural conditions from our three-dimensional world. Here the question of visual depth perception is tackled through visual attention, as a first process to perceptual consciousness.

The goal of this chapter is threefold. First, we study how one central element of stereoscopic vision, the binocular disparity, acts and affects the deployment of visual attention as a preliminary task to consciousness and perception. Second, regarding the sparse literature in the human vision community on computational models of visual attention considering the stereoscopy, we assess if this ability requires a reassessment -and then a potential gain- of existing visual attention models. Third, following this analysis, we aim to propose and include different modelling of depth-related visual features in a consistent visual attention model. The emphasis is on the evolution of the different visual features that interact over time.

### 2.1.1 Consideration of Binocular Disparity

As the depth processing is known to follow the bottom-up processing in the ventral pathway, it is interesting to assess how depth features might modulate and contribute to the visual attention over time.

As we have seen in the last chapter, the problem of recovering the distance to objects and surface in a scene is ambiguous and the visual system relies on a combination of different depth cues. While the monocular cues -available from one eye- give **relative** depth information on how far objects are relative to each other, (except for the accommodation and the familiar size cue, that give absolute information on the distance to object), the

binocular cues inform the absolute distance to objects. And contrary to convergence that gives distance signal with low depth resolution at short distances (up to 2 meters), the binocular disparity is useful at short and medium distances with a high discrimination of the depth thresholds [27]. Depth perception thus involves a combination of multiple but possibly conflicting depth cues to estimate the 3D structure of the surrounding visual scenes.

## 2.2 Past Works on Integration of Different Level Visual Features

### 2.2.1 Past Works on Depth Integration

There have been different suggestions to consider either the global depth, the stereo disparity or the stereo vision as individual features of a computational model of visual attention.

Maki et al. [80],[81] first proposed a computational model based on image flow, depth and motion detection. The depth is used to prioritize the targets so that the closer the objects are, the higher priorities they are given. The main limitation comes from this assumption, as the closest object is not necessarily the most salient.

Ouerhani et al. [99] also included the raw depth and some depth related features into Itti's model [56]. Depth was integrated as an additional feature and transformed into a conspicuity map based on center-surround mechanisms. The principle and consistency of depth integration were qualitatively illustrated.

More recently Zhang et al. [151] proposed to handle the stereoscopic visual attention. The raw depth is combined with motion and static saliency map (from Itti's model). The fusion of these three attributes with arbitrary weights is then performed. It is unfortunate that neither the comparison of model's performances with human observers, nor stereoscopic perception consideration was given.

Actually, one of the rare attempts to account for the stereoscopic perception is the stereo visual attention framework from Bruce and Tsotsos [14]. The selective tuning model of Tsotsos has been extended to address the binocular rivalry occurring in stereo vision. Unfortunately the model's performance was not given.

To summarize the past works on integration of depth in visual attention model, the main characteristics of each model are illustrated in Table 2.1.

| MODELS | CHARACTERISTICS | | |
|---|---|---|---|
| | based on low-level features | competition in fusion | binocular rivalry |
| Maki et al | yes (phase) | no | no |
| Ouerhani et al | no | yes | no |
| Zhang et al. | no | no | no |
| Bruce and Tsotsos | yes | yes | yes |

Table 2.1: Summary of the characteristics of existing models including depth or stereo disparity or stereo vision

### 2.2.2 Past Works on Integration of Central Bias

Among the numerous visual factors that might significantly influence our visual deployment, the central bias is one of the most prominent and early activated. It is often assumed that this effect results from motor biases in the saccadic system or from the central distribution of image features. However, Tatler [127] showed that the central fixation bias is irrespective of observer's task or image features distribution. The distribution of image features and fixation distributions from the original paper shown in Figure 2.1 illustrates this independence.

More recently, Bindemann [10] assesses the importance of laboratory setting -the bias to the center of the screen- by distinguishing between the central viewing to the **screen** center and the central viewing to the **scene** center. The observation on eye movements revealed a consistent central fixation bias immediately after scene presentation driven by both the location of the visual scene center and a complimentary tendency to direct the gaze to the center of the screen. Then his findings demonstrate that the central bias is both due a central viewing tendency in scene analysis, and a potential artifact in visual perception experiments on screen. The scene center effect suggests an advantageous position for extracting visual information from the scene, while the tendency to fixate the screen center might be purely an experimental artifact due to the onscreen presentation. This has proven to be difficult to remove in laboratory conditions, either by offsetting scenes from the screen center, or by varying the onscreen location of a preceding fixation marker, or by manipulating the relative salience of the screen, or by varying the distribution of visual features in a scene.

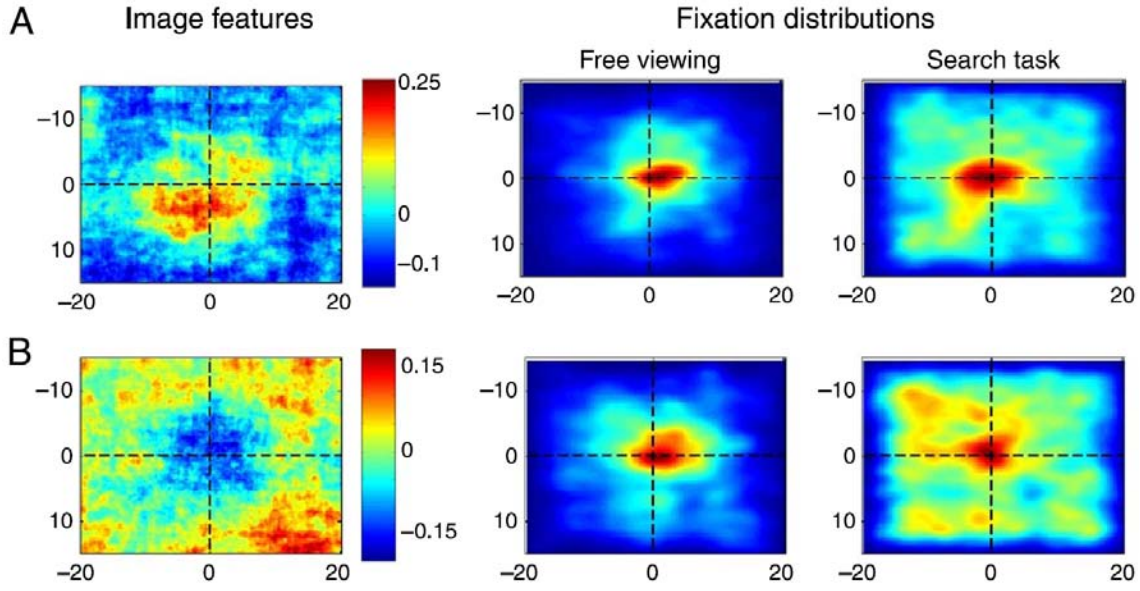



Figure 2.1: (A) Distributions of image features and fixations (in both free viewing and search conditions) for images in which there was a bias toward centrally distributed image features. The color bar for the image features plot shows the modulation in the image features across the distribution as a proportionate difference from the mean in the distribution. Fixation distributions are kernel density estimates. (B) Distributions of image features and fixations (in both free viewing and search conditions) for images in which image features were biased toward the periphery of the scenes. (Courtesy of Tatler, [127])

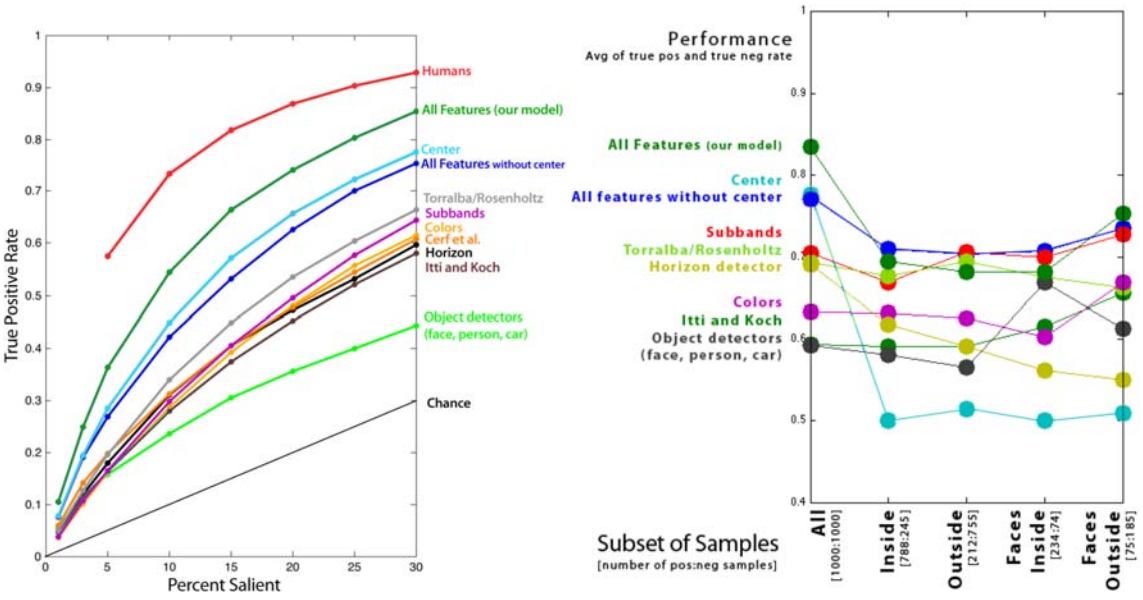
Be that as it may, the inclusion of the center bias in existing saliency models signifi-

cantly increases the performances [62],[152]. These models involve additional features to the model of the center bias and will be described in next section.

### 2.2.3 Past Works on Integration of Multiple Hierarchical-Level Visual Features

Multiple proposals and extensions of traditional low-level feature based visual attention models have been done since the emergence of visual attention computational models.

Judd et al. [62] proposed a global visual attention model on still image mixing the contributions of different cognitive level visual features. Over the intensity, orientation and color contrast low-level features plus subband based features, higher level features related to top-down vision mechanisms are proposed. A center and horizon prior are included as external a priori and mid-level features respectively. Face and person detectors are added as high-level features, necessitating a higher-resource cognitive task for human and face recognition. These features are then normalized to have zero mean and unit variance. A support vector machine trains a model on a selection of top salient and bottom salient locations on 903 training images. The ROC curves of performances for different sets of features are illustrated in Figure 2.2 (left).




(a) (b)

Figure 2.2: Left: the ROC curve of performances for SVMs trained on each set of features individually and combined together. Right: Average rate of true positives and true negatives for SVMs trained with different feature sets on different subsets of samples (Reproduced from [62]).

These results highlight and confirm the predominating role of center bias in prediction of visual attention. But one should be careful because it does not prove that this center prior is a sufficient feature to model visual attention. It effectively matches the central fixation distribution of humans (in [62] 40% of fixations lie within the center 11% of the image, 70% within the center 25%), but does not suffice to explain some of these fixations happening following the initial central fixation tendency. Indeed, Judd et al. performed additional tests on inside and outside center regions of the images. Results, illustrated on Figure 2.2 (right), show that in both inside and outside regions the center prior feature

performs as well as chance, while using all-features-without-center performs more robustly. Thus, it proves that the center prior can't model nor explain what is happening inside and outside of the center region.

Recently, [152] proposed a simpler saliency model based on low-level feature plus high-level face feature and center bias. The center-bias is modelled as a multiplication of any time-dependent center bias - a 2D Gaussian centered at fixation - and any time-independent center bias - a 2D Gaussian centered at screen center. Since the observed covariance matrix converges after 3-5 fixations, a single Gaussian filter is finally justified and used to model the central bias. As in Judd et al. [62], the results highlight the importance of mixing a set of optimal weighted features with a center prior.

A first proposal to consider together the relative contributions of depth information and central bias has been done recently by Vincent et al. [136]. In addition to potential high-level factors like lights and sky, the foreground and central bias contributions are quantitatively studied. Following a binary manual marking of regions containing these "object" features (foreground, sky, light, near light), two viewing experiences were conducted, the first with 9 images during 20 sec. of viewing times, the second with a shorter viewing time (5 sec.) but with much more natural images (68), larger sized light sources and additional features("look at edges" and "extreme light"). Results highlighted the potential role of foreground and central bias in saliency prediction, as shown in the figure 2.3, for the experience 1 (left) and 2 (right) respectively. The predominance of foreground contribution over central bias in experiment 1 and its opposite in experiment 2 might be explained by the small number of images used in experiment 1. Importantly, and contrary to [62],[152], the central feature is not centred to the screen center, but fitted to the fixation distribution. Then, the resulting averaged central bias is centered about two thirds of the way down from the top of the image. This might interfere with the foreground feature, usually located at the bottom in a scene.

Then the contributions of these different visual features were fixed over time. This study doesn't consider the potential time-dependent implication in visual attention of these features.

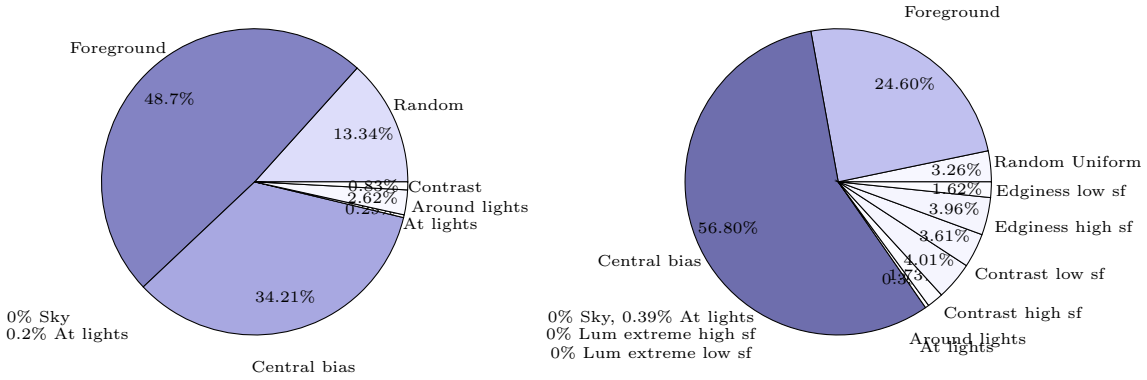



Figure 2.3: Decomposing the observed fixations into a number of proposed generators (diagram plotted from tables of Vincent et al.[136]). The two diagrams show the most likely probabilities for the seven (left) and twelve (right) possible fixation generators together with their 95% confidence limits assessed using a bootstrap technique, in experience 1 (left) and 2 (right). sf: spatial frequency

Ho Phuoc et al.[50] followed a similar methodology, but to study over time the role of

some low-level visual guiding factors. Following their statistical analysis of the evolution of feature weights, an adapted saliency model was proposed. The pooling of feature maps was based on a set of learned weights. However, as in Vincent [136], these weights were fixed over time.

In this chapter, we propose to design a time-dependent computational model of visual attention in order to predict where observers look on still pictures for 2D and 3D conditions.

The next section presents the materials as well as the eye tracking dataset from [58]. Behavioral and computational studies are presented in sections 2.4 and 2.5. The section 2.6 describes the proposed time-dependent saliency model as well as its performances. Thus, this chapter aims at answering 5 questions:

- Is the inter-observer congruency influenced by the binocular disparity?
- Does the binocular disparity affect spatial locations of: fixated areas, center and of a **hypothetical** depth biases?
- Is the predictability of state-of-the-art bottom-up saliency models affected by stereo disparity?
- How to model these center and depth-related bias effects as individual features?
- How to include these features into a time-dependent model? A new proposal for time-dependent saliency models.

## 2.3 Experimental Conditions and Methods

An eye-tracking database provided by Jansen et al. [58] is used in this paper. The experimental conditions, i.e. materials and methods to construct this database in 2D and 3D conditions, are reminded here. Stereoscopic images were acquired with a stereo rig composed of two digital cameras. In addition, a 3D laser scanner was used to measure the depth information of these pairs of images. By projecting the acquired depth onto the images and finding the stereo correspondence, disparity maps were then generated. The detailed information relative to stereoscopic and depth acquisition can be found in [124].

### Stimulus acquisition

The acquisition dataset is composed of 28 stereo images of forest, undistorted, cropped to 1280x1024 pixels, rectified and converted to grayscale. An illustration of an acquired forest images is given in Figure 2.4 (left).

(a) (b)

Figure 2.4: Original picture (left) and its disparity map (right) (black areas stand for the closest areas whereas the bright areas indicate the farthest ones).

### Stimulus Generation

A set of six stimuli was then generated from these image pairs with disparity information: 2D and 3D versions of natural, pink noise and white noise images. Our study focuses only on 2D and 3D versions of natural images of forest. In 2D conditions two copies of the left images were displayed on the left and right view of an auto-stereoscopic display. In 3D conditions the left and right image pair was displayed stereoscopically, introducing a binocular disparity to the 2D stimuli.

An illustration of the binocular disparity, as introduced by the stereoscopic screen, can be visualized in Figure 2.4 (right).

From [58]: "The 28 stimulus sets were split-up into 3 training, 1 position calibration and 24 main experiments sets. The training stimuli were necessary to allow the participant to become familiar with the 3D display and the stimulus types. The natural 3D image of the position calibration set was used as reference image for the participants to check their 3D percept."

#### Stimulus Presentation

A 2 view auto stereoscopic 18.1" display (C-s 3D display from SeeReal technologies, Dresden, Germany) was used for stimuli presentation. The main advantage of this kind of display is that it does not require special eyeglasses. A tracking system adjusts the two displayed views to the user position. A beam splitter in front of the LCD panel projects all odd columns to a dedicated angle of view, and all even ones to another. Then, through the tracking system, it ensures the left eye perceives always the odd columns and the right eye the even columns whatever the viewing position. A "3D" effect introducing binocular disparity is then provided by presenting a stereo image pair interlaced vertically. In the 2D condition, two identical left images are vertically interlaced.

#### Experimental Design

The experiment involved 14 participants. It was split into two sessions, one session comprising a training followed by two presentations separated by a short break. The task involved during presentation is of importance in regards to the literature on visual attention experiments. Here, instructions were given to the subjects to study carefully the images over the whole presentation time of 20 s.

They were also requested to press a button once they could perceive two depth layers in the image.

Finally, participants were asked to fixate a cross marker with zero disparity, i.e. on the screen plane, before each stimulus presentation. It has been shown in [127] that a random location of the prefixation marker do not allow to remove the central bias: the central location of the prefixation marker do not suffice to explain the central bias. The fixation corresponding to the prefixation marker was discarded, as each observer started to look at a center fixation cross before the stimulus onset and this would have biased the fixation to this region at the first fixation.

An "Eyelink II" head-mounted occulometer (SR Research, Osgoode, Ontario, Canada) recorded the eye movements. The eye position was tracked on both eyes, but only the left eye data were recorded; as the stimulus on this left eye was the same in 2D and 3D conditions (the left image), the binocular disparity factor was isolated and observable. Observers were placed at 60 cm from the screen. The stimuli presented subtended 34.1° horizontally and 25.9° vertically. Data with an angle less than 3.75° to the monitor frame were cropped.

### 2.3.1 Methods for Evaluation of Models

Different metrics or criteria exist to assess the performance of saliency models and their output saliency map relatively to the human fixation pattern. In this chapter the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) criterion is firstly used. Two ways exist to compute the AUC.

The first by comparing the human fixation distribution to a saliency map. Let $\gamma$ be the decision threshold such that the saliency value of a spatial position will be salient if it is over this value, and non-salient if it is under. By comparing the fixations to random points on the thresholded saliency map, a "true-positive' and "true-negative" classification rates are obtained. By varying the $\gamma$ threshold between the maximum and minimum value, a ROC curve is obtained representing the rate of true positive (when a fixation is positioned in the salient area) versus the false positive (when a random point is inside the salient area).

The second method compares directly the "fixation" density map (or human saliency map) with a predicted saliency map. The human saliency map is obtained by convolving the fixation distribution with a two-dimensional Gaussian of 1° to account for the fovea size and the eye-tracker accuracy. It is then thresholded by varying as previously a $\gamma$ between the maximum and minimum value. The predicted saliency map is thresholded in order to keep 20% of positive salient areas. Then, for all the points of the human and predicted saliency maps, a classification of true positive and true negative values is obtained. This second method will be used when the fixation points are too few (and the density of fixation point consequently too sparse) to reliably calculate an AUC score.

The Normalised Scanpath Salience (NSS) criterion [107] will also be used as a complimentary metric to assess the performance of saliency models. The saliency map is first normalized (centered, reduced) so that its mean is null and its standard deviation is unity. A NSS value for each fixation then corresponds to the value on this centered reduced saliency map at that fixation position. The NSS is calculated for each fixation and then averaged over all observer fixations on a dedicated image. The higher the NSS score, the closer the saliency map is corresponding to the human fixations. Contrary to AUC criterion, the NSS is not bounded, does not use random fixations and can give a score for each fixation.

In the following sections, either the spatial coordinates of visual fixations or ground-truth i.e. human saliency map is used. The human saliency map is obtained by convolving a 2D fixation map with a 2D Gaussian with full-width at half-maximum (FWHM) of one degree. This process is illustrated in Figure 2.5.

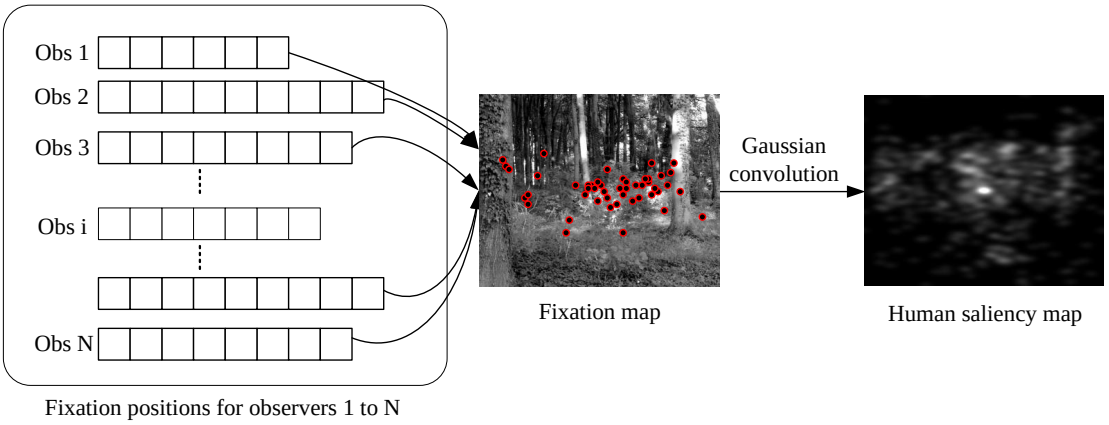



Figure 2.5: Illustration of the human saliency map computation from N observers.

## 2.4 Behavioral study: Impact of the Binocular Disparity

Jansen et al. [58] gave evidence that the introduction of disparity altered the basic properties of eye movement such as rate of fixation, saccade length, saccade dynamics, and fixation duration. They also showed that the presence of disparity influences the overt visual attention especially during the first seconds of viewing. Observers tend to look at closer locations at the beginning of viewing. By using a complementary approach, we examine this trend as well as the position of the salient areas. Four questions are here asked: first we investigate whether the disparity has an influence on the inter-observer variability. Second, we examine whether the disparity modifies the spatial locations of salient areas.

The third and fourth questions are related to the disparity influence on the center and the depth bias.

### 2.4.1 On the Inter-Observer Variability

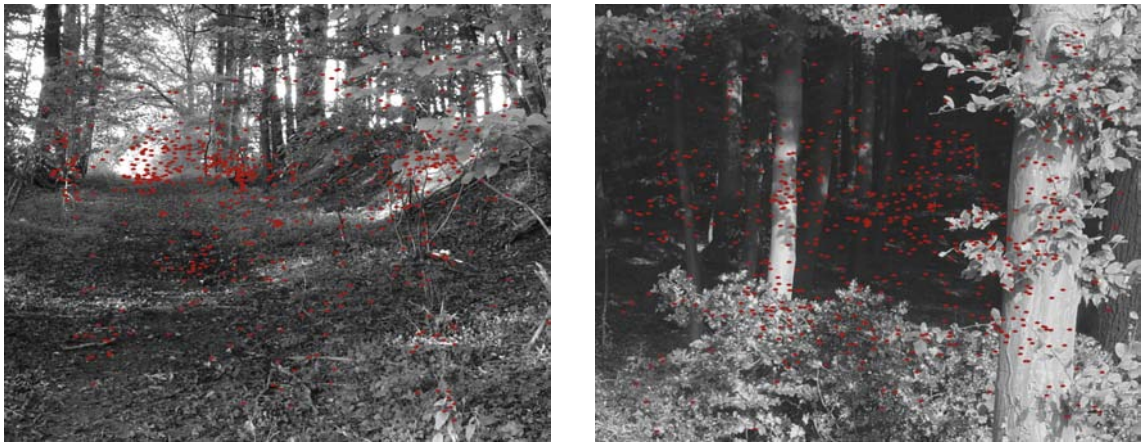
In [58], the congruency of fixation locations between 2D and 3D images was not investigated. However, it might be reasonable to think that the presence of disparity could modify the inter-observer congruency. It is indeed tempting to posit that a nearby object will attract more easily and more continuously our visual attention than a distant object. However the role of the disparity in the deployment of overt attention is still difficult to define. It is indeed worth remembering that, even when we look at a picture, we are able to infer the depth in an effortlessly manner. This inference is based on the use of different monocular depth cues (see a review in [101], p.204). The disparity might be considered as another cue that might be used to confirm or not, a preliminary depth perception based on monocular cues. To pursue the investigation related to the role of disparity, we evaluate the dispersion between observers.

To test whether the introduction of disparity has an effect on the inter-observer congruency, we use a one-against-all approach (also called leave one out) as in [130]. The first step consists in computing a 2D fixation distribution from the fixation data of all observers except one for a given picture. Each pixel of this map represents the probability to be fixated. The fixation distributions were then convolved with a two-dimensional Gaussian. The standard deviation of the Gaussian kernel is set at one degree to reflect estimates of fovea size. As the viewing distance is 60 cm and the height of the screen is of 23 cm, one degree of visual angle represents 40 pixels. This map is then thresholded to select an image area having the highest probability of being fixated. The threshold is adaptively set in order to keep 25% of the image. The goal is now to compute the percentage of the visual fixations of the remaining observer that fall within salient parts of the thresholded saliency map. This process was iterated for all observers. For a given picture, the congruency between observers is the average of the aforementioned percentage for all participants. A high value (equal to 1) would indicate that observers tend to fixate the same areas. Conversely, a low value would suggest that the scan patterns are uncorrelated meaning a strong variability between subjects. Before presenting the results, we would like to underline that this measurement does not take into account the fixation order nor the fixation duration.

Three analyses are performed. The first one is performed by using all visual fixations collected for the whole presentation time. The two others involve the first 20 and first 10 visual fixations. These three analyses are performed to disengage bottom-up contributions from top-down ones. These contributions are indeed time-dependent. The most common hypothesis is that the contribution of bottom-up mechanism is maximal just after the stimuli onset and is progressively overridden by the top-down mechanism. In particular, this was the conclusion of Parkhurst et al. study [104]. Another hypothesis defended by Tatler et al. [128] is that this is not the contribution of low-level saliency that decreases with viewing time but rather the top-down contributions that increase. Recent studies support this idea [70, 84, 37]. These studies, using different databases, have indeed shown that interesting objects correlate well with low-level saliency in natural scenes. The databases are composed of natural scenes in which objects of interest were manually marked. Although that participants have made a conscious choice to mark the object as being of interest or not, purely bottom-up models of visual attention succeed in predicting

these objects significantly above chance. This suggests that the bottom-up mechanism influences cognitive processes even after several seconds of viewing. In the margin of this question, there is a consensus on the consistency in fixation locations between observers. This consistency decreases with viewing time.

**Over the Whole Presentation Time** The mean inter-observer congruency for 2D condition (Mean (M)=0.61 Standard Deviation (SD)=0.135 Number of observations (N)=300) is not significantly higher than the mean inter-observers congruency for 3D condition (M=0.59, SD=0.130, N=300) using the two sample t-test for equal variances ($F(299)=1.08$, $p<0.24$, $t(598)=1.3$, $p<0.19$). The highest congruency is of 0.72 (for the picture (a) of Figure 2.6) whereas the lowest is equal to 0.50 (for the picture (b) of Figure 2.6). A random observer was also simulated: his scan path presented on average similar properties of eye movements to those of real observers. The congruency between the random observer and real observers is about 0.25.



(a) (b)

Figure 2.6: Pictures for which the congruency is maximal (2D condition (a)) and minimal (2D condition (b)). The red dots represent the human visual fixations.

**For the first twenty fixations,** the mean inter-observer congruency for 2D condition (M=0.63, SD=0.151, N=300) is not significantly lower than the mean inter-observer congruency for 3D condition (M=0.64, SD=0.157, N=300) using the two sample t-test for equal variances ($F(299)=0.93$,$p<0.28$), $t(598)=-0.3$, $p<0.76$).

**For the first Ten fixations,** the mean inter-observer congruency for 2D condition (M=0.71 SD=0.175 N=300) is not significantly higher than the mean inter-observers congruency for 3D condition (M=0.71 SD=0.164 N=300) using the two sample t-test for equal variances ($F(299)=1.13$,$p<0.13$), ($t(598)=-0.17$, $p<0.86$).

In conclusion, the disparity of the scene does not increase or decrease in a significant fashion the congruency between participants. This outcome is consistent over time. Results also indicate that the inter-observer variability increases over time. This is consistent with previous studies [128],[104]. Observers look more at the same locations just after the stimulus onset than after several seconds of viewing. The congruency over time

diminishes from 0.71 to 0.61. This decrease is statistically significant for the three tested configurations (10, 20 and all fixations) for both conditions.

### 2.4.2 Impact on the Fixated Areas

The previous results suggest observers are more consistent just after the stimulus onset than after several seconds of viewing. This first outcome is observed for both 2D and 3D conditions. The fact that the inter-observer congruency is very similar for both conditions over time does not indicate that the allocation of attention is the same for both conditions. The previous study [58] showed that the "presence of mean disparity changed the allocation of attention at the beginning of stimulus presentation".
The AUC curve is used to quantify the degree of similarity between 2D and 3D human saliency maps. The thresholded 3D human saliency map is then compared to the 2D human saliency map. For the 2D human saliency maps taken as reference, the threshold is set in order to keep 20% of the salient areas. For 3D human saliency maps, the threshold varies linearly in the range of 0 to 255.

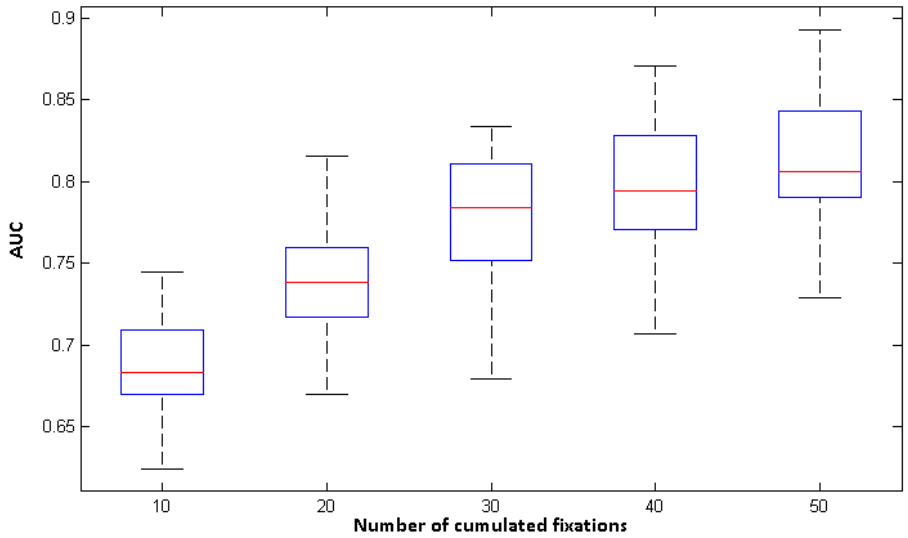



Figure 2.7: (a) Boxplot of the AUC values between 2D and 3D human (experimental) saliency maps as a function of the number of cumulated fixations (the top 20% 2D salient areas are kept).

Figure 2.7 shows the AUC scores between these 2D and 3D human saliency maps obtained with different viewing times (the first 10 fixations (1-10), the first 20 fixations (1-20), etc). Cumulated fixations over time allow here to deal with the increase of inter-observers dispersion over time and to put emphasis on salient areas due to the re-fixation trend over time [69].

The median value is equal to 0.81±0.008 (mean ± standard error of the mean). When analyzing only the first fixations, the similarity degree is the lowest. The similarity increases from 0.68 to 0.77 in a significant manner (paired t-test, $p<0.01$). Results suggest that the disparity influences the overt visual attention just after the stimuli onset.

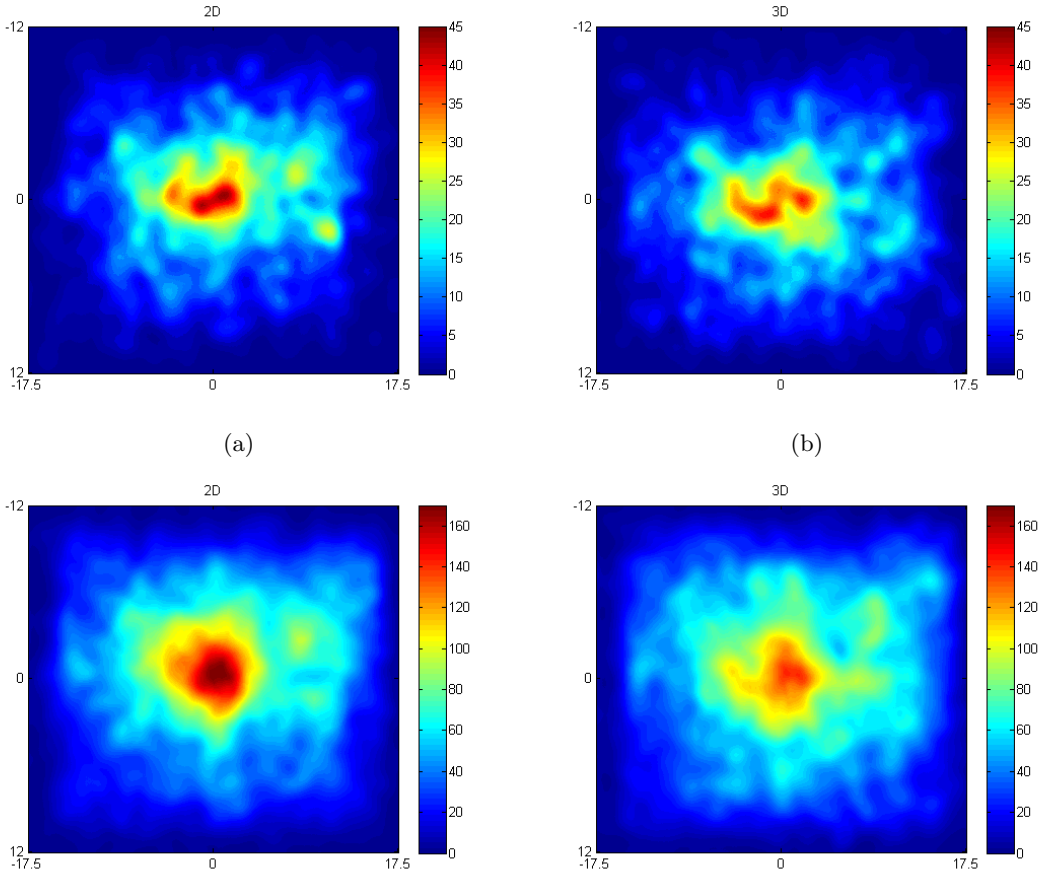
Although the method used to quantify the influence of stereo disparity on the allocation of attention is different from the work of Jansen et al. [58], we draw the same conclusion. The presence of disparity on still pictures has a time-dependent effect on our gaze. During the first seconds of viewing (enclosing the first 30 fixations), there is a significant difference

between the 2D and 3D human saliency maps.

### 2.4.3 Impact on the Center Bias

Previous studies have shown that observers tend to look more at the central regions than at the peripheral regions of a scene displayed on a screen. This tendency might be explained by a number of reasons (see for instance [127]). Recently, Bindemann [10] demonstrated that the center bias is partly due to an experimental artifact stemming from the onscreen presentation of visual scenes. He also showed that this tendency was difficult to remove in a laboratory setting. Does this central bias still exist when viewing 3D scenes? This is the question we address in this section.

When analyzing the fixation distribution, the central bias is observed for both 2D and 3D conditions. The highest values of the distribution are clustered around the center of the screen (see Figure 2.8).




(a) (b)

(c) (d)

Figure 2.8: (a) and (b) are the distributions of fixations for 2D and 3D condition respectively, from the first to the 10th fixation. (c) and (d) are the distributions of fixations for 2D and 3D condition respectively, for all the fixations.

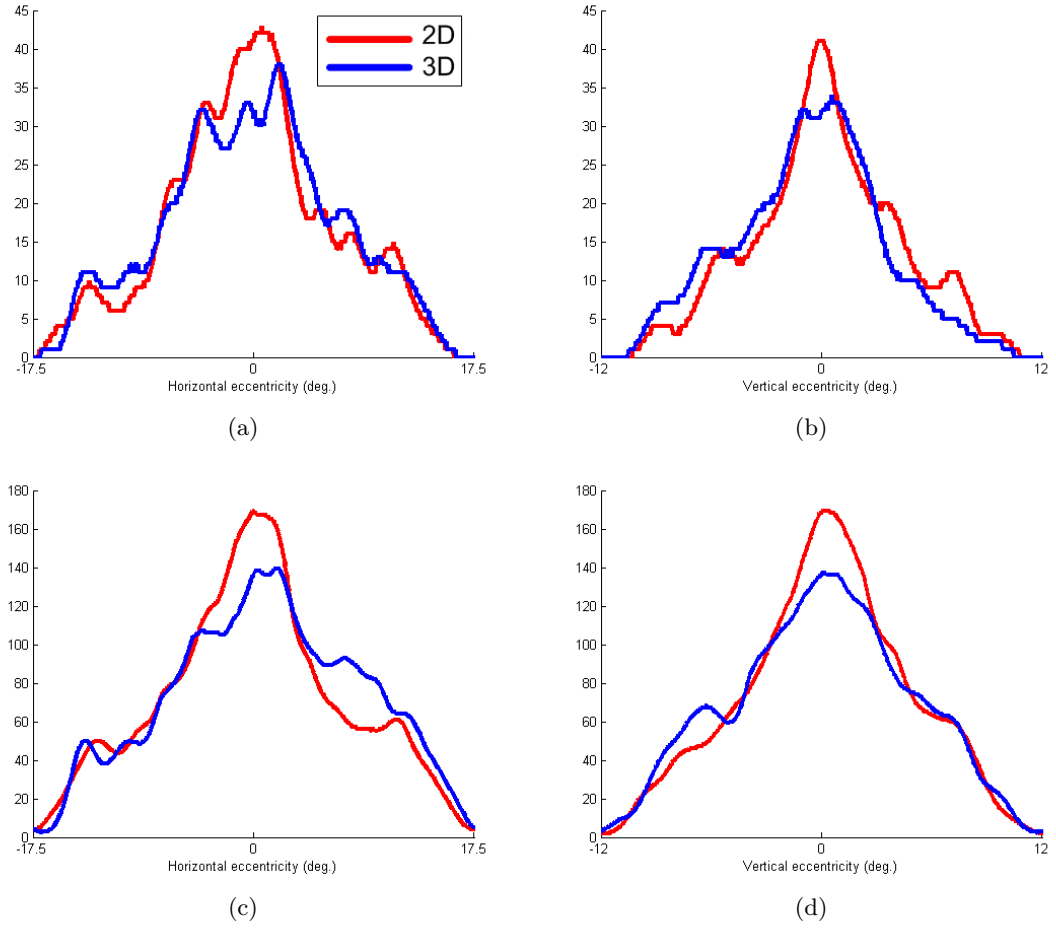



Figure 2.9: (a) and (b) represent the horizontal and vertical cross sections through the distribution shown in Figure 2.8 (a) and (b).
(c) and (d) represent the horizontal and vertical cross sections through the distribution shown in Figure 2.8 (c) and (d)

As expected, this bias is more pronounced just after the stimuli onset. To quantify these observations further, a 2x3 ANOVA with the factors 2D-3D (stereoscopy) and three slots of viewing times (called early, middle and late) is applied to the Euclidean distance of the visual fixations to the center of the screen. Each period is composed of ten fixations: early period consists of the first ten fixations, middle period consists of the next ten fixations and the late period is composed of the ten fixations occurring after the middle period (note that this is different from the previous analysis where cumulated fixations over time were used. This is here less appropriate since the center bias is time-dependent).

A 2x3 ANOVA shows a main effect of the stereoscopy factor $F(1, 6714) = 260.44$ $p<0.001$, a main effect of time $F(2, 6714) = 143.01$ $p<0.001$ and an interaction between both $F(2, 6714) = 87.16$ $p<0.001$. First the viewing time is an important but already known [127] factor, influencing the center bias. Just after the stimuli onset, the center bias is more pronounced than after several seconds of viewing. Second there is a significant difference of the central tendency between 2D and 3D conditions, for the three considered time periods.

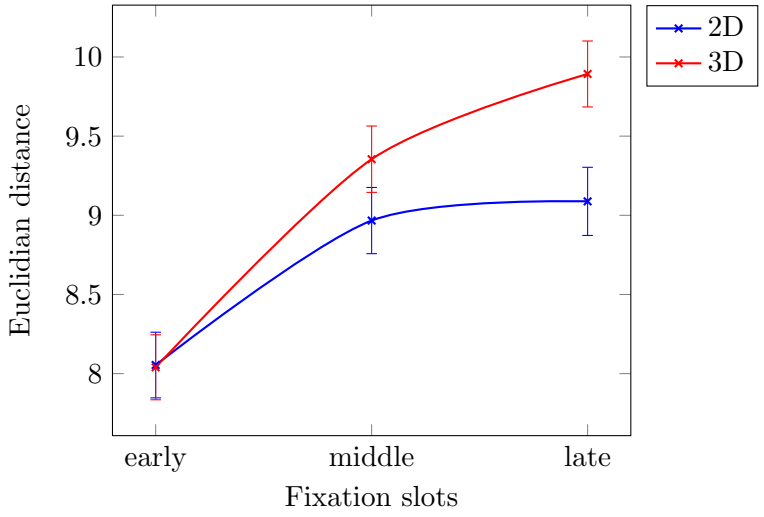



Figure 2.10: Average Euclidean distance between the screen center and fixation points. The error bars correspond to the SEM (Standard Error of the Mean)

Bonferroni t-tests, however, showed that the central tendency between 2D and 3D conditions is not statistically significant for the early periods as illustrated by Figure 2.10. For the middle and late periods, there is a significant difference in the central bias ($p<0.0001$ and $p \ll 0.001$, respectively).

### 2.4.4 Impact on the Displayed Disparity of Fixated Areas

In [58], a depth bias was found suggesting that observers tend to look more at closer areas just after the stimulus onset than at farther areas. A similar investigation is conducted here but with a different approach. Figure 2.4 (right) illustrates a disparity map: the lowest values represent the closest areas whereas the farthest areas are represented by the highest ones. Importantly, the disparity maps are not normalized and are linearly dependent on the acquired depth.

The mean fixated depth is measured from disparity maps for each fixation point in both conditions (2D and 3D). A neighborhood of one degree of visual angle centred on fixation points is taken in order to account for the fovea size. A 2x3 ANOVA with the factors 2D-3D (stereoscopy) and three slots of viewing times (called early, middle and

late) is performed to test the influence of the disparity on the gaze allocation. First the stereoscopy factor is significant $F(1, 6714) = 8.8$ $p<0.003$. The factor time is not significant $F(2,6714)=0.27$ $p<0.76$. Finally, a significant interaction is observed between both factors $F(2,6714)=4.16$ $p<0.05$. Bonferroni t-tests showed that the disparity has an influence at the beginning of the viewing (called early), ($p<0.0001$). There is no difference between 2D and 3D for the two other time periods, as illustrated by Figure 2.11. The observers effectively tend to allocate their attention to closer areas and significantly for the first ten fixations.

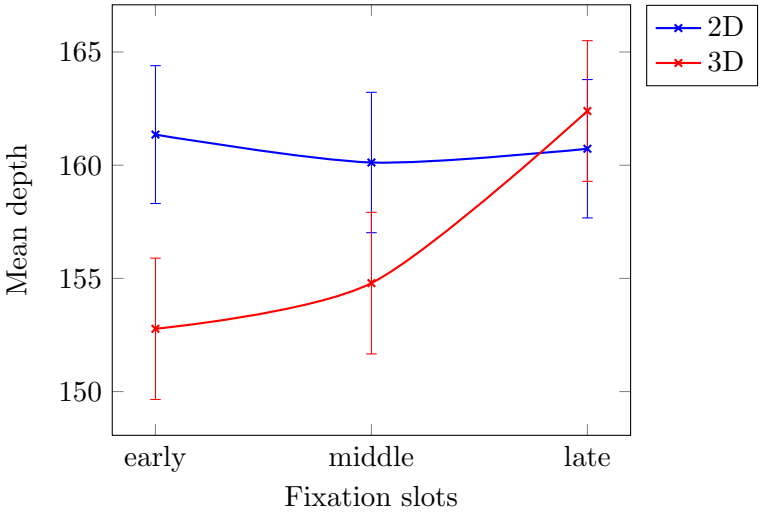



Figure 2.11: Mean fixated depth (on 8 bits) as a function of the viewing time (early, middle and late). The error bars correspond to SEM (Standard Error of the Mean)

## 2.5 Computational study: predictability of visual attention models in stereoscopic conditions

As shown in previous sections, the introduction of binocular disparity significantly impacts our gaze on still images especially on the first fixations. Indeed, the last analysis (see section 2.4.4) indicates that the disparity induced by the stereoscopic condition effectively impacts the visual deployment: in the stereo condition, we tend to look at closer locations on the first fixations. Beyond the impact of binocular disparity on eye movement properties, it is interesting to assess the extent to which computational models of visual attention are able to predict where observers look at in stereoscopic conditions.

### 2.5.1 Selected state-of-the-art models

In this study, three state-of-the-art models are used, two belonging to the biological inspired models and one to the statistical models:

- The model of Itti [56] was among the first to propose a method to compute a topographic saliency map from a color image. From the input image, some early visual spatial features (color, intensity, orientation) are extracted, filtered out and then normalized and fused together to generate a final saliency map.
- A second model, Bruce and Tsotsos's model [14] is based on the assumption that a rare event is probably more salient than a non rare event. Saliency is then obtained by using the self-information of image's patches.

- Le Meur et al. [72] proposed an extension of Itti's model by adding human perception properties such as Contrast Sensitivity Function (CSF), hierarchical decomposition and visual masking mechanisms.

The Figure 2.12 illustrates for a given picture the saliency maps computed by these three saliency models.

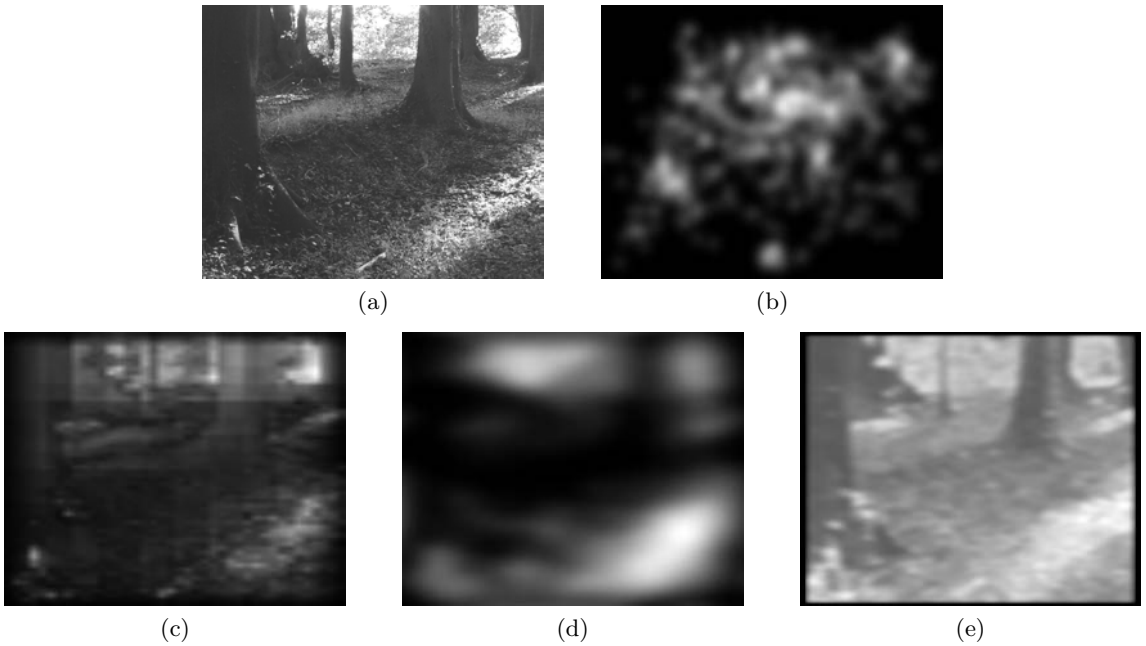


Figure 2.12: An original left luminance image (a), corresponding human (b), and predicted Itti (c), Le Meur (d) and Bruce (e) saliency maps.

### 2.5.2 Performance of models in both 2D and 3D conditions

As in section 2.4.1, the AUC (Area Under ROC Curve) is used, but to quantify the degree of similarity between human saliency maps (either in monoscopic i.e. "2D" or in stereoscopic i.e. "3D" conditions) and predicted saliency maps. As previous analyses showed that the impact of depth is time-dependent, it is interesting to test the accuracy of model's prediction on the same time periods. Figure 2.13 illustrates the performance of the models on three independent slots of viewing times. By considering the human saliency maps on the first 10, 20 and 30 fixations, the only model showing a mean saliency prediction in 2D conditions significantly different, and higher than in 3D conditions ($t(95) = 2.41$, $p<0.05$, $p=0.0088$) is Itti's model. By considering each fixation slot separately (i.e. we analyze the statistical difference for the first 10 fixations on the 24 AUC values, or for the 20 following etc), none of the models presents a mean saliency prediction in 2D condition significantly different than in 3D. One reason might be due to the small population involved in the test (24 pictures).

Based on previous observations, we would expect that, due to the depth bias occurring on the first fixations, these models would show a loss of accuracy in 3D conditions on the first fixations in term of saliency prediction. Moreover the performance gap between conditions would tend to reduce with the viewing time. Results are however contrasted: Itti's model results confirm the aforementioned hypothesis, (but the difference is not statistically significant) while the models of Le Meur and Bruce show no clear distinction.

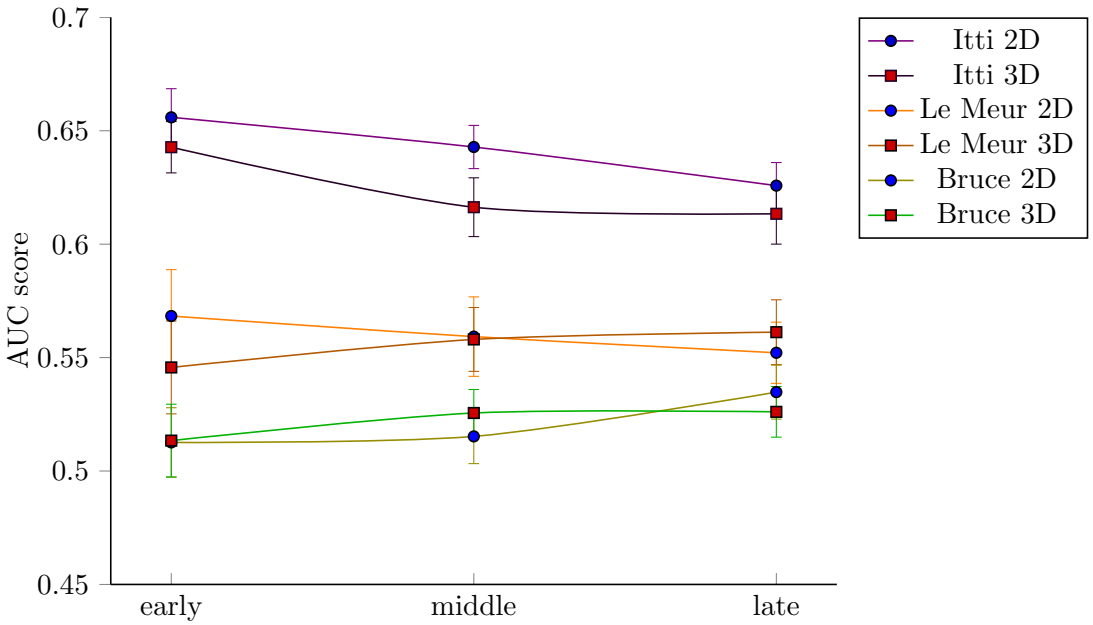



Figure 2.13: AUC values between predicted (i.e. model) saliency maps and 2D or 3D human saliency maps taken as reference (the top 30% salient areas are used) on the first 10, 20 and 30 fixation intervals, respectively early, middle and late.

Finally, in order to validate the fact that predicted saliency maps do not contain low-order features that would artificially increase the AUC values of each model, the predicted saliency maps are block-randomized for each model by a block size of 64 by 64 pixels. Mean AUC values for randomized saliency maps versus human 2D or 3D saliency maps confirm that the degree of similarity between randomized saliency maps and human ones is at the chance level. (Itti : 0.497 in 2D and 0.501 in 3D, Le Meur : 0.489 and 0.490 in 3D, Bruce : 0.496 and 0.499 in 3D).

### 2.5.3 Conclusion

In the behavioral first part of this study, we investigated whether the binocular disparity significantly impacts our gaze on still images. It is, especially on the first fixations. This depth cue induced by the stereoscopic condition indeed impacts our gaze strategy: in the stereo condition and for the first fixations, we tend to look more at closer locations. These confirm the work of Jansen et al. [58], and support the existence of a **depth** bias. The ability of existing models to predict where observers look at in stereoscopic conditions is relatively lower, especially for Itti's model. To improve their performance, these models are extended by taking into account the time-dependent depth bias. This is described in next sections.

## 2.6 Toward a Time-Dependent Saliency Model

Recent behavioral [127] and neuropsychological [153] studies have shown the importance and the influence of the « external biases » in the deployment of the pre-attentive visual attention. In itself, the degree to which visual attention is driven by stimulus dependent properties or task-and-observer dependent factors is an open debate [104], [128], [26]. But their mixing has proven to be relevant in order to improve the model predictability. [136, 62, 152]. Here we propose to integrate the learned interaction of these external features **over time** in a time-dependent visual attention model.

### 2.6.1 Statistical Analysis

Following the temporal behavioral study, we include the center and depth biases as potential guiding factors to existing visual attention models. In order to quantitatively evaluate the contribution of these factors, we followed a similar approach to Vincent's et al. one [136]. A statistical model of the fixation density function $f(x,t)$ is expressed in terms of an additive mixture of different features or modes, each associated with a given probability or weight. Then, each mode consists of an a priori guiding factor over all scenes. The density function is defined over all spatial fixation positions represented by the bi-dimensional variable $x$ so that:

$$f(x,t) = \sum_{k=1}^{K} p_k(t)\phi_k(x) \tag{2.1}$$

with $K$ the number of features, $\phi_k(x)$ the probability density function for each feature $k$ and $p_k(t)$ the contribution or weight of feature $k$ with the constraint that $\sum_{k=1}^{K} p_k = 1$ for a given time $t$. The statistical analysis aims at separating the contribution of the bottom-up saliency feature (itself based on low-level features) from additional features observed in the previous sections. To perform this analysis, each fixation is used separately to characterize the temporal evolution of contribution weights $p_k(t)$. An "Expectation-Maximization" (EM) method estimates the weights in order to maximize the global likelihood of the parametric model [33]. Before explaining this method, we describe the center and depth modeling.

#### Model of the Center Bias

The strongest bias underlined by laboratory experiments is the central bias, as underlined in section 2.4.3. Tatler [127] gave evidence that the central fixation tendency persists throughout the viewing in the free viewing condition, while rapidly dissipating in a search task. Indeed from the third fixation, the central bias is hardly noticeable. In our case of depth-layer detection task, the observers were asked to press a button as soon as they distinguished at least two depth layers in the image. Whatever the images, observations show a strong central fixation tendency on the earliest fixations followed by a sparser fixation distribution. As in the case of a search task in [127], there is little evidence for a central fixation bias from the third fixation. Considering the results of the literature and our observations, the central bias is modeled by a single 2D Gaussian. The use of a single Gaussian filter is empirically justified by the convergence property of the fixation distribution [152]. As proposed in [50], the parameters of the Gaussian function are predefined and are not estimated during the learning. On the present dataset, this choice is justified by the strong central fixation distribution on the first fixation that

goes into fast spreading and then tends to converge. The central bias is then modeled by a time-independent bi-dimensional Gaussian function, centred at the screen center as $N(0, \Sigma)$ with $\Sigma = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$ the covariance matrix and with $\sigma_x^2$ and $\sigma_y^2$. We fit the bidimensional Gaussian to the fixation distribution on the first fixation only. Whatever the viewing conditions (2D or 3D), the fixation distributions are similarly centred and Gaussian distributed ($\sigma_{x2D} = 4.7°$, $\sigma_{y2D} = 2.5°$, $\sigma_{x3D} = 4.3°$,$\sigma_{y3D} = 2.3°$)

### Model of the Depth Bias

Results presented in section 2.4.4 show that the perceived mean depth depends on the viewing conditions. At the beginning of viewing (early stage), the perceived mean depth is significantly lower in the 3D condition than in the 2D condition. Observers show a tendency to fixate more on the closest locations at the beginning of visualization (from the first fixations) than the farthest ones. This suggest that there exists an ability of the Human Visual System to attend selectively to dedicated closer regions.

This process of segregation of the closer things from the farther ones is called Figure-ground organization. This was first discovered by the Danish psychologist Edgar Rubin (1921) [114]. He based this distinction on a phenomenological analysis of the difference between subjective experiences for Figures and grounds. The Figure appears closer to the observer and the contour "belongs" to the figural region rather than to the ground. (At the opposite, the ground appears farther away and extends behind the contour.) It is indeed ecologically and evolutionary valid that people would attend to Figures rather than (or before) ground, because the Figure is closer and then of more interest, in the sense of perception for action.

Because the goal of perceptual organization is to construct a hierarchy consisting of parts, objects and groups, the Figure/ground process might be resulting from region segmentation, itself resulting from local edge detection. Figure/ground organization can then be understood as a sub-element of the edge interpretation depth cue. [101]. The existence of such perceptual organization support our proposal of Figure-ground implementation by a segregation of depth maps as individual foreground and background components. We employ a simple method of depth map segmentation without any fitting: these foreground/background maps have been obtained by thresholding at half the depth magnitude through a sigmoid function. Pixel values smaller and higher than 128 (on 8-bit disparity maps) rapidly cancel out on background and foreground respectively. Background values are inverted such that the farther a point is in the background, the more it contributes to the background feature. At the opposite end, the closer a pixel is to the foreground, the more it contributes to foreground feature. Two resulting foreground and background maps are illustrated on Figure 2.14(a).

### Proposed combination

The proposed model aims at predicting where we look at in 2D and 3D conditions. The prediction is based on a linear combination of low-level visual features, center and depth biases (see equation 2.1). However, other contributions much more complex than those mentioned above are likely to occur over time. For instance, a top-down process could interact with them, especially in the late time of fixation. To deal with this issue, an additional feature map whose fixation occurs at all locations with same probability is then used to model the influence of other factors such as prior knowledge, prior experience, etc. Obviously the contribution of the uniform map has to be as low as possible meaning that

other features (low-level saliency map, center and depth biases) are the most significant to predict where we look. In summary five feature maps are used as illustrated in 2.14(a):

- A first one is obtained by using one of the state-of-the-art bottom-up models (Itti, Bruce and Le Meur). This represents the "low-level saliency";
- one for the central fixation bias;
- two related to the depth cue, i.e. the foreground and background features;
- a uniform distribution feature

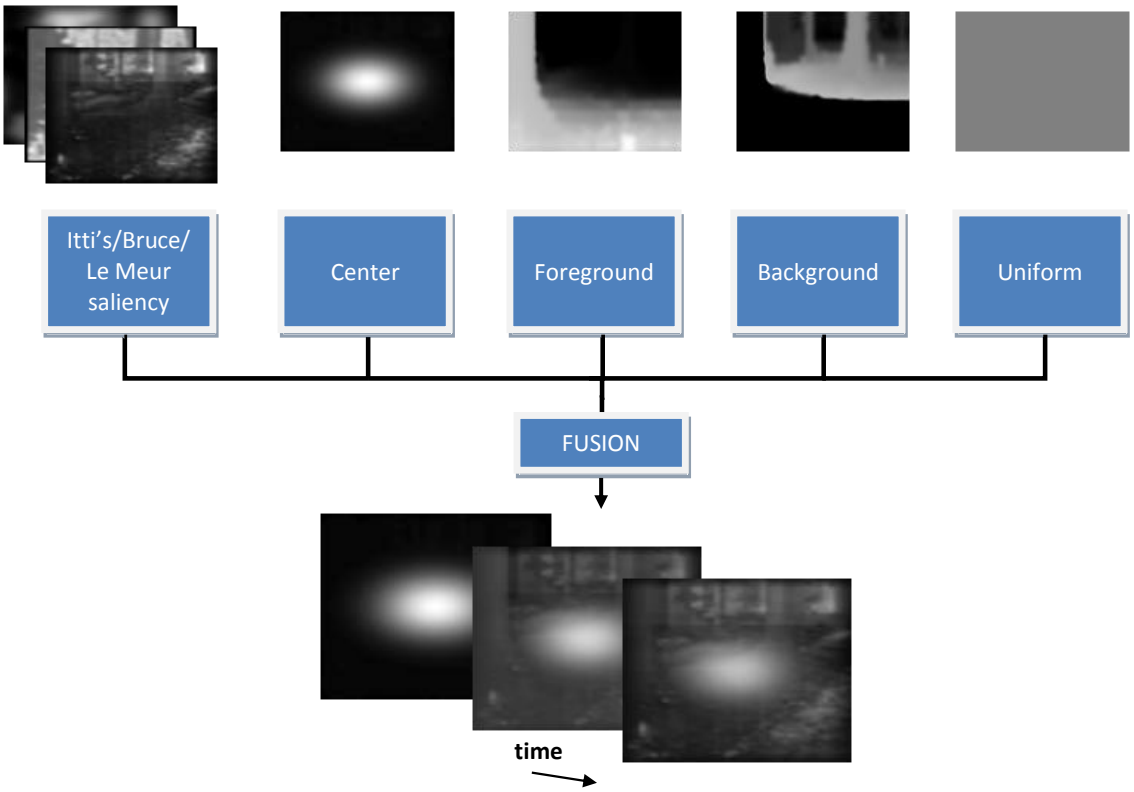



Figure 2.14: (a) Upper Row: Illustration of Itti's saliency map obtained from image originally presented in Figure 2.6 (a), center bias in 2D condition, corresponding foreground and background feature maps. (b) Middle row: Description of the proposed time-dependent model. (c) Lower Row: Illustration of the resulting time-dependent saliency map for the first, 10th and 20th fixation in 2D condition (when Itti's model is used to predict the bottom-up saliency map).

Low-level saliency, foreground and background features are dependent on the visual content. The center and uniform map represent higher-level cues. They are fixed over time and identical for all stimuli. The additive mixture model is then given by:

$$f(x,t) = p_{Sm}(t)\phi_{Sm}(x) + p_{Cb}(t)\phi_{Cb}(x) + p_{Fg}(t)\phi_{Fg}(x) + p_{Bg}(t)\phi_{Bg}(x) + p_{Un}(t)\phi_{Un}(x) \quad (2.2)$$

with $t$ the time variable incremented for each new observer's fixation, $\phi_{Sm}$ the saliency maps of one of the 3 models, $\phi_{Cb}$ the central Gaussian function, $\phi_{Fg}$ and $\phi_{Bg}$ the foreground and background map respectively and $\phi_{Un}$ the uniform density function. Each feature is

homogeneous to a probability density function. $p_{Sm}, p_{Cb}, p_{Fg}, p_{Bg}$ and $p_{Un}$ are the time-dependent weights to be estimated, their sum being equal to unity. The pseudo-code 2.1 describes the EM algorithm. The weights $p_k^{(m)}(t)$ are the only parameters estimated for each iteration $m$. In practice, a fix number $M$ of 50 iterations is a good trade-off between estimation quality and complexity.

**Algorithm 2.1:** Pseudo-code of the used EM algorithm

With $t_k = \{Sm, Cb, Fg, Bg, Un\}$ the estimated missing probability for each features
Initialization of the weights $p_k^{(0)}(t) = 1/K \quad \forall k$;
**for** *each fixation rank from 1 to 24* **do**
  **for** *each iteration* $m = 1..M$ **do**
    **for** *each feature* $k = 1..K$ **do**
      **for** *each participant* $i = 1..N$ **do**
        Expectation step: Given a current estimate of the parameters $p_k(t)$, $t_k$ is computed:
        $t_{i,k}^{(m)} = P\{x_i \text{ comes from the feature } k\}$
        $t_{i,k}^{(m)} = \frac{p_k^{(m-1)} \phi_k(x_i)}{\sum_{l=1}^{K} p_l^{(m-1)} \phi_l(x_i)}$;
      **end**
      Maximization step: The parameters $p_k^{(m)}(t)$ are updated for the iteration $m$:
      $p_k^{(m)}(t) = \frac{\sum_{i=1}^{N} t_{i,k}^{(m)}}{N}$;
    **end**
  **end**
**end**

The temporal contributions of the proposed features to visual attention are evaluated. The EM-based mixture model is run on half of the image dataset at each fixation rank (from the first to $24^{th}$ fixation): each fixation per observer is projected on all the feature maps associated with a given stimulus image. There are 14 participants and consequently at most 14 fixations per fixation rank per image. The process is repeated at each fixation rank, and with fixations in 2D and 3D conditions

### Results

The EM algorithm gives at convergence an estimation of the mixture weights maximizing the linear additive combination of different features with respect to the original human fixation distribution. The resulting temporal contributions of all the visual guiding factors are illustrated in 2.15.

The best predictor for both viewing conditions is the predicted low-level saliency (from Itti's model and called "Sm" in 2.15). As expected, the central fixation bias shows a strong contribution on the two first fixations but rapidly drops to an intermediate level between saliency (Sm) and other contributions. The contribution of the center bias (Cb) is significantly (paired t-test, $p<0.001$) more important in 3D condition than 2D condition, while the foreground (Fg) is significantly (paired t-test, $p<0.001$) more important in 2D condition than in 3D. Indeed the center bias is partially compensated first by the high foreground contribution from the $3^{rd}$ to the $18^{th}$ fixation, second by the progressive saliency

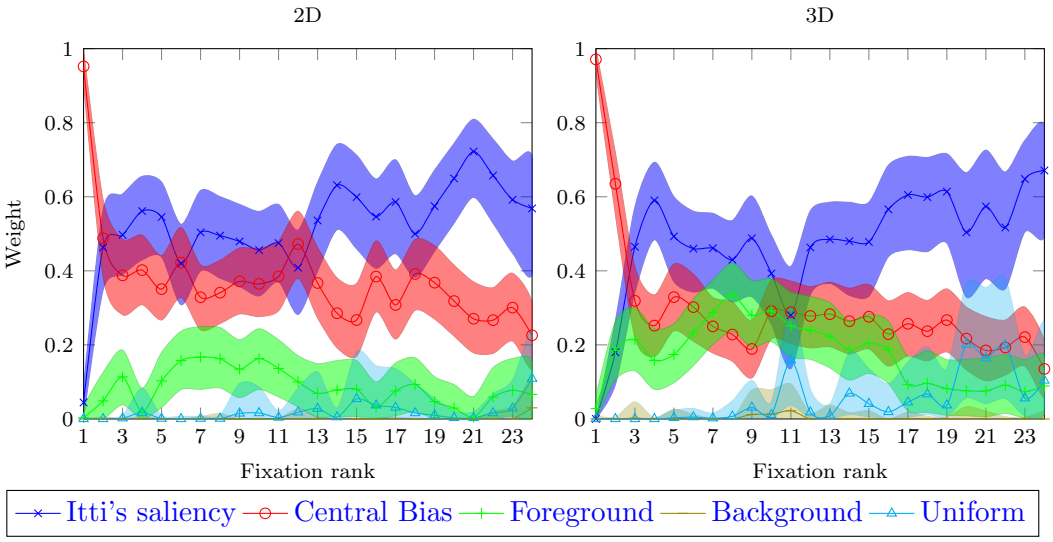



Figure 2.15: Temporal contributions (weights) of 5 features on 2D (left) and 3D (right) fixations to eye movements as a function of the fixation rank. Low-level saliency feature ("Sm") here comes from Itti's model. The error areas at 95% are computed by a "bootstrap" estimate (1000 replications).

increase. Finally, the background and uniform contributions remain steadily low in the 2D case, but increase progressively in the late period in 3D condition.

### Discussion

The temporal analysis gives a clear indication of what might guide the visual exploration on a fixation per fixation basis. We have considered different plausible features linearly combined with time-dependent weights. The temporal evolution of central bias, foreground and low-level saliency is highlighted.

According to our observation, the central bias is strong and paramount on first fixation, and decreases to a stable level from the third fixation. As shown by Tatler's experiments [127] and in accordance with [50], the central fixation point at the beginning of visualization is very probably not due to the central fixation marker before stimuli onset, but to a systematic tendency to recenter the eye to the screen center. Indeed, this tendency exists even with a marker positioned randomly within a circle of 10° radius from screen center [127]. Also, in these central bias observations and Tatler's findings (in search task), center bias was not evident from the third fixation. In our context, the contribution of center feature from third fixation is effectively lower but not negligible.

The binocular disparity introduction promotes the foreground feature up to the $17^{th}$ fixation. Results suggest that foreground helps to predict salient areas in the 2D condition but all the more in the stereo condition where its contribution is much more important. This is consistent with our previous conclusions (cf. section 2.4.4). It is known that different depth cues interact to drive the visual attention pre-attentively. Our results show that a depth-related feature like the foreground contributes to predict salient areas in monoscopic conditions, because depth can be inferred from many monoscopic depth cues (like accommodation, motion parallax, familiar size, edge interpretation, shading etc.). But our results also show that the binocular disparity greatly increases the contribution of foreground to visual attention deployment and indeed might participate to the Figure-

ground organization.

In contrast, the background feature does not contribute to visual attention deployment, or when it does (from the $22^{th}$ and $19^{th}$ fixation in 2D and 3D conditions respectively), it is combined with a contribution of uniform distribution. We could expect that observers tend to direct their gaze globally to background plane after viewing the foreground area at the very beginning of viewing. This is not the case: fixations can occur in the background, but observers do not show a common tendency of looking at the background from a certain fixation rank.

Finally, the contribution of the uniform distribution term remains low up to the late period of visualization. It models the influence of potential high-level factors possibly due to top-down mechanisms that are not accounted by our proposed factors. Results show these factors contribute few with temporal saliency construction on the first 20 fixations. Afterwards, the uniform distribution contribution increases over time suggesting that the existing features are not sufficient to explain the eye movements.

The temporal analysis is also reiterated with the low-level saliency maps of Bruce and Le Meur models. Results are very similar. In the following section, we use the learnt time-dependent weights to predict where observers look. Performance of the time-dependent saliency models is evaluated on the remaining half image dataset. The performance analysis is carried out from the first to the 19th fixations, a time slot for which the contribution of the uniform distribution is stable and low in all conditions.

### 2.6.2 Time-dependent Saliency Model

In the previous section, we have learnt through an EM algorithm the linear combination of five visual guiding factors matching the ground-truth visual saliency. The following step consists in using these weights to compute a saliency map taking into account the low-level visual features, the depth and the center bias. The same additive pooling of equation 2.2 is used.

For each fixation, the learned weights vary, leading to a time-dependent adapted saliency map. The time-dependent saliency model is then compared to corresponding original saliency model in 2D and 3D conditions. Three methods are evaluated in both 2D and 3D conditions:

- The original saliency model: the saliency map is the output of state-of-the-art models.
- The equally weighted model: the final saliency map is the average of the five feature maps. The weights $p_k(t)$ are not time-dependent and are set to $1/K$, where $K$ is equal to 5 in our study.
- The time-dependent saliency model: the time-dependent saliency map is the linear combination (cf. formula 2.2) using the learned and time-dependent weights $p_k(t)$.

In the second and third case, each feature is at first normalized as discrete probability density functions, (so that the sum of the whole values is equal to one) before weighting and summing all features.

The experimental dataset contained a reduced number (24) of images with different attributes of orientations, depth and contrast. The learning of the weights by EM algorithm was performed on half of this dataset, and the test with adapted saliency models on the remaining half images.

Thereafter, we use two comparison metrics to assess the performance of saliency models, i.e. their quality of fixation prediction.

Again, the ROC analysis is used. However, two saliency maps were compared in section 2.5.2. Here, to assess the performance for each fixation rank, the analysis is performed between a distribution of human fixations and a predicted saliency map. Then for each couple "image x fixation" (with each participant's fixation for a given fixation rank), an AUC score is obtained. Results are then averaged over all test pool images (12 images) for a given fixation rank.

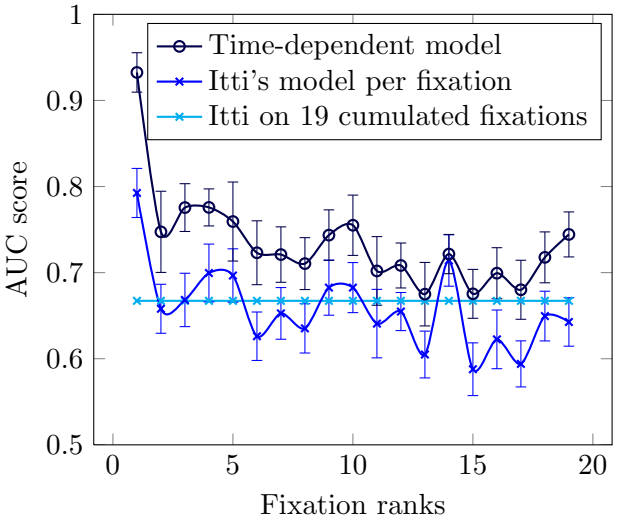



Figure 2.16: Temporal evolution of the performance of the time-dependent model based on Itti's, versus the Itti's model per fixation, and versus the Itti's model on 19 cumulated fixations in 2D conditions.

The AUC values of original Itti's model fixation per fixation are plotted in Figure 2.16 and compared to the performances of the time-dependent model. For reference, the AUC value between Itti's model and the first 19 cumulated fixations, as it is usually computed, is also plotted (light blue horizontal line). Results show a constant positive performance difference over time and emphasize the importance of time in the computational modelling of visual attention.

To strengthen the analysis, the "Normalized Scanpath Saliency" (NSS) [107] is also used to assess the performance of the normalized predicted saliency maps at the fixation positions. A NSS value is given for each couple "image x fixation/participant/fixation rank". Results are also averaged over all participants and all images for each fixation rank.

The Figure 2.17 illustrates the NSS and AUC performance for the 3 state-of-the-art and the proposed models, in 2D and 3D conditions, averaged over time. First we note that results are all much higher than the chance level (0 for NSS and 0.5 for AUC). Not surprisingly, models including the 5 visual features low-level saliency, center bias, foreground and background (plus the uniform feature) significantly outperform existing models for both metrics. The differences are all statistically significant (paired t-test, $p<0.05$) for both criterion in both conditions and for all saliency models (except in two cases marked "NS" on the Figure 2.17).

The time-dependent model based on Itti's low-level saliency ranks first, with a NSS score of 0.98 in 2D and 0.91 in 3D condition, and an AUC of 0.74 in 2D and 0.73 in 3D conditions. The final proposed method has greatly improved but also balances the

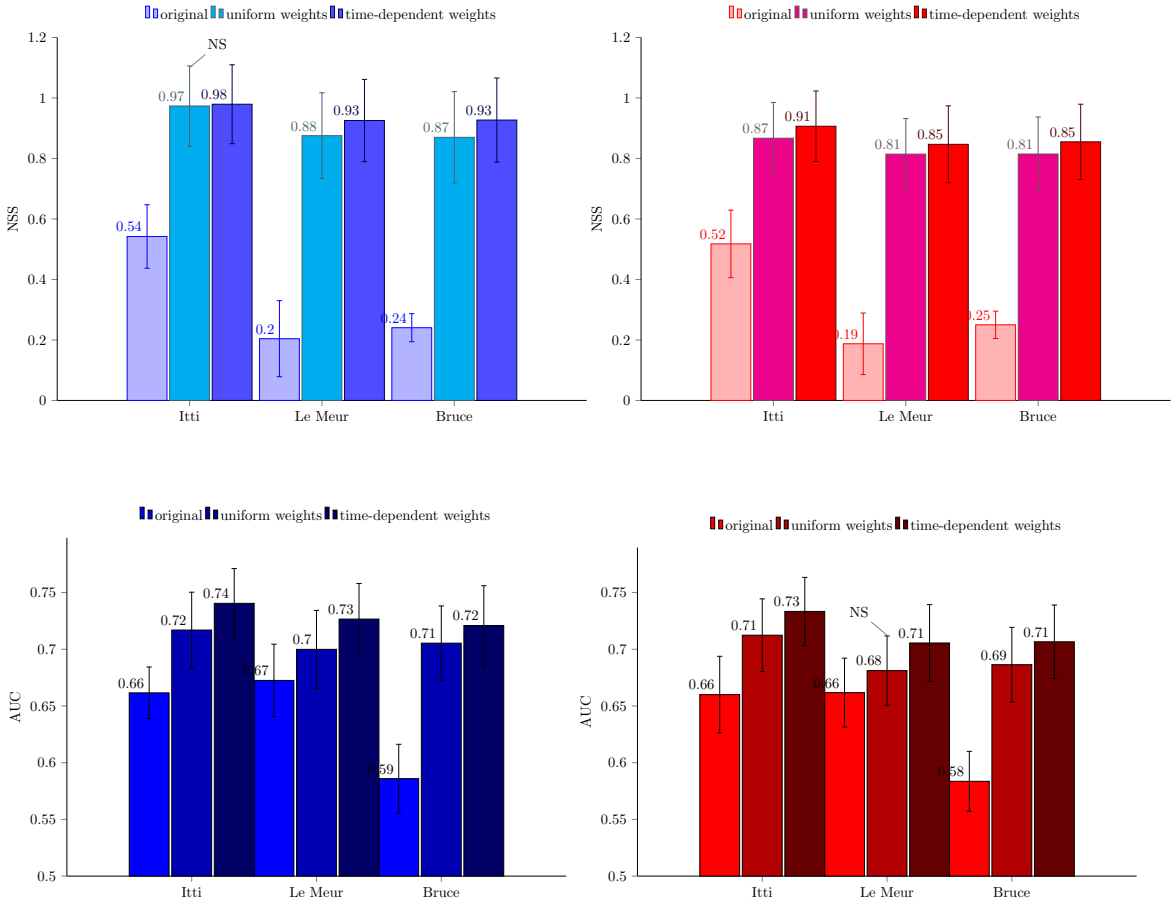



Figure 2.17: Comparisons of the performances of original, uniformly weighted and the time-dependent models (from the three selected models) in 2D (left) and 3D (right) conditions. Upper row: NSS score. Lower row: AUC score. The error bars correspond to the SEM. NS corresponds to Non-Significant. When the term NS is not indicated, results are significantly different ($p<0.05$).

performance between saliency models, both for NSS and AUC values. While the model using uniform weights without time adaptation leads to significant improvement, the time-dependent weighting increases the performance even more.

#### Discussion

The proposed approach based on time-dependent weighting improves the performance of existing attention models. By integrating different external and higher level feature contributions to three different existing models based on low-level visual features, the relevance of the saliency map has been increased in all viewing conditions and over time. There are however two limitations.

First of all, luminance only stimuli have been used for experiments. Even if colour might be a weak contributor to attention deployment relatively to luminance, it is however known that saliency models including color features improve their predictability [61]. From these statements and because low-level saliency models were run without the color component, we can argue that the contributions of low-level saliency features could be more important [70]. A second limitation is due to the content of the image itself. Natural scenes of forests were only presented to participants. Thus the depth perception, and foreground contribution in particular, might be influenced by the content of the scene itself, as well as by its geometry. A scene containing a single close object might induce a stronger foreground contribution on the early and middle period. However these remarks do not involve a reconsideration of our framework. Even if the importance of low-level saliency and foreground features might be modulated, the consideration of a pooling of low-level saliency with the foreground and the central feature is plausible and proved to be efficient on this dataset of images.

Importantly, the foreground feature might contribute significantly more to visual deployment when binocular disparity was presented to observers. Indeed binocular disparity constitutes an additional binocular depth cue to existing monocular ones to infer the depth from 2D retinal images. In the presence of this cue, observers do not only look closer in the first fixation instants. The findings also show that the foreground itself constitutes a good predictor and a plausible visual feature that contribute to a second stage of Figure-ground organization in the bottom-up visual attention.

## 2.7 Extended Results on a New Database: Perfomances and Limits of the Time-Dependent Model

### 2.7.1 Purposes

A supplementary study has been conducted to validate and extend the proposed time-dependent saliency model in different conditions of viewing and for different contexts of images. Thus the purpose of these experiments was to confirm and prove the validity of the model, its reliability with different experimental factors on various stimuli images.

Firstly the stimuli consist mainly of a mix of forest and urban scene images in contrast to the previous experiment restricted to forest images. Then, the learned time-dependent weights from the previous database could be tested on the same kind of forest (but colour) stimuli first, before being extended to different kinds of urban scene stimuli. All these stimuli are displayed in monoscopic conditions (2D images on a monoscopic screen), so

that the only changing factor is the context introduced by images: natural outdoor forest scenes and natural outdoor urban scenes.

Secondly, the tests were realized on colour stimuli in contrast to previous tests with graylevel images presented to the viewers. The low-level visual saliency models of Itti and Le Meur relied on a colour feature that could not be used and activated with the previous graylevel images.

Thirdly, the viewing conditions have been modified to ensure of the invariability of the results to different experimental conditions. The images stimulus are displayed natively vertically rather than horizontally, the screen is turned 90° to display stimuli in portrait format rather than in landscape format. This should allow the evaluation of the validity of the center-bias to different screen positions.

Also, the central prefixation marker displayed between stimuli images is replaced with a uniform white image. This to confirm that the center bias is not due to the presence of a central marker but to a systematic re-centering tendency of the HVS.

Finally, the duration of stimuli presentation is reduced from 20 to 7 seconds. This duration remains relatively long regarding most of the eyetracking experiments, but it was required to obtain a significant number of fixations.

### 2.7.2 Experimental Conditions

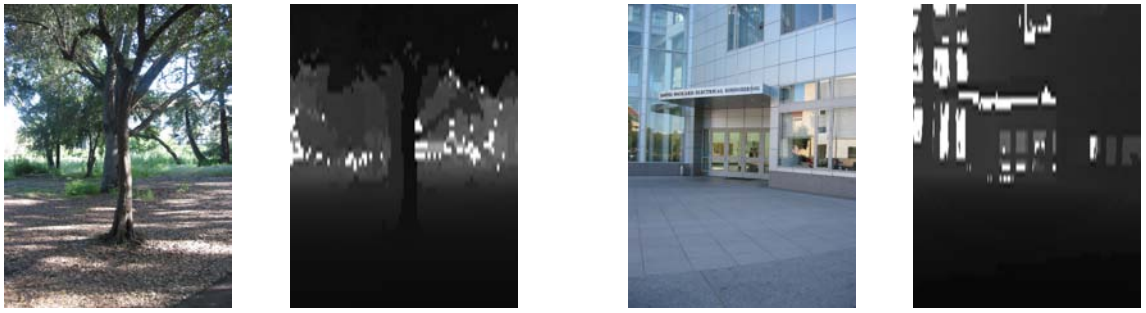
The stimulus acquisition and presentation are briefly described.

#### Stimulus Acquisition

The image and depth maps were acquired by Pr. Saxena et al. [115, 116] on the campus of Stanford University. The original purpose of the joint image plus depth acquisition was to learn and retrieve depth from single monocular images. A comparison of this estimation to the acquired ground truth depth enabled an objective measure of performance.

Various viewpoints of outdoor scenes have been acquired, involving either forest or park scenes, park scenes with urban buildings in the background or pure urban scenes. 40 forest and 40 urban scene images were manually selected as belonging to one or the other context. Two illustrations are given on Figure 2.18 with their corresponding depth maps. Please note that the glass in building illustrated in Figure 2.18(c) reflect the laser emitted to measure the depth; the corresponding depth in Figure 2.18(d) are believed to be very far while this is not the case in practice. The depth maps haven't been post-processed.

A Canon PowerShot S40 camera was used to shoot color images in portrait orientation. As for the preceding experiments, the depth maps where also acquired by a laser range scanner. A sparse depth map of 55x305 pixels was obtained for a corresponding 1600x1200 pixel resolution of the colour image. As with the previous database, an interpolation led to a smooth but blurred depth map at native displayed image resolution. However, the depth maps here are real depth maps, i.e. each pixel value expressed the distance in meters. The maps used by Jansen were disparity maps because they expressed the disparity between a pixel in the left view and its corresponding pixel in the right view. For such a typical dataset, the disparity is inversely proportional to the depth however (see equation 5.1.3 in section 5.1.3). The knowledge of depth in meters is important, because it may help to extend the simple figure/ground segregation at fixed threshold to a threshold in meters.



(a) (b) (c) (d)

Figure 2.18: An original color forest (a) and urban (c) image. Corresponding depth maps, normalized for clarity.

#### Stimulus Presentation

The acquired images were resized to the native resolution of a dell 20" display in portrait orientation (1600 pixels of heights times 1200 pixels width). The eighty images of forest and urban scenes were randomly mixed and displayed in two sessions of forty images. Each image stimulus was presented during 7 s. and separated from the next by a white image stimulus of 2 s.

#### Eyetracking

A "Facelab" eyetracker (Seeing Machines, Canberra, Australia) was used during the stimuli presentation. This system uses two infra-red cameras to acquire the eye scanpath during the stimuli presentation. The accuracy of the eyetracker is 1°. The position of the two eyes was recorded and led to a single on-screen position calculated from the eye vergence.

#### Experimental Design

17 observers with normal or corrected to normal vision participated in the tests. The experiment was split into two sessions of freeviewing a mix of forest and urban images. Instructions were given to the observers to study carefully the images over the whole presentation time of 7s. Observers were also told that a question paper follow the experiments, this to ensure the correct inspection of the visual scenes. Observers were placed at 70 cm from the screen, then the stimuli presented subtended 25°horizontally and 30°vertically.

A summary of the main differences between the experiments is presented in Table 2.2.

### 2.7.3 Results

The purpose of these extended tests is to assess the portability of our model to new image contexts and alternative viewing conditions. The time-dependent saliency model is then tested on this new database with the combination and weights learnt from the previous database of Jansen et al.

The performance over the time of the proposed model is assessed with the AUC and the NSS criterion. This is illustrated in Figure 2.19.

The results here come from the Itti's model with the colour feature disabled as in the previous experiments. The same investigation has been ran with the Itti's saliency maps with colour feature, and consequently with the adapted saliency maps with colour. Both

| EXPERIMENTAL CONDITIONS | DATABASE | |
|---|---|---|
| | **Jansen** | **Proposed** |
| Stimuli context | forest | forest & urban |
| Stimuli colour | grayscale | colour |
| Conditions | 2D & 3D | 2D |
| Screen orientation | horizontal | vertical |
| Prefixation marker | yes | no |
| Stimulus duration | 20 s. | 7 s. |
| Available depth | normalized (8bits) | in meters (16 bits) |

Table 2.2: Summary of the main differences in experimental conditions between both image, depth, and eye-tracking databases.

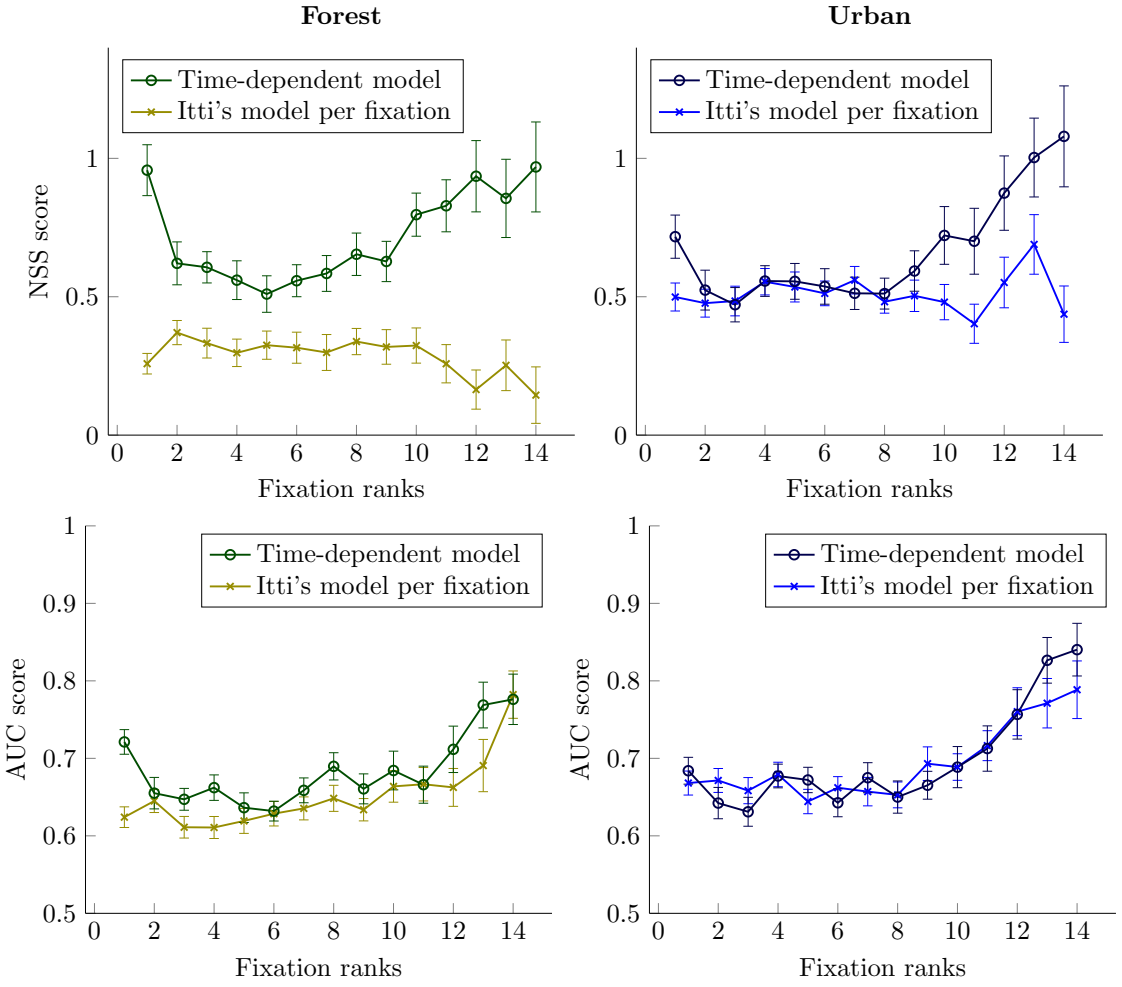



Figure 2.19: Temporal evolution of the NSS (top) and AUC (bottom) performances of the time-dependent model based on Itti's, versus the Itti's model per fixation for the forest scenes (left) and the urban scenes (right)

the AUC and NSS performances are very close, only the results without the color are displayed here.

First the results confirm the pertinence of the approach on the forest images. Both the AUC and the NSS scores are improved. However the AUC scores can merge with the original Itti's model score for some fixations.

The proposed approach shows strong limitations on the stimuli of urban scenes. The improvement exists for the first fixation and then for the fixations in the middle period, but it seems relatively weak when considering the confidence intervals of the AUC score.

Also, from the 9th fixation, there is a contradiction between the AUC and NSS scores and the potential improvement of the proposed model. The NSS scores shows significant gain of the method, but not the AUC. The variation of differences of NSS and AUC score are not correlated. This might be due to the small number of participants involved in the test (16), which give a reduce number of fixation per fixation rank and does not allow disambiguation.

The modelled central bias seems not adapted to catch the tendency to re-fixate the center of the screen or scene after presentation of urban scene stimuli and we might even wonder if it takes place. Different explanations will be discussed in the next section.

### 2.7.4 Discussion

The proposed approach confirms its capabilities of prediction on natural forest content. We could expect that the improvement might be weaker due to these unique 2D conditions but previous experiment results show that the method improved predictability both in 2D and 3D conditions. The weaker improvement, but also not regular over fixation rank- might be explained by other reasons such as the impact of the viewing conditions, the unadapted central bias to the new viewing distance and display orientation, etc.

The limited impact of color model performance proves the robustness of our model to the introduction of color feature. The performances of the time-dependent model are also improved for forest conditions relatively to the original Itti's model used with the colour feature. On one hand the supposed impact of colour on model performances is limited. On the other hand the proposed model appears very sensitive to viewing conditions and scene context. The previous conclusion must be reconsidered: the color is very probably not a limiting factor of the first experiment but the viewing conditions significantly impact the performances and must be integrated in the central bias and foreground calculation.

Finally, the context influences the gaze deployment and very probably from the first fixation. This does not amend the approach but indeed invites the integration of more visual features in a more complex visual attention model. In the urban scene, the gaze might be guided by the perspective depth cues up to a converging vanishing point. The orientation of the surfaces in depth might also hypothetically direct the saccades, as observed by Wexler et al. [143] on synthetic stimuli. This opens perspectives on the consideration of the scene and its spatial layout as a crucial element of attention and interaction.

## 2.8 Conclusion

The purpose of this study was to assess the differences in the visual deployment in monoscopic and stereoscopic viewing conditions, and to evaluate the contributions of relevant features that might participate in visual attention. In addition, we propose a new saliency model in which a time-dependent pooling of relevant features is used to predict where we look at on natural scenes.

Behavioral observations first underline that visual exploration in a depth layer detection task significantly differs with the introduction of binocular disparity. This lasts up to the 30th fixation. If the influence of viewing time on center bias is already demonstrated, our result suggests that this central tendency significantly differs between 2D and 3D conditions. This is particularly true in the middle and late time. Moreover a depth bias is also observed: participants tend to look at closer areas with the introduction of binocular disparity, and significantly in the early time.

Following these observations on external center and depth biases, some corresponding features are proposed. Low-level saliency, center, foreground and background visual guiding factors are integrated into a time-dependent statistical parametric model. These parameters are learnt from an experimental eye fixation dataset. The temporal evolution of these features underlines some successive contributions of center, then foreground feature with a constant implication of low-level visual saliency (from the third fixation). The strong contribution of the foreground feature, reinforced in the more natural presence of binocular disparity, makes the foreground a reliable saliency predictor in the early and middle time. Then, foreground integration constitutes a simple but biologically plausible way to incorporate a complex mechanism of Figure-ground discrimination for Figure selection as processed in V2 area [109]. A systematic recentering tendency followed by foreground selection are dedicated processes that might play an active role in the first instants of the human visual attention construction.

Finally, an adapted time-dependent saliency model based on an additive mixture and the pooling of 5 features is proposed. This model significantly outperforms three state-of-the-art models. Nevertheless, the additive pooling in itself in the integration of high level visual features is a strong hypothesis. As mentioned by [50] in the case of low-level feature combination, this hypothesis is very simple with regard to the complexity of visual attention construction [135] and with regard to other computational proposals of fusion [20]. However, it constitutes an attempt of integrating V1 low-level features with external and higher-level features that are known to occur later along the ventral pathway. All these high-level processes are complex and not well understood however. The extended results underline however the importance of considering the context, such as the GIST [98], into a more complex model of feature integration. Importantly, this adaptive methodology is applied at a stage where bottom-up and top-down factors are known to interact.

Final results highlight the importance of a temporal consideration of individual visual features, which are known to be processed specifically over time in the visual system. Integrating different features independently over time into a time-dependent saliency model is a coherent but also plausible way to model the visual attention.

# Part II

# 3D Video Coding and Applications to View Synthesis

CHAPTER 3

# Introduction to 3D Multi-View Representation and Coding

## Introduction

Today, a "3D" video system is a system able to render an additional depth sensation, often by the introduction of the binocular disparity to the viewer's perceived video. Such a system can provide 3D content in different forms, depending on the input source, its acquisition, the scene geometry, the desired level of quality, the type of application and the bandwidth. Then, the 3D-scene representation and compression are the key technologies between acquisition (or content generation), transmission and display (see Figure 3.1).

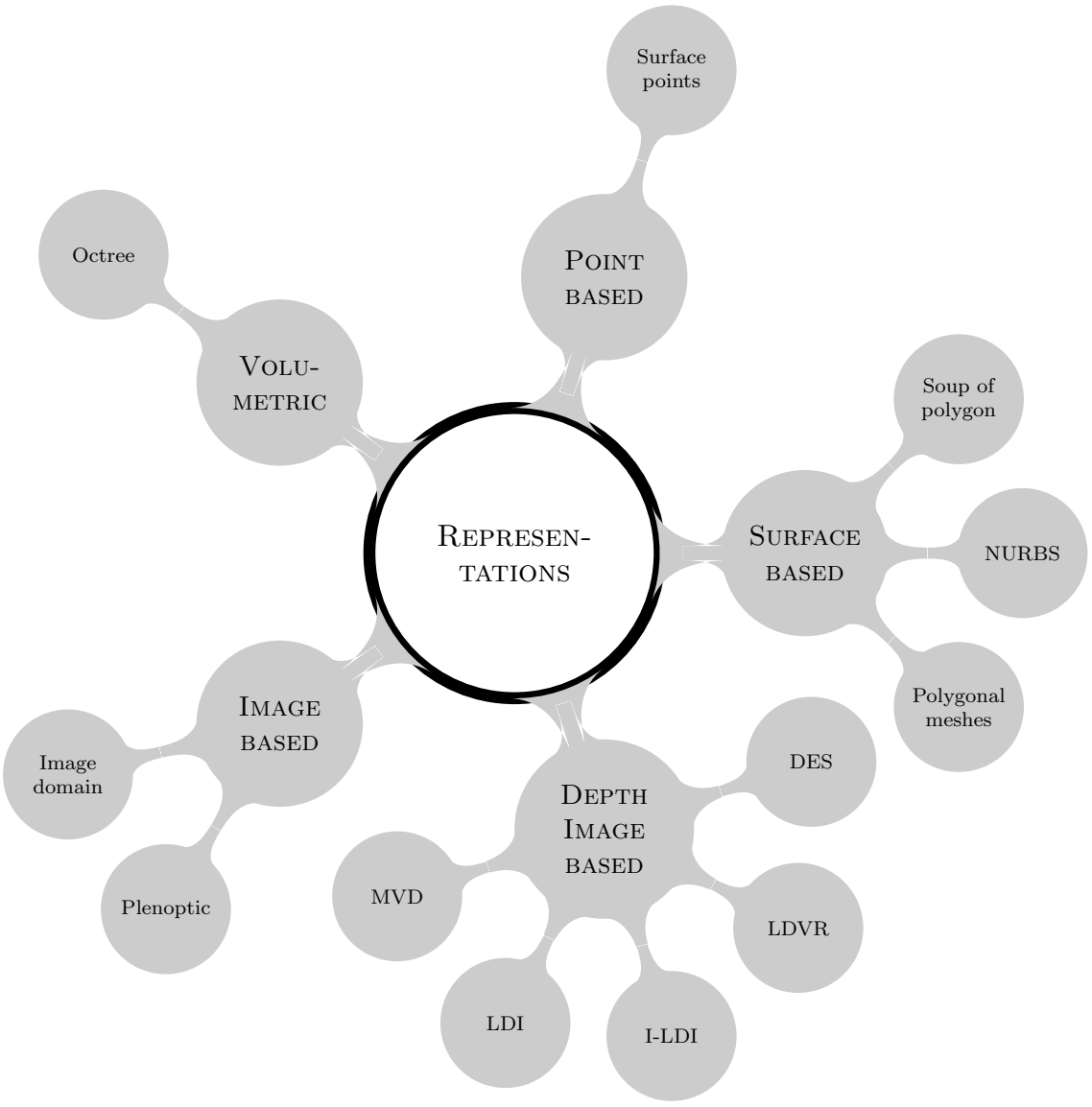
Especially, the compression stage will be highly dependent on the chosen representation, so both stages have to be jointly considered. Moreover the requirements for each stage may vary and may not always be compatible. Representation will be partly conditioned by the acquisition, but the choice of representation will depend not only on the compression but also the transmission, rendering, degree and mode of interactivity.

In this chapter an overview of the main state-of-the art representations for efficient compression of various 3D data is first drawn. These intermediate representations cannot be considered without their input data type, their complexity nor their possible output compression and rendering, which will be also described.

The second part is devoted to the subsequent 3D video coding standards. The history of the different 3D video standards is traced with the mechanisms they introduce, the representation they aim to support and the methods they inherit from corresponding 2D video standards.

Finally, a particular focus is done on the last multiview-plus-depth video codec being currently standardized. This format promises an efficient compression rate at perceptually good quality by the use of innovative techniques. Such coding and rendering techniques have been proposed as contribution to the scientific community, these are presented in chapter 4 and 5.

## 3.1 3D video representation




A survey of existing representations will be introduced. Keeping in mind the 3DTV and FTV applications and their requirements, they will be considered in terms of feasibility, compression efficiency or compactness, interactive rendering, level of detail and perceptual quality.

### 3.1.1 Requirements

**3DTV** implies that, whatever the type of display, a set of multiple videos should be displayed in real time. The stereoscopic effect is basically reproduced by the display of two slightly shifted views to each viewer's eye. The usage of more views guarantees the accessibility of more viewers on certain display technologies, but also a better viewing comfort (the user can move freely without necessity of glasses), thus increasing the viewing experience.

**Quality** and **consistency** between the multiple views has to be guaranteed over a higher level of quality on existing standards. Indeed, it is now widely admitted that 3DTV

could not be accepted if the quality perceived by viewer does not exceed the nowadays 2D quality standards of High Definition. The choice of the representation and compression methods, should also consider the **progressivity**, regarding the broadcasting methods in limited or noisy environment (cable, satellite, antennas, Internet). Finally, **backward compatibility**, i.e. the capacity of the codec -software- to include a former one, or the capacity of the hardware -set top box now widely used- to decode this new bitstream, is necessary for deployment. Typically, the end-to-end broadcast architecture should support for one channel the diffusion of multi-view videos and guarantee the display in real-time without jiggs nor any block effect and whatever the terminal computation capabilities.

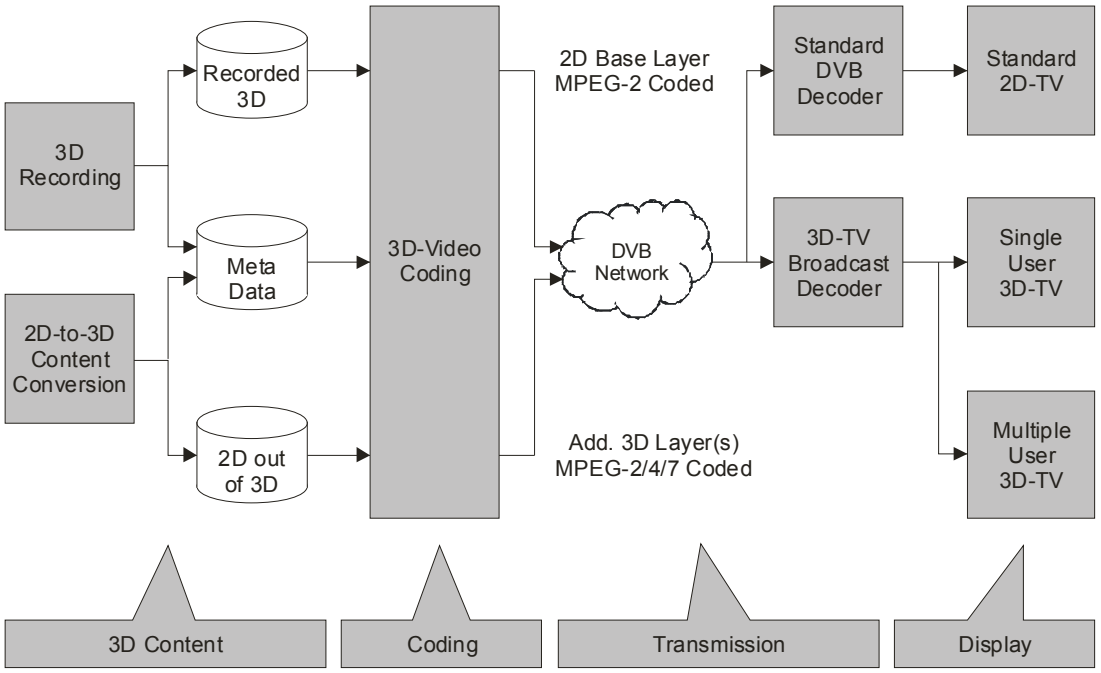



Figure 3.1: The ATTEST 3-D video processing chain

**Free Viewpoint TV** (or FTV, also called FVV) is a functionality for viewing and interactively controlling the viewpoint in a natural video scene. (For synthetic video, it is commonly called virtual reality). The viewer chooses where to move in the scene, and the new desired viewpoint -virtual camera- is displayed interactively.

So considered, the focus of attention will be partly controlled by the viewers rather than by a movie director, so each viewer may observe a unique viewpoint. Thus, the wider and denser the acquisition cameras are placed on an axis, the better the reconstruction quality will be. However, a correct trade-off depending on the application should be defined between the amount of camera -data- at capture, transmission capacity and complexity guaranteeing real-time rendering.

Seen from this point of view, the FTV requires especially **level of detail** (LoD) scalability: the viewer may want to move freely in the scene and zoom on parts of the texture. Despite this fact, different real time techniques such as interpolation by inpainting (see chapter 5) on the user side may be preferred. This means that high computational capabilities realized today by Graphics Processor Units should be supported on rendering devices

if FTV functionalities are expected. Efficient **compression** of this accurate 3D scene is also needed for transmission and broadcasting. **Flexibility and capability** of FTV are also important issues, space-time manipulation by the viewer implying real-time rendering at the price of high computation capabilities. Thus a **low complexity** 3D model might be necessary.

**Mutual requirements** Multi-view video acquisition can range from partial (about 30 degrees) to complete (360 degrees) coverage of the scene. Stereoscopic views can then be rendered and used both for 3DTV and FTV applications, once virtual views have been synthesized. 3DTV and FTV functionalities are then compatible but mutually utilizable.

In the end 3DTV and FTV share common requirements of flexibility, compression efficiency, and quality. In the case of FTV as an extra application of 3DTV where the free viewpoint navigation is realized commonly with stereo display, high levels of requirements have to be expected : progressivity, scalability, quality, and flexibility.

### 3.1.2 Image based representations

The representations based only on color images can be divided in two types, either describing the data in the image domain, through an array of pixels, or representing it as a light flow.

#### Image domain

Two existing solutions represent the information in the image domain only. The first one displays directly from cameras to back-end the same original images, either in a stereoscopic or in a multi-view manner.

**Conventional Stereoscopic Video** The stereoscopic mode consists of providing a binocular stereoscopic perception with a pair of displayed left and right videos i.e. with a stereo camera system for the acquisition and with a stereoscopic screen for the display, as illustrated in figure 3.2:

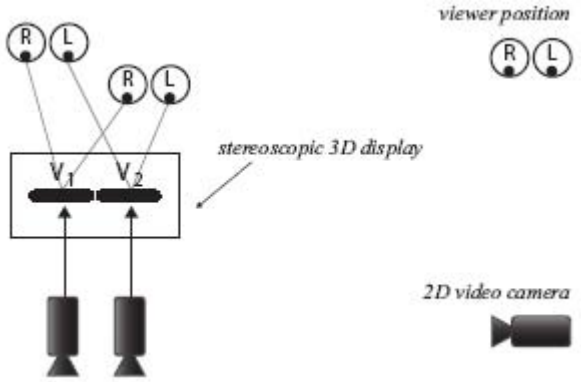



Figure 3.2: Efficient support of stereoscopic display based on stereo video content (from [30]).

A common way to represent and transmit these two video streams is to multiplex them temporally or spatially. Within the time multiplexed format, left and right pictures are interleaved temporally, as alternating frames. Whereas with spatial multiplexing, left and

right pictures are squeezed along the horizontal or vertical axis to fit in the original picture dimension, at the cost of a spatial resolution divided by two along this axis.

The main limitation of the stereo representation is the hardware acquisition dependency. The conditions of acquisition, especially the fixed baseline between the two cameras, are optimized for one type of stereoscopic display (regarding its size, type). Without much information other than two 2D views, occlusion-disocclusion and new view synthesis can hardly be supported without much intensive computation.

**Multiview Video** As described in the 3DTV requirements, we focus on the representations dedicated to multi-view rendering and its promising results of viewing quality and immersion. A first general representation is the Multi View Video (MVV) [85] representation. It commonly describes a set of consecutive views which acts like local stereo pairs to guarantee stereoscopy to the viewer (see Figure 3.3). The Head motion parallax, as described in Section 1.3.1 can then be supported within practical limits. Without any intermediate representation, MVV minimizes the image transformation, but suffers from flexibility and its high capacity channel requirements.

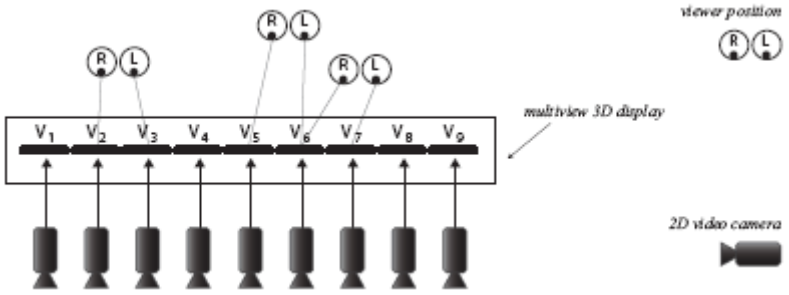



Figure 3.3: Efficient support of multiview autostereoscopic displays based on MVV content (from [30]).

The second one is based on a real-time estimation of an approximate geometry of the scene in order to generate virtual viewpoints [Zhang 04, Nozick 06, Taguchi 08]. In the first case, the acquisition and display processes are linked, flexibility is impossible. In the second case, the estimated geometry is likely to contain inaccuracies, resulting in rendering artefacts if the density of the sampling is sparse.

### Plenoptic function and light field

An alternative to the input 2D images -pixel array- from the camera consists of describing the flow of light : the plenoptic function [3]. It describes the light rays received in different direction of space, generally in a 7 dimension space. The light intensity $I$ is received at a 3D space viewpoint $(x, y, z)$, under a certain viewing direction $(\theta, \phi)$, with a certain wavelength $\lambda$ and at a considered time $t$. The information acquired by camera gives certain discrete values of this unknown function and an interpolation could then be applied on the known values. But acquiring the raw plenoptic function is not feasible in practice due to the heavy processes and the huge amount of data required. The question is then on how to reduce the dataset while keeping the rendering quality.

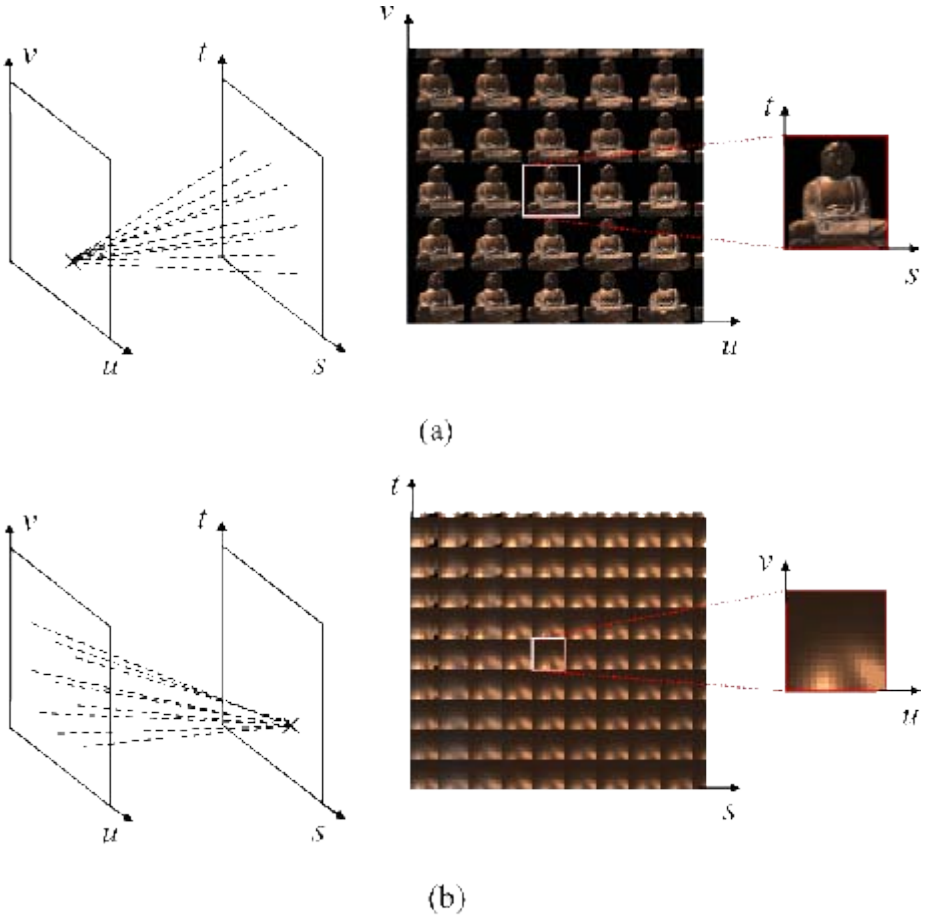



Figure 3.4: Two visualizations of a light field: $(a)$ each image in the array represents the rays arriving at one point on the $uv$ plane from all points on the $st$ plane, as shown on left.$(b)$ each image represents the rays leaving one point on the $st$ plane bound for all points on the $uv$ plane. The images in $(a)$ are off-axis perspective views of the scene, while the images in $(b)$ look like reflectance maps. The latter occurs because the object has been place astride the focal plane, making sets of rays leaving point on the focal plane similar in character to sets of rays leaving points on the object.

Levoy et al. [73] proposed making different assumptions, by ignoring wavelength and time dimensions first, which reduce the function to 5 dimensions. The light field, the radiance, is expressed as a function of position and direction, in regions of space assumed free of occluders. Consequently it does not change along a line in free space, and the

light field in this *free space* is then a 4D and not a 5D function. An image is then considered as a two dimension slice of this 4D light field: creating a light field from images corresponding to inserting each 2D slice into the 4D light field representation. At the opposite, generating new views corresponds to extracting and resampling a slice. Figure 3.4 presents two visualizations of a light field from two slices called the camera plane $(u, v)$ and the focal plane $(s, t)$.

Precisely, light rays are stored by the intersection of one plane with coordinates $(u, v)$, the other with coordinates $(s, t)$. During the rendering, each ray $r_i$ passes through the two planes and generates a particular sample $(u_i, v_i, s_i, t_i)$. If this sample already exists in the database, the color value is applied, if not, the nearest ones are selected and interpolated.

The advantages of plenoptic function and light field are their capacity to render photo-realistic images, but at the cost of a high camera density and a necessary high bitrate.

### 3.1.3 Depth-image based representations

A representation of a point in 3D-space can consist in a three dimension vector (or four in homogeneous coordinates). The depth, or distance to the referential -here the camera- expresses this third coordinate often called z. Now considering the projection of all the z coordinates of an object in space into an image plane viewed from a given camera, we obtain a 2D image called "depth map" or "Z-map". There exist various ways to obtain those maps. Either by real sensor acquisition, like a laser scanner (see section 2.3) or a Z-cam: it is possible to get a set of luminance and color points of a real object in 3D space, with their coordinates x,y,z. Either by stereo calculation (or stereo correspondence): the projection of pixel displacement from one view to another reflect, under assumptions and uncertainty, the disparity between these two views and then the depth map from one view. The depth map is said estimated.

The common usage of a depth map with its associated 2D color (or sometimes called "texture") image in the same coordinates enables the building of a coherent 3D-like representation in one view: a 2D+Z representation.

Knowing the position of a pixel in a 3D space, it can then be projected to another location in an image to render an arbitrary view of the scene through image warping. Occlusions would appear that need post-processing (see chapter 5). But, this reduced amount of transmitted data is particularly useful for free-viewpoint TV and 3DTV scenarios where novel views can be generated with this depth information. This is called **Depth Image Based Rendering** (DIBR). Then, compared to a configuration with a number of cameras equal to the number of possible views, the density of cameras over an axis can be subsequently reduced, or the viewing angle enlarged. We can intuitively consider that the larger the viewing angle, the better the 3D experience. In the case of the depth map is not transmitted but computed at the receiver side, this computational cost for rendering will limit the real time capabilities. Again, using depth maps in the representation constitutes another advantage. Finally the format of depth map - 2D one component map at the same resolution as video - makes the depth image based representation backward compatible to existing 2D TV digital coders.

In the next subsections, the evolution of the depth based representation will be described, from the 2D+Z format to the DES multi-layered multi-view based one, thanks to the gain of additional occlusion information and views.

### 2D+Z

A 2D+Z representation consists, for a given viewpoint, of a 2D image and its associated depth map. This pair allows the generation of a novel viewpoint for stereoscopy but in a relative short range limited by the disocclusion appearance.

Two further approaches are given to cope with the intrinsic limitation of a single texture view plus a depth map, the first one by multiplying the number of views and depth maps to transmit, the second approach using additional layers to transmit a part of the occluded regions and limit the disocclusion visibility.

### MVD

Considering the necessity of a larger viewing angle, both for free viewpoint and for 3DTV DIBR applications, a Multi-View video + Depth (MVD) representation can be considered (see Figure 3.5). This is a combination of the previous 2D+Z with MVV representations: multiple 2D videos are used with their associated depth video.

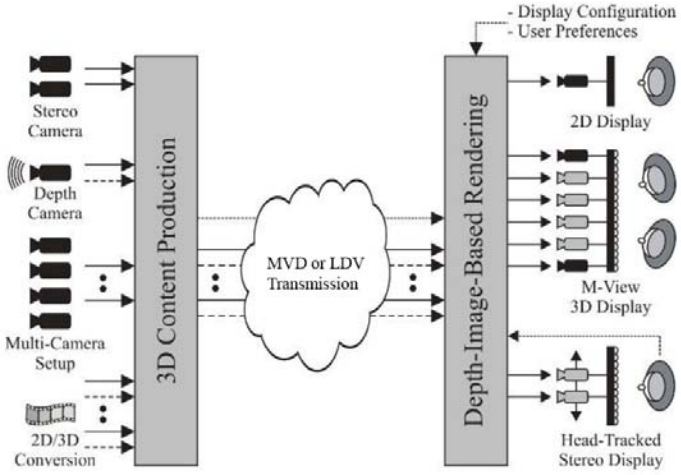



Figure 3.5: Possible scenario of a future 3DTV service, relying on MVD representation and transmission (from [64]).

Multitexturing - i.e. multiple camera views combining texture of the 3D scene - permits theoretically to increase the resolution and so the quality of the rendered images. The drawback is the relative rendering complexity and the high correlation of information between different views, leading to huge input view data-volumes to be compressed. Depth video has to be acquired or estimated for N views , then N-2D videos and N-depth videos have to be transmitted, and finally multiple virtual views have to be rendered, depending on the device, as illustrated in Figure 3.6 with an autostereoscopic display. Scalability and progressivity in MVD can however be considered, where a base layer is accessible for low complexity devices.

Concerning the rendering quality of MVD, some efforts have been devoted to improve it either during acquisition (camera calibration issues), representation, coding (like MVC, please refer to section 3.2) and displays. In the proposal of Zitnick et al. [155] a MVD representation is enhanced with matting information at depth discontinuitues. This matting information increases the rendering quality at the objects boundaries where pixel color values are usually mixed between background and foreground.

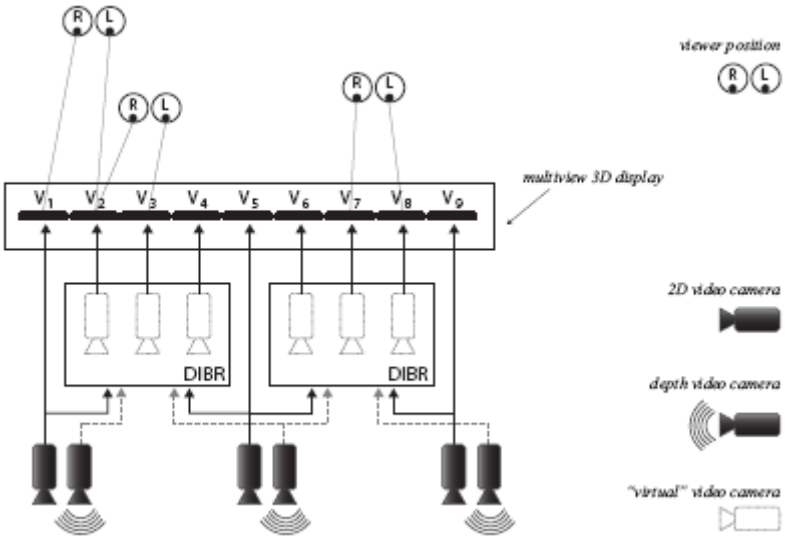



Figure 3.6: Support of multiview stereoscopic displays based on MVD content.

### LDI

The Layered Depth Image (LDI) representation [120] consists of representing the color and associated depth pixels in their consecutive positions along some depth layers. Then, a set of layers, a layered depth image, could store the distribution of relevant texture onto layers. Then it avoids the MVD limitation of storing redundant identical textures obtained from different views: the common textures are all fused and expressed in a single common view.

Concretely, a LDI is then a 3D matrix of visible and occluded pixels viewed from a reference camera. Each LDI pixel, i.e. Layered Depth Pixel (LDP), is composed of different Depth Pixels (DPs) carrying both color and depth information. Then, the main advantage of LDI lies in the reduction of the correlated data over the multi-view video sequences, at the expense of a computational cost of projection. Historically, the idea was to filter the depth values of the warped LDP, using a depth threshold $\Delta_z$, both for avoiding warping inaccuracy during construction [120] or for compression efficiency [150]. Cheng et al. [21] introduced a clustering over the depth pixels to avoid matting or ghosting effects.

### LDI extensions

In the next sections, different variants and extensions of the LDI over the time, space (among the views), -or over the ways to organize those layers- are presented.

**I-LDI** An alternative approach to reduce pixel redundancy between layers - hence to reduce filling rate - was presented by Jantet et al [60]. The layers were decorrelated using an incremental construction: a logical exclusion between a real view and a virtual view obtained from a temporary LDI enabled the computation of occluded areas that could be added to the reference viewpoint in the I-LDI.

While the LDI can contain many but partially empty layers, the I-LDI incremental approach, based on a mutual-exclusive construction, is supposed to carry the necessary information, especially at the occluded areas. Secondary texture and depth layers contain a better effective pixel distribution. Studies show that a clustered-based segmentation of

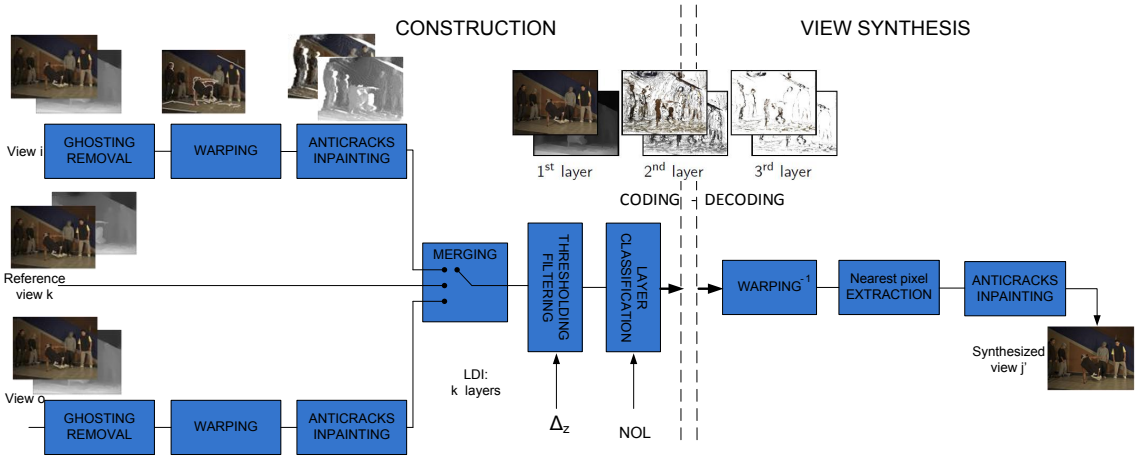



Figure 3.7: LDI construction and rendering scheme.

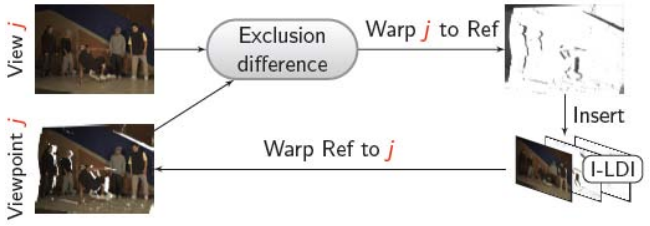



Figure 3.8: I-LDI exclusion-based construction scheme.

classical LDI can also help to reduce the layers completion rate and decrease the spreading of the layer's pixels.

**DES** Another extension to those LDV has been proposed by Smolic, Mueller and Merkle in [122]. To overcome the high variety of 3D video formats, they proposed "a generic, flexible and efficient format" combining the capabilities of the basic 3D video formats : the concept of what they call Depth Enhanced Stereo (DES).

This format, that can be seen as a container format, extends the conventional stereo with the LDI capabilities. This double-LDI representation provides stereo backward capabilities, but also enables depth-based view synthesis for autostereoscopic rendering with an improved quality over simple LDI. No implementation nor results have been proposed yet, but we can bet this generic solution has a promising future in standardization.

### 3.1.4 Surface-based representation

In this section we discuss three different representations: polygonal meshes, NURBS subdivision surfaces and polygon soup based on both 3D primitives and 2D textures.

#### Polygonal meshes

Polygons are widely used in the computer graphic community (from entertainment to manufacturing), as they are the primitives of hardware rendering technologies. Today's graphic card processes more than millions of polygons, enabling the rendering of realistic scenes containing complex objects. But such complex meshes are expensive to store, transmit and render. Many mesh simplification and compression techniques have been

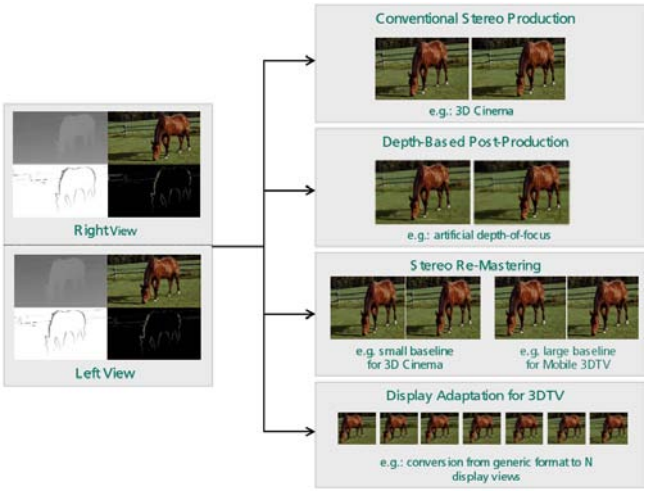
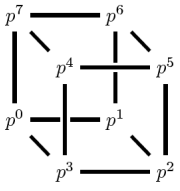



Figure 3.9: Depth enhanced stereo (DES), extending high quality stereo with advanced functionalities based on view synthesis. Reproduced from [122].

proposed to cope with this complexity and lead to different flexible representations with different levels of detail.

As the mesh consists of a set of polygons usually connected by their edges (but not always, see section 1.4.7 about the depth edges), the VRML format proposes storing separately into two tables the geometry and the connectivity as illustrated in Figure 3.10. The geometry of a mesh consists of a list of the vertices of the mesh with their 3D coordinates. The connectivity of the same mesh is stored in a second circular list of connected vertices.




(a) Meshes.

| id | X | Y | Z |
|---|---|---|---|
| $p^0$ | 0 | 0 | 0 |
| $p^1$ | 0 | 0 | 1 |
| $p^2$ | 0 | 1 | 1 |
| $p^3$ | 0 | 1 | 0 |
| $p^4$ | 1 | 1 | 0 |
| $p^5$ | 1 | 1 | 1 |
| $p^6$ | 1 | 0 | 1 |
| $p^7$ | 1 | 0 | 0 |

(b) Geometry.

| id | Nodes |
|---|---|
| $a^0$ | $p^0\ p^1\ p^2\ p^3\ p^0$ |
| $a^1$ | $p^0\ p^1\ p^6\ p^7\ p^0$ |
| $a^4$ | $p^0\ p^3\ p^4\ p^7\ p^0$ |
| $a^2$ | $p^1\ p^2\ p^5\ p^6\ p^1$ |
| $a^3$ | $p^2\ p^3\ p^4\ p^5\ p^2$ |
| $a^5$ | $p^4\ p^5\ p^6\ p^7\ p^4$ |

(c) Connectivity.

Figure 3.10: VRML representation of a simple cube in meshes.

Further techniques exist that rely on the scalability allowed by the meshes, the level of detail being adapted to the context of rendering.

The *progressive meshes* is a seminal technique, where an arbitrary triangular mesh can be stored as a coarser mesh with a sequence of mesh refinement operations called *vertex splits*. It consists of a local elementary mesh transformation that adds a single vertex to the mesh. Then a continuous sequence of meshes can be represented with increasing accuracy depending on the viewpoint. Indeed, polygonal artifacts can appear along the silhouette boundaries, especially for a close viewpoint or for low resolution representation. A view-dependent rendering combined with an associated transmission strategy can selectively refine a progressive mesh along object boundaries for a given viewpoint by using such vertex split operations [51].

The *progressive time-varying meshes*: A static connectivity along the time, i.e for all frames of the animation, with transmission of the vertex positions, is an efficient and space-saving technique, but which often leads to inadequate modeling of a deformable

surface. A progressive scheme based on edge splits (contractions) to refine (or simplify) the geometry of a given mesh has been recently proposed [65]. The edge contractions are clustered according to a base hierarchy that gives a LoD (Level of Detail) approximation for the first frame. The hierarchy is then incrementally adapted to the geometry of the next frame by using edge swap operations (see Figure 3.11). The whole deforming surface is thereby coded by initial vertex positions with the base hierarchy, the swap sequence and the vertex displacements for each frame.

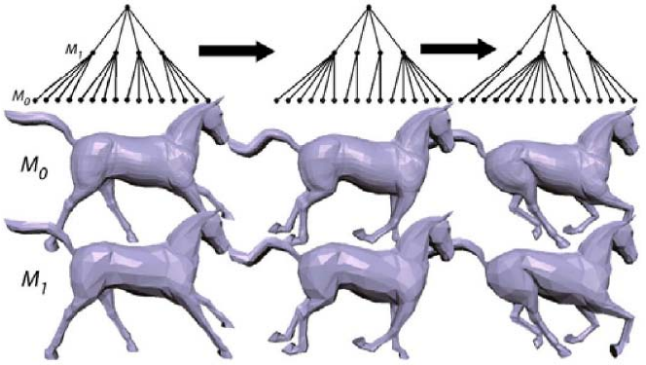



Figure 3.11: Progressive Time-Varying Meshes. The 3D Horse is presented at two levels of detail. ©ACM Inc.

### NURBS

Non-Uniform Rational B-spline Surface (NURBS), is the most common representation of parametric surface based representation, especially in CAD and CAM domains. A B-spline surface is a continuous piecewise polynomial surface defined as the union of surface patches of fixed degrees.

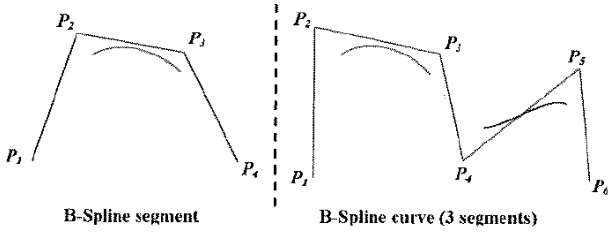



Figure 3.12: B-spline curve. © 3D image processing.

The NURBS are a generalization of these B-spline functions (illustrated in Figure 3.12), where a weight $w_{ij}$ is associated for each control point (see Figure 3.13). This coefficient is a tension parameter: increasing the weight of a control point pulls the surface toward that point. A point on a NURBS surface $S$ is defined by:

$$S(u,v) = \frac{\sum_{i=0}^{n}\sum_{j=0}^{m} N_{ip}(u)N_{jq}(v)w_{ij}\mathbf{B}_{ij}}{\sum_{i=0}^{n}\sum_{j=0}^{m} N_{ip}(u)N_{jq}(v)w_{ij}B_{ij}}, 0 \leq u,v \leq 1$$

where $\{N_{ir}\}$ denotes the B-spline basis functions, $p,q$ the degrees (order) of the surface in $u$ and $v$ directions, $\mathbf{B}_{i,j}$ a mesh of $n \times m$ control points. The two knots $U$ and $V$ are vectors specifying the domain over which the B-spline basis functions is defined. It consists of two

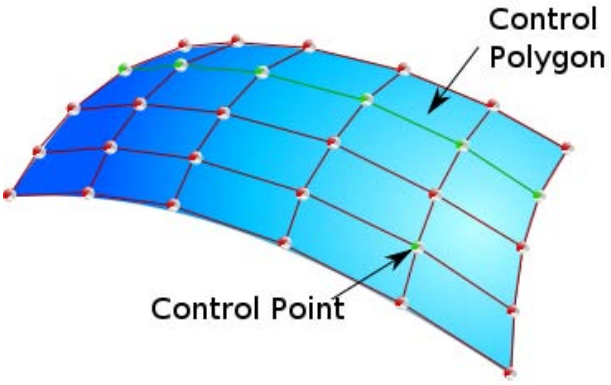



Figure 3.13: Illustration of a NURBS surface with control point and control polygon.

non-decreasing sets of real numbers (knots) that partition the parameterization domain into subintervals : $U = u_1, ...u_n$ and $V = v_1, ...v_n$.

The NURBS representation -as a tensor product surface- can represent planar surfaces or quadrics (spheres, cylinders) but also surfaces having sharp edges. The benefits of these surfaces are that they are mathematically complete and easy to sample or to digitize in voxels or triangles. As they don't depend on a scale factor, NURBS have theoretically an infinite resolution.

In practice, it is sampled into triangular or quadrilateral representation for GPU-based rendering. The Level of Detail control of NURBS surfaces can be constructed only if the complete representation is available, which is rarely the case in a compact representation goal.

The second limitation to represent fine details comes from the NURBS construction itself. Local refinement of a NURBS surface necessitates large-scale modification. To add a single control point within a patch, an entire column or row of control points must be split to preserve the desired quadrilateral grid structure. Two solutions consist of either using displacement maps, where a model stores fine details as if there were a kind of texture information, and then map it onto the surface during rendering. The other relies on the use of the hierarchical B-spline, but are not sufficiently generalized to work on arbitrary complexity.

### Polygon soup representation

From the input MVD data, Colleu et al. [23] introduced a polygon soup representation, which takes advantage of polygonal primitives: compactness, surface continuity, and graphic processor compatibility. The textures remain defined in 2D for compression efficiency and possible backward compatibility, but a polygon soup replaces the depth maps.

The 3D polygon soup is composed of polygons stored in 2D with their depth values at each corner. These 2D polygons are extracted from the depth maps using a quadtree decomposition method for each view (fig.3.14).

This decomposition per view provides an accurate model of the associated depth map, while preserving discontinuities : created quads are recursively divided if they contain a depth discontinuity or a bad approximation of the original depth.

From this preliminary soup of polygon, an inter-view redundancy reduction step eliminates identical or inadequate quads, while increasing the compactness (as illustrated in fig.3.15).

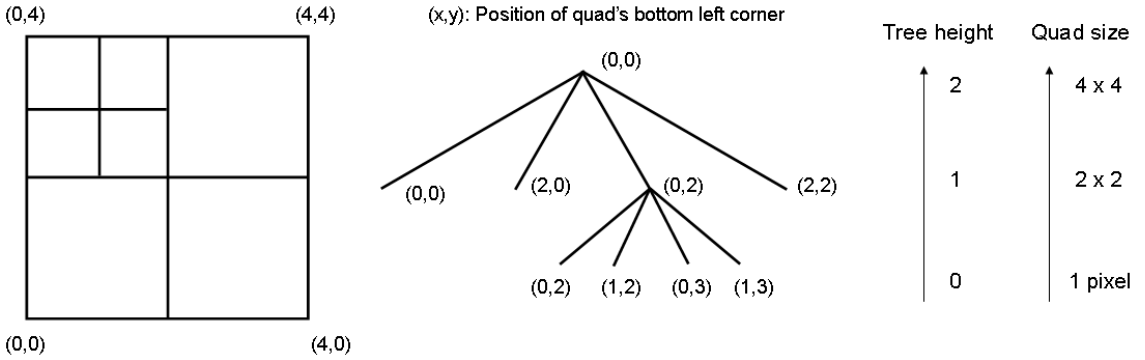



Figure 3.14: Example of image decomposition and quadtree structure. Each level of the quadtree fives the size of the quads and each node gives the position of the bottom left corner of the quads (courtesy of T.Colleu).

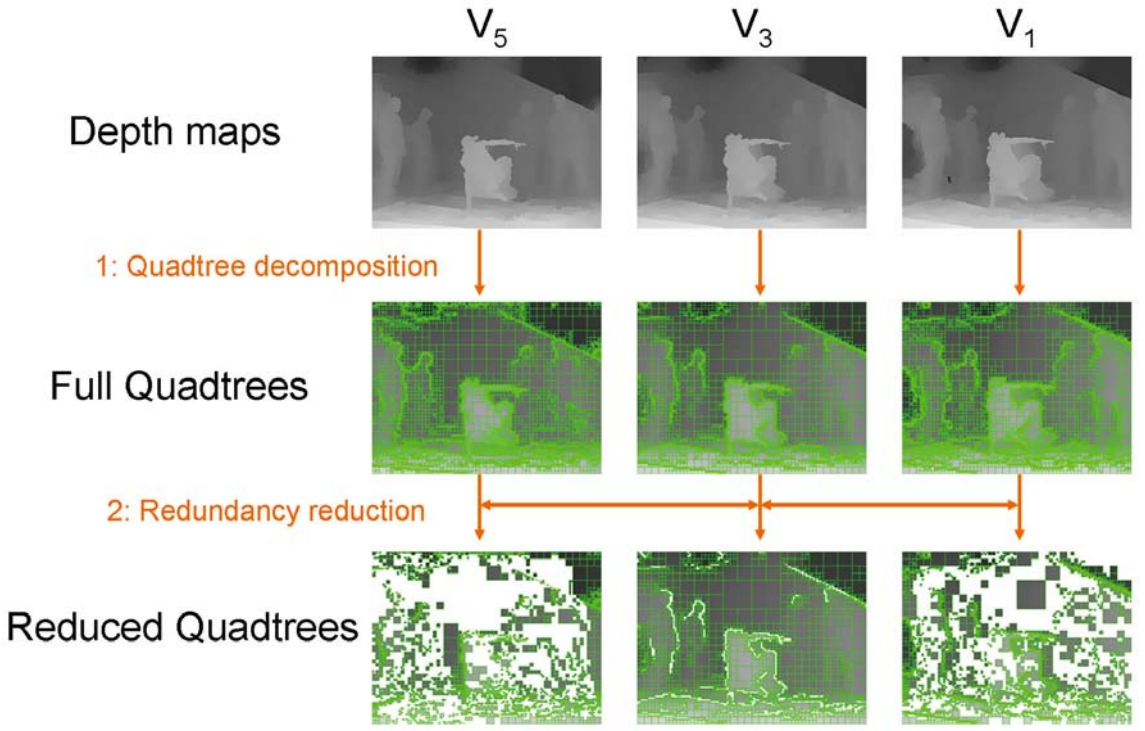



Figure 3.15: Overview of the Polygon soup construction method. (courtesy of T. Colleu).

At the rendering stage of synthesized view, a triangulation and quadtree restriction enables crack elimination, while a multiview adaptive blending (based on view-dependent texture mapping) smooths color and geometry inconsistencies. Ghosting artifacts having been eliminated during the quadtree reduction process, the virtual view is finally inpainted on unknown pixels and edge-filtered on object boundaries to provide a natural appearance. The polygon soup representation while preserving rendering quality, transfers the complexity to the construction of the representation. The compactness and low complexity synthesis make this representation a good candidate for 3DTV.

### 3.1.5 Point-based representation

We have seen that triangles (or polygons) can be used in an efficient way for surface representation. Surface points, particles or surfels can also be used as a more basic display primitive for surface representation.

The topology or connectivity isn't explicitly stored, but instead a set of points sampled from the surface and their surface normals and colors are recorded. Thus the point-based schemes are not limited by the topology, and can easily be used to represent any complex 3D scene. The first point-based representation has been proposed by [74], but it recently regained attention due to its rendering complexity capacity. As the abilities to acquire more and more meshes put the classic polygonal models to the limit of graphic card capabilities, the rendering of individual points instead of polygon rasterization becomes much more efficient.

Also, the splatting is a common technique to avoid visual artifacts like holes (that appears due to projection and grid positionning), aliasing and undersampling effects. Each surface point is associated with an oriented tangential disc: a surfel. The size and shape are changing depending on the type of surface and the local density. The shade or color of the point is warped so that intensity decays in the radial direction from the center. When a single image pixel is influenced by several overlapping splats, the shade of the pixel results in the intensity-weighted average of the splat colors. Different 2D Gaussian-based filtering finally achieve a high quality rendering of point-based surface models at an interactive frame rate.

A recent example of a 3D video framework relying on point-based representation has been presented by [141]. The acquisition part is composed of multiple 3D video bricks containing a projector, two grayscale cameras and a high-resolution color camera (as illustrated in Figure 3.16).

The depth calculation is aided by structured light patterns projected on the surfaces of the scene. Textures images and pattern-augmented views of the scene are acquired simultaneously by time multiplexed projections of patterns and camera exposures.

Then, the depth maps are extracted using stereo matching on the acquired pattern images. Each surface sample corresponding to depth value is merged into a view-independent point-based 3D data structure. Each point is modelled by a 3D Gaussian ellipsoid, and the resulting point-cloud is post-processed to remove outliers and artifacts. At the rendering stage, enhanced probabilistic EWA volume splatting and view-dependent blending lead to high quality synthesized views.

### 3.1.6 Volumetric representation

Volumetric representations consist of a decomposition in volume units or primitives of the 3D space, the discrete model and primitive-based model respectively. The primitive-based models are the extension of surface parameterization to volume, and rely on cylinders, superquadrics, supershapes or hyperquadrics and other polynomial models combined with different operations such as graphs to form a coherent representation. Despite their advantage in compactness, it is inherently difficult to model arbitrary and natural objects through primitives; even if flexibility has been considered (deformation by torsion, etc.), this primitive based representation is more dedicated to CAD, object extraction or object modelling.

The discrete model decomposes and segments the 3D space into volume units called voxels. The set of voxels constitutes the reconstructed volume in a world reference frame, but there exist different ways to organize this set. Each voxel contains the properties

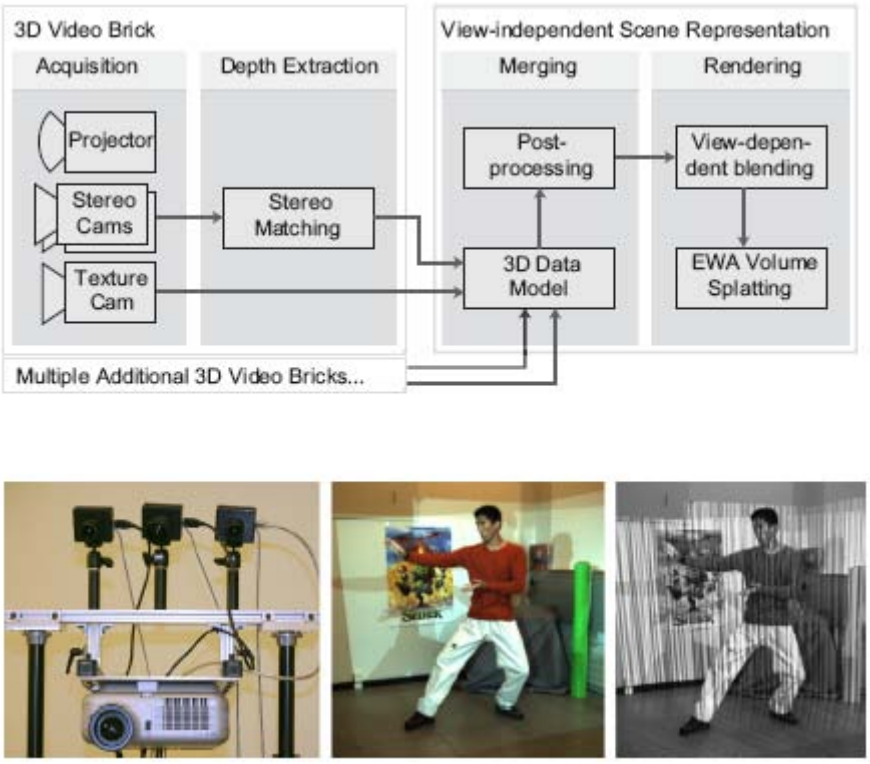



Figure 3.16: Overview of the Waschbüch 3D video framework (top) and illustration of a brick (bottom left), simultaneously acquiring textures (middle and structured light patterns (right), from [141].

of a surface segment within it, and empty voxels can be either flagged as empty, or not represented at all. This last method results in the **octree** structure [39],3.17 . Each octree codes the areas that contain surfaces only. The data structure is organized as a tree growing in depth at the nodes that correspond to occupied voxels. Then octrees show efficiency in term of memory space and are a common method to store voxels in term of compression and ease of implementation.

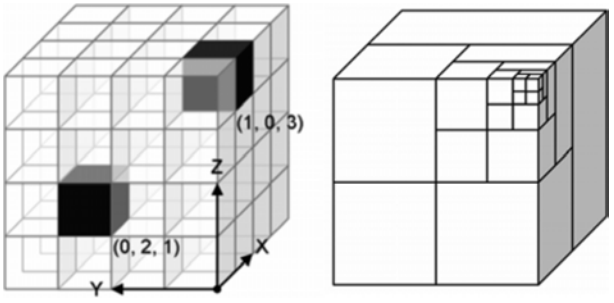



Figure 3.17: Illustration of data structures implementing volumetric representations, voxel buffer (left) and octree (right).

The main advantages of the voxel representation in the multiview video context is that it provides a common reference frame where the surface information obtained from different views are combined. It also enables direct - and then fast - access to neighboring voxels for the rendering, as this neighborhood is coded into the tree. This facilitates

the computation of voxel visibility, the absence of potential voxels between camera and image surface. This also helps in 3D convolution operations and the detection of connected components, used for local geometric properties calculation, and noise-filtering operations.

In conclusion, advantages of voxel-based representations are numerous, linear access time to the structured data, and then complexity-independent rendering. But the cube unity results in low quality rendering when the camera is positioned too close to the cube based surface, which is not the case with polygons. The conversion from a volumetric representation to a mesh-based representation at the rendering stage is under investigation by the computer vision community. Radial basis functions (RBF) [17] could be applied on voxel representations, as it shows promising results on surface interpolation.

### Summary and conclusion

Different representations have been developed in order to fulfil specific requirements. Pure image-based representations are well-suited for stereoscopic visualisation or multi-view display at constant viewpoints. Their compression formats dedicated to stereo and multi-view video will be detailed in the next section. In order to introduce baseline reduction, baseline extension or navigation, higher complexity representations are however needed.

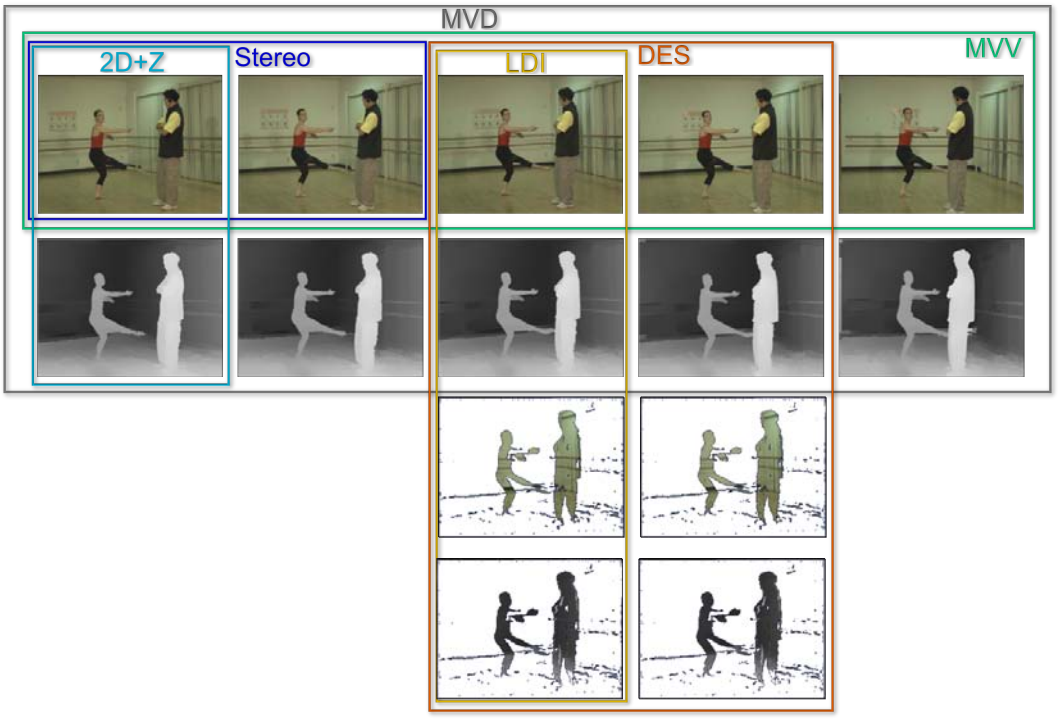



Figure 3.18: Summary of the different image and depth-image based representation with their different video or video + depth datasets. Note that in practice the stereo and 2D+Z representations are positioned on the central viewpoint.

These other complex representations convey the geometry information to allow the synthesis of virtual viewpoints inside a larger viewing area. First, with the geometry information, arbitrary views can be synthesized at the rendering stage to adapt to the display or to the required displacement between views (which also depends on the viewing distance). The display configuration is then independent of the acquisition configuration.

Second, the geometry brings a data-reduction over a very redundant high number of views to be transmitted. Flexibility and compactness are then their main quality.

Among the existing geometry based representations the surface-based, point-based and volumetric representations suffer several disadvantages. While these are adapted to efficient coding of single objects, they can render arbitrary detailed scenes at the cost of complex long view synthesis or at the price of a low image quality.

A good trade-off can be found with depth image based representations (see Figure 3.18). While the MVD keeps all the redundancies induced by an extensive number of views, the LDI attempts to remove most of them. The LDI and related DES thus allow a good trade-off compactness/synthesis quality, but the complexity has to be partially supported at the decoder side.

The MVD however allows backward compatibility with existing 2D and stereo format -and consequently with 2D and stereo displays- because it relies on an image structure widely supported by the existing block-based video standards. The high inter-view spatial redundancy is tackled by mechanisms similar to the existing temporal and spatial redundancy reduction tools. This will be detailed in the next section.

## 3.2 3D video coding standards

Different 3D video representations have been presented in the first section. This section aims at linking these representations with their associated codec proposals.

We can distinguish two types of video standards for 3D services, the standards that rely on the current state-of-the-art 2D video codec at their time and add a 3D functionality while remaining backward compatible, and the "non-restrictive" standards that do not specify particular coding algorithms but instead a framework proposal.

3D standards can also be distinguished by the representation of information they convey. The conventional stereo representation codecs will first be described, before the various video-plus-depth based codecs are detailed. Finally, the advanced codecs for both multiview video and multiview video-plus-depth will be introduced.

### 3.2.1 Conventional stereo video coding

The main idea in the past stereo video encoders has been to exploit the high inter-view redundancy between the left view and right view video sequences.

**MPEG-2 MultiView Profile** (or MVP) has been proposed as an extension to the MPEG-2 standard to allow the transmission of two video signals for stereoscopic TV applications. It relies on scalable video coding tools to guarantee the backward compatibility with the MPEG-2 Main profile.

One view is defined as the *base* layer and the second view as an *enhancement* layer. It features a monoscopic coding of its base layer with same tools as the main profile -for compatibility- with a hybrid prediction of both motion and disparity between views for compression efficiency. The temporal scalability tools are used for coding the enhancement layer. The resulting temporal and inter-view prediction structure is illustrated in in Figure 3.19.

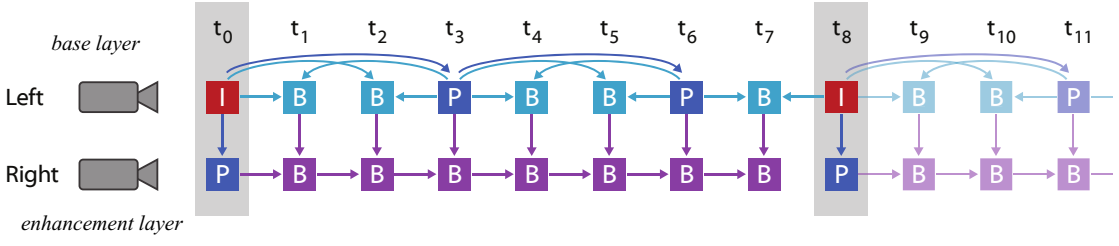



Figure 3.19: MPEG2-MVP prediction structure at the Group Of Picture (GOP) level. The GOP is composed here of "IBBP" pictures.

Thus, the temporal prediction is only applied on the base layer, and the enhancement layer relies simultaneously on a past temporal and on an adjacent view prediction. Note that the camera parameters (focal length, etc...) are also included. The resulting bitstream remains backward compatible to be decoded by legacy 2D MPEG-2 decoders.

**H.264/MPEG-4 AVC stereo SEI message** An amendment to the H.264/MPEG-4 AVC has been proposed in 2004 with the Fidelity Range Extension (FRExt) to support a wide variety of purposes: alpha composition, bit depth precision increase and stereo video etc. Instead of transmitting a second view with additional pictures at the GOP level, the stereo video may be interleaved or multiplexed in time or in space.

The stereo Supplemental Enhancement Information (SEI) then informs the decoder to distinguish the left and right view inside the multiplexed bitstream to extract separately the two views. Different 3D interleaving arrangements have been defined such as "0" or "checkerboard", where pixels alternate from Left (L) and Right (R) view or "1" where L and R views are interlaced by columns etc.

The H.264/MPEG-4 AVC coder then encodes the interlaced sequence in a field coding mode in a single bit-stream or transport-stream respectively BS/TS, which is later de-interlaced into two distinct view video sequences. Because of the interleaving and the necessity at the decoder side to be able to read the SEI message, the backward compatibility is not supported by traditional 2D devices.

### 3.2.2 2D+Z coding

**MPEG-4 MAC** The MPEG-4 Part 2 defines Multiple Auxiliary Components (MAC) that can also be employed for a large variety of applications. The MAC are single or multiple grayscale components that can either describe the transparency ("alpha") of a video object, but also its shape or any other texture etc. This alpha channel can then carry the depth video additionally to the original texture video. Because the MPEG-4 Part 2 absorbed the features of MPEG-2, the encoding of texture and auxiliary components employ motion compensation and DCT transforms.

Cho et al. also [22] also proposed that this standard be used for stereo video coding: the disparity vector field, the luminance and chrominance data of the second view can be assigned as three auxiliary components of MPEG-4 MAC.

**MPEG-4 AFX** (Animation Framework Extension) or MPEG-4 Part 16 specifies models for representing 3D graphics content. Higher-level synthetic objects extend MPEG-4 by specifying geometry, texture and compression algorithms. Among them, we can highlight the depth-image-based-rendering (DIBR), point rendering and view dependent multi-texturing.

MPEG-4 AFX is compatible with the LDI representation because it specifies a Depth Image (DI) structure composed of a SimpleTexture (DI) or a PointTexture (LDI). This format can also contain various animated objects, i.e. sets of compressed video and depth streams. As with MPEG-4 MAC, the MPEG-4 compression techniques should be re-used on texture and depth videos.

**MPEG-C Part 3** or ISO/IEFC 23002-3 standard specified a representation format for video-plus-depth. The depth maps are encoded as conventional 2D sequences plus additional parameters for interpreting the depth on the decoder side. In itself, the MPEG-C Part 3 does not specify the transport and compression techniques. Then the texture video could be encoded in MPEG-2 or H.264/MPEG-4 AVC, as the depth video. Instead this standard provides interoperability of the content, both display and capture technology independence, compression efficiency and backward compatibility by specifying high-level syntax to the decoder side.

Finally, these three standards are either high-level format or extensions of existing MPEG-4 video codec, so they all guarantee a backward compatibility with existing decoders.

### 3.2.3 3D MultiView video Coding: MVC

The first format to support the representation of two and more views has been developed by the JVT as an extension to H.264/MPEG-4 AVC: the MVC extension. It aims at encoding multiple views and exploiting their inter-view redundancy through traditional mechanisms of motion vector prediction. A base view forms the main reference for the other views. However this base stream is independently coded in H.264/MPEG-4 AVC to guarantee backward compatibility.

#### Inter-view dependencies

The MVC prediction structure relies on both hierarchical temporal prediction and inter-view prediction as illustrated in Figure 3.20.

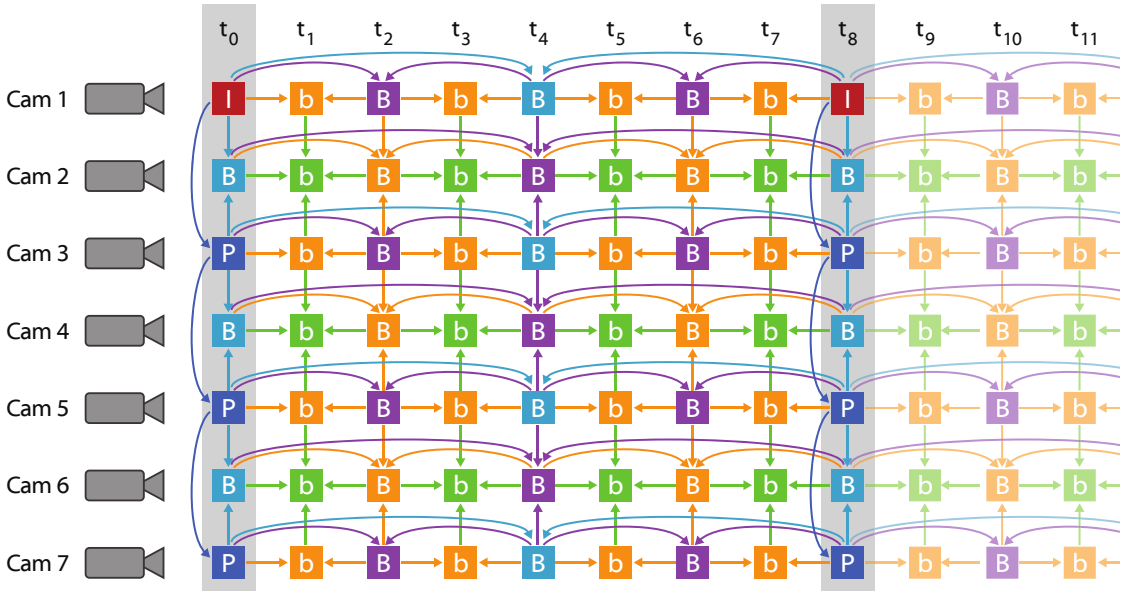



Figure 3.20: MVC prediction structure for seven views.

The inter-view dependencies varies on the camera position and on the choice of the camera defined as the base view.

**The base view** A single view is coded as the *base* view in simulcast H.264/MPEG-4 AVC. This base view guarantees random access features by *intra* mode coding of the key picture: the *anchor picture*. The predictions are basically the same H.264 motion compensated predictions over time.

**The non-base view** The non-base views are encoded by both motion and disparity compensation to cope with the temporal and inter-view redundancy. Then the compression rate is improved over multiple simulcast streams but the random access is lost on each non-base view. This is partly overcome with V-pictures having no temporal dependence

on other temporal pictures of the same view. Thus it can only be predicted by the pictures of other views at the same timestamp.

#### Inter-view predictions

The inter-view redundancy between cameras is exploited by the disparity estimation and compensation. A mechanism similar to the temporal prediction of H.264/MPEG4-AVC is used to compensate the disparity at a block level: the Disparity-Compensated Prediction (DCP), as shown in Figure 3.21.

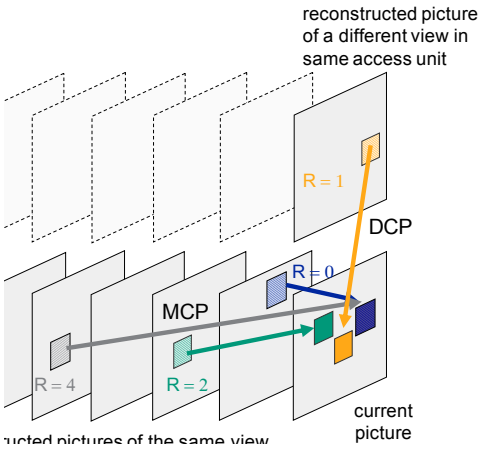



Figure 3.21: Disparity-compensated prediction as an alternative to motion-compensated prediction. From HHI [48].

Also illustrated by the Figure 3.20, a bi-directionally predicted B-picture from a non-base view can be in fact predicted from both temporally neighbouring frames and from adjacent views. A non-base view P-picture is obtained by disparity compensation from the anchor picture of the base view.

**Specific Inter-view SKIP mode** As the SKIP mode in H.264/MPEG4-AVC aims at exploiting the often similar movement of spatial neighbourhood macro-block (MB), the Inter-view SKIP mode assumes that motion vectors between neighboring views will be very similar. The motion vector of a current MB is thus obtained from a corresponding -disparity compensated- MB in the picture of a neighbouring view at the same timestamp.

### 3.2.4 3D Video Coding Standard for MVD: 3DVC

MPEG/ISO issued a Final Call for Proposal (FCfP) in April 2011 on 3D Video Coding (3DVC) technology to provide efficient compression and high quality synthesis of an arbitrary number of dense views, i.e. multiple standard videos or "texture view" videos plus multiple depth videos.

Up to now, several solutions emerged in parallel for multiview plus depth coding, one as an extension of H.264/MPEG4 AVC, one as an extension to H.264/MPEG4 MVC and one based on HEVC. No choice has been made by the Joint Collaborative Team on Video Coding (JCT-VC, ex: JVT, formed by MPEG/ISO and VCEG/ITU) committee yet.

This section will focus on the latest HEVC based solution because it is supposed to be the most efficient and promising base to provide evolutionary coding efficiency. The

coding of the multiple texture views will first be described before a particular focus be made on the different tools use for the prediction of depth content. The techniques and modes used for intra coding of depth maps directly compete with a depth map compression contribution of this thesis, these techniques are regrouped in chapter 4.

Finally, the proposed method for view synthesis in 3DVC is also closely related to the contribution made on inpainting-based view synthesis described in the last chapter 5. Consequently, the descriptions of all the state-of-the-art view synthesis techniques will be done in this chapter.

### Data format and System description

The 3DVC video is represented in the MVD format where the captured views and associated depth maps are encoded and multiplexed into a single 3D video bitstream. At the decoder side, the video and depth video data enables the generation of intermediate views (i.e. interpolate) suitable for the display of 3D content on a variety of screen technologies, such as autosteroscopic displays. This is illustrated in the Figure 3.22.

Several similarities exist with MVC, the encapsulation of camera parameters in the bitstream, the header information to signal the view identifier, a distinct base view coded independently in simulcast for backward compatibility, etc. Also, the standard can provide like MVC an MVV representation without the depth data, in that case video pictures can be decoded independently of the depth data.

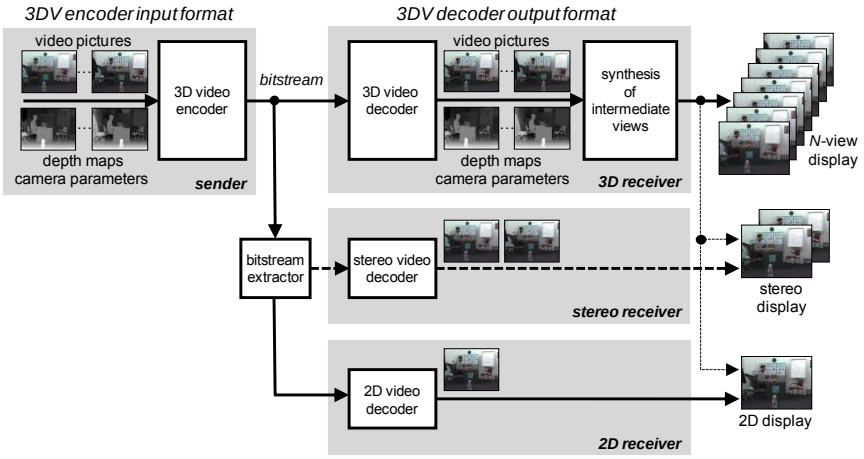



Figure 3.22: Overview of the system structure and the data format for the transmission of 3D video. From HHI [48].

The 3D video bistream is arranged in a way that its sub-bitstream containing the base view can easily be identified by a packet header and thus simply extracted. Also, the encoder can be configured so that the sub-bitstream of stereo views could be directly extracted on a stereo decoder. As we have seen, the encoder can finally be set in a way that videos are independently decoded of the depth videos. All this distinct multi-views configuration remain compatible with the embedded intermediate view renderer.

The basic structure of 3DVC is illustrated in Figure 3.23, where each component appears to be coded on a HEVC base. The output bistream packets or Network Abstraction Layer (NAL) units are finally multiplexed in a single bitstream. These modified HEVC codecs include various coding tools and inter-component prediction techniques relying on

already coded data inside the same access unit (red arrows). These techniques will be detailed in the following sections: first the techniques exploiting the inter-view redundancy, and second the techniques for the depth video coding.

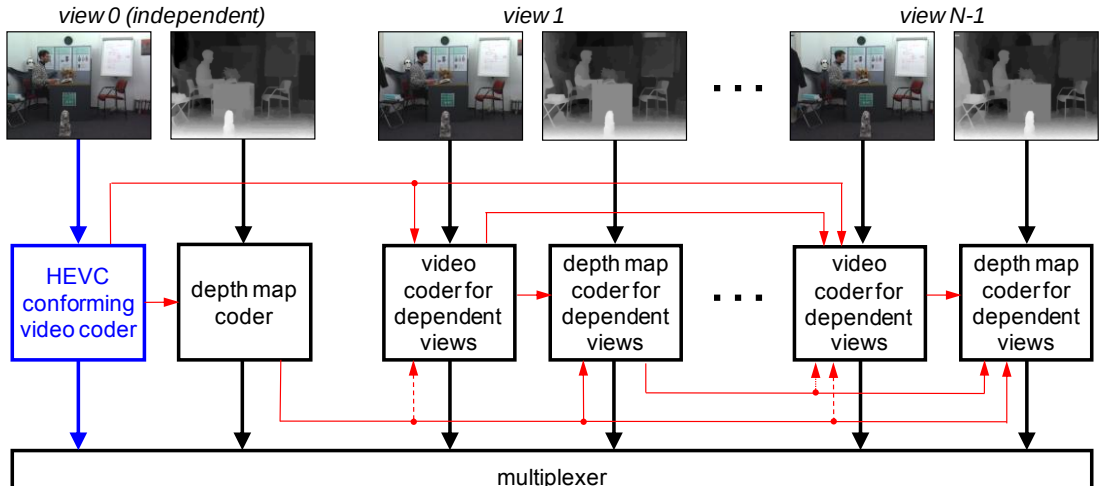



Figure 3.23: HEVC-based codec with additional coding tools for dependent views and depth maps (red arrows). From HHI [48].

### Coding of Texture Views

**Disparity-compensated prediction** As in the previous MVC, HEVC based 3DVC (shortened as "3DVC" in this section) relies first on the concept of disparity-compensated prediction or DCP. The process is already described for MVC in section 3.2.3. The macroblock syntax and decoding process remain similar to the usual Motion Compensated Prediction (MCP). Motion vectors of motion-compensated blocks are only predicted from neighbouring blocks in the temporal reference pictures. In contrast disparity vectors of disparity-compensated blocks are only predicted from inter-view reference pictures neighbouring blocks.

**View synthesis based inter-view prediction** At both the encoder and the decoder sides, an identical inter-View Synthesis Prediction (VSP) algorithm is implemented. From the adjacent views, a new virtual view is synthesized in the position of the currently processed view.

To cope with the disoccluded regions (appearing in newly synthesized views) but occluded in previously coded views, a binary availability map is defined. The availability map then identifies the disoccluded regions so that both encoder and decoder can determine whether a Coding Unit (CU, equivalent of MB in HEVC) is coded or not. In practice, only small parts of the current view are disoccluded in the other coded views, thus few CUs are selected by the rate-distortion-optimization to be coded in the side view.

In order to reduce the texture artifacts in the synthesized view, two final post-processings are implemented within the VSP loop.

The Depth-Gradient-based Loopback Filter (DGLF) aims at reducing the synthesis artifacts in the regions of abrupt depth changes. The synthesized image is adaptively filtered according to the depth gradient intensities. Strong depth edges strongly low-pass filter the synthesized texture at those positions, while flat depth regions are not filtered at all. We will see this postprocessing technique is commonly used in view synthesis algorithms (see chapter 5).

The latter Availability Deblocking Loopback Filter (ADLF) smooths the artifacts resulting from a CU block coding. Because the shape of coded regions, a grid of block, rarely matches the shape of the binary availability map, artificial edges can appear between those regions during the VSP. The ADLF then smooths by interpolation the transition i.e. the seam between the coded and synthesized region.

**Inter-view motion prediction** As in MVC with the Inter-view SKIP mode, the idea is to exploit the fact that motion vectors between corresponding blocks of neighbouring views are similar. It requires a depth map estimate for the current picture to extract a candidate motion vector or disparity vector in the neighbouring reference picture.

The prediction block in the already coded reference picture is used as the reference block. If this reference block is MCP coded (see above), the associated motion parameters can be used as candidate motion parameters for the current block in the current view. The derived disparity vector can be used in the same way as candidate disparity vector for the DCP.

Two methods are proposed to estimate the depth map of the current picture based on already transmitted information. They can be selectively activated at the encoder side.

The first depth map estimate is based on already coded depth maps, so it requires the transmission of them. Then a decoder must first decode the depth maps of previously coded views for decoding dependent views.

The second method doesn't require the depth maps transmission because the depth map are estimated from the disparity between a reference view and its neighbouring dependent view.

Finally we have seen that two modes for signalling the motion parameters have been specified in HEVC: AMVP and Merge modes. 3DVC proposes to extend them and to include the inter-view motion parameter prediction in an extended AMVP candidate list (at the first position of the list). Similarly, the candidate list of motion parameters in the merge mode is extended by a motion parameter set obtained from inter-view motion prediction.

**Inter-view residual prediction** This concept consists of a prediction of the prediction errors themselves. As for the inter-view motion prediction, the inter-view residual prediction is based on a depth map estimate for the current picture. The same prediction is used in practice.

A disparity vector is calculated for a current block and the residual block in the reference view. As for motion compensation, the block of residual samples in a reference view (located at the derived reference location) is subtracted from the current residual. Only then the weaker difference signal is transformed. Practically, a flag alerts of the usage of inter-view residual prediction in the CU syntax. When it equals 1, the residual is predicted from reference residual signal: only the difference is transmitted using transform coding. When it equals 0, the residual is usually coded by the HEVC transform.

**Adjustment of QP of texture based on depth data** A perceptual tool of interest in the scope of this thesis is the adjustment of texture's Quantization Parameters (QPs) based on depth data. The idea is to simply increase texture quality of objects in the foreground while decreasing the quality - and then increase the compression factor - of background objects. A variable QP is assigned to CUs depending on the depth values. This is done at both sides of the coding chain in order to avoid additional information transmission. The new $QP'$ is defined by the equation:

$$QP' = QP - 2.6 + 8 * \left( \frac{255 - max_{x,y \in CU} d_{x,y}}{256} \right)^2$$

with $QP'$ the adjusted $QP$ value for a CU of disparity $d_{x,y}$.

### Coding of Depth Views

On one hand, most of the usual coding tools for efficient texture video coding are transposed for the coding of depth video. The same usual intra-prediction, motion-compensated prediction, disparity-compensated prediction and transforms are used. On the other hand, the inter-view motion and the inter-view residual prediction are not implemented for the depth coding. The motion parameters are then derived from the texture video pictures. Other tools are described in the following paragraphs.

**A perceptual tool for depth maps coding: Non-linear depth representation** First, because a depth map consists of a single matrix of distance points expressed in a camera viewpoint, the picture format of the depth maps is in 4:0:0 where the chrominance components are disabled.

Second, a non-linear depth mapping is applied on the depth maps for perceptual reasons. From the observation that the human depth sensitivity depends on the relative distance rather than on the absolute distance of viewed objects, the MPEG contributor states that the internal depth representation should be non-linear: closer objects should be represented with more accuracy that distant ones to improve quality of synthesised views. A power-law expression, similar to the gamma correction law for luminance intensity has been adapted to the coding of the internal depth sample values:

$$Z_{internal} = \left( \frac{Z_{external}}{Z_{max\ external}} \right)^{exponent} \cdot Z_{max\ internal}$$

The inverse equation is stated to retrieve the $Z_{external}$ value. The exponent is automatically chosen by the encoder depending on the QP of the depth and finally sent to the decoder in the encoded bitstream:

$$exponent = \text{clip}\left[ (QP_{depth} - 30) * 0.0125 + 1.25; 1.0; 1.66 \right]$$

Thus the depth map samples are represented on a wider bit range with the use of the Internal Bit Depth Increase (IBDI) tool.

**Z-near Z-far compensated weighted prediction** The $Z_{near}$ and $Z_{far}$ parameters are respectively the minimum and the maximum distances of an acquired depth map for a given camera viewpoint and for a given time instant. These are transmitted in the bitstream. However because the distances are changing between viewpoints and then

between views, or over the time and then between frames, a given depth value can be represented with a different gray-scale value on a depth map. These inter-depth picture variations will lead to poor prediction between a reference and other depth maps.

The idea behind the $Z_{near}$-$Z_{far}$ Compensation (ZZC) is then to rescale each depth map before it is used for prediction. Then this is processed before any inter-frame depth prediction: each depth map on the codec reference picture list is scaled in order that gray-scale depth values in both scaled and currently coded image refer to the same actual depth. The compensated disparity is calculated as follow:

$$L_T = L_S \cdot \frac{Z^S_{far} - Z^S_{near}}{Z^T_{far} - Z^T_{near}} + 255 \cdot \frac{Z^S_{near} - Z^T_{near}}{Z^T_{far} - Z^T_{near}}$$

$L_T$ being the compensated disparity between $Z^T_{near}$ and $Z^T_{far}$, and $L_S$ the original disparity within the respective $Z^S_{near}$ and $Z^S_{far}$ depth ranges. The resulting compensated depth maps are thus used for later prediction.

**Other ad-hoc modifications for depth map prediction** As explained in chapter 4, depth maps possess sharp edges that must be preserved. However, the usual sub-pel accuracy interpolation used for predictions could lead to ringing artifacts at edges in depth maps. To avoid this issue, the eight-pixel interpolation filter used for MCP and DCP interpolation is disabled for depth maps. The inter-picture prediction is then realized at full pixel accuracy, which also reduces the encoder and decoder complexity. The transmitted Motion Vector (MV) residuals are thus coded at full-sample precision.

Additional complexity reduction has been found by deactivating the in-loop filter dedicated used for natural texture video. Thus, the de-blocking filter, the Adaptive Loop Filter (ALF) and the sample-ALF are disabled.

**Depth coding intra prediction modes** Because the depth maps have intrinsic properties that differ from the textures, four new intra-prediction modes have been added for the coding of depth maps at the block level. These techniques are further described in Chapter 4 because they refer to depth map coding techniques. Four modes *Wedgelet_ModeIntra, Wedgelet_PredIntra, Wedgelet_PredTexture and Contour_PredTexture* are added to the set of block coding intra modes of 3DVC.

**Motion parameter inheritance** A point in 3D coordinate space $(x, y, z)$ is both acquired and projected by a traditional camera and a corresponding depth camera on a $(u, v)$ grid to extract its luminance and chrominance (often in $R, G, B$ values) and its distance ($Z$ value). These are both projections of the same 3D point at the same 2D coordinate point from the same viewpoint and time instant. Because a 3D point maps to the same coordinate in the texture and depth map, its own motion is also similar between these maps.

The Motion Parameter Inheritance (MPI) aims to avoid the recalculation of both the quadtree and motion parameters by making the depth video inherit the treeblock subdivision into CUs and PUs and their motion from the texture video. The quarter-pel motion vectors are quantized to their nearest full-sample position to be compliant with the depth map full-sample motion vector accuracy.

As shown in Figure 3.24, the encoder chooses for each depth map block between the inheritance of the motion data from the co-located region in the texture picture and the transmission of a new motion data.


Figure 3.24: Illustration of the concept of motion parameter inheritance. From HHI [48].

Because the MPI is based on the motion, it can only be used if the whole region of the texture video from where the motion and partitioning data are inherited is coded by Inter prediction.

**Encoder Control: View Synthesis based Rate-Distortion Optimisation** As within previous MPEG-encoders, when coding a MB or a CU, many options such as motion vector, reference frame, direction of prediction in B-pictures etc. have to be chosen to optimize the encoding of the current picture. Optimize the encoding in the compression sense means to **minimize** the cost of coding a current CU under the constraint of maintaining at best possible or **maximizing** its quality.

As in H.264/MPEG-4 AVC and HEVC, this mechanism is done by the Rate-Distortion Optimization (RDO): a Lagrangian cost function is run for each option, i.e. for each mode and parameter combination. The option giving the smallest cost is selected. The search of the best combination of coding options is formulated as minimizing the Lagrangian cost function:

$$J = \mathbf{D} + \lambda \cdot \mathbf{R}$$

where the $D$ is the Distortion obtained by quantifying the quality of a given block after coding/decoding, $R$ the Rate i.e. the number of bits required to code that block with this particular combination of coding options.

The idea has been to apply this decision on the coding of depth maps but where the distortion measure of the depth map as been replaced with a distortion measure of the synthesized intermediate views. Concerning the view synthesis used, the reader is invited to refer to chapter 5 for a description of the view synthesis algorithm used by 3DVC.

The geometry information of the depth maps is not used directly for the display but indirectly by the view synthesis process: lossy coding of depth data provokes distortions in the synthesized views. But the distortion in the synthesized view cannot hardly be deduced by just considering the local distortion of a depth map: a pre-synthesis step is thus needed at the encoder side. A view synthesis optimization is then run in the encoder.

However, the disocclusions and occlusions appearing between a depth map view and its synthesised view prevent a bijective mapping of the distorted areas in depth maps to

distorted areas in the synthesized views. To solve this issue, the change of the overall distortion in a synthesised view picture implied by a single depth map change with a block B is determined. The Synthesized View Distortion change (SVDC) is then defined as the distortion difference $\Delta D$ between two synthesized textures $s_T^{'}$ and $\tilde{s}_T$ as:

$$\Delta D = \tilde{D} - D = \sum_{(x,y)\in I} [\tilde{s}_T(x,y) - s'_{T,Ref}(x,y)]^2 - \sum_{(x,y)\in I} [s'_T(x,y) - s'_{T,Ref}(x,y)]^2$$

with $s_{T,Ref}^{'}$ a reference texture rendered from original video and depth data and $I$ the set of all samples in the synthesized view.

Finally, a renderer model based on rate-distortion optimization using the SVDC is integrated in the encoding process. Conventional distortion measure is replaced with computation of the SVC in all distortion computation steps related to :

- mode decision
- CU partitioning
- intra- and inter- residual quadtree coding
- motion parameter inheritance (MPI)
- merging

In contrast the renderer model is not used for motion estimation and RDO quantization. The Lagrangian cost function is thus adapted to include the View Synthesis optimization done by the renderer model. The Lagrange multiplier is then adjusted by a constant factor $l_s$:

$$J = \Delta D + l_s \cdot \lambda \cdot R = \Delta D + \lambda_l \cdot R$$

with $\Delta D$ the change of synthesized view distortion on the whole picture provided by the renderer model.

### 3.2.5 Conclusion

The diagram illustrated in Figure 3.25 clarifies the evolution of the 2D and 3D video coders and summarizes their interdependence. The inheritance of 2D to 3D coders along time clearly appears, while the different representation choice appears to alternate and to be indeed complex: user- and technology-dependent.

The 2D coders have become more and more complex, but always gained in efficiency to maintain a large quality/rate ratio and adapt to the growing interest on video at higher and higher resolution.

Several solutions inherit from these 2D codecs for different representations of 3D videos. These also become more and more powerful with the addition of geometry information transmission, the increase of possible view to convey etc.

Finally it is very plausible that 2D and 3D video codecs continue to raise lots of interest and issues for the entertainment and TV market.

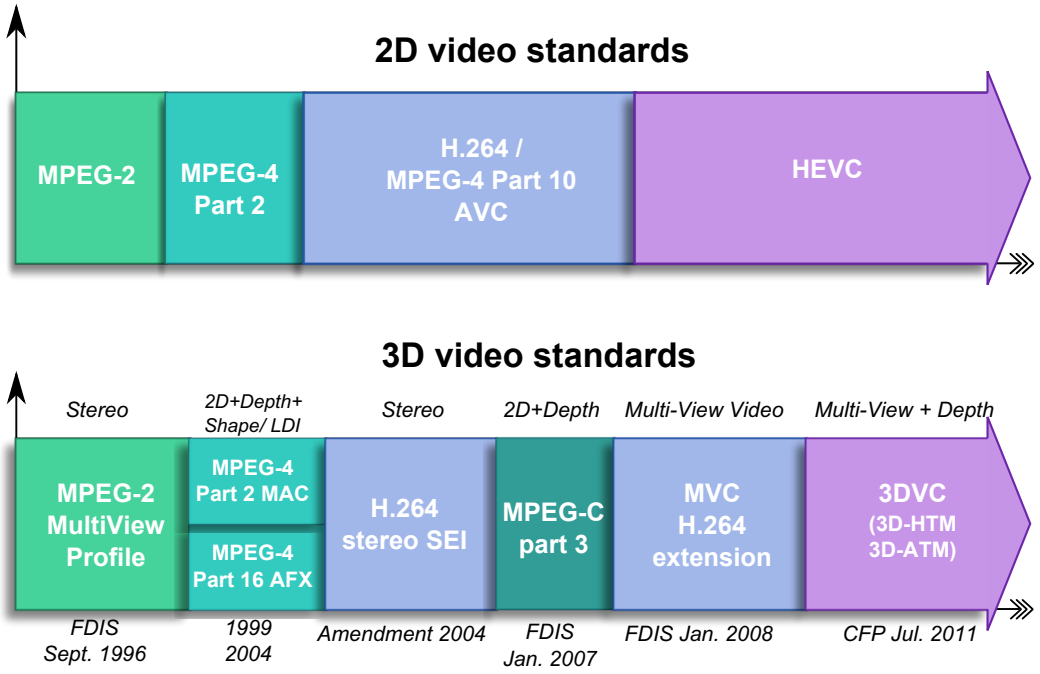



Figure 3.25: 2D (top) and 3D (bottom) related evolution of video coding standards. The colors indicate the re-use and backward compatibility of 3D standards with their compatible 2D formats.

## Conclusion

This chapter focused on the existing 3D representations and coders dedicated to FTV and 3DTV in particular.

The multiplicity of the representations as well as their associated coding formats illustrate the various issues of the image and video usage in our information era.

The image-plus-depth based representations appear to be a good trade-off between complexity, rendering quality and compression rate while maintaining a backward compatibility with the existing -but highly widespread- set-top box decoders.

To this purpose, the MPEG and VCEG community successfully developed a powerful encoder to cope with the growing demand for immersive 3D content on various type of single and multi-view displays.

Definitely, the adoption by the user of these 3D technologies will not only depend on the deployment of this technology but will also be solved by the capacity of TV channels and cinema, film studios and director, to accept and provide a 3D content of quality.

CHAPTER 4

# Edge based Depth Map Compression for 3D Coding

## 4.1 Introduction

At the coding stage along the 3DTV framework, the depth map compression methods have recently gain interest in the video coding community. The transmission of geometry joined to the texture information is necessary to Depth Based Image Rendering (DIBR) for interactive multi-view display on 3DTVs. The geometry provided by the depth of the acquired scene will allow the generation of new virtual viewpoints adapted to the end-user's display requirements: additional viewpoints inside or outside of the range of transmitted views to offer more views and then extend the range of stereoscopic vision on auto-stereoscopic displays for example.

It is indeed essential to efficiently encode and reconstruct a depth map in a way that preserves the distance properties of objects within each other because this will be precisely used for later view synthesis. In contrast the depth map smoothly varying surface leads to very predictable pixel values along the spatial dimension that might not need to be perfectly reconstructed if they don't affect the rendering.

A depth map is basically representing different depths of objects in the scene and more exactly different distances to object surfaces. These surfaces are delimited by the object borders. A distortion into the object depth might result in a local deformation of the rendered object, but a distortion at borders between the object and its background depths might lead to worse artefacts: mixed foreground/background object textures.

In this chapter an existing depth map coding method based on platelet and wedgelet will be recalled. Because this work has recently been embedded into the 3DVC HEVC-based proposal, the corresponding implemented modes and new functionalities will be described in the last section. Then we will present an efficient depth map compression introducing lossless edge coding as an alternative to approximating piecewise linear functions whose coefficients need to be encoded.

## 4.2 Past Depth Map Compression Method

The idea proposed by [92] is to approximate the depth map content with modeling functions. Two classes of modeling functions are defined: piecewise-constant functions to

model flat surfaces of smooth regions and piecewise-linear functions to approximate planar surfaces of the scene with gradually changing grey levels like ground plane and walls.

These surfaces are located within a quadtree decomposition (Figure 4.1 (c)) dividing the map into variable-size blocks. A block is subdivided until no suitable approximation is found. Also to avoid too many small blocks along a discontinuity, they propose to divide the block into two regions separated by a straight line, each coded by an independent function. This is illustrated in Figure 4.1 (b) and (d).

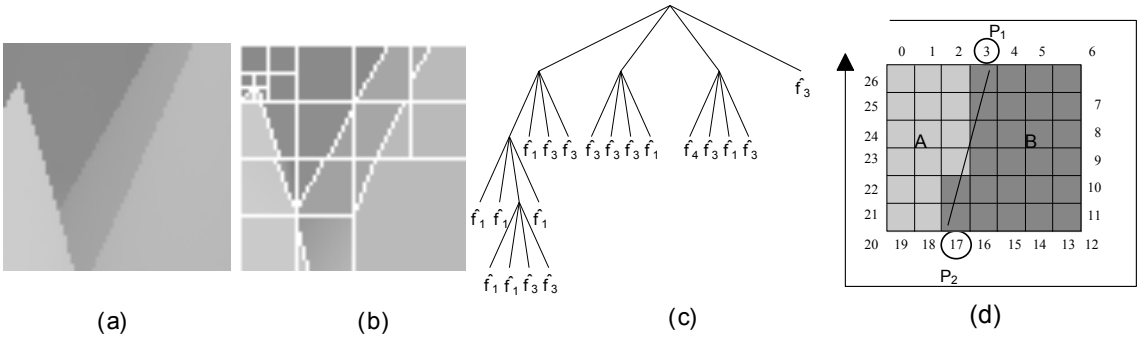



Figure 4.1: Example of quadtree decomposition. Each block (b), i.e. node of the quadtree (c) is approximated by one modeling function. From [92].

Four block modeling functions are proposed for each leaf of the quadtree :

- Function $\hat{f}_1$: approximation with a constant function.
- Function $\hat{f}_2$: approximation with a linear function.
- Function $\hat{f}_3$: subdivision of the block into two regions A and B separated by a straight line and approximation with a constant function: a **wedgelet**.

$$\hat{f}_3(x,y) = \begin{cases} \hat{f}_{3A}(x,y) = \gamma_{0A} & (x,y) \in A \\ \hat{f}_{3B}(x,y) = \gamma_{0B} & (x,y) \in B \end{cases}$$

- Function $\hat{f}_4$: subdivision of the block into two regions A and B separated by a straight line and approximation with a linear function: a **platelet**.

$$\hat{f}_4(x,y) = \begin{cases} \hat{f}_{4A}(x,y) = \theta_{0A} + \theta_{1A}x + \theta_{2A}y & (x,y) \in A \\ \hat{f}_{4B}(x,y) = \theta_{0B} + \theta_{1B}x + \theta_{2B}y & (x,y) \in B \end{cases}$$

The Figure 4.2 depicts some example of patterns for each modelling function.

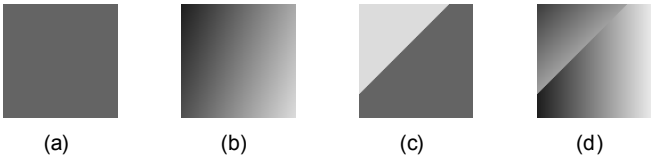



Figure 4.2: Example patterns well modelled by $\hat{f}_1, \hat{f}_2, \hat{f}_3 and \hat{f}_4$ respectively. From [92].

An estimation step is necessary to model coefficients minimizing the approximation error. For $\hat{f}_1$ one coefficient has to be computed which minimizes the error between the model and mean value of the original data. For $\hat{f}_2$, a linear function approximates the

gradient of a block by : $\hat{f}_2(x,y) = \beta_0 + \beta_1 x + \beta_2 y$ through a least-squares optimization. For the wedgelet $\hat{f}_3$ and the platelet $\hat{f}_4$ functions, both the model coefficients but also the separating line have to be found. The coefficient estimation is tested on every possible line orientation/translation dividing the block into two areas. The wedgelet/platelet coefficients are computed over the candidate regions A and B using the average pixel values and a least-squares minimization respectively.

The authors finally proposed to optimize the compression in a rate-distortion sense through the optimal choice of the best modelling functions. Three parameters are successively optimized: an independent selection of modelling functions, a quadtree structure optimization and a quantizer step-size selection. The usual approach is then to define a Lagrangian cost function that combines both rate $R_z$ and distortion Dz of the depth map $z$:

$$J(R_z) = D_z(R_z) + \lambda R_z$$

Since the rate and distortion are additive functions over all blocks of the depth maps, they propose an independent optimization within the blocks.

For each block, the algorithm first selects the best modelling function $\hat{f}$ in a R-D sense: $\hat{f} = \arg\min_{\hat{f}_1 \in [\hat{f}_1, \hat{f}_2, \hat{f}_3, \hat{f}_4]} (D_m(\hat{f}_i) + \lambda R_m(\hat{f}_i))$ with $R_m(\hat{f}_i)$ and $D_m(\hat{f}_i)$ the rate and distortion due to one modeling function $\hat{f}_i$ respectively.

Then the optimal quadtree decomposition is obtained by a bottom-up tree-pruning technique that consists of pruning the four children nodes of a common parent node whenever the sum of their four Lagrangian cost functions is higher than the one of the parent node. When not pruned the sum of the four coding costs is assigned to the parent node. This process is recursively performed in a bottom-up manner.

The selection of an appropriate quantizer $\tilde{q}$ out of a given set of possible scalar quantizers $q_2, ...q_8$ at 2 to 8 bits per level re-uses the Lagrangian cost function such as: $\tilde{q} = \arg\min_{q_1 \in q_2, ...q_8} D_j(R_j, q_j) + \lambda R_j(q_l)$ where $q_l$ is added to represent the quantizer selection. The image $j$ is encoded with all possible quantizers to select the quantizer $\tilde{q}$ that minimizes the coding cost $J_i(R_j, \tilde{q}) = D_j(R_j, \tilde{q}) + \lambda R_j(\tilde{q})$. The parameter $\lambda$ is calculated using an usual bisection search yielding the highest image-quality at a specific bitrate.

To resume, an input depth image and a weighting factor $\lambda$ are required by the algorithm. The depth map is subdivided into a full quadtree decomposition. All nodes are approximated by the four modeling functions before their coefficients are quantized using one scalar quantizer $q_l$. An optimal modeling function $\hat{f}$ is then chosen. The full-tree is finally pruned in a bottom-up fashion. The coefficient quantization step is repeated for all quantizers in order to select the quantization minimizing the global coding cost.

The quantized zero-order coefficients are predicted from their neighbourhood and the residual is coded with an adaptive arithmetic encoder. The prediction scheme coupled to an arithmetic encoding yields between 0.5 and 1.5 dB of PSNR gain at fixed bitrate.

Experimental results are then given in [87] on two Microsoft Research (MSR) dataset sequences. The results show Rate-Depth map PSNR distortion curves for multiview depth video compression based on H.264/MPEG-4 AVC Intra, H.264 MVC which both perform better than the Platelet-based coding for an average of 8 cameras and 25 frames. This is because the first methods additionally rely on temporal and inter-view reference picture prediction. However the Platelet-based coding outperforms those methods on the Breakdancers sequence when considering their Rate-Distortion on synthesized view which, indeed, really matters.

The efficient RD optimizes the depth map coding method with the supplementary advantage to be block-based and to rely on a quadtree decomposition. Because the new

HEVC encoder relies on the same structure, some parts of this described method have recently been proposed as depth coding modes for the 3DVC-HEVC based solution. This will be described in the next section.

## 4.3 Recent work: ad-hoc depth coding modes in 3DVC

Some concepts proposed by [92] (see last section 4.2) have been recently implemented into the 3DVC-HEVC-based proposal: the sub-block partitioning into two arbitrary regions, the approximation of these two regions with constant values (the wedgelet) and the prediction of a region from its adjacent samples of neighbouring blocks.

However, the constant function, linear function, and the subdivision of a block into two regions approximated with a linear function (i.e the platelet functions) are not directly implemented, but an existing similar mode already exists (Intra DC prediction mode etc...)

However four different depth-modeling modes based on partitioning have been added. They differ in their way to partition the blocks:

1. The Explicit Wedgelet signaling
2. The Intra-predicted Wedgelet partitioning
3. The Inter-component-predicted Wedgelet partitioning
4. The Inter-component-predicted Contour partitioning

### 4.3.1 Mode 1: Explicit Wedgelet signalization

The explicit Wedgelet Signalization consists of finding and transmitting the best approximation function through a Wedgelet partition. The wedgelet block partition information is stored in the form of a binary partition pattern signalling which pixel belongs to P1 and which to P2, as illustrated in Figure 4.3. At the encoder, an extensive search of the best wedgelet partition using the original depth of the current block to be coded is carried out. Once the partition minimizing the distortion with this original block has been found, the prediction signal is evaluated with the conventional 3DVC mode decision process.

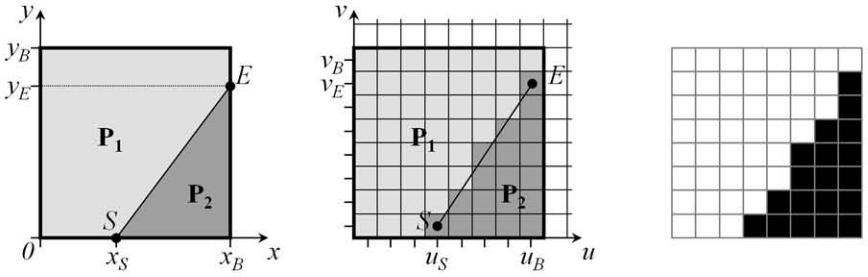



Figure 4.3: Wedgelet partition of a block: continuous (left) and discrete signal space (middle) with corresponding partition pattern (right). From [118].

For efficient coding and fast search of the best matching partition, the patterns of all possible combinations of line partitions are stored in a look-up table (LUT). The resolutions for the start and end line positions depend on the block size: 2 sample accuracy for 32x32 and 16x16 blocks, full sample accuracy for 8x8 blocks and 1/2 sample accuracy for 4x4 blocks.

### 4.3.2 Mode 2: Intra-predicted Wedgelet partitioning

In this mode the separation line of the current block is predicted from its neighbourhood. There can be a prediction from a neighboring wedgelet reference block by continuing the separation line in the current block. This is illustrated in Figure 4.4 (left). If the reference block is of type intra, the gradient is derived from the intra prediction direction and the start position -"$S_p$" on Figure 4.4- by selecting the adjacent sample position along the maximum slope. The resulting $E_p$ is calculated from this start point and from the gradient.

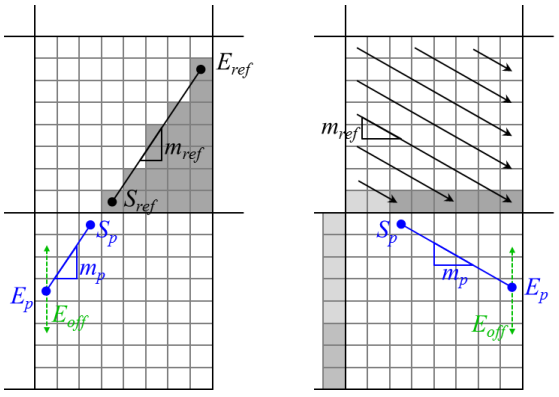



Figure 4.4: Intra prediction of Wedgelet partition (blue) in the scenario where the above reference block is either of type Wedgelet partition (left) or regular intra direction (right). From [118].

### 4.3.3 Mode 3: Inter-component-predicted Wedgelet partitioning

The idea behind this mode is to predict the Wedgelet partition of the current depth block from a co-located texture block in the video picture, as illustrated in Figure 4.5 in blue. The partition in wedgelet is not transmitted but signalled so that the inter-component prediction uses the reconstructed video as reference for the partitioning.

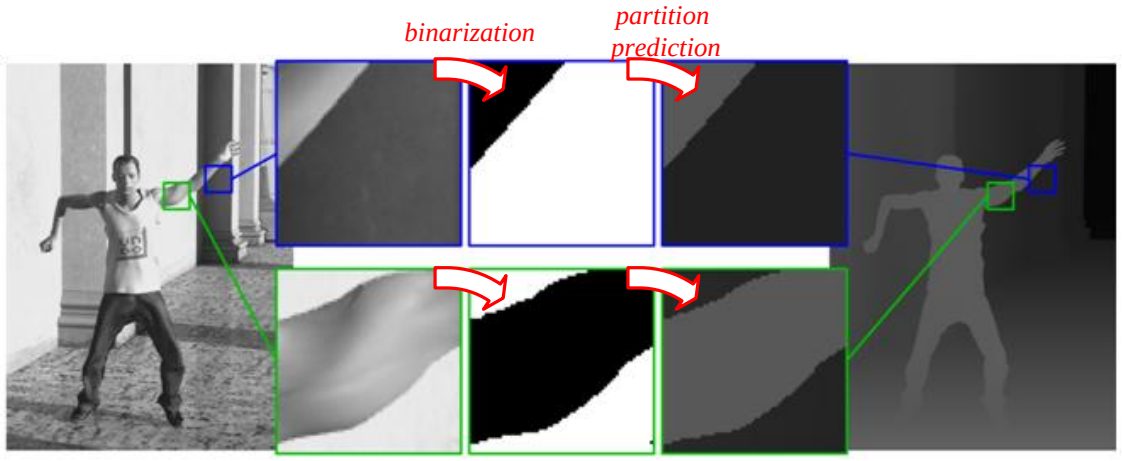



Figure 4.5: Prediction of Wedgelet (blue) and Contour (green) partition information from texture luma reference. Modified from [118].

### 4.3.4 Mode 4: Inter-component-predicted Contour partitioning

The last added mode aims at predicting a contour partition instead of a line partition. Inter-component prediction is also used; the co-located texture block is used as reference, as illustrated by Figure 4.5 in green. However, the partition depends on a threshold fixed as the mean value of texture reference block. If the sample value is above or below this threshold, it is assigned as a part of one or two regions in the contour partition pattern.

### 4.3.5 Constant Partition Value and Mode Pre-selection

The four modes rely on constant values for each region at the difference of [92]. This limitation may be overcome by the small possible size of block. It also allows the prediction of these constant partition values (CPV) from their adjacent samples at the left and top neighbouring blocks.

One of the four modes or one of the conventional intra prediction modes is finally signalled for each intra-coded CU. Each of the four modes can be signalled with or without the differential CPVs, resulting in 8 mode identifications. To reduce the complexity at the encoder, a mode pre-selection excludes depth modeling modes unlikely to be selected (such as Inter-Compononent Wedgelet and Contour partitionings derived from uniform texture blocks for instance).

### 4.3.6 Last contributions on edge orientation coding

As we will see, our proposal consists in separately coding the edge location and edge values along a predefined oriented path along the edges.

Recently, [57] proposed coding explicitly the edge location and more precisely the path direction along the edge. This is done at a block level by encoding the chain code i.e. the successive index of one orientation among the 8 possible orientations. This contour partition coding might be an effective alternative to the "Inter-component-predicted Contour partitioning" mode, but its implementation is not known yet. It achieves between 0.2% and 0.4% Bjontegaard Distortion (BD)-rate gains compared to 3DV-HTM version 3.1.

## 4.4 A Lossless Edge based Depth Map Coding Method

Depth maps have two main features that must be preserved but can also be relied on for efficient compression. The first one is the sharpness of edges, located at the border between object depths. Distortions on edges during the encoding step would cause highly visible degradations on the synthesized views, that may require depth map post-processing. The second one comes from the general smooth surface properties of objects whose depth is measured.

While Merkle et al. first proposed that smooth regions could be approximated by piecewise-linear functions separated by straight lines, we indeed assume that these smooth surfaces could be entirely reconstructed by interpolating the luminance values from their boundaries. Then the coefficients of the piecewise-linear functions would not be transmitted but instead the pixel values on both side of the edges.

To this end, we can observe that depth maps share similarity to cartoon-images. Mainberger et al. [79] proposed a dedicated cartoon-image encoder, that - in low bitrate conditions - beats the JPEG-2000 standard. After a Canny edge detection, the edge locations are encoded with a lossless bi-level encoder, and the adjacent edge pixel values are lossy

quantized and subsampled. At the decoding stage, a homogeneous diffusion or an interpolation are used to retrieve the inside unknown areas from lossy decoded edges. Indeed, the demonstrated performances -while beating state of the art codecs- reach the limit of 30dB.

We revisited this edge-based compression method by proposing improvements to fit the high quality, low bitrate, and specific requirements of depth maps. Finally, we increase the diffusion-based depth map encoding performance, which might be generalized to all kinds of images.

In the next sections the encoding process is described. Then the decoding, diffusion and interpolation methods are explained. Results, performances and comparison with state-of-the-art methods based on a traditional objective metric are then given in the next section, before subjective experiments draw more meaningful results on the effectiveness of our proposal.

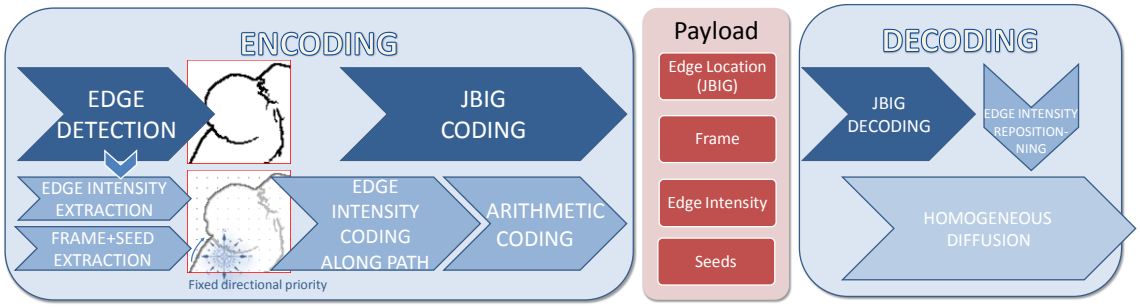



Figure 4.6: Diagram of the proposed depth map compression method.

### 4.4.1 Encoding

The encoding is a 5-step process: first is the detection of edges, then encoding of the edge location and finally encoding of the edge, border and seed pixel values, as illustrated by Figure 4.6.

#### Edge detection

Different operators exist to extract the contour of an image. An optimal edge detector should provide:

- a good detection: the algorithm should find as much real edges as possible.
- a good localization: the edges should be marked as edges as close as possible to the real edges.
- a good robustness: as much as possible, the detector should be insensitive to noise.

In our context of depth map edge coding, several requirements are added. The quality of reconstruction by diffusion should be maximized, while minimizing the number of edges to encode. To avoid diffusion from bad positioned edges causing "leakages", the localization of contours should be quasi-perfect. The detection of contours should be good but avoiding an over-detection. Up to a certain limit, weak contours (i.e. with a low gradient) might

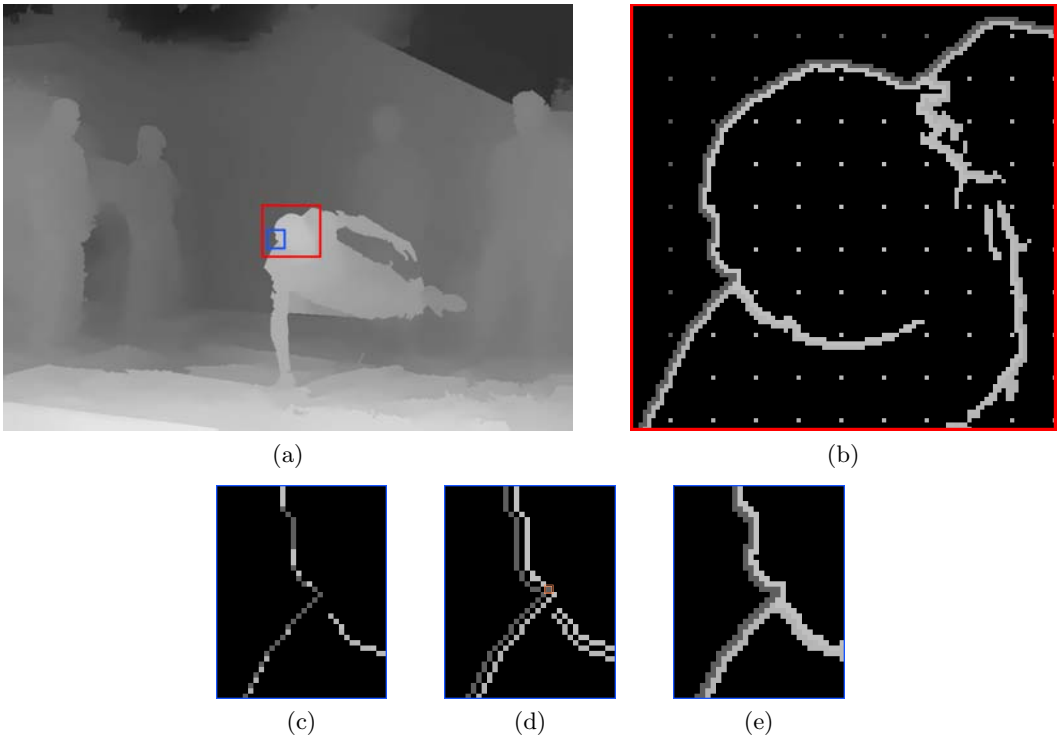


Figure 4.7: (a) A "Breakdancer" depth map, (b) the encoded and decoded Sobel edge and seed pixels (red selection on (a)), (c) a zoom (blue selection) on Canny edges, (d) the selection of corresponding pixel adjacent to Canny edges (c) as in [79], with an intruder edge pixel (orange-framed) that will lead to bad diffusion, (e) the proposed Sobel selection of edge pixel values, exactly located from both side of the frontier edge.

be useless for the reconstruction and might unnecessarily increase the edge coding cost. Also, noisily detected pixels should be avoided for the same reason.

The Marr-Hildreth edge detector combined with Canny-like hysteresis thresholding is used in [79], but suffers from errors of localization at curved edges. The widely used Canny edge detector has also been benchmarked. It relies on a 5x5 gradient prefiltering to cope with noise before local maxima edge detection. But this prefiltering step also makes this detector vulnerable to contour localization errors, as illustrated in Figure 4.7(c), where inexact selection of adjacent edge pixels leads to improper diffusion. In contrast Sobel has the advantage of an accurate contour localization -as shown in Figure 4.7(d)- at the cost of a noisy, edge over-detection. To cope with these over-detected edges, contours $c$ shorter than a certain value ($c < 14$) are excluded. Pixels with a bi-dimensional gradient amplitude larger than a threshold $\lambda$ are extracted. Used with sharp depth maps, this gives well-localized contours.

### Encoding the contour location

As in [79], a bi-level edge image containing the exact location of previously detected edges is first encoded using the *JBIG (Joint Bi-level Image Experts Group)* standard. This is a context-based arithmetic encoder enabling lossless compression of bi-level images. We use the JBIG-Kit, a free C implementation of the JBIG encoder and decoder. The progressive mode is disabled to reduce the required bitrate.

### Encoding the contour values

Once the edge pixel locations have been encoded, the pixel luminance values have also to be losslessly encoded following our initial requirements. The authors in [79] proposed to store the pixel values on both sides of the edge, instead of the pixel values lying on the edge itself. Indeed, for blurry contours, this might be valuable to interpolate the inner part of the edge and code the luminance values on both sides. However, with sharp depth maps, the pixel values lying directly on an edge, as illustrated in Figure 4.7(b), alternate between one side or another from this edge and couldn't be interpolated correctly.

With the Sobel edge detection not thinned to a single edge pixel, we ensure retaining at least one pixel value from each side of the frontier edge as shown in Figure 4.7(d).

We keep the idea of storing the pixel values by their order of occurrence along the edge to minimize signal entropy. A path with fixed directional priorities (E, S, W, N, NE, SE, SW and NW) is used. As the intrinsic properties of pixels along an edge or "isophote" are their small luminance variation, then we propose to compute the differential values of edge pixels in a Differential Pulse Code Modulation (DPCM) way. From this optimized path encoding method, the stream of DPCM values is then encoded with an arithmetic coder.

Additionally to these edges we also encode two kinds of information. The pixel values from the image border are stored to initiate the diffusion-based filling from borders. Inspired by the work of [7] on "dithering" for finding optimal data for interpolation, we propose to sparsely deploy, at regular intervals, some seeds of original depth pixels as shown in Figure 4.7(b) (The interval $s = 10$ in practice) While having low overhead, we discovered that this helps accurate reconstruction by initializing and accelerating the diffusion in large missing areas.

Thus, these extra border and seed pixels are coded in DPCM and added to the differential edge values. The resulting file is thus composed of the payload of the JBIG data and of the arithmetic encoded bitstream of the DPCM border, edge and seed pixel values. A typical PCM payload and its subsequent DPCM payload (from a depth map encoding of the "breakdancers" sequence) are illustrated in figure 4.8.

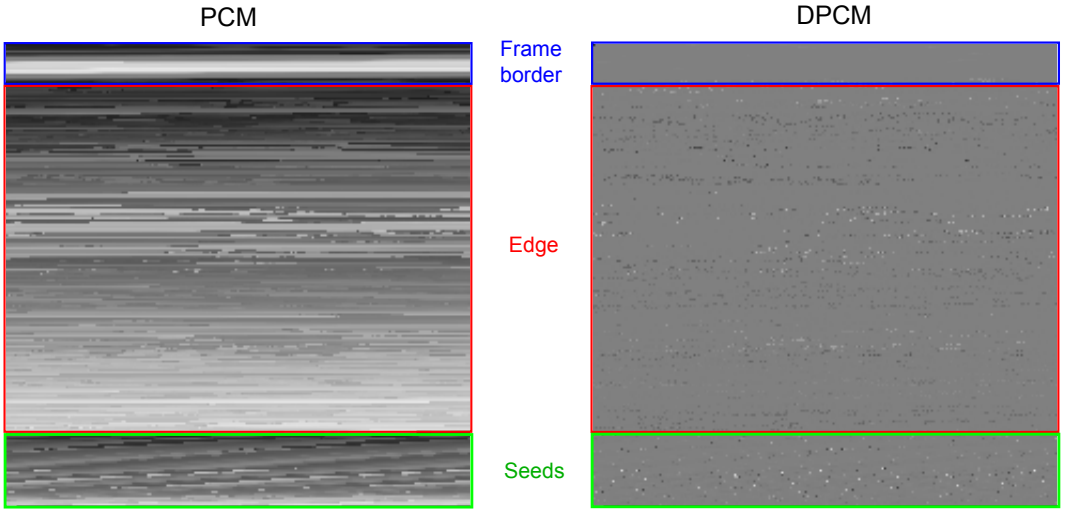



Figure 4.8: Illustration of the payload of the encoded pixels intensity values ("breakdancers" depth map). The PCM (left) pixel intensity and the DPCM (right) residuals are displayed in raster scan order.

It can be seen on the PCM values (4.8 left) that the pixels vary very smoothly along the frame border: the frame border pixel entropy is very low after DPCM on upper part of 4.8 right. Interestingly , the upper frame border is black while the lower is bright: it comes from the scene configuration: the upper part is the farther part, the background whole, and lower part of the scene is the ground closer to the camera.

The original pixel values of the edge pixels coded with a walk along the edge appear more cluttered: edges pixels contribute the most to the global entropy of the image. The edges are found and coded in a raster -scan order on the whole image: even if the edges have low-varying values along their edges, once an edge is terminated, the iterative process follow the encoding from the next vertex whose values can be very different. Each jump to the next edge can be illustrated by the higher or lower residual error than the middle gray value in 4.8 right.

Finally the seeds values appear regular: their luminance is encoded in a raster scan order so the highest residual pixels come from a jump from seeds on background to foreground and reversely.

### 4.4.2 Decoding, diffusion vs interpolation

A lossless decoding of the border, edges and seeds is performed. Then two methods are proposed and evaluated to reconstruct the depth map surfaces: a lossy diffusion or an interpolation from the decoded edge pixels. Finally, a quadtree approach is proposed to place the seeds.

#### Decoding contour location and pixel values

Once the edge positions from JBIG payload are decoded, the edge pixel values are decoded and positioned following the same order in which they were encoded: the path along contour location respecting directional priorities. The border and seed values are also re-positioned following a predefined location.

#### Reconstructing the Missing Values by Diffusion

We now have a sparse depth map containing only the edge, border and seed pixel values. A homogeneous diffusion-based inpainting approach is used to interpolate the missing data. This method is the simplest of the partial differential equations (PDEs) diffusion method, and has the advantage of low computational complexity. It directly stems from the heat equation:

$$\begin{cases} I_{t=0} = \widetilde{I} \\ \frac{\delta I}{\delta t} = \Delta I \end{cases}$$

where $\widetilde{I}$ is the decoded edge image before diffusion that will constitute the Dirichlet boundaries of the equation. The diffused data then satisfies the Laplace equation $\Delta I = 0$. The diffusion process is run in a hierarchical manner, each diffusion step being in addition helped with seeds and appropriate initialization. These three improvements have been introduced in the classical diffusion approach to limit the number of iterations required to converge, hence to speed up the entire process.

**diffusion** A Gaussian pyramid is built from $\widetilde{I}$. The diffusion process is first performed on a lower level of the pyramid and the diffused values are then propagated to a higher level (3 levels are used and show good performances). The propagation of the blurred

version of the diffused pixel values from a lower level to an upper one helps to initialize the diffusion in unknown areas.

**Middle range initialization** On the highest level, instead of starting from unknown values of $\widetilde{I}$ set at 0, we propose to initialize unknown values to the half of the possible range: 128 for an 8 bit depth map. This facilitates and speeds up the process of diffusion by limiting the number of required iterations to converge.

**Seeding** As explained in section 4.4.1, some seeds are chosen from a regular pattern both to accelerate the diffusion process and to provide accurate initialized values in large unknown areas. Indeed, this definitely achieves a fast and accurate diffusion -with a gain of 10 dB- for a quasi-exact reconstruction of the depth map.

#### Reconstructing the Missing Values by Interpolation

A bi-linear interpolation is also tested to render the missing values between border, edges and seeds. A linear interpolation in both directions from the border, edge and seed pixels is simply realized. It is equivalent to a two-table lookup with linear interpolation in both horizontal and vertical directions.

This approach is more simpler than the diffusion one and gives better results: no iterative process of diffusion are required up to an asymptotic point. This will be assessed in the section 4.5.

### 4.4.3 An Extension by a Quadtree Approach

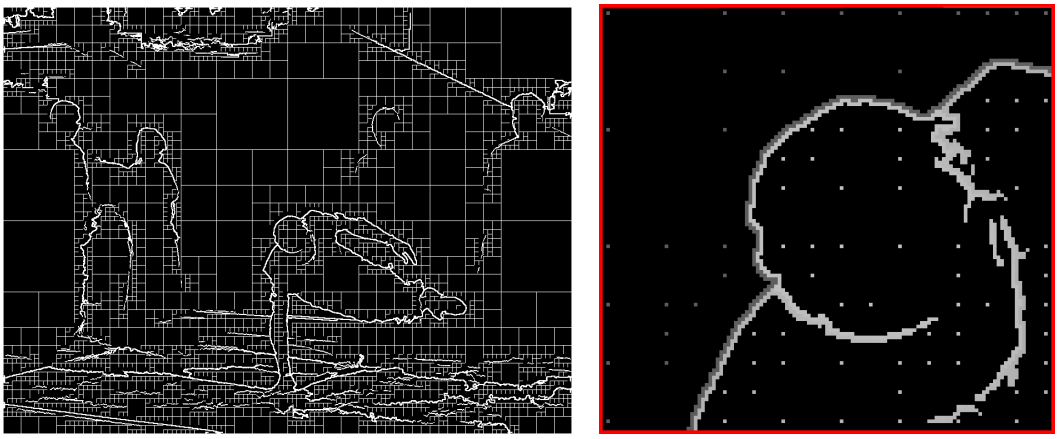
Another approach has been finally proposed to add flexibility to the coding of seeds. Instead of placing the seeds on a regular pattern, the seeds are positioned on each vertex of blocks obtained from a quadtree decomposition. The original depth map is recursively divided into four equal-sized square blocks based on the presence of edges within the block. If an edge is present within a given block, this block is subdivided into smaller blocks, up to the limit of the smaller size of block.

Because the edge positions are transmitted and the smaller size of block is fixed and known at the decoder side, no quadtree decomposition needs to be conveyed: only the border, edge, and seed -positioned according to the quadtree- values are transmitted in addition to the pixel location.

A quadtree decomposition of a “breakdancers” depth map is illustrated in 4.9(a). The quadtree is well decomposed into smaller blocks along the edges, while big blocks remain in the large smooth areas. However, the interpolation in these areas might be poor because only four seeds at vertices will help the interpolation (except at frame border). Small variation of depth surfaces not extracted by edge detector will not be reproduced on large areas, while they would be partially reproduced with seeds placed on a regular pattern. This will be assessed in the next section. The hypothetical advantage to use seeds placed according to a quadtree rather than on a regular interval is illustrated on Figure 4.9(b), to compare with 4.7(b). The seeds are more dense in the neighbourhood of detected edges which might lead to a better interpolation and reconstruction of the small surfaces.

## 4.5 Results based on objective quality evaluation

The performances of the proposed compression method and its extensions are first evaluated on an objective quality metric ground. An original resolution depth map from a MVD



(a) (b)

Figure 4.9: Ilustration of the quadtree decomposition on the criterion of edge presence. (a): Quadtree decomposition blocks are illustrated in white. The minimum size of block is 8x8. (b) A zoom on 4.7(a) with the seeds positioned according to the quadtree.

sequence "Breakdancers" from [155] is used. It was accurately estimated through a color segmentation algorithm. The original good quality of the depth map will enable a precise evaluation of the impact on texture synthesis reconstruction of the depth distortions.

### 4.5.1 Depth map objective quality evaluation

The reconstruction quality of our PDE-diffusion-based method is investigated and compared with the JPEG-2000 and HEVC-HM-4.1 Intra compressed versions. First, to visually illustrate the difference of quality reconstruction on edges, the three methods are compared at equal *Peak-Signal-to-Noise-Ratio* (PSNR),(45 dB, JPEG-2000 with a Quality factor Q=25, HEVC-Q=40).

A zoom on the head of a character presenting initially sharp edges highlights the difference of edge quality depending on the compression type (Figure 4.10). While at high PSNR, the JPEG-2000 (a) and HEVC (b) versions of the depth map tend to blur the edges. This is commonly referred to as ringing artifacts. It appears with JPEG-2000 because of the lossy quantization following wavelet transformation. It might appear with HEVC because of deblocking filter limitation. Then both JPEG-2000 and HEVC cannot efficiently reconstitute the smooth gradient on uniform areas while preserving the edges. In contrast, our proposed approach stores the exact edges and diffuses regions between these edges, resulting in a smooth gradient restitution on slanted surfaces and non-distorted edges.

Thus we evaluate the global depth-map rate-distortion performances of the three previous encoding methods plus the interpolation and the interpolation-plus-quadtree methods. Figure 4.11 shows that the diffusion approach outperforms JPEG-2000 except in very low or high bitrate conditions, while being under HEVC. No dedicated adjustment was performed in our method, only the threshold $\lambda$ was varied to adjust its bitrate (an interval of 10 pixels between seeds was chosen for the tests).

The interpolation approach with the same regular pattern of seeds gives substantial improvements in terms of depth map PSNR (with the original depth map as reference). An average gain of 4dB in term of PSNR is obtained in practice between the interpolated-decoded depth map and the diffused-decoded depth map.

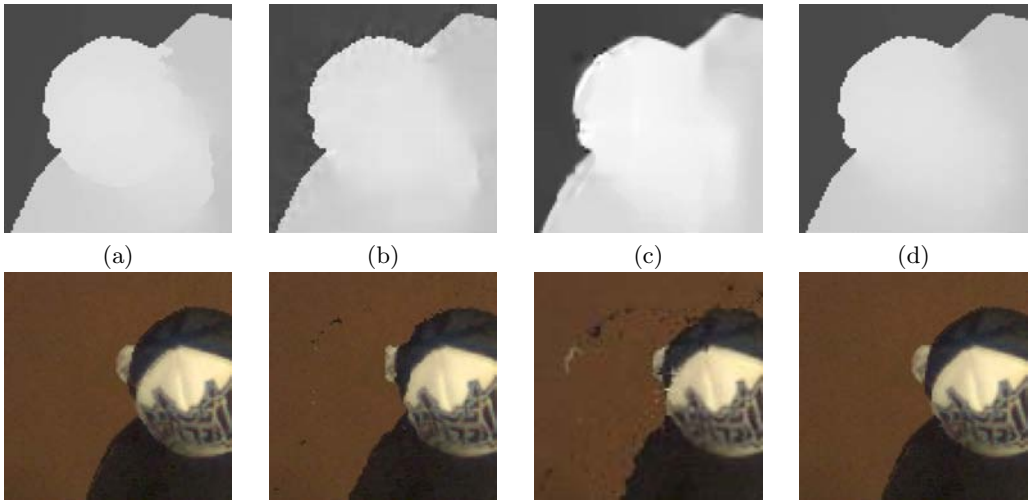



Figure 4.10: Upper row: zoom on the head of a dancer on original View #3 ($V_3$) depth map (a) highlights -by comparison at equal depth map PSNR (45dB) referenced to (a)- the ringing artifact on JPEG-2000 (b) and the blur effect with HEVC (c). Our method (d) based on exact edges and homogeneous diffusion prevents this effect (contrast has been increased on depth maps for distortion visibility). Lower row: zoom on corresponding synthesized view $V_4$ without (e) or with JPEG-2000 (f) and HEVC (g) compressions and our diffusion-based method (h).

With the quadtree approach, the supposed advantage of reduced seed number on the bitrate is counterbalance by the decrease of depth map PSNR quality. For different minimum sizes of quadtree blocks, 8 and 32 pixels, the average fall of quality is of 6dB and 7.5dB on average respectively. The maximum size of quadtree blocks was 512 pixels and might also have been limited to a small size. As proposed before, it is precisely on these large uniform areas that the loss of PSNR quality is important. But have these falls repercursions on the objective visual quality of the synthesized - and then displayed - view? This will be presented in the next section.

### 4.5.2 View synthesis quality evaluation

The impact of depth compression methods on rendering is measured by calculating the PSNR of a synthesized view (from a pair of uncompressed textures and compressed depth maps), with respect to an original synthesized view (from a pair of uncompressed textures and depth maps). The corresponding synthesized view from two original depth maps is then the reference. VSRS 3.0 [78] is used for view interpolation from this 2-view dataset.

The R-D synthesis performance, illustrated in Figure 4.12, justifies the edge-coding approach over wavelet based encoders: undistorted edges permit an accurate and efficient view coding and rendering. The PSNR quality of synthesized view is better than JPEG-2000 with the edge-based method with both diffusion or interpolation filling from regular seeds. The interpolation method even beats the HEVC intra coded method for 0.1 to 0.2 bpp. However, the PSNR measure shows its limitation of objective evaluation on perceived quality. Our method does not always outperform in term of rate-distorsion the existing methods (Figures 4.11, 4.12), but still can improve the perceived quality of the synthesized view, especially around critical edges (see Figure 4.10).

The non-linearity of the depth map and synthesized view PSNR quality deserves some explanation. The rate-distorsion PSNR curves are all monotonic but a drop appears

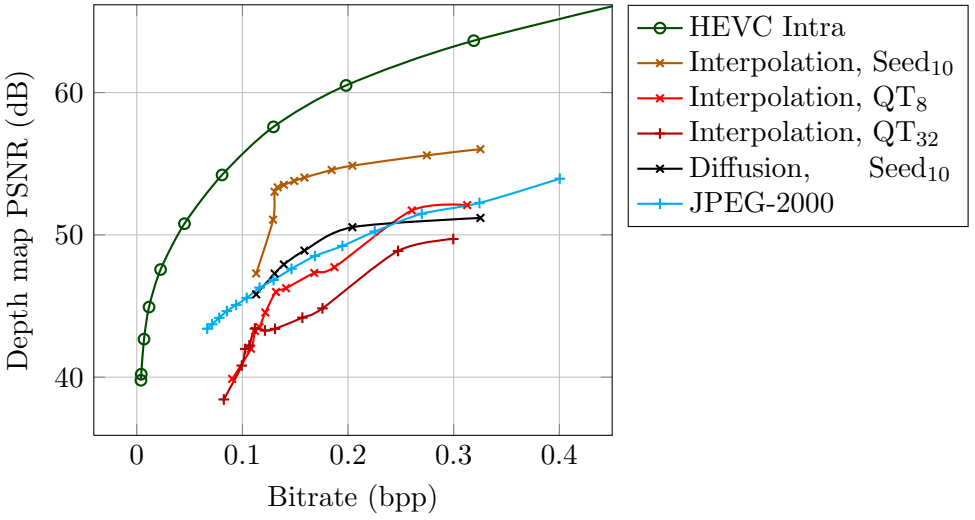



Figure 4.11: Rate-Distortion performance of the $V_3$ first "breakdancers" depth map with different quality factors of JPEG-2000 and HEVC and different Sobel detection thresholds $\lambda$ for the three proposed methods. The proposed methods differ in the interpolation from decoded edges and seeds: diffusion or bi-linear interpolation from regular seeds and bi-linear interpolation from quadtree-placed seeds.

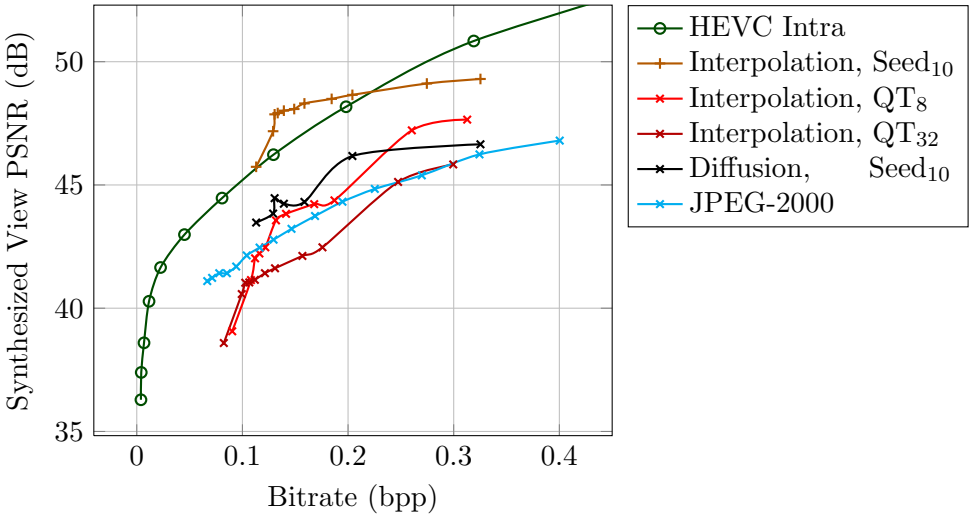



Figure 4.12: Rate-Distortion performance of synthesized $V_4$ with the bitrate of $V_3$ depth map, for different quality factors of JPEG2000 and HEVC, and different Sobel detection thresholds $\lambda$ for the three proposed methods.

with diffusion-based and especially with interpolation-based solutions when the bitrate is reducing under 0.12-0.14 bpp. This can not be explained by the constant cost of seeds along a regular pattern that could become not negligible at low bitrate. The quadtree solution -which effectively reduced the cost of seeds and then the total cost of the methods- still introduces this non-linearity.

This fall might be due to detected edges non-connected anymore. When reducing the bitrate and then the edge detection threshold, the edges around object become non-connected or their edge pixels values along the edge are not well preserved. In those cases with our walk-along-edge technique, when a hole around an edge is encountered, a new vertex non-correlated with the preceding edge pixel has to be transmitted. When this happens over the whole image, the total quality of depth map reconstruction is affected because entire surfaces become poorly interpolated. A solution could be to reconnect the edge pixels at decoding side before interpolating their values along the edges and then between the edges.

## 4.6 Subjective Results

Subjective assessment of the influence and impact of compressed depth maps on 3D view synthesis have been conducted in the IVC lab of IRRCyN in Nantes. The experiments were in line with the MPEG on-going activities on 3D video coding and then relied on similar test conditions and view rendering techniques and the same sequences as input tested data. These were realized within the PERSEE project context and involved different depth coding methods of four french labs.

The idea was to observe, test and measure the influence of the proposed depth compression method on the perceived visual artefacts of rendered views and then on the consequent perceived visual quality by observers.

### 4.6.1 Experimentation

The first experiment presented below involves the rendering of multiple spatially-translated but temporally-fixed intermediate frames from the same two adjacent views obtained from uncompressed textures and compressed depth images. An intermediate rendering from a pair of frames was processed, then another shifted in space rendering from the same pair was processed and so on. The resulting shifted intermediate frames were then concatenated into a single monoscopic video displayed on a monoscopic screen. This leads to a video where the viewpoint is changing over time while the acquired time doesn't evolve: a bullet-time effect video.

#### Depth Map Coding Methods

Two state-of-the-art simulcast and one multiview-plus-depth video coding methods are selected for the benchmark: H.264/MPEG-4 AVC, HEVC and 3DVC respectively. The depth coding method is also compared to the recent JPEG-2000 still image coding standard which is based on wavelet-transform and is then supposed to efficiently encode the 8-bits uniform area while limiting the ringing artefacts. The selected method from our previous work was based on interpolation and quadtree seed distribution with a quadtree block minimum size of 32 pixels. This choice was made before the objective quality comparison with previous interpolation methods presented in 4.5.2. This method was originally

selected according to a bitrate criterion to compete with the state-of-the-art 3DVC technique. Finally, filtered and edge-filtered versions of the original depth map are tested for extensive comparisons.

#### Preselection of Coded Depth Map Versions According to the Perceived Synthesized Quality

Because the objective quality scores -with their associated bitrate- of depth maps could not give a good opinion on the range of subjective quality of the synthesized views, three experts first pre-selected different qualities of coded depth maps according to three classes of perceived quality of synthesis. This to ensure that the rank were given in roughly the same range of quality. Then the quality of synthesized views could be compared between each other into a dedicated class and evaluated by an observer.

#### Competing with 3D-HTM

The main issue when we want to compare a method to another one is how to do that fairly. This issue arises with the comparison of our depth map coding technique to the depth coding method embedded into the 3DVC video standard.

First, while our method does not involve any temporal prediction, it seems fair to compare the rate of the depth map coding technique to the distortion in view synthesis, expressed on a single frame instead of on the whole video. The intra mode of H.264 and HEVC are then used for the single frame comparison. The bitrate of the single Intra coded depth image is then used, while the distortions are evaluated on the resulting rendered temporally-shifted video.

Second, while our method uses in input only a single depth map, the 3DVC depth map coding uses a set of different inter-component prediction techniques to efficiently encode jointly the texture and the depth data (see sections 3.2.4 and 4.3), such as the inter-prediction Mode 3 and 4, but also the View Synthesis Prediction.

#### View Synthesis Method

The view synthesis method is the same as used by MPEG: the View Synthesis Rendering Software (VSRS). The same software version as currently used for normalization of MPEG 3D-HTM is retained, and the same rendering is used: view interpolation from two adjacent views. Two modes of interpolation are used, either with view blending or without.

The view synthesis is run with a pair of non-compressed texture images and the corresponding pair of compressed depth maps coded with one of the depth coding methods. Then, the impact of the depth map coding is isolated and can be measured independent of texture coding.

#### Resulting bullet-time video stimuli

All the tested methods are compared at three levels of quality of the resulting synthesized view, predefined by a video quality expert. For each depth map coding the blending mode of VSRS is either activated or deactivated. Thus, for each method, 3 levels x 2 blending modes are tested, so 6 bullet-time video stimuli were displayed, evaluated and annotated by the viewer.

Between each frame, the intermediate view was shifted $^1/_{50}$ from the left toward the right view. When the extreme intermediate right view before the original right view was synthesized, the inverse camera movement was done. Then, 50 resulting frames shifted

toward the right plus 50 frames in inverse movement were displayed successively. This allows the viewer to clearly identify the potential artefacts appearing in a moving video -as it will be the case in practice- while the pair of depth maps is the same over time.

#### Test material

Six videos as proposed by MPEG have been retained. Two Class-A Full HD video and four Class-C 1024x768 pixels resolution videos were used.

| Sequence | Encoded views | Displayed views |
|---|---|---|
| Undo Dancer | 1-9 | $(1 +^{1}/_{50*(9-1)-2}) \rightarrow (9 -^{1}/_{50*(9-1)-2})$ |
| GT Fly | 1-9 | $(1 +^{1}/_{50*(9-1)-2}) \rightarrow (9 -^{1}/_{50*(9-1)-2})$ |
| Kendo | 1-5 | $(1 +^{1}/_{50*(5-1)-2}) \rightarrow (5 -^{1}/_{50*(5-1)-2})$ |
| Balloons | 1-5 | $(1 +^{1}/_{50*(5-1)-2}) \rightarrow (5 -^{1}/_{50*(5-1)-2})$ |
| Newspaper | 2-6 | $(1 +^{1}/_{50*(5-1)-2}) \rightarrow (5 -^{1}/_{50*(5-1)-2})$ |
| Book Arrival | 6-10 | $(1 +^{1}/_{50*(5-1)-2}) \rightarrow (5 -^{1}/_{50*(5-1)-2})$ |

Table 4.1: Selected multi-view video sequences with their respective encoded and displayed views.

### 4.6.2 Viewing conditions

The assessment of the quality of the synthesized view was realized on a 2D conventional LCD display (Panasonic BT-3DL2550) in a controlled environment (following the recommendations ITU-R BT500-11). For the HD sequences, the observation distance was of 3H (ratio between height of the screen (310mm) and observation distance (93cm)).

### 4.6.3 Participants

27 subjects with normal or corrected-to-normal vision participated in the experiment. The experiment was split into two sessions. Each subject evaluated all the video stimuli.

### 4.6.4 Test protocol

The subjective assessment was conducted with the Absolute Category Rating with Hidden Reference (ACR-HR) methodology, as set forth in ITU-T recommendation P.910. The test sequences were presented one at a time and rated independently on a category scale. Among the displayed stimuli, a hidden reference was included to prevent the assessment values from being affected by differences in the video content used for assessment. The assessment results obtained by the ACR method are "normalized" by using the following formula to calculate the difference in scores between the assessment video and reference video, expressed as DMOS (Differential Mean Opinion Scores):

$$\text{DMOS} = \{\text{assessment video score}\} - \{\text{reference video score}\} + 5$$

with the quality of the reference video judged to be from "1: Bad" to "5: Excellent" by the subject between each video presentation.

### 4.6.5 Results: DMOS

The average DMOS of the 27 observers for the "Balloons" and "Book Arrival" bullet-time video synthesized sequences are illustrated in Figures 4.13. Additional results for the rest of MPEG sequences of 3DVC corpus are presented in Annex B.1. Different trends can be observed for the set and for each video sequence. These will be presented before the limitations and perspectives be discussed.

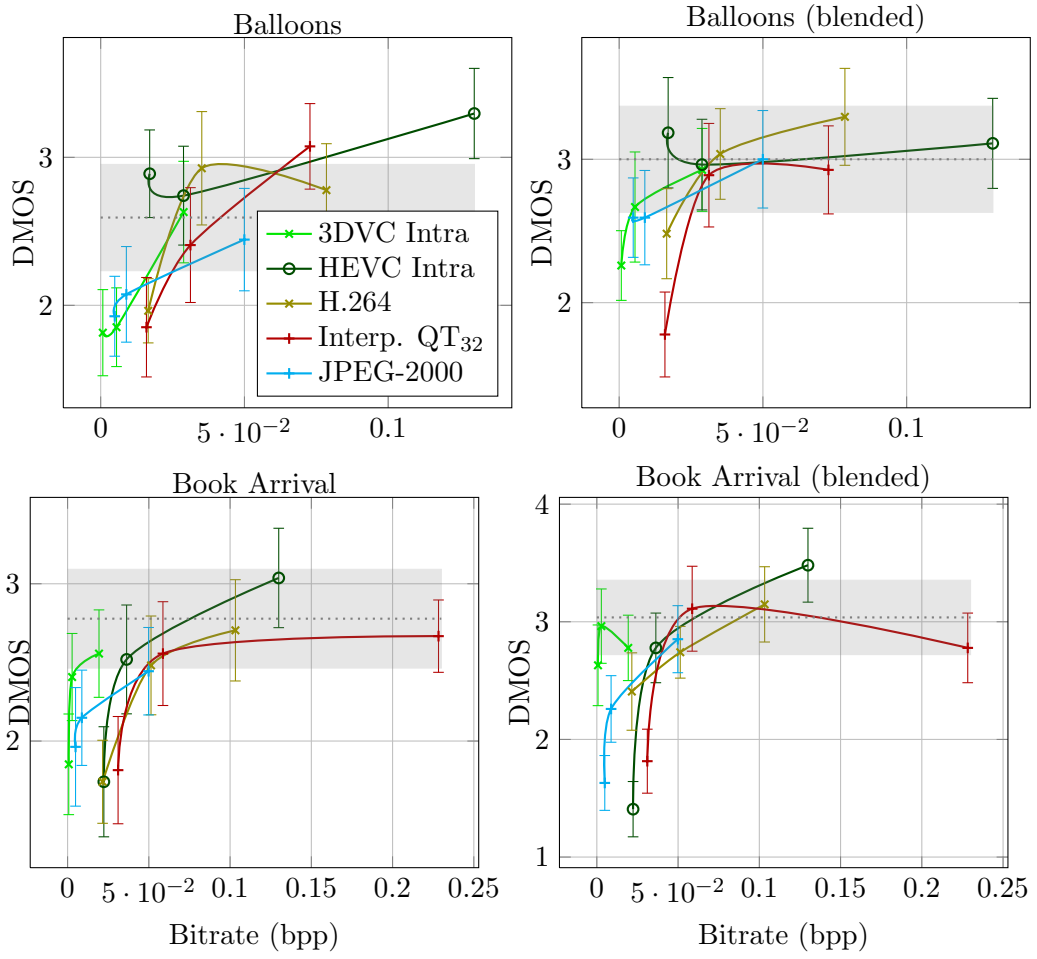



Figure 4.13: Average DMOS reported by 27 observers on Class-C "Balloons" and "Book Arrival" bullet-time synthesized videos for 3 classes of quality (high, middle and low quality) with 2 modes of rendering by the VSRS interpolation software (blending on/off to the right/left columns respectively) and for five depth map encoding methods. The DMOSs are obtained from a normalisation of the viewers' MOS by the MOS of the hidden non-coded reference picture. This HR-MOS is overlaid on the figure with a gray dashed line surrounded by its confidence interval illustrated by a gray area.

Globally speaking, the evolution of DMOS appears similar between the versions of VSRS rendering with or without blending, and this whatever the sequences.

Also, the 3DVC Intra coding of depth maps generally performs best among the low quality class of video. However, the 3DVC intra coded methods sometimes also present by far the lower subjective quality (Balloons and Kendo not blended). This raises the question of the best **trade-off** between **perceived quality** and **bitrate requirements**.

The compressions of the five methods selected in the middle quality class gives globally a subjective score which is inside the confidence interval: the five methods manage to compress effectively the sequence in a relatively low range of bitrate without affecting the perceived visual quality (between 0.025 and 0.04 bpp for class C videos). These two observations are important because they tend to show that -without considering the texture compression- a good compromise can be found between perceived quality and bitrate up to a certain limit. Under this limit, the bitrate is effectively decreased but so is the quality.

It is hard to conclude on the best solution for high quality reconstruction. H.264, lossless-edge coding method (with seeds placed at vertex of block of 32*32 minimum size) and HEVC manage to increase the perceived visual quality over the original depth map, but among this high quality class, their bitrates are often important. This interesting property of increase of quality from the depth map compression might be hard to use in practice because of the joint texture compression and because of the very low allocated depth bitrates.

Concerning each sequence, the "Balloons" sequence DMOS (without blending) of the different methods appears close to the HR-MOS for the middle quality class, but our method exceeds the others in term of rate-distortion ratio only for the high quality class.

On the "Book Arrival" sequence our method suffers from high bitrate at high and low quality class of videos. It exceeds the other method for the middle class in blended mode, but not significantly.

The "Kendo" is known to have poor quality depth map, so most of the method performances remain in the confidence interval of the original HR-MOS (without blending). The blending deeply increases the perceived video quality of the original reference, so that its performance and confidence interval are higher and lower respectively. Our method does not perform well in this setting for all classes of quality.

The high quality "Kendo" and "Newspaper" compressed with our method show either lower or higher bitrates respectively than the other methods. It shows the limit of the approach. No linear regressions were implemented to limit the size of the depth map according to a bitrate range rather than to an edge detection threshold.

The Full-HD Class A videos are synthetic videos and it seems hard to conclude on a global tendency. The artefacts appeared difficult to evaluate on the "GT Fly" sequences because most of the methods are within the confidence intervals of the original HR-MOS. It means either that the distortions are not visible enough, that the baseline between cameras is too small or that all the methods perform well on this content. This last hypothesis is very unlikely however.

In contrast, the "Undo Dancer" compressed versions are much lower than the hidden original version. This might be explained by the numerous planes at various depth that make the distortions between views very noticeable. Except with 3DVC, all the methods show a DMOS score under the rank of 2 points. Another experiment (with reduced camera baseline, slower movement, etc..) might help to clarify the performance of the five tested methods on this sequence.

### Discussion

First, the relative small amplitude and high confidence intervals of DMOS for the different tested methods limit the capabilities of interpretations and differentiations of the method performances. Second, the fact that these DMOS are most of the time inside the confidence intervals of the Hidden Reference MOS (HR-MOS) also affects the possible conclusions on the supposed impact of depth compression on perceived quality in 2D. The tested methods

have close subjective quality to the original reference, but it seems hard to differentiate them.

However, two clear trends appear. The 3DVC depth map coding gives globally the best "rate-subjective-distortion" performances for the class of low quality, with the lower bitrate but also with the lower DMOS. To evaluate precisely these implications, additional tests must be realized. But it is very likely that in practice the allocated depth bitrate will imply these distortion effects. However, the distortions impacting the perceived quality might be limited by the View Synthesis Optimization.

Also, the impact of the blending on the perceived quality is negligible except for "Kendo" whose original depth maps are very distorted. Additional tests with joint depth and texture compression could confirm or deny the limited impact of blending.

It seems very important to precisely adjust the baseline of views for synthesized videos by pre-tests. This has been done however and baselines were doubled for nearly all videos following an initial rendering. According to the results, the "GT Fly" rendering introduces a too short baseline while the "Undo dancer" baseline configuration is too large. Typically, a large baseline induces large disoccluded areas for each view before merging. It is on these merged areas that potential artifacts will appear. From one sequence to another, the size of these regions might vary, and influence the marks by the viewers. In other words, viewers' marks for a given sequence might be influenced by another preceding sequence and its synthesized quality.

Two important findings are highlighted from this first campaign of experiments in 2D conditions. First, all the methods manage to perform within the same range of perceived quality of synthesis from the original depth map within the middle and high quality classes (see the second and third points along each curves).

Second, and consequently, there seem to exist a **critical threshold** of distortion visibility where the DMOS is dropping, especially for the low quality classes of videos with bitrates under 0.025 bpp. This threshold might be decreased however in 3DVC by the use of advance compression predictions techniques such as the View Synthesis Optimization (VSO).

### Perspectives

These tests were conducted with the lossless-edge version with the quadtree configuration that led to the worst objective results. The comparison based on an objective metric was not realized before the subjective tests and so the quadtree configuration was retained only according to the low bitrate criterion. In the light of the objective results, the lossless edge encoding method with an **adapted quadtree** could give much better subjective results within the low and middle quality classes of videos, for slightly higher bitrates however. Because additional experiments are planned in stereoscopic conditions, this quadtree configuration giving the best objective scores will be tested.

The experiments raise the question of the best trade-off between quality and bitrate allocated to the depth map. Up to which **limit** can we decrease the depth map bitrate without affecting the synthesis quality so it remains interesting to transmit the depth maps rather than transmitting supplementary texture views (with or without their depth maps)?

Behind this question, two other fundamental questions are asked. How to determine a **correspondence law** between the quality factors of texture and depth maps, this depending on the baseline? How to evaluate precisely, objectively and subjectively the **impact of distortions** appearing on the **disoccluded areas** with the interpolation or

extrapolation rendering methods? This last issue is tackled in the next chapter with the use of objective quality metrics applied on extrapolated synthesized views and on their reconstructed disoccluded areas.

## 4.7 Conclusion

Different depth map coding methods have been presented in this chapter. One of the first block-based and block-partition-based proposed methods in the video coding community has been recently implemented into the new 3DVC encoder into one coding mode: the explicit wedgelet signalling. The two other intra-predicted and inter-component predicted wedgelet partitioning modes also rely on this idea of partitioning.

An alternative method has been presented that consists of separately coding the edge location at the picture level through a binary JBIG arithmetic encoder and the pixel values on both side edges in a predefined directional order.

An extension of this method has been proposed. It relies on the idea of an adaptive placement of seeds denser in the edge neighbourhood. But what is gained around the contours is also lost in large areas without contours. These surfaces become poorly reconstructed and then penalize the final synthesised quality. Then a good trade-off has to be found between a maximum size of block not too large that would decrease the synthesized quality and a minimum size of block not too small that would increase the bitrate without gain of quality.

The subjective results confirm the pertinence of the approach, but also show its limitations at very low quality and very high bitrate when using a non-adapted quadtree approach.

Thus the idea of coding the edge location, its partition by predefined pattern or chain code are both relevant. The 3DVC finally simplifies the edge values on both sides of the edge as constant ones. The complexity of the encoder is instead put on the refinement of the quadtree and of the depth block quantization optimizing an RD criterion on the resulting view synthesis: the view synthesis optimization.

Finally, the depth map coding methods may be substantially improved in the near future by considering the ecological structure of the scene; object based and context based approaches might induce relevant improvements on the depth map coding performances.

# CHAPTER 5

# Inpainting based View Synthesis for 3DTV and FTV

## 5.1 Introduction to View Synthesis

The view synthesis techniques are used at the last stage of the 3D framework before the display of 3D contents. It aims to render new viewpoints thanks to the previously encoded and decoded video plus depth dataset.

A new virtual viewpoint can be generated either by interpolation from multiple views or extrapolation from a single view. Both methods rely on a projection of image plus depth data - or warping - to a new viewpoint: the DIBR (see section 3.1.3).

In order to introduce the contributions on view synthesis aided by inpainting, the general formulation of view projection and of 3D image warping will first be presented with their problematics. The state-of-the-art inpainting methods in the general and particular case of hole-filling for 3D view synthesis will then be described. A formulation of what should involve an inpainting for perceptually correct synthesized view will be drawn before the proposals and results are finally presented.

### 5.1.1 Formulation of view projection

#### Projective Geometry

Human eyes and camera 2D images both result from the mapping of the 3D world. This lawful projection can be appropriately described by the mathematics of projective geometry. Projective geometry is the study of how a higher-dimensional space is mapped onto a lower dimensional space. Then it can explain where a given 3D scene composed of objects will project onto a given 2D image camera or retina plane.

The projective geometry offers the advantage to model points at infinity unlike finite Euclidean geometry. It can be observed that parallel lines in the 3D world do not project along parallel lines, but instead project on a 2D image towards a single vanishing point (as already illustrated in Figure 1.14). We have seen in chapter 1 that this element is furthermore used by the HVS to infer the depth and orientation of surface from the converging properties of lines in the visual environment: the appropriately called a perspective cue.

Projective geometry is thus a convenient mathematical framework to operate geometrical operations.

**Pinhole Camera Model** The pinhole camera enables the formation of a 2D-image of the 3D world by the use of a pinhole in front of an image plane in a lightproof box. The light falling on the image plane travels in straight lines from reflecting or emitting points in space. This situation gives the basic geometry of perspective projection and its pinhole camera model [47].

The pinhole camera model is mathematically described by an *optical center* or camera projection center **C** and an *image plane* **I** (see Figure 5.1). The distance from the optical center **C** to the image plane **I** is the focal length $f$. The line $Z_C$ crossing the optical center and orthogonal to the image plane is the *optical axis* of the camera. Finally the plane containing the optical center and which is parallel to the image plane is called the focal plane.

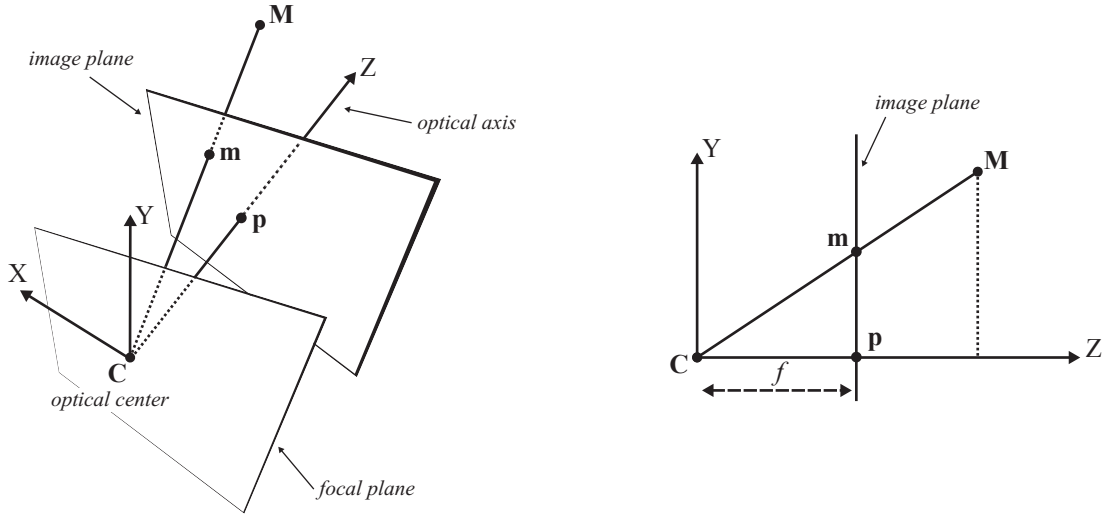



Figure 5.1: The pinhole camera coordinate system with its Image plane. From [30].

The projection from the 3D-world coordinate system $\mathcal{R}_0$ to a 2D-image plane in the camera coordinate system $\mathcal{R}_C$ is realized in four steps:

1. Coordinate transformation: the 3D coordinate system (3D-world) $\mathcal{R}_0$ is changed to the 3D-camera coordinate system $\mathcal{R}_C$.

2. 3D-to-2D projection: the 3D point now expressed in $\mathcal{R}_C$ is projected in the image plane.

3. Image plane shifting: displacement of the origin of the image.

4. Unit change: the distance metric coordinate units are changed to pixel units.

The camera model parameters distinguish between the extrinsic -external- parameters and intrinsic -internal- parameters [47, 105]. The extrinsic parameters describe the position and orientation of the cameras with respect to the $\mathcal{R}_0$ coordinate system, so these are required to pass to the coordinate system $\mathcal{R}_C$ (see Figure 5.2). The intrinsic parameters describe the properties of the camera itself, its lenses and its sensor: the focal length, the pixel size, the center of projection.

**Extrinsic camera parameters** The transformation from $\mathcal{R}_0$ to $\mathcal{R}_C$ is made of a 3x1 position vector **C** and the 3x3 rotation matrix **R**, as illustrated in Figure 5.2. The relation

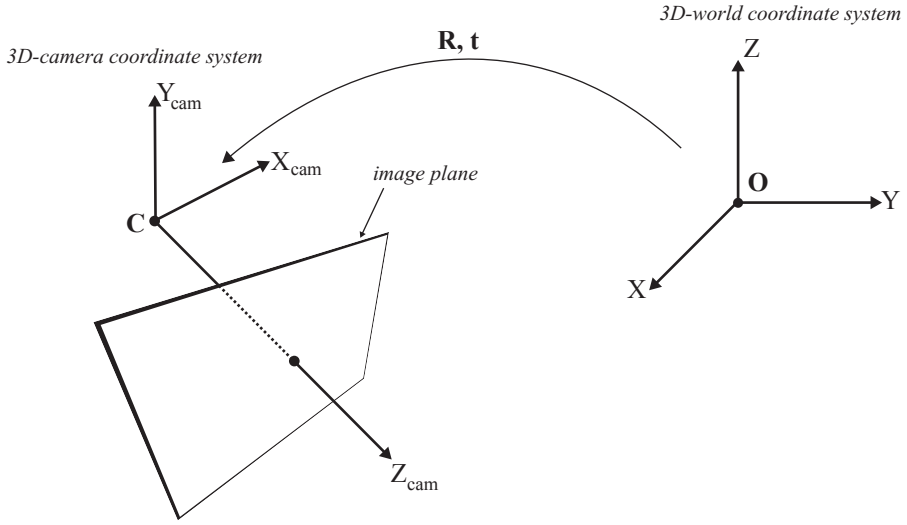



Figure 5.2: Projection from 3D-world to 3D-camera coordinate. From Daribo

between coordinates of 3D point expressed in $\mathcal{R}_0$ represented by a 3-element vector $\mathbf{M_0} = (x, y, z, w)^\intercal$ and $\mathbf{M_C} = (x', y', z', w')^\intercal$ expressed in $\mathcal{R}_C$ is:

$$\begin{pmatrix} x' \\ y' \\ z' \\ w' \end{pmatrix} = \begin{bmatrix} \mathbf{R} & \mathbf{0_3}^\intercal \\ \mathbf{0_3} & 1 \end{bmatrix} \begin{bmatrix} \mathbf{I_3} & -\mathbf{C} \\ \mathbf{0_3} & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} \tag{5.1}$$

with vector $C$ in non-homogeneous coordinates, all zero element denoted by the 1x3 vector $O_3$ and the 3x3 identity matrix by $I_3$. The equation can be reformulated as :

$$\begin{pmatrix} x' \\ y' \\ z' \\ w' \end{pmatrix} = \begin{bmatrix} \mathbf{R} & -\mathbf{RC} \\ \mathbf{0_3} & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{bmatrix} \mathbf{R} \\ \mathbf{0_3} \end{bmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} - \begin{bmatrix} \mathbf{RC} \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{R} \\ \mathbf{0_3} \end{bmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} + \begin{bmatrix} \mathbf{t} \\ 1 \end{bmatrix} \tag{5.2}$$

with $t$ being a $3 \times 1$ translation vector: $t = -RC$. Then the coordinates of $M_0 = (x, y, z, w)^\intercal$ transformed in the $M_C = (x', y', z', w')^\intercal$ 3D camera coordinate system will become:

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{bmatrix} \mathbf{R} | \mathbf{t} \end{bmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} \tag{5.3}$$

**Intrinsic camera parameters** The intrinsic camera parameters are then necessary for the perspective projection of a 3D-camera point expressed in $\mathcal{R}_C$ to a point in the image plane. They describe the properties of the lenses and of the sensor chip: the focal length $f$ the pixel size ($s_x$,$s_y$) and the center of projection ($u_0$, $v_0$). These are usually regrouped in the 3x3 camera calibration matrix $K$:

$$K = \begin{bmatrix} f/s_x & 0 & u_0 \\ 0 & f/s_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{5.4}$$

Knowing these intrinsic matrix, the 3D to 2D perspective projection in homogeneous coordinates of a 3D-camera point $M_C = (x', y', z', w')^\intercal$ to a 2D-image point $m = (u, v, \eta)^\intercal$ can be expressed as:

$$\begin{pmatrix} u \\ v \\ \eta \end{pmatrix} = \begin{bmatrix} f/s_x & s & u_0 & 0 \\ 0 & f/s_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{pmatrix} x' \\ y' \\ z' \\ w' \end{pmatrix} = \left[\mathbf{K}|\mathbf{0_3}\right] \begin{pmatrix} x' \\ y' \\ z' \\ w' \end{pmatrix} \tag{5.5}$$

where $\eta = z'$ is the homogeneous scaling factor.

**Summary**

Finally, a 3D-world point $\mathbf{M_0} = (x, y, z, w)^\intercal$ is given in the 2D-image plane by point $m$ of coordinates $m = (u, v, \eta)^\intercal$ such as:

$$\begin{pmatrix} u \\ v \\ \eta \end{pmatrix} = \mathbf{P} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} \tag{5.6}$$

$\mathbf{P}$ is a 3 x 4 camera projection matrix, generally invariant by scale factor, such as:

$$\mathbf{P} = \mathbf{K}[\mathbf{R}|\mathbf{t}] \tag{5.7}$$

with $\mathbf{K}$ the 3x3 camera calibration matrix defined in equation 5.4, $\mathbf{R}$ the 3x3 rotation matrix and $\mathbf{t}$ the 3x1 translation vector.

Then the camera projection $\mathbf{P}$ is a 3x4 matrix having 11 degrees of freedom:

- 5 from calibration matrix, $\mathbf{K}$
- 3 from rotation matrix, $\mathbf{R}$
- 3 from translation matrix. $\mathbf{t}$

### 5.1.2 Formulation of a 3D image warping

The 3D image projection -or warping- is one Depth Image Based Rendering (DIBR) technique to generate -or synthesize- a virtual new view. This view is said to be virtual because in the case of a 3D system the view in question is not transmitted but needs to be reconstructed from existing raw texture(s) and corresponding geometry(s) (depth maps in the case of MVD).

For each new view, the points from the reference texture image plane need to be mapped to a targeted image plane of the virtual view, as illustrated in Figure 5.3.

A **forward projection** can be realized in two distinct steps:

1. Back-projection of the reference image points into the 3D-world. The 3D points constitute a cloud of points of the scene in the 3D space.
2. Re-projection of the previous back-projected scene points in the targeted image plane [86].

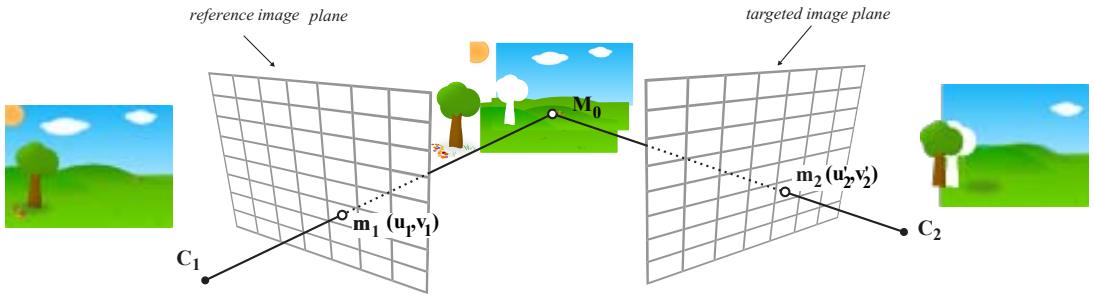



Figure 5.3: Projection of an image point $m_1$ to a 3D-world point $\mathbf{M}_0$ and then to an image point $m_2$.

First, the back-projection of a point $\mathbf{M_0} = (x, y, z, w)^\intercal$ from a reference view image $I_1$ at pixel grid coordinates $I_1(u_1, v_1)$ can be expressed as the product of the inverse matrices of the reference view $I_1$ such as:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \lambda \mathbf{R}_1^{-1} \mathbf{K}_1^{-1} \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} - \lambda \mathbf{R}_1^{-1} \mathbf{t}_1 \tag{5.8}$$

where $\mathbf{K_1}, \mathbf{R_1}$ and $\mathbf{t_1}$ are the camera intrinsic, rotation and translation matrices respectively. $\lambda$ is the positive scaling factor defining the position of the 3D point along the ray of possible back-projection (as defined by the inverse problem of back-projection in Appendix A).

The intrinsic, rotation and translation matrices of the targeted virtual view camera, respectively $\mathbf{K_2}, \mathbf{R_2}$ and $\mathbf{t_2}$ will then allow to re-project the back-projection 3D world point $\mathbf{M_0} = (x, y, z, w)^\intercal$ to the target 2D-image plane $I_2$ as :

$$\begin{pmatrix} u'_2 \\ v'_2 \\ w'_2 \end{pmatrix} = \mathbf{K}_2 \mathbf{R}_2 \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \mathbf{K}_2 \mathbf{t}_2 \tag{5.9}$$

The equations 5.8 and 5.9 can then be combined to express the pixel location of a point $(u_2, v_2$ in $I_2)$ from a corresponding point in $I_1$ at $(u_1, v_1)$. The world coordinate system referential is expressed in the first camera system so there is no rotation nor translation of the first camera: $\mathbf{R_1} = \mathbf{I_3}$ and $\mathbf{t}_1 = \mathbf{0}_3$. Then a point in $I_2$ can be retrieved by:

$$\begin{pmatrix} u'_2 \\ v'_2 \\ w'_2 \end{pmatrix} = \lambda \mathbf{K}_2 \mathbf{R}_2 \mathbf{K}_1^{-1} \begin{pmatrix} u_1 \\ v_1 \\ 1 \end{pmatrix} + \mathbf{K}_2 \mathbf{t}_2 \tag{5.10}$$

The homogeneous coordinates of the point $m_2$ can then be expressed in pixel positions as $(u_2, v_2) = (u'_2/w'_2, v_2/w'_2)$.

### 5.1.3 The Rectified Camera Case

A bench of cameras situated along an horizontal axis (with collinear optical centers) is often used in practice. Identical cameras are used consequently with identical intrinsic camera parameters, and controlled extrinsic parameter differences between cameras. Theoretically, the idea is to restrain the camera movement to a single translation. This is achieved in practice after camera calibration and correction.

Several simplifications will then lead to updated equations:

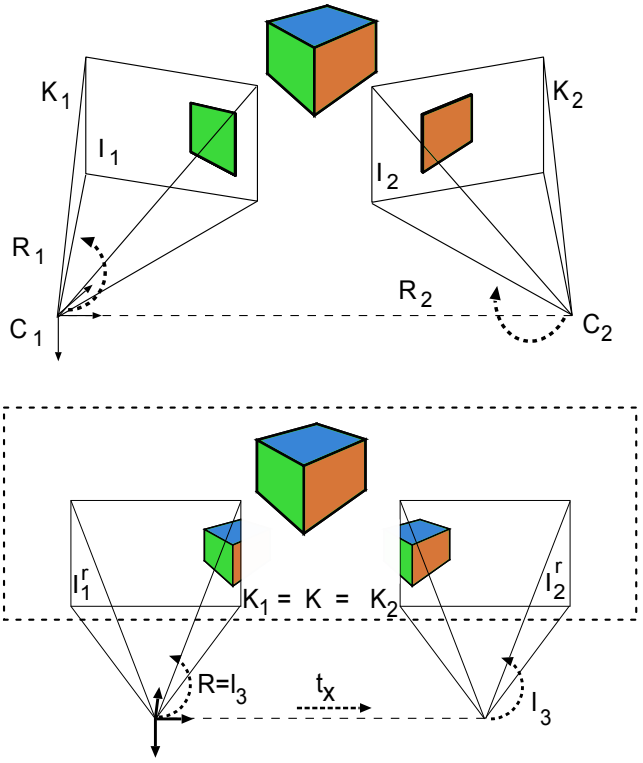



Figure 5.4: Illustration of differences between non-rectified (top) and rectified (bottom) cameras set-up. Note that the camera parameters in rectified configurations share the same $\mathbf{K}$,$\mathbf{R} = \mathbf{I}_3$ and are just shifted by a $t_x$ translation along the x axis.

1. First, because there is no rotation between cameras, the other virtual camera has its rotation matrix identical to the first reference camera, so $\mathbf{R_2} = \mathbf{I}_3$.
2. Second, because the translation is done along a single, horizontal, axis, the translation vector can be simplified as : $\mathbf{t} = (t_x, 0, 0)^\intercal$.
3. Third, because the cameras are supposedly identical, their intrinsic parameters are also identical: $\mathbf{K_1} = \mathbf{K_2} = \mathbf{K}$.

The equation 5.11 can be simplified in such a case to:

$$\begin{pmatrix} u_2' \\ v_2' \\ w_2' \end{pmatrix} = z \begin{pmatrix} u_1 \\ v_1 \\ 1 \end{pmatrix} + \mathbf{K} \begin{pmatrix} t_x \\ 0 \\ 0 \end{pmatrix} \tag{5.11}$$

The scaling factor $\lambda$ then simplifies as $\lambda = z$. Finally, the new horizontal coordinate of $m$ in the targeted rectified view image becomes:

$$u_2 = u_1 + \frac{f \cdot t_x}{z} \tag{5.12}$$

where in non-homogeneous coordinates $(u_2, v_2) = (u_2'/w_2', v_2'/w_2')$, $t_x$ is the horizontal camera translation and $f$ the focal length of both cameras. Because the translation is purely horizontal, there is no other translation: $v_2 = v_1$.

The horizontal displacement of $M_0$ from pixel coordinate $u_1$ to $u_2$ is the disparity $d$. Then we obtain:

$$d = u_2 - u_1 = \frac{f \cdot t_x}{z} \tag{5.13}$$

In conclusion, the disparity is equal to the product of focal length times the translation between cameras divided by the depth of each acquired point in the rectified cameras case. Thus the disparity is inversely proportional to the distance in rectified cameras configuration.

### 5.1.4 Warping Problems: disocclusions, cracks and ghostings

#### Disocclusions

The partial geometry of the scene - acquired by a set of depth cameras or estimated - at different viewpoints does not suffice to fully render a new image onto an additional virtual viewpoint. Indeed, the 3D image warping from a single texture and depth map will make appear previously occluded regions as disoccluded in the virtual view, as illustrated in Figure 5.5(a).

We have seen in chapter 1 that the HVS precisely relies on these occluded regions to infer the distance between objects. Statically, occluded regions from one eye but visible from the other eye are a stereoscopic source of information to infer the relative distance between the object surfaces: the Da Vinci Stereopsis. Dynamically, the accretion and deletion of texture due to a moving observer viewpoint reveal in the same way -but over time- some previously occluded regions. Because the HVS makes usage of these regions, it is all the more important to correctly render them. We will propose in the next section some perceptual rules that will be as much as possible involved by a new method in order to guarantee a plausible reconstruction.

Several techniques exist to fill-in the disoccluded regions:

- At the representation stage: by the choice of a representation that handles and will transmit the occluded region information. We have seen in previous chapter 3 that LDI representations address specifically this problem, at the price of a structure not easily compressible.

- At the rendering stage with a multi-view interpolation: by using a multi-view image-plus-depth representation with a relatively close distance between camera and views, a blending of multiple views synthesized in the common virtual viewpoint enables the retrieval of the disoccluded areas. This technique of interpolation is precisely used by the renderer of 3DVC in the VSRS-based rendering stage [126].

- At the rendering stage by inpainting: a 3D-warping from a single view is realized, so the virtual view is extrapolated. The inpainting is performed to fill-in the disoccluded holes. This is often used in practice when no additional view is available, such as in the View Synthesis Prediction (VSP) at the encoder side in the 3DVC codec, or for extrapolation of an additional view for stereo rendering of movies from a unique monoscopic video source. A solution will be described in the following section.

#### Sampling artifacts: cracks

Due to the regular camera sampling that does not match the exact image pixel grid of the virtual camera, sampling artifacts named “cracks” might appear both on the texture and depth map in the virtual view. Because they appear sparsely on the texture image and have small size, they are often filtered by median or mean windows.


(a) (b)

Figure 5.5: (a) Original view $V_3$ from "Ballet" sequence warped to $V_4$. (b) Resulting artefacts without post-processing: disocclusion, ghosting and cracks artefacts.

#### Ghosting

Finally the ghosting artifacts are due to mixed foreground-background color pixels at object borders in the reference view that project individually either to foreground or background. This leads to annoying ghost effects on virtual views and needs to be removed.

Zitnick et al. [155] proposed to separate the boundary layer detected at depth edges from the main layer. The synthesis consists of a projection of the main layer followed by a projection of the boundary layer where it becomes visible. Müller et al. [93] extended this idea by subdividing the boundary layer into distinct foreground and background boundary layers. After the main layer projection, the foreground boundary layer is added where it is visible in the rendered view and the background layer is used to fill in the remaining holes.

### 5.1.5 Backward Projection

In section 5.1.2, the forward projection was described as the process of a 2D-to-3D back-projection followed by a reverse 3D-to-2D re-projection from 3D-world to the virtual view 2D-image. Every known pixel from the reference view were then warped in the virtual view.

At the opposite, the **backward projection** is a technique where every pixel from the virtual view image are warped back or "backward-projected" onto the reference view in order to interpolate their color [91]. The backward projection has the advantage to avoid the anti-cracks post-processing on texture.

For the virtual view image onto the reference view back-projection, one needs to know the depth value of every pixel in the virtual view. The depth map in the virtual is then first computed with a usual forward projection. Because of the forward projection, the virtual depth map contains lots of artifacts described previously such as cracks and disocclusions. To cope with these artifacts, the virtual depth map is filtered and the holes are simply interpolated. Note that this process is acceptable and efficient on depth maps because they contain smoothly varying surfaces. In contrast, simple texture interpolation would lead to visible artifacts.

Thus the backward projection is composed of three distinct steps:

1. Forward projection of the reference view depth map onto the virtual new view to compute a virtual view depth map.

2. Virtual depth map filtering: because the forward projection generates sampling artifacts -or cracks-, those are filled in with the usual median interpolation technique.

3. Backward projection of the virtual view texture image: each texture pixel is warped back into the reference viewpoint thanks to their previously filled-in virtual depth map. This results in floating coordinates of back projected pixels in the reference view image plan, non-aligned on the reference view coordinates. The texture values of these floating points are thus bi-linearly interpolated and assigned onto the virtual view texture image, which avoid the error-prone anti-cracks post-processing.

## 5.2 Introduction to Inpainting

The inpainting consists of correcting or modifying an image in an **undetectable** way [8]. This notion of undetectability or invisibility in the perceptual sense is important because we will later refer to perceptually correct view synthesis. Its applications are various from the restoration of damaged paintings and photographs to the recent digital techniques of object removal, image coding, error concealment etc. For example it can be used for removing scratches or stains in an image or for removing overlaid text and logos etc. It is also used in image coding to limit the transmission for instance to a dictionary of the image that could be latter reconstructed at the decoder side.

The image and video inpainting methods are often classified as pixel-based inpaintings and template-based inpaintings. These techniques will be described before more recent hybrid methods are detailed.

### 5.2.1 Pixel-based Inpainting

The pixel-based or diffusion-based approach of Bertalmio et al. [8] is motivated by the intensive work on Partial Differential Equations (PDEs) in image processing and computer vision. The inpainting relies on isotropic or anisotropic diffusion to propagate the boundary pixels first along the **isophote** -level line of equal pixel intensity- direction. The a priori consists of assuming that the possible structure is delimited by edges that must be reconnected first before hypothetical smooth regions on the inside are filled in.

This approach propagates the pixel intensity values by diffusion and then effectively fills-in supposed uniform regions by smoothing, but fails to reconstruct textured regions.

### 5.2.2 Template-based Inpainting

Soon after the Bertalmio proposal, based on the works of Igehy and Pereira [55] and Efros [36] on texture synthesis, Harrison [46] highlighted the importance of ordering when selecting texture patch(es) from the known neighbourhood instead of single pixels. The exemplar-based or template-based techniques emerged from this idea of copying from the neighbourhood and pasting into the hole some elements of textures or patches according to a certain priority.

Criminisi et al. [24] combined the structure ordering into a textural synthesis: the supposed textured missing regions are inpainted along the isophote direction according to its strength. This method will be described in detail because it constitutes the base of our proposed method.

**Criminisi algorithm** Criminisi et al. supposed that exemplar-based inpainting contains the essential process to replicate both texture and structure, and confirmed that the structure propagation depends on the order in which the patches are filled-in.

For a given input image $\mathcal{I}$, the source region $\Phi$ is defined as the entire image minus the hole region to fill in: $\Phi = \mathcal{I} - \Omega$. A priority $P(p)$ is calculated for all the patches $\Psi_p$ centered at points $p$ along the hole border also called fill front ($p \in \delta\Omega$). This is illustrated in Figure 5.6. The priority is defined as the product of a confidence term $C(p)$ and a data term $D(p)$:

$$P(p) = C(p)D(p) \quad \text{with} \quad C(p) = \frac{\sum_{q \in \Psi_p \cap (\mathcal{I}-\Omega)} C(q)}{|\Psi_p|}, D(p) = \frac{|\nabla I_p^{\perp} \cdot n_p|}{\alpha} \tag{5.14}$$

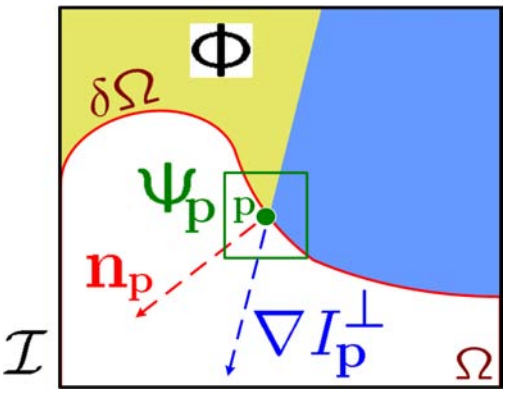



Figure 5.6: Notation diagram. In the image $\mathcal{I}$, given the patch $\Psi_p$, $n_p$ is the normal to the contour $\delta\Omega$ of the targeted region $\Omega$ and $\nabla I_p^{\perp}$ is the isophote (gradient and intensity) at point p. From [24].

with $|\Psi_p|$ the area of $\Psi_p$, $\alpha$ a normalization factor, $n_p$ a unit vector orthogonal to the front $\delta\Omega$ in $p$. $C(p)$ is a measure of the amount of reliable information surrounding the pixel $p$ i.e. the reliability of the current patch. $D(p)$ introduces higher priorities to patch having strong isophotes hitting the front $\delta\Omega$ at each iteration. Then this term boosts the patch having isophote flowing into, thus it favours linear structures to be synthesized and propagated first into the hole region.

Once all the priorities on the fill front have been computed, the patch $\Psi_p$ is selected and the most similar patch $\Psi_{\hat{q}}$ among candidates $\Psi_q$ is derived by:

$$\Psi_{\hat{q}} = \arg\min_{\Psi_q \in \Phi} d(\Psi_{\hat{p}}, \Psi_q) \tag{5.15}$$

with the distance $d$ defined as the Sum of Square Differences (SSD).

Once the best source examplar $\Psi_{\hat{q}}$ has been found, the value of each remaining pixel into the patch to be filled $p' \in \Psi_{\hat{p}} \cap \Omega$ is copied from its corresponding pixel in $\Psi_{\hat{q}}$ and the confidence term $C(p)$ is updated as follows:

$$C(p) = C(\hat{p}), \quad \forall p \in \Psi_{\hat{p}} \cap \Omega \tag{5.16}$$

### 5.2.3 Hybrid methods

Bertalmio et al. proposed in 2003 [9] a new method for simultaneous filling-in of texture and structure, not based on an iterated priority term as in [24] to respect the structure, but instead relies on a separated reconstruction of the structure and texture. The regions of bounded variation are reconstructed by a diffusion based technique defined in [8]: the propagation by PDE is achieved along the isophote directions. The textural regions are reconstructed separately by a common texture synthesis such as the one proposed in [36].

More recently, Liu et al. [75] proposed a complex but more powerful edge-based inpainting method applied to image compression. Edge extraction and region removal are done accordingly at the encoder side. At decoder side, the exemplar and edge information are first decoded. The non-exemplar regions are classified into structures and textures according to their distances to the edge in the encoder. Structures are generally propagated first by a pixel-wise structure propagation, followed by a common texture synthesis. Indeed, a confidence map similar to that presented above by [24] enables the guidance of the order of structure propagation as well as texture synthesis.

## 5.3 Depth-based View Synthesis by Extrapolation

The presented inpainting technique relies on the search and propagation of the most similar pixel or template based on its colorimetric similarity. In the context of hole-filling for view synthesis from a single view, additional constraints could be added to help to a correct rendering. The past methods and their principles will be first presented before we formalize three statements in order to form a perceptually correct view synthesis. The proposed method will then be detailed.

### 5.3.1 Past methods

**Oh et al.** [97] based their method on depth thresholds and boundary region inversion. The foreground boundaries are replaced by the background one located on the opposite side of the hole. Despite the use of two image projections, their algorithm relies on an assumption of connectivity between disoccluded and foreground regions, which may not be verified for high camera baseline configurations. Indeed, over a certain angle and depth, the foreground object does not border the disoccluded part anymore.

**Daribo et al.** [29] proposed an extension to the Criminisi's [24] algorithm by including the depth in a regularization term for the priority and patch distance calculation. First the patch priority is updated with a level regularity term $L(p)$ defined as the inverse variance of the depth of the current patch $Z_p$ as follows:

$$P(p) = C(p) \cdot D(p) \cdot L(p) \qquad \text{with } L(p) = \frac{|Z_p|}{|Z_p| + \sum_{r \in \Psi_p \cap (\mathcal{I} - \Omega)} (Z_r - \overline{Z_p})^2}$$

with $|Z_p|$ and $\overline{Z}_p$ the area and the mean value of the depth values $Z_p$ respectively. More priority is then given to the patch at the same level, but this doesn't prevent the patch propagation from the foreground. The patch matching was also updated as follows:

$$\Psi_{\hat{q}} = \arg \min_{\Psi_q \in \Phi} d(\Psi_{\hat{p}}, \Psi_q) \; + \; \beta \cdot d(Z_{\hat{p}}, Z_q) \tag{5.17}$$

with $\beta$ the factor of importance of depth distance minimization and $d$ the Sum of Square Difference (SSD) distance.

In order to perform the above inpainting operations on the warped view texture containing holes, a knowledge of the corresponding holes $\Omega$ in the warped depth is assumed. In practice however, the holes appear both on the texture and depth of the warped view. The Daribo's depth map hole filling proposal consists of using pixel-based Bertalmio [8] inpainting. This approach provides convincing results but still results in visible artefacts (see next section for comparison).

In the rest of this chapter the proposed method and comparison of existing ones will be provided either with the original depth map of the targeted virtual view, or with the virtual depth map. The depth map inpainting is out of the scope of this thesis, despite that the interpolation method presented in chapter 4 could be applied for depth map hole filling.

### 5.3.2 Formulation of a Perceptually Correct View Synthesis Based on Inpainting

In the light of the first human vision processes (see a description in Appendix A) and recent proposals on template-based inpainting we can tackle the issue of disocclusion filling by inpainting techniques that aim to give perceptually correct rendering of virtual views.

Nakayama and Shimojo [95] specified what they called the "Da Vinci Stereopsis" source of depth information where the "amount" of occluded regions in one eye appearing disoccluded in the other informs about the relative depth between the object surfaces (see section 1.1.3). Importantly, they state that the monocularly viewed region is always part of the distant surface (relatively to the object edge, but not always the distant surface in the scene) i.e. the local background. Also, if this monocularly viewed region is present in the right eye/image/viewpoint it will always lie to the right of the occluding edge of the foreground, and inversely for a left viewpoint.

Then we can state a first condition necessary for a correct perceptual rendering: the disoccluded region must always render an artificial region belonging to the connected background surface.

Regarding the four stages of visual processing proposed by Marr [83], we can also state that the region-to-fill-in must necessarily involve the image, surface, object and categorical properties of the background. This statement is important, it means respectively that:

- the edges, luminance and colour of the filled-in region must match the background.
- the inner textural elements (or texels) of the region to fill in must be coherent: they must not only reproduce but propagate the connected structural and textural elements of the background according to their orientation and convergence.
- the previous conditions do not suffice. The reconstructed region must be a plausible reproduction of what would be in the place of the background object disoccluded region: it must form a coherent background object.
- the object must be coherent in itself and also ecologically valid within its neighbourhood so that the human visual system perceives the inpainted reconstructed object as valid in its environment.

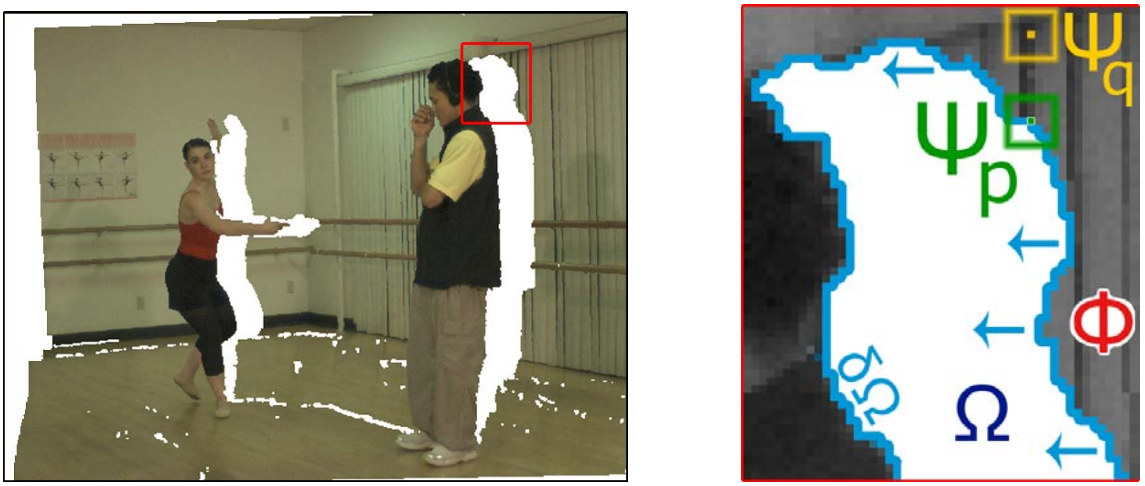
The last conditions are hard to achieve in the context of hole-filling view synthesis without a complex algorithm of scene and object recognition. However, we can add the following statement: the disoccluded region must be filled in by propagating the lines, edges, i.e. structure, but also the textures from and along the neighbouring background object so that it appears as its extension.

Regarding the perceptual figure/ground organization, Rubin [113] describes the subjective distinctions between figure and ground with a phenomenological analysis. The background appears distant to the viewer, without its shape defined by the foreground contour but instead extending behind it and appears not "thinglike". Then we can add a third statement: the disoccluded regions must appear farther to the viewer than the foreground regions and must appear as extending behind the contour (the same idea of propagation of the background structure and texture).

The fact that people tend to not perceive the background as a thing is true when presenting in experiments a uniform background, but this is more hypothetical in a practical situation with a highly textured background surface.

## 5.4 Proposed algorithm

The motivation to use a Criminisi-based algorithm resides in its capacity to organize the filling process in a deterministic way. As seen previously and illustrated in Figure 5.7, this technique propagates similar texture elements $\Psi_{\hat{q}}$ to complete patches $\Psi_p$ along the




(a) (b)

Figure 5.7: Illustration of inpainting principle. (a) a warped view, (b) a zoom on the disoccluded area behind the person on the right with the different elements overlaid.

structure directions, namely the isophotes. This technique then employs the condition of structure and texture propagation defined in last section.

In the context of view synthesis, some of the conditions defined as necessary to a correct perceptual rendering can be directly added as constraints to perform the inpainting. Because the projection in one view will be along the horizontal direction, for a toward-right camera movement the disoccluded parts will appear on the right of their previously occluding foreground (Figure 5.7a), and oppositely for a toward-left camera movement. Whatever the camera's movement, we have seen that these disoccluded areas should always be filled in with pixels from the background rather than the foreground. Based on this a priori knowledge, we propose a depth-based image completion method for view synthesis based on robust structure propagation.

### 5.4.1 Tensor-based priority

First, the data term $D(p)$ of the inpainting method proposed by [24] involving the color structure gradient is replaced with a more robust structure tensor. While Criminisi's method favours the patches having isophotes that are "hitting the front" i.e. isophote orthogonal to the orientation of the hole border, we propose to favor the propagation of patches with the strongest isophotes. The approach is different because we aim at propagating the strongest structure first, whatever its position relative to the border. The tensor term is inspired by partial differential equation (PDE) regularization methods on multivalued images and provides a more coherent local vector orientation [131]. The Di Zenzo matrix [34] is given by:

$$J = \sum_{l=R,G,B} \nabla I_l \nabla I_l^T = \sum_{l=R,G,B} \begin{pmatrix} \frac{\partial I_l}{\partial x}^2 & \frac{\partial I_l}{\partial x}\frac{\partial I_l}{\partial y} \\ \frac{\partial I_l}{\partial x}\frac{\partial I_l}{\partial y} & \frac{\partial I_l}{\partial y}^2 \end{pmatrix}$$

with $\nabla I_l$ the local spatial gradient over a 3x3 window. This tensor can also be smoothed with a Gaussian kernel $G_\sigma$ to be more robust to outliers, without suffering from cancellation effects. We call it $J_\sigma = J * G_\sigma$. Finally, the local vector orientation is computed from the

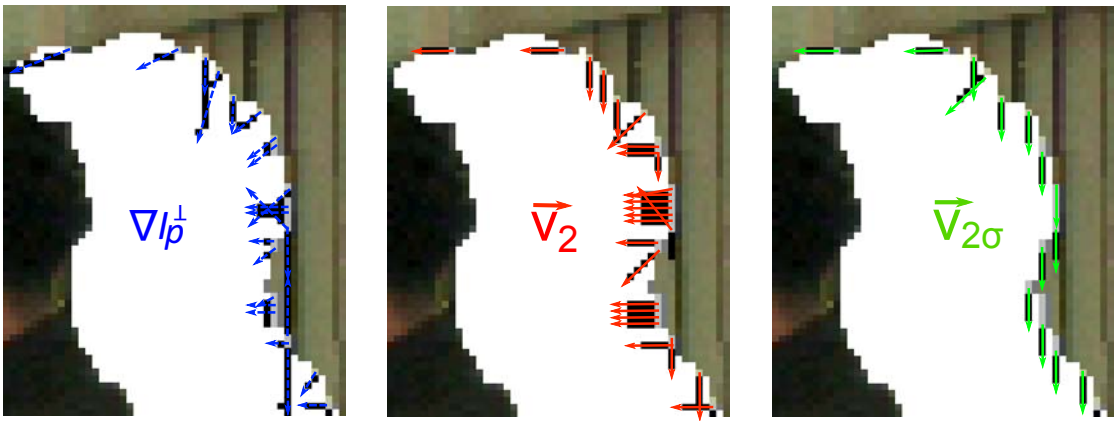



Figure 5.8: Illustration of the different methods to determine the isophote. For clarity, the black vectors on a pixel grid obtained from the software are overlaid with color vectors. Left: Isophote direction $\nabla I_p^{\intercal}$ as computed by [24], as the perpendicular vector to the local gradient. Middle: eigenvector $v_2$ computed without pre-smoothing of the tensor. Right: $v_2$ computed with pre-smoothing of the tensor.

structure tensor $J_\sigma$. Its eigenvalues $\lambda_{1,2}$ reflect the amount of structure variation, while its eigenvectors $v_{1,2}$ define an oriented orthogonal basis. Of particular interest is $v_2$ the preferred local orientation and its "force" $\lambda_2$. The reliability of this eigenvector $v_2$ relative to the gradient to predict structure orientation is illustrated in Figure 5.8. The eigenvector from a non-smooth tensor is also illustrated. The direction and orientation of $v_2$ appear to be more reliable than the term $\nabla I_p^{\intercal}$ for calculating the isophote intensity that drives structure propagation. In this example, the structure of the curtain at the background texture will be propagated first because the isophote well matches the direction of the structure to begin propagation. The result leads to a convincing rendering that respects the structure connectivity principles defined from vision psychology principles in 5.3.2.

More than the eigenvalue $\lambda_2$, it is the difference between $\lambda_2$ and $\lambda_1$ that reflects the intensity of the isophote. Based on the coherence norm proposed in [142], the data term $D(p)$ is then defined as:

$$D(p) = \left[\alpha + (1-\alpha) exp\left(\frac{-C}{(\lambda_1 - \lambda_2)^2}\right)\right] \cdot ||v_2||^2$$

with $C$ a constant positive value and $\alpha \in [0,1]$ ($C$=8 and $\alpha$=0.01). Flat regions (when $\lambda_1 \approx \lambda_2$) do not favour any direction, it is isotropic; while with strong edges ($\lambda_1 >> \lambda_2$) the propagation begins along the isophote.

### 5.4.2 Depth-aided and direction-aided priority

The priority computation has been further improved by exploiting the depth information, first by defining a 3D tensor product, secondly by constraining the side from where to start inpainting.

#### 3D tensor

The 3D tensor allows the diffusion of structure not only along color but also along depth information. It is critical to jointly favour color structure as well as geometric structure. The depth-aided structure tensor is extended with the depth map taken as an additional


(a) Zoom from 5.7 with $\delta\Omega$ (b) C(p) (c) D(p) (d) P'(p)

Figure 5.9: Illustration of the different terms used for priority calculation P'(p) along the hole border (in blue in (a)). The intensity of each map in reversed: a darker pixel value means an higher intensity.

image component $Z$:

$$J = \sum_{l=R,G,B,Z} \nabla I_l \nabla I_l^T$$

**One side only priority**

The second improvement calculates the traditional priority term along the contour in only one direction. Intuitively, for a camera moving to the right, the disocclusion holes will appear to the right of foreground objects, while out-of-field area will be on the left of the former left border (in white on the left of Figure 5.7a). We then want to prevent structure propagation from foreground by supporting the directional background propagation, as illustrated in Figure 5.7b with the blue arrows.

For a projection to the right of the reference view, the patch priority is calculated along the hole border but the top, bottom and left border priorities are set to zero. Globally, the priority term is reset such as:

$$P'(p) = \begin{cases} C(p) \cdot D(p) & \text{if } p \in \delta\Omega_{Dir-projection} \\ 0 & \text{if } p \in \delta\Omega_{Top}, \delta\Omega_{Bottom}, \delta\Omega_{Opp-dir-projection} \end{cases}$$

with $\delta\Omega_{Dir-projection}$, the border pixels on the hole side of direction of projected view, $\delta\Omega_{Opp-dir-projection}$, the opposite side border pixels, $\delta\Omega_{Top}$ and $\delta\Omega_{Bottom}$ the top and bottom side border pixels respectively.

Then for disoccluded areas, the border of the side opposed to direction of view projection is possibly connected to the foreground. It will be filled at the very end of the process. For out-of-field areas, even if left (or right) borders are unknown, we will ensure to begin from the right (or left) border rather than possible top and bottom ones.

These two proposals -extension of the tensor term and non-linear direction inhibition of foreground connected contours- have been included in the prioritization step and are illustrated in Figure 5.9. For a given iteration, the confidence values $C(p)$ along the hole border appear stronger (darker in 5.9(b)) in the concavity of the border, because more pixels are known: the confidence is higher. The data term $D(p)$ 5.9(c) is high everywhere in the neighbourhood of structure, while weak where the local patch is uniform. The resulting multiplication followed by directional inhibition is the priority term $P'(p)$ in 5.9(d). The top, bottom, and left side of $\delta\Omega$ are effectively reset to a null priority.

### 5.4.3 Patch matching

Once we precisely know from where to start in a given projected image, it is important to favour the best matching candidates in the background only. Nevertheless, starting from a non-foreground patch does not prevent it from choosing a candidate among the foreground, whatever the distance metric used. Thus, it is crucial to restrict the search to the same depth level in a local window: the background. We simply favour candidates in the same depth range by integrating the depth information in the commonly used similarity metric $L_2$-norm Square Sum of Differences (SSD):

$$\Psi_{\hat{q}} = \underset{\Psi_q \in \Phi}{\arg\min}\, d(\Psi_{\hat{p}}, \Psi_q) \quad \text{with } d = \sum_{p,q \in \Psi_{p,q} \cap \Phi} \alpha_l \left\| \Psi_{\hat{p}} - \Psi_q \right\|^2$$

with the same notation as in 5.15. The depth channel is chosen to be as important as the color one ($l \in R, G, B, Z$ with $\alpha_{R,G,B} = 1$ and $\alpha_Z = 3$). Then it will not prevent the search in foreground patches, but will seriously penalize and unrank the ones having a depth difference greater than i.e in front of the background target patch.

Once the best matching candidates have been found, it is important to combine several of them into a single element to fill in the hole. Directly copying one candidate pixel would lead to a trivial solution that might be not satisfying.

A K-Nearest-Neighbour (KNN) search and combination algorithm is used to compute a robust final patch to fill in iteratively the holes. The approach proposed by [144] is retained: the patch candidates are not equally reliable and then should not contribute equally to a final combined patch solution. The distance $d$ is translated to a similarity measure $s$ to favour the most similar patch:

$$s(\Psi_{\hat{p}}, \Psi_q) = e^{-\frac{d(\Psi_{\hat{p}}, \Psi_q)}{2\sigma^2}}$$

with $\sigma$ the factor that will control the smoothness of the patch and $d$ the usual SSD similarity measure. Wexler et al. [144] proposed to choose it as the 75-percentile in the current locations. We calculate it as the squared mean distance of the first and second closest patch, $\sigma = \left[{}^{(d^1+d^2)}/_2\right]^2$. then the combined patch $\Psi_c$ is a linear combination of the $K$ best candidates such as:

$$\forall q \in K \qquad \Psi_c = \frac{\sum_q \Psi_q \; s(\Psi_{\hat{p}}, \Psi_q)}{\sum_q s(\Psi_{\hat{p}}, \Psi_q)}$$

The number of candidates to combine are set in practice to $K = 5$.

**Iteration** After having filled in a patch, Criminisi et al. [24] updated the priority. For each filling iteration, the priority P(p) was then previously recalculated all along the border of the holes and the patch with the pixel $p$ of highest priority was selected to be filled-in. This method is time-consuming and might lead to favour too much the propagation of the structure.

Our proposal consists in starting the filling from the $N\%$ of pixels along the border with the highest priority, what is called the Percentile Priority-based Concentric Filling (PPCF). Once the patch of the $N\%$ highest priority pixels have been filled in, the priority along the reduced hole border is recalculated. Practically, PPCF allows the maintenance of the propagation of image structure followed by image texture, but to a lesser extent than what is proposed by [24]. The first percent of border pixels of highest priority is used to start the filling in practice.

## 5.5 Results

### 5.5.1 View Synthesis for Extrapolation: the 3DTV case

In a MVD context with reduced baseline between cameras, we can consider that the extrapolation from a single view located at one or two camera steps from the original camera could provide 3DTV functionality: a subset of views can be rendered within the limited camera range or even one camera step farther. For instance, for a set of acquired views $V_1$ to $V_8$, the view in the middle of $V_1$ and $V_2$, i.e. $V_{1.5}$ could be easily rendered, but also new extreme virtual views such as $V_0$ and $V_9$. Beyond this distance, both the warping and inpainting methods for an extrapolated view synthesis become much more challenging and ambitious.

Figure 5.10 illustrates the results obtained with the proposed method, compared with methods from the literature [24], [29], when rendering views located at varying distances from the reference viewpoint: one $V_{5\to4}$ at one camera step and one $V_{5\to2}$ at three camera steps from the transmitted reference camera $V_5$.

The three versions take in input the same warped view texture and the same original view depth information, except for the approach in [24] relying on texture only.

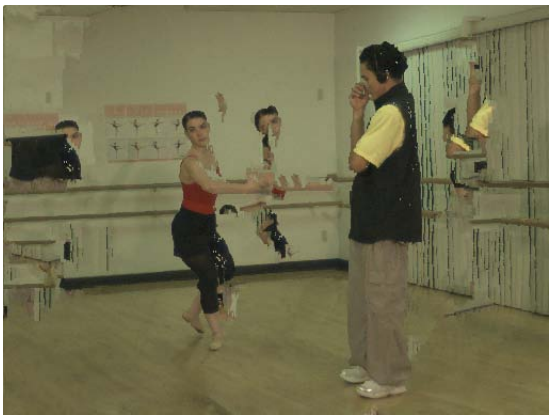
Our method not only preserves the contour of foreground persons, but also successfully reconstructs the structure of missing elements of the disoccluded area (i.e. edges of the curtains and bars behind the person on the right, background wall behind the left one).

Thanks to our combination term, we can even extend the synthesis to very distant views, without suffering of aliasing effects. As illustrated, the view 5 is projected to view 2 ($V_{5\to2}$) and the out-of-field blank areas occupying one quarter width of the warped image are reconstructed. The counterpart of the patch combination is the blurring effect appearing on the bottom part of this area. By taking different numbers of patches for combination, it is possible to limit this effect.

(a) $V_{5\to4}$ after warping & background antighosting

(b) $V_{5\to2}$ after warping & background antighosting

(c) $V_{5\to4}$ inpainted with Criminisi's method



(d) $V_{5\to2}$ inpainted with Criminisi's method

(e) $V_{5\to4}$ inpainted with Daribo's method

(f) $V_{5\to2}$ inpainted with Daribo's method

(g) $V_{5\to4}$ inpainted with our method

(h) $V_{5\to2}$ inpainted with our method

Figure 5.10: Illustration of different methods of inpainting. Our approach relying on 3D tensor and directional prioritization shows efficient filling.

### 5.5.2 Importance of the prior depth reconstruction

Following our contribution, Jantet et al. [59] recently proposed improving the Daribo's method by updating the patch matching with the actual full depth knowledge of the patch to fill in with color. The depth map is first projected and the disoccluded pixels are copied only from the background. These depth pixels corresponding to the hole in the texture image are important because they help to find a corresponding similar patch both in texture and in depth in the neighbourhood.

The equation 5.17 to search for the most similar patch was then updated such as:

$$\Psi_{\hat{q}} = \arg\min_{\Psi_q \in \Phi} d(\Psi_{\hat{p}}, \Psi_q) \; + \; \beta \cdot d_{\mathcal{I}}(Z_{\hat{p}}, Z_q)$$

where the $d$ is the SSD similarity metric performed on the known texture pixel $\Phi$ and on the fully known depth pixel of the patch i.e. $\mathcal{I} = \Psi_q \cup \Phi$.

The impact of a correct depth reconstruction and its influence when taking into account in the Daribo's modified method, while not updating our proposal -without full-depth knowledge- is illustrated in Figure 5.11. This illustrates that an update of the patch matching method significantly improves the original Daribo method and leads to similar results to ours. In perspective it would then be very promising to update the patch matching term in our method with the full depth knowledge per patch.

### 5.5.3 Importance of the patch and window size on visual quality

In the literature, recent proposals have been drawn to improve the general inpainting technique on still images. Considering our proposed method applied to view rendering, different aspects of the inpainting method can lead to significant improvements.

We proposed in this section to study the impact of the size of search windows and of the size of patch on the quality of the inpainted synthesized views. The issue of optimal patch and search window size is important and can lead to significant results. Recently in [71] a search for the best candidate patches among different patch sizes has led to qualitative improvements. We propose here to study the influence of these two inpainting parameters on the quality of the synthesized views.

#### Limits of Objective Qualitative Results

In the last chapter, subjective quality evaluation tests were conducted to measure the influence of difference depth map compression on quality of the synthesized views through a bullet-time video.

Nevertheless, tests were not conducted to assess the quality of different inpainting techniques on view synthesis. However, assessing the visual quality of a given view synthesis method in another way than with subjective quality experiments is attractive but risky. First because the state-of-the-art objective quality metrics are precisely dedicated to evaluate the distortion and quality of 2D images or videos but not of synthesized images or on videos. Second our inpainting method for view rendering does not handle the temporal redundancy and the motion compensation of inpainted hole necessary to smooth video with imperceptible filling. This will be discussed in a later section.

Keeping in mind the limitation of the objective metrics applied to rendered view quality assessment, we will first illustrate some parts of hole filling for different patch sizes and window sizes, before to drawing objective quality results.

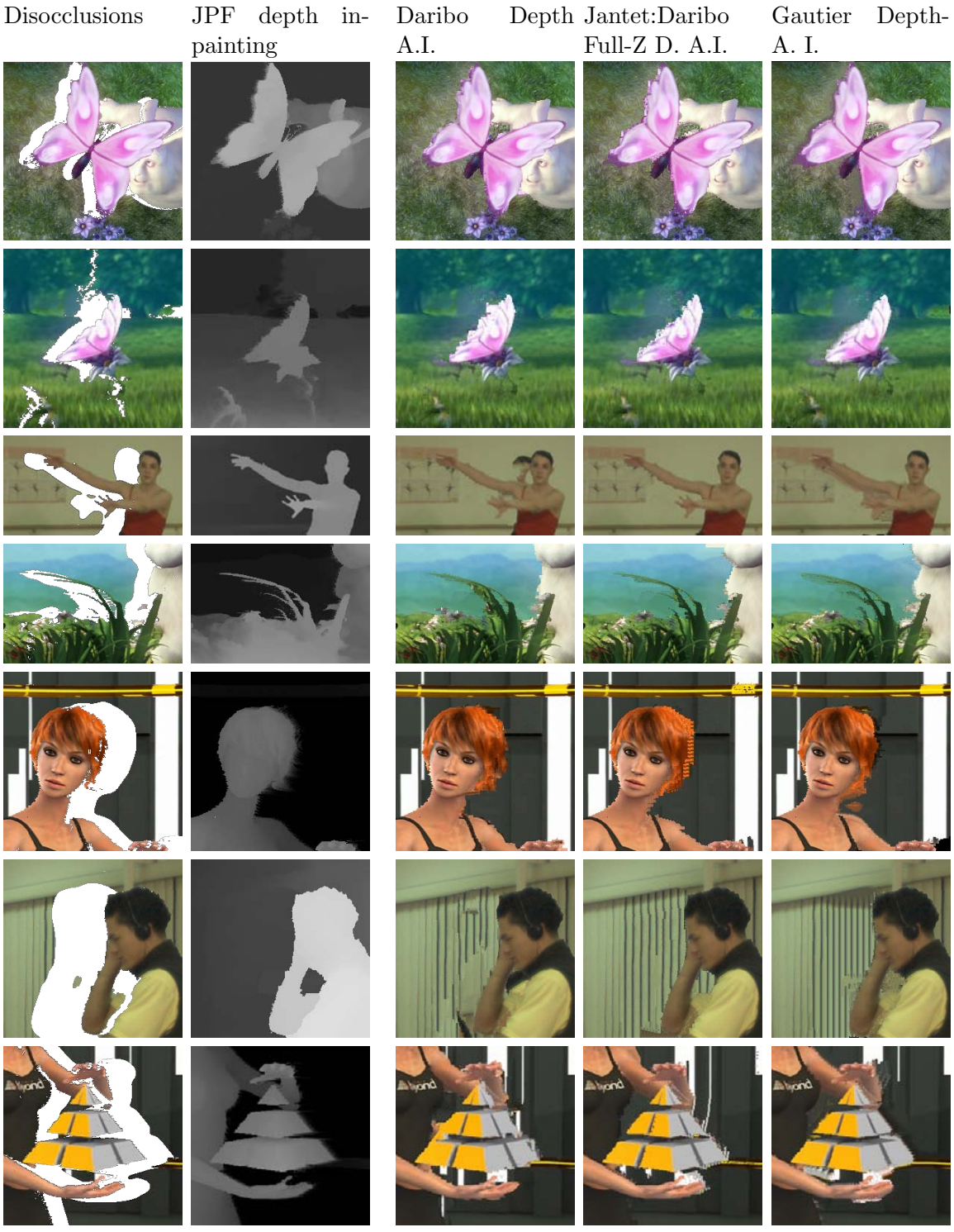



Figure 5.11: The first column shows a synthesized view with disocclusions. Column 2 illustrates the synthesized depth maps, obtained with the depth inpainting algorithm by backward projection proposed by Jantet et al. [59]. Columns 3, 4 and 5 show the results of three template-based inpaintings based on the original texture shown in column 1 and guided by the depth map presented in column 2. From [59].

### Choice of Objective Quality Metrics

Four commonly used state-of-the-art 2D quality metrics are selected.

With the PSNR, the Structure Similarity Index (SSIM) is nowadays widely used as an alternative. Based on the assumption that human visual perception is highly adapted for extracting structural information from a scene, Wang et al. [139] proposed predicting the quality from the degradation of structural information. Three components of luminance, contrast and structure comparisons are computed within a local 8x8 window, weighted and multiplied in a local SSIM score and averaged in a global image mean SSIM score. The Multi-Scale SSIM is an improvement proposed by the authors of SSIM [140]. This metric is selected because it considers image details at different resolutions better accounting for the perceptual capability of the observer's visual system.

The Visual Information Fidelity index [121] is another Full Reference (FR) metric proposed to assess the quality of rendered views. It is derived from a statistical model for natural scenes, a model for image distortions and a HVS model in an information-theoretic setting. It is supposed to perform better than the SSIM in the conditions of test in [121].

Finally, a fourth metric is used, the more recent Visual Signal to Noise Ratio (VSNR). First, wavelet-based models of visual masking are summed to determine the distortions. If the distortions are below the threshold of detection, the image is of perfect visual fidelity. If not, a second-stage based on low-level property of perceived contrast and mid-level property of global precedence is applied. Because VSNR operates on physical luminances and visual angle, it is supposed to better accommodate to different viewing conditions.

### Results

The influence of patch and window sizes is illustrated in Figure 5.12. The best search configuration appears to be with the smaller size of patch and the larger size of window where to search for the best patch. The same observation can be drawn with the other tested metrics. This conclusion is true within the limit of the sizes observed, the limits induced by the depth map distortions and the limits of the 2D objective metrics used to assess the quality of rendered views.

It can also be observed that over a search window of 30 pixels of radius, the gains on objective visual quality are limited. Also the step around 8 or 9 pixels of radius indicates that it is important to keep patch size within the range of 2 to 9 pixels of radius, but that a smaller patch only slightly increases the objective quality. It is important to keep in mind that increasing the size of search window and decreasing the size of patch both increase the complexity, so the search for an optimal or adapted tradeoff between patch and window size is important.

Tests have been conducted on the other sequences of the MPEG 3DV video corpus and give surprisingly similar observations. The slope of visual quality is slightly or abruptly decreasing with the increasing size of patch for all the videos. The slope decreases more strongly from the radius of patch larger than 8 to 9 pixels especially for lower size of search windows.

When considering subjectively the visual patches, the same global conclusion can be drawn (see the Figure 5.13 illustrating a zoom on an inpainted background from the "Balloons" sequence for different sizes of patch and different sizes of search window). However some amendments may be necessary when looking closer at the most plausible inpainted holes.

The correct reconstruction of the background appears to be performed at best for a 3x3 patch size, as predicted by the PSNR and VSNR quality score within a very limited scale.

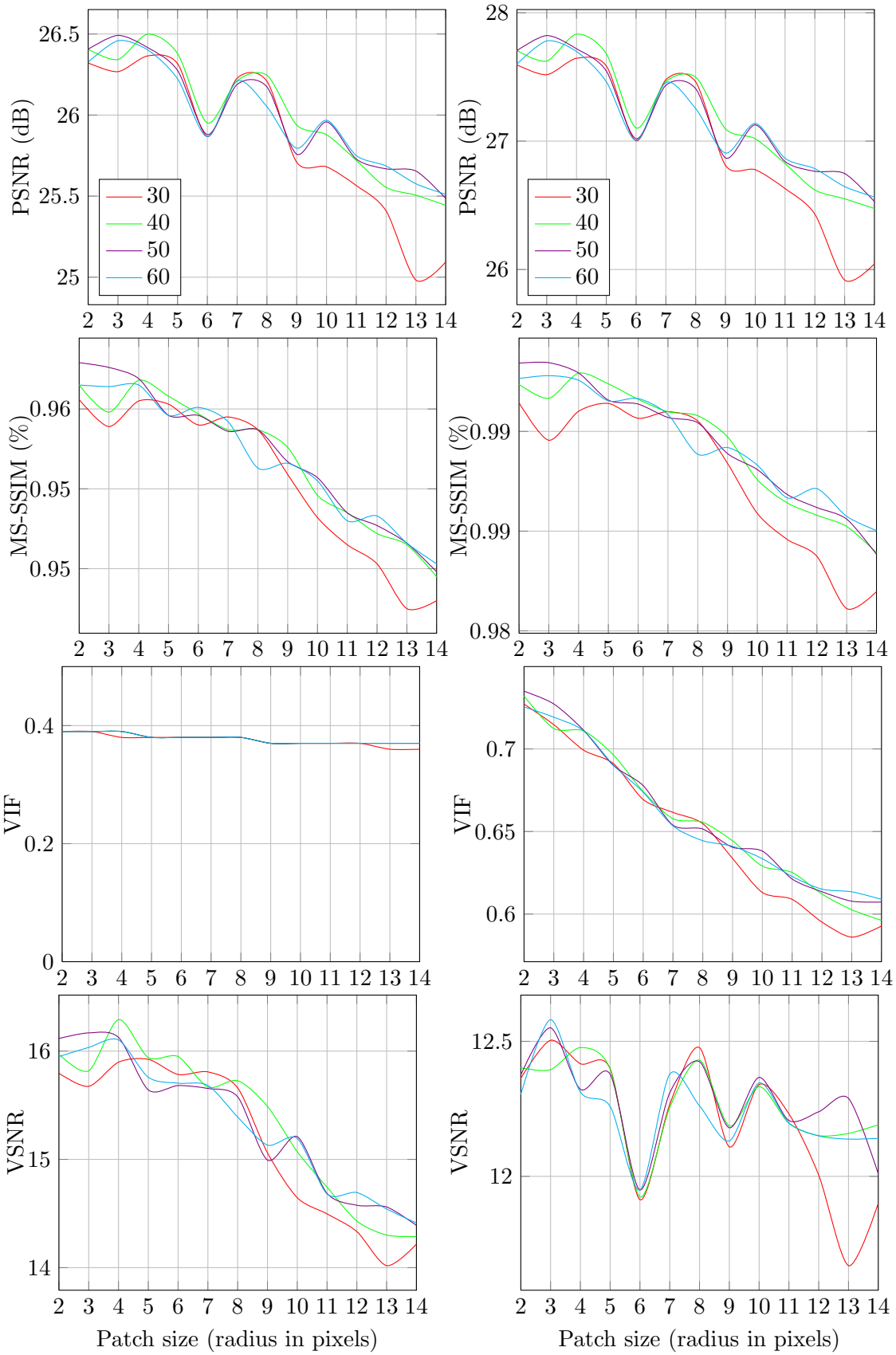



Figure 5.12: Influence of varying patch size and window search size (radius in pixels) on the objective visual quality scores. Comparison between full synthesized images (left) and between occluded part of the synthesized images (right) the rest being reset.

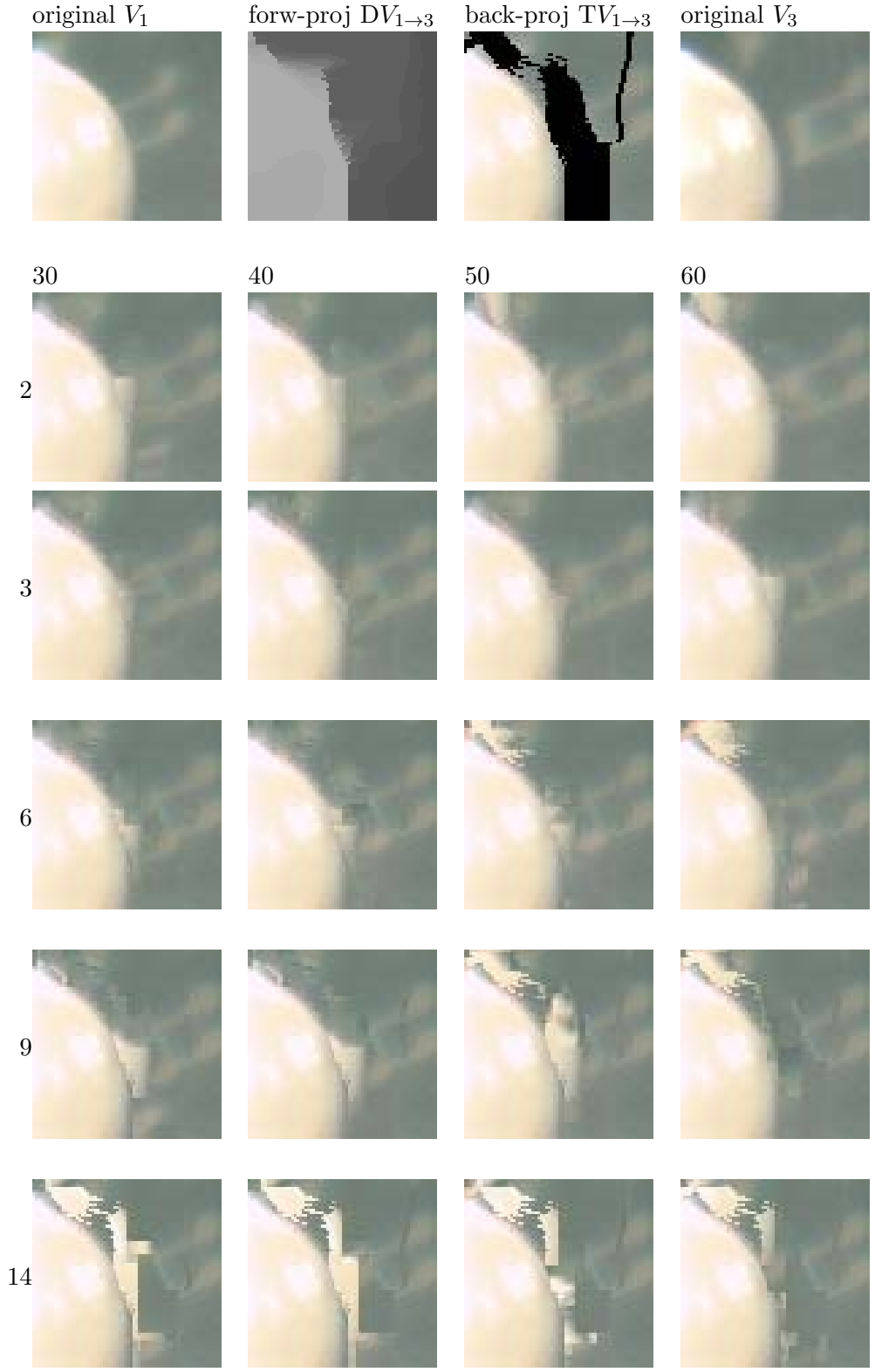



Figure 5.13: Visual impact of an inpainting with varying patch sizes (vertical position) and window sizes (horizontal position) on a hole of "Balloons" sequence due to projection (zoom window of 70x70 pixels). A zoom on the original view $V_{1\rightarrow}$, the forward projected depth map after filling, the resulting backward projected texture to inpaint and the original result to target are illustrated on the top row.

It is hard to conclude on the best size of search window for visually correct inpainting, but the larger (60 pixels radius window) is clearly not the best, as also predicted by the PSNR, SSIM and VSNR.

The proposed depth-aided inpainting method was initially developed without considering the depth map projection and its inpainting: the original depth map in the projected view was used on the basis of a separate geometry representation transmission and synthesis. Here the results were obtained in a MVD context with projected depth maps. The influence of the preliminary depth map reconstruction on the reconstruction artefacts clearly appears: the upper part of the left balloon object has depth "leakages" on the background that are propagated on the texture as visual artefacts. The depth map reconstruction is in fact very sensitive to the foreground antighosting step that is supposed to prevent these leakages. The search for an optimal antighosting parameter that actually fits to the scene geometry content is thus promising.

Without considering the depth artefacts, the inpainting methods give visually convincing results on the rest of the image and for small size of patch and high size of search window. The depth reconstruction appears to be a preliminary but highly important stage to the proposed inpainting technique.

### 5.5.4 Importance of the time and limit of an image inpainting method

The goal of this section was to assess the visual quality per frame of a synthesized view over a whole video to ensure the limited range of distortion over time. However, because the inpainting method is not extended to the temporal dimension, it is obvious that flickering will appear and distort the perceived visual quality of the video, but the video inpainting is out of the scope of this thesis.

The rendered view is compared to the original view captured by the camera over each whole frame. The four objective quality metrics used to assess the impact of patch and window size are kept. The range of measured quality over time for the "Balloons" MPEG sequence is illustrated in Figure 5.14.

The theoretical PSNR range is not bounded, but it is often accepted that a decrease of few decibels (<5dB) would be acceptable. In practice the variation of the PSNR over time appears to be limited. The variations of MS-SSIM, VIF and VSNR of synthesized video frame also appear limited relatively to their corresponding source video frames. The same tests have been conducted on different kind of video sequences. Similar ranges of variation are observed on the other sequences.

The results confirm that the impact of a still image rendering method is bounded over time. A comparison of the objective quality scores of rendered views inpainting with or without the original depth map of the rendered view was run. No fall among the objective quality scores was observed commonly with the four metrics, but this does not suffice to prove the limited impact of depth on the synthesis quality over time. The results of the last section indicate that depth artefacts indeed have an impact in term of subjective visual quality.

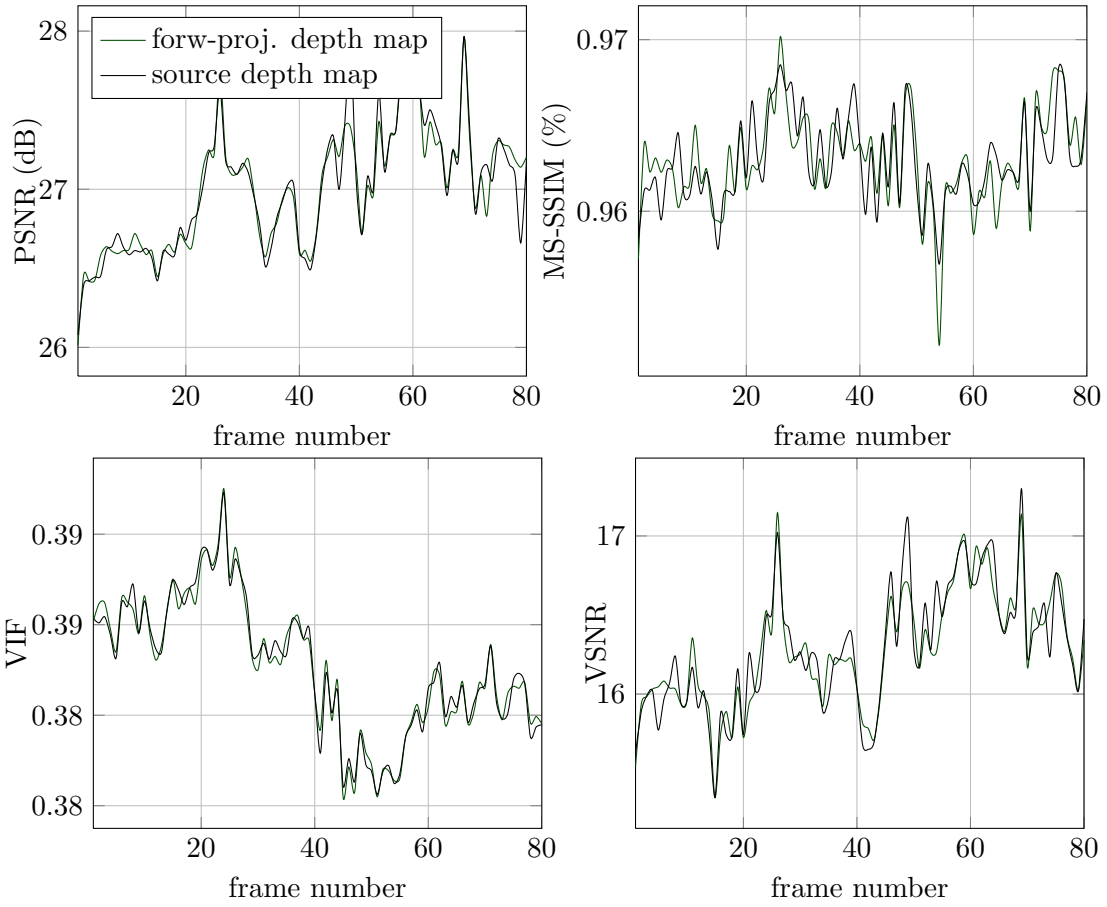



Figure 5.14: Objective visual quality score of "balloons" sequence ($V_1 \rightarrow_3$) obtained from four state-of-the-art quality metrics on 80 frames with the original depth map (black) or forward projected and inpainted depth map (green).

## 5.6 Conclusion

In this chapter a review of the past and recent inpainting techniques dedicated to object removal, texture synthesis and especially hole-filling for DIBR has been presented. Then a formulation of what should involve an inpainting method for perceptually plausible view synthesis has be drawn. Clearly the basic notion of undetectable inpainting becomes a notion of coherent background object reconstruction.

A depth-aware foreground-background segmentation based inpainting has then been proposed. It introduces threefold contributions to traditional template-based method to manage an efficient and plausible reconstruction of disoccluded regions. First during the search for highest priority patch to begin with, the coherent depth and color structures are favoured through a robust tensor-based isophote calculation. Directional inhibition also prevents starting from foreground borders. Finally a combination of closest geometric and photoconsistent candidates manages an effective natural filling.

Including the recent work on virtual view depth inpainting, this algorithm could be used for View Synthesis prediction at an encoder stage of a MultiView-plus-depth coder. The proposed method leads to convincing results and visually plausible reconstructed areas, but remains very sensitive to the prior depth reconstruction. The directional inpainting method

could probably be improved by limiting the depth artefact effects in the priority term. The combination of best patches could finally be replaced by multi-resolution patches whose choice or combination is based on a structural similarity metric.

# Conclusion and perspectives

This thesis proposed several contributions to the understanding of spatial and binocular perception and its applications to 3D systems.

The first part of this work gives some proposals to extend the predictability of visual attention models to both monoscopic and stereoscopic viewing conditions by considering the time-dependent contributions of visual features.

The second part focuses on the computer vision and coding community subjects of multiview-plus-depth transmission and view synthesis. A lossless edge depth map compression method for qualitative synthesis is proposed. In addition a new hole-filling method both for view interpolation and extrapolation is described that allows the rendering of perceptually correct virtual views.

## Synthesis of Thesis Contributions

### Perceiving the Depth

The perception of our visual world and its spatial layout is realized thanks to inferences and knowledge but also through combinations and fusions of multiples sources of depth information. The binocular or stereoscopic disparity is known to have strong influence on this depth perception due to early visual mechanisms.The potential influence of stereo disparity for depth estimation but also selective attention is studied within a visual attention model.

#### Integrating the Binocular Disparity in Visual Attention: of the Importance of Foreground and Time

Over the last decade, the existing saliency models have became more and more complex by the integration of bottom-up features aided by top-down contributions related to a priori, knowledge, task or action etc.

We have proposed addressing this issue by combining over time low-level visual features with central bias and higher-level attributes known to interact in the HVS within V2. The figure-ground segregation is modelled by a foreground-background discrimination from the depth. A statistical analysis confirms the influence of the foreground in the visual deployment in both 2D and 3D conditions and especially in 3D from the middle interval of fixation. The stereoscopy might indeed facilitate the visual deployment to figural positions within the visual field. Finally, the performance of the proposed model confirms the validity

of integrating dynamically different features which are known to be process specifically over time by the visual system.

### 3D Video Coding and Application to View Synthesis

The Multiview-plus-Depth videos promise new rendering capabilities and flexibilities for a new generation of 3D display and for a new quality of experience. However in the attempt to distribute video contents of high quality, the existing block-based coding video community failed to consider, predict or model the perceptual subjective quality of 2D videos, and all the more with the introduction of stereoscopic videos implying new artefacts but also a totally new image perception. Without involving the typical distortions due to stereoscopic displays, our two contributions related to MultiView-plus-Depth Videos target qualitative rendering. We based our two contributions on the video quality through the incorporation of the visual properties of the scene: image, surface and object entities that must be preserved at the end-user side.

#### Edge-based Depth Map Compression for 3D Coding

A selective encoding of both the depth map edge pixel positions and intensity values allows a correct decoding of the geometry of the scene. Because this decoded geometry participates in the rendering of new virtual viewpoints within the range of camera baselines for 3DTV or outside for FTV, we propose to encode the depth map losslessly.

In addition to the placement of seeds according to a quadtree decomposition based on edge presence, the decoded edge and seed pixels allow a correct interpolation of the depth map content.

The visual results, the objective results using 2D quality assessment metrics and the subjective tests confirm the pertinence of the approach but raise questions about the best bitrate to allocate to texture and depth videos. The subjective experiments also highlight a critical threshold of distortion visibility that should not be crossed.

#### Inpainting based View Synthesis for 3DTV and FTV

This last contribution found solutions to the problem of Depth Image Based Rendering or DIBR for extrapolation both for 3DTV and FTV. A theoretical formulation of what properties of the visual representation should be used by an inpainting method for DIBR is given. The typical problems of warping i.e. disocclusions, cracks and ghosting are dealt separately by an hole-filling method applied after cracks and background ghosting removal. A depth-aware inpainting method is then proposed to manage a plausible reconstruction of disoccluded regions by a new tensor-based directional structure propagation from the background surfaces.

## Perspectives

This thesis covers a wide scope of the promising 3D technologies for improved visual immersion. We have studied and proposed several contributions dedicated to a better comprehension of the spatial perception by the human visual system and to a more efficient transmission and synthesis of 3D Multiview-plus-depth video contents. This thesis covers multiple stages from the transmission to the perception of 3D contents; consequently some topics have been identified as deserving further investigation and developments:

### Model of Depth Cue Combination

The modelling of the combination of various depth cues in a single coherent framework has been addressed in the 90's in some theoretical models. It would be relevant to be able to identify the sources of information into a given scene of a video to estimate the amount of depth perceived globally and locally within an image (displayed both in 2D or 3D conditions). This could be used for accurate stereo video acquisition in studios in order to consider the spatial perception properties of the HVS. Also, it could be interesting for video edition and display manufacturers in order to better limit and manipulate the variations of stereo disparity but also the variations of overall depth feeling over time or space.

### Context based Depth Map Representation and Coding

As the human visual system relies on strong hidden assumptions to estimate the distance to objects in space (i.e. source of illumination at top inducing convex surfaces, gravity implying that most objects at bottom are often closer), the next 3D video representation and coding standards could use these properties to increase the rendering accuracy of regions close to viewer while diminishing the bitrate allocated to the predictable spatial layout and geometry of the scene.

### Stereoscopic Objective Quality Metrics

Recent 2D quality metrics manage to include some visual properties of the human visual system into some index of perceived visual quality. The development of new stereoscopic quality metrics both accounting for the typical distortions of the stereosopic displays (cross-talk etc.) and for the human visual system properties (binocular suppression, binocular rivalry etc.) might finally lead to significant advances.

# APPENDIX A

## The Spatial Vision

## Introduction

From the beginning of time humans have tried to explain the process of their perception of image and color though eyes and brain: the Vision. Oldest recorded studies of human vision date back to the time of ancient Greece and Plato (4th century BC), when vision was thought to be the process of an "inner fire" that gave rise to rays emanated from the eye toward perceived objects. In the Middle Ages popular theory also suggested that the viewer's eye sent out emissions to the object being viewed and that those emissions enabled vision. Because these theories were based on observation of scholars without extensive experimental scientific materials, they might sound unsatisfying today. The Arabic philosopher Alhazen (10th century AD) proposed the first that the eye is like a pinhole camera, where light from external sources is reflected from object surfaces and enters into the eye.

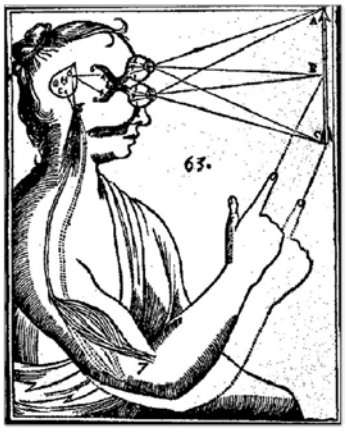



Figure A.1: Woodblock cut from designs found in Descartes's manuscript: *Le Monde ou traité de la lumière.*

Once it has been clear during the development of geometric optic in the 17th century that vision occurs from light entering the eyes through lenses, scientists seek to understand how a knowledge from the environment can arise from this light. Different classical theories

of vision emerged to state mechanisms and principles that not only organize and explain observed facts, but also make predictions about new facts.

With the deceptively simple question "Why do things look as they do ?", the Gestalt psychologist Kurt Koffka [66] framed in 1935 the problem for theories of visual perception. Three theories emerged opposing either environment versus organism, empiricism versus nativism, atomism versus holism or introspection versus behavior.

The structuralism approach was that perception occurs from basic sensory atoms - elements of sensory experience - evoking memories of other associated sensory atoms. These localized sensations were thought to be combined in perception by concatenation. Thus perception was assumed to be an associative and cumulative process of memories accessed through experience. Because "atoms" of experience rely on an analogy to chemistry and because the structuralist method relied on trained introspection instead of behavioral techniques, structuralism can be seen as a transition period between an ancient philosophical period and a more complex psychological one.

The Gestalt (meaning "whole form" in German) school arose by opposition to structuralism. As opposite to perception as a sum of atoms, they believed that experiences were intrinsically structured more than a set of parts and piecewise relations. The gestaltism successfully described the opponent process theory of color perception, believing there should be an opponent structure in the neural processes of color perception. As the Structuralists made an analogy of sensory "atoms" with chemical atoms, Gestaltists made an analogy of mental processes to force fields in physics (i.e. electromagnetism), where a charged particle can change the structure of the entire field extending over space. Wolfgang Köhler [67] goes further and assumes first that the brain was a physical gestalt i.e. a dynamic physical system that converged to an equilibrium of minimal energy. He also proposed that underlying neural mechanisms of perception involved electromagnetic fields generated by events of neurons. The failure to prove experimentally these ideas led the Gestalt theory to be abandoned by scientific community.

The ecological optics theory developed by J.Gibson also opposed structuralism, but unlike the Gestaltists also rejects the hypothesis that the structure of the organism is the basis for perceptual theory. He instead proposed that it is through the structure of the organism's environment i.e its ecology that the perception can be understood. This approach then relies on the informational basis of environment perception rather than on its brain mechanistic basis. This approach contrasted to previous theories and led the way to modern works in vision. But two controversial proposals raise critics: the "information pickup" process by which the brain perceives the environment without ambiguity just through a retinal image and an exploring organism [40, 41]. Also his "direct perception" stated that the visual perception of the environment is fully specified by the optical information at the retina without mediating internal processes. He was opposed to the "unconscious inferences" proposal of constructivists, as we will see in next paragraph.

Constructivism is the modern and dominant approach to visual theory. Beyond the four dichotomies that distinguished the three last theories, this combines the best elements of each. Being a theory focused on the internal processes of perception, it nevertheless assumes that these are based on the extraction of environmental information from the retinal stimulation. Global percepts should be constructed from local information, but an intermediary step of emergent properties of lines, figures,- as claimed by Gestaltists -might exist. It also claims that some aspects of perceptual processing are certainly innate, but others are learned by experience, and this should be settled by behavioural measures rather

than by introspective analysis as claimed by structuralists and Gestaltists. Von Helmotz proposed a process of unconscious inference [49] that definitely bridged the logical gap between 2D optical information and 3D interpretation of the environment that previous theories never succeeded to do.

Vision science enters in its modern era in the 1950s with three fundamental developments that definitely changed the vision comprehension: the emergence of computer to model cognitive processes, the application of information processing to psychological theories and the idea that the brain is a biological processor of information. This information processing psychology rises through three developments: the filter theory of attention by Broadbent [12] in 1958 that the first describes a psychological theory in the form of an information processing flowchart. Sperling [123] (1960) then discovered a form of visual memory - the iconic memory - that confirms that vision is worth exploring through the avenue of information processing. The third contribution that acts the information processing paradigm as the dominant approach to understand human cognition results from the invention of physiological techniques. These techniques enabled the study of the neural activity in the visual system and to understand the visual mechanisms in the retina and visual cortical areas. They are numerous: the single-cell recording, the autoradiography and the brain imaging techniques (X-rays, computer-aided tomography "CT" scans, magnetic resonance imaging "MRI" and positron emission tomography (PET) scans.

The next section will detail the visual perception as we know it today, but inherited from many vision theories, from the structuralists to the modern constructivists. The following section will then give a brief overview of the anatomy of the visual system, before we describe where and finally how the mechanisms of spatial vision occur.

### A.0.1 The Visual Perception

The Visual perception is not only the ability of acquiring visual information from the external world. It is specifically the process of construction of knowledge about the viewed scene, objects and events through information extracted from the light they emit. And it is precisely on this complex issue that modern theorists opposed and advanced.

This acquisition-construction of knowledge as a cognitive activity distinguishes the visual system from the simple analogy of a camera catching light. Instead, people perceive the scene and have a knowledge of the objects and events in their environment: the perception is an experience that gives a meaning about the nature of external reality.

Visual theorists agreed on the fact that the optical information is the origin of all vision and that the visual system somehow extracts environmental information from the pattern of retinal stimulation. After having described what visual perception basically is and before defining how it might be done (what are the structures to do this), we will shortly answer what visual perception is for.

### A.0.2 The Evolutionary Necessity of Vision

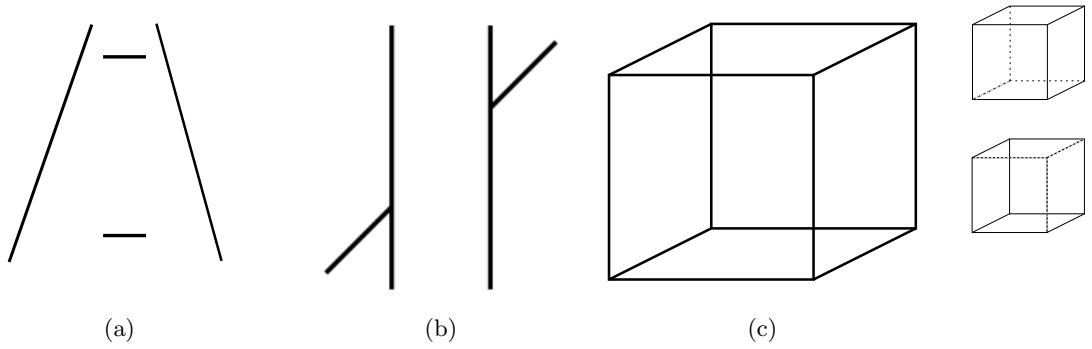
For homo sapiens and many other species, vision is prominent, because it gives to the perceiver highly reliable information about the location and properties of objects in the environment. Identification and locations are the key words of visual perception but also of the survival of species. Desirable objects and situations must be found and approached, while dangerous ones avoided and escape from. And vision has evolved only because it is reasonably accurate to succeed in these tasks, and precisely because light is a rich source of environmental information.

It is so true that people often think that the vision is a clear window onto reality. Indeed our perception is most of the time consistent with the actual reality of things in the environment, what we called a "veridical perception". But this is not always the case for fundamental reasons that will be explained below.

### A.0.3 The Constructive Act of Vision

Different phenomena occur in vision that will prevent us stating that visual perception is a clear window onto reality, essentially because it rather can change over time or according to the situation, past experience or knowledge:

- Adaptation: the vision can adapt to specific conditions and then change over time. For example in a dark room you need time to adapt your vision before you can see the whole room and objects inside.
- Aftereffects: in contrast when you receive a flash from a camera in your eye, you first see a blinding blaze of light, this is a veridical, truthful perception but then you experience a dark spot where you perceived the initial flash.
- Illusion: these are systematic perception that appear to be nonveridical. The visual perception then come less accurate or even false, as illustrated by the Figure A.2(a), where the horizontal line segments are identical in size, or by the Figure A.2(b) where the oblique lines are indeed collinear.
- Ambiguous figures: a given image can create distinct multiple perceptions, as shown by the Necker cube in figure A.2(b), where one can perceive two different orientations. The change of perception implies that the interpretation of the interposition between the lines -their depth relations- also changes.



(a) (b) (c)

Figure A.2: Example of visual illusions. The Ponzo illusion in figure (a): the two horizontal lines share the same length but do not appear to be so. The Poggendorf illusion in figure (b): the oblique lines appear offset. The Necker cube of figure (c) can be either a cube seen from above (top cube on the right) or below (bottom cube on the right).

The last phenomenon indicates that different interpretations can occur, but only one of them at a time: these are mutually exclusive. Then perception must imply the construction of different representations of the sensory data, different interpretive models. If we continue to observe the image, the two possibilities alternate back and forth: this multistable perception suggests that these two representations compete with each other. The winner alternatively relax and lets the loser gain the advantage, and so forth.

Thus more than one interpretation from the retinal stimulation are sometimes possible, proving the constructive nature of perception by response to the ambiguous nature of information conveyed by light stimulating the eye. But Gibson and his ecological optic approach opposed this idea of internal representation and claimed such phenomena occur in conditions that are ecologically invalid, seldom or never present in everyday living conditions. We will follow the approach of constructivists in the sense that a visual theory must account for all phenomena of visual perception. Therefore a complex process must occur in vision to create knowledge from the information strictly given by the light stimulating our retinae.

### A.0.4 The Inverse Problem Solved by Vision

The main issue in understanding vision comes from a single question: once light projects from a three-dimensional world onto the two-dimensional surface of the two retinae, how to get back from these optical images to the knowledge of the objects in the three-dimensional world?

The fact is that the mathematical relation between the environment and its projected image is not symmetrical. The projection from the 3D world to the 2D retinae image is a well-defined reduction of dimensionality by projection. Each world point maps into a unique point on the retina. But the inverse projection from image to environments goes from two dimensions to three and is a ill-posed problem: each point in the image can be re-projected into an infinite number of positions in the environment, as illustrated by the figure A.3. Theoretically, that means that infinite distinct 3D environments could have

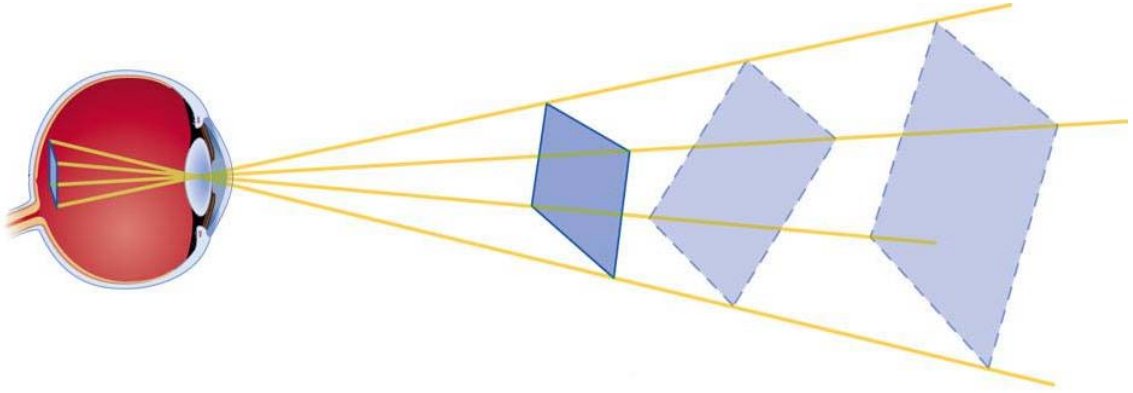


Figure A.3: As a single point or a single line, a single plane on the retina could be the projection of an infinite number of planes from the real world.

given rise to a single 2D retina image. But that is not the case, the visual system manages to give a veridical perception most of the time, the 3D perception is precisely possible. Gibson and his theory of ecological optic claimed that this perception is direct and active over time and does not require any heuristic process or internal representations to recover the 3D information. But this does not suffice to recover a unique 3D (+time) solution from unique 2D retinal images over time.

Unlike Gibson, constructivists state that vision requires a process of inference to extract a perceptual 3D interpretation from the 2D optical information. Helmoltz [49] admitted the indeterminate nature of the 2D vision and proposed that the inverse problem could be solved by adding hidden assumptions to reach perceptual conclusions about the environment. This process of perceptual inference is unconscious because it does not require awareness of people to make visual inference. But on which base these inferences are

made ? He later stated that vision comes from an interpretation of the most likely state of affairs in the environment having caused the retinal images. This likelihood formulation of interpretation of vision is now widely accepted by scientists.

Vision is thus a heuristic process relying on assumptions and unconscious inferences to define the most likely environmental condition that could have produced a given retinal image. This process is heuristic and led sometimes -as seen with illusions- to a wrong solution and then a wrong conclusion. However most of the time these assumptions are true and we experience a highly veridical perception. The window onto reality is not so perfect but clear enough for humans. How does it open, how does vision occur ?

## A.1 The Visual System: Anatomy and Functions

The structure of the visual system will be shortly described by its anatomy before exploring the physiological functions that underlie the processes of vision.

### A.1.1 The Human Eye

As everybody knows, the eyes act as the sensors receiving the incoming light from the external world. The two human eyes are spherical except for a bulge at the front. They are located in the horizontal midline of the head and are separated by approximately 6.3 cm. Six extraocular muscles, controlled by specific areas in the brain, move the eye to scan different areas of the visual field without needing to turn the head.

The frontal placement of the eye results in a large area of overlap of the two visual fields. These overlapping visual fields allow a binocular i.e “two-eyed” vision and an accurate depth perception. While many other species have their eyes located at opposite side of their head, hunters like humans have an overlapping field advantageous to gauge the distance to prey. In contrast, prey generally have laterally placed eyes to monitor as much of the world for danger.

The analogy of the eye to a camera is not only limited to the light collection from the environment onto a surface. The eye also has the ability to focus the incoming light in a clear image at the back of the eye. The Figure A.4 illustrates the different anatomical parts of the eye. Different optical functions are realized. First, the light enters the cornea, a little transparent protuberance that stands in front of the eye and that surrounds a cavity with a clear liquid, the aqueous humor. Next the light crosses a kind of camera diaphragm: a pupil, whose variable opening in the opaque iris controls the amount of light to let enter. Then the remaining light passes through the lens, whose thickness is controlled by ciliary muscles. The light then crosses the vitreous humor in the central part of the eye before it strikes the curve surface at the back of the eye, the retina. This back surface of the eye contains photoreceptors whose function is to transform light into neural activity. The eye is not perfect due to its optical elements (some light get absorbed by lens, vitreous humor and blood vessels) but most of the light entering the eye reaches the retina.

Thus the amount of light is controlled by the iris and pupil, the focus is realized by ciliary muscles and the lens, and the acquisition and quantization of photons are realized by the retina and its photoreceptors.

### A.1.2 The Retina

The optical functions have brought the light from a distant object into proper focus and quantity on the retina. The next function of the eye will be to quantize the light photons

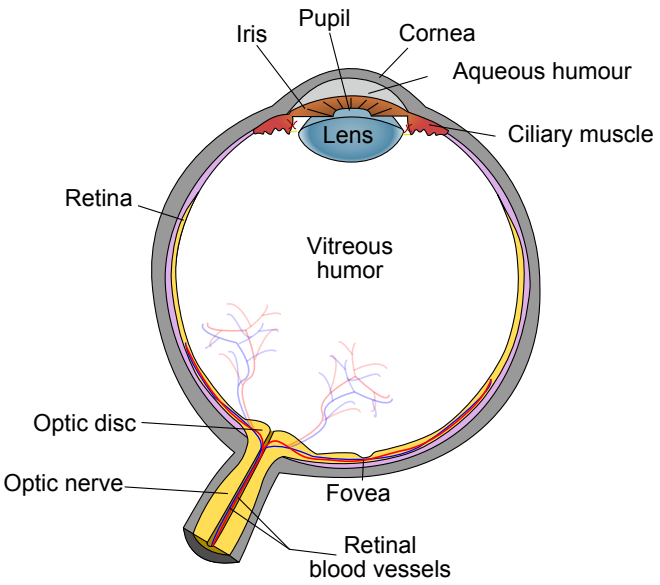



Figure A.4: Cross section of a human eye. The light comes into the eye through the cornea, aqueous humor, lens, and vitreous humor before striking the photoreceptors of the retina. (Modified from [112])

and convert them into neurochemical responses later processed by the brain. The main components of the brain will first be described before the explanation of light to neural signal conversion.

### Neurons

The neuron is the basic cellular entity of the brain. This specialized cell integrates the input activity of other linked neurons and propagates the integrated output activity to other neurons. A complex process of biochemical events occurs within the neuron to realize these functions of integration and transmission.

The dendrites collect chemical signals from other neurons and convert them into electrical signals that travel along the thin membrane of the neuron. This signal is a graded electrical potential between the inside and the outside of the dendrite, that depends on the initial stimulation by other neurons.

The cell body wraps the nucleus in a membrane that integrates the input electrical impulses coming from all the dendrites. The graded potentials from the dendrites are then converted in a series of all-or-none electrical potentials (called action potentials, electrical impulse or spikes) output to the axon.

The axon is a fine cable-like projection of the neuron along which electrical impulses are transmitted to other neurons. The power of the integrated signal is encoded in its firing rate: the number of spikes generated in a period of time. The insulating myelin sheath enables the electrical impulses to travel faster whilst using less energy. While most neurons have only one axon, it may undergo extensive branching and communicate with different target cells. Also, the axon can carry back information to the cell.

The axonal terminal contains the synapse where the electrical activity is converted back into a chemical signal. The neurotransmitters are released into the terminal and dendrite gap of the next neuron. The signal strength depends on the quantity of neurotransmitters released.

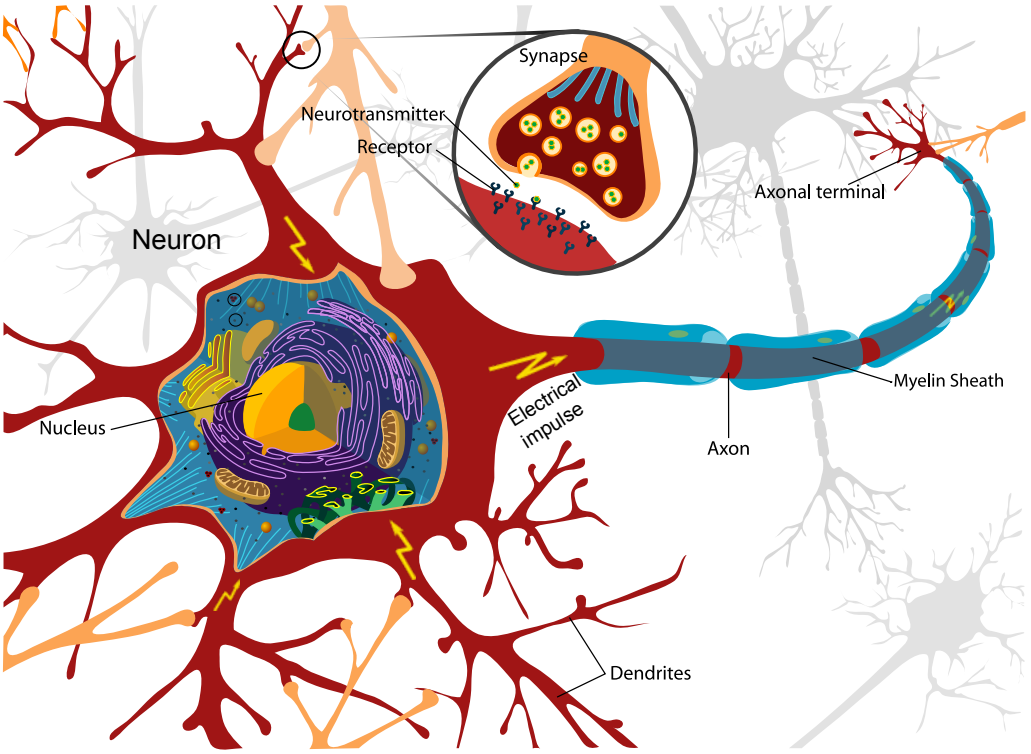



Figure A.5: A neuron in its environment of other neurons. A neuron cell consists of a cell body integrating graded electrical signals from its dendrites. The result is transmitted in discrete action potentials along an axon, covered by a myelin sheath, to axonal terminal. There the synapse drops neurotransmitters to stimulate the dendrites of next neurons and so on. Modified from [148].

In order that some neurons could receive input from previous ones and could transmit their output to next ones, an initial energy must be provided from the environment to the neurons in the good form. In the visual system, this function is realized by photoreceptors located in the retina. As we will see, the output response from photoreceptors is successively treated by horizontal and bipolar cells, then amacrine and ganglion cells. This initial processing happens inside the retina, as illustrated by Figure A.6 and A.7.

### Photoreceptors

A photoreceptor is a specialized cell in the retina that converts the light stimulation into a neural response. There exist two types of photoreceptors: the rods and cones. The rods are longer (see Figure A.8) and more numerous, sensitive to light and located everywhere in the retina except at its center . The cones are shorter and have a conelike ends. They are much less sensitive to light but instead sensitive to specific ranges of the light spectrum, under normal daylight lighting conditions. These are located in high density in the center of the retina: the fovea. This area covers only 2°of the visual field, spread on less than 1% of the retinal size but takes up over 50% of the visual cortex in the brain.

There exist three types of cones in the normal trichromat's retina, whose sensitivity to photons wavelength differs. The short-wavelength (S) cones absorb maximally the light of at 440 nm wavelength, the medium-wevalength (M) cones at 530 nm and the long-wavelength (L) cones at 560 nm. Their number and spatial distribution are not uniform: the

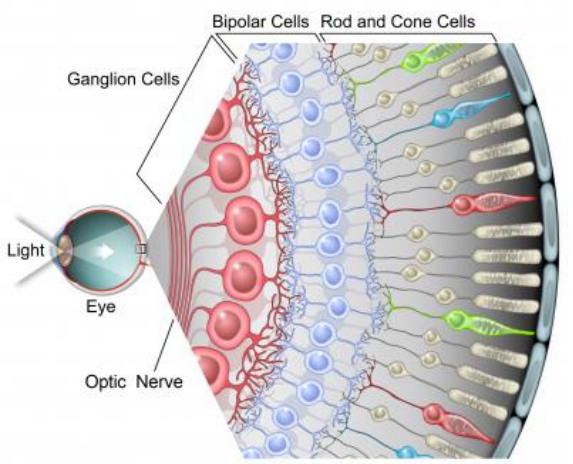



Figure A.6: The human retina. The retina is composed of five major types of neurons: receptors (rods and cones), bipolar cells and ganglion cells (horizontal and amacrine cells are not represented). There exist three types of cones (in colour) that differ in their sensitivity to photons wavelength.

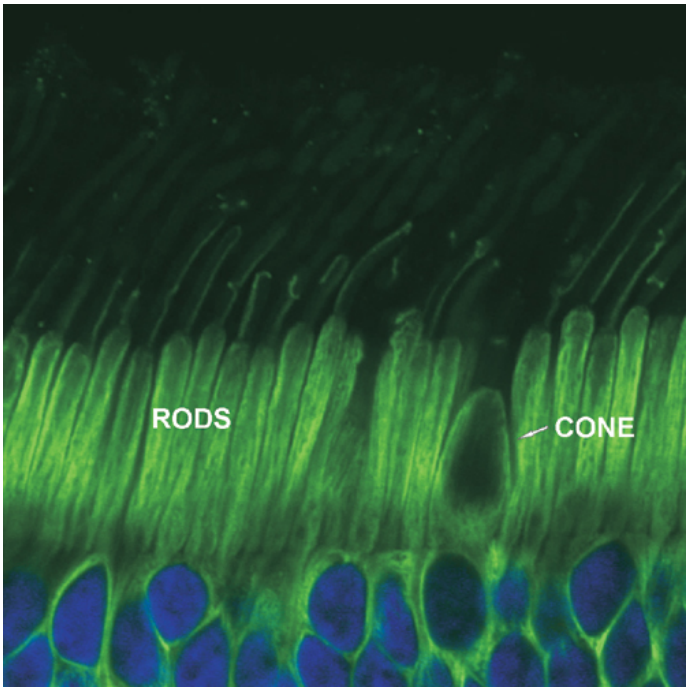



Figure A.7: Illustration of rods and cones in the receptor pigmented layer.

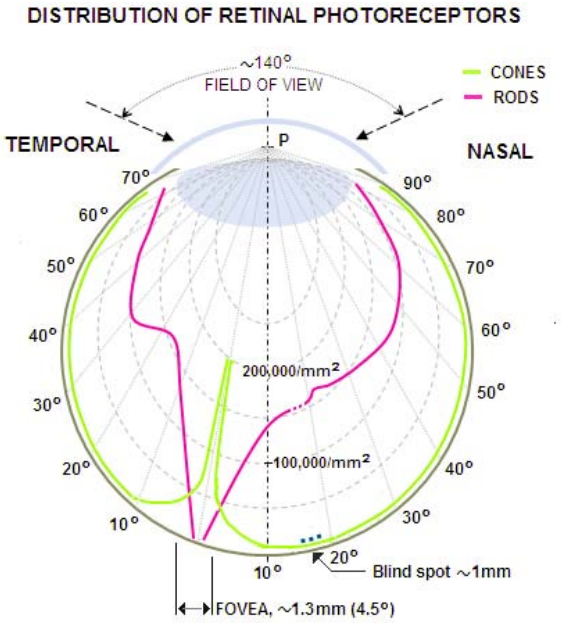



Figure A.8: Distribution of rods and cones in the human retina. Notice that there is no cone or rod (dashed lines) in the region covered by the blindspot. Diagram modified from [16].

ratio of L to M to S is about 10:5:1 and the fovea is almost exclusively composed of M and L cones. These are misleadingly called blue, green and red cones, as illustrated in Figure A.6, because of their selective light absorbance at dedicated wavelengths.

The photoreceptors transform this electromagnetic energy into neural signal through embedded pigment. When a photon strikes a pigment molecule, it absorbs the photons and changes its shape, modifying the electric current flow around this pigment molecule. This electric change is then propagated down to the synapse of the photoreceptor, where neurotransmitters chemically transfer the information to the next neuron. The electrical output from input photons absorbed within a photoreceptor are then integrated. As we have seen previously, most of neurons receive graded potentials from the dendrites and respond by a series of electrical impulses. But the photoreceptors and bipolar cells instead respond by producing graded potentials, i.e. continuous -rather than discrete- change in electrical potential. For a photoreceptor neuron, this graded response is a logarithmic function of the number of absorbed photons.

As illustrated in the Figure A.6, a particularity of the retina is to have its photoreceptors at the lowest layer and not at the first layer that incoming light encounters. It might probably be due to the location of enzymes, needed for pigment regeneration, that remain in the pigment ephitelium at the opaque external part of the retina. One could also expect, since one of the roles of the eye is light-to-neural-signal conversion, that photoreceptors would respond to light by an increase in synaptic activity. The opposite actually occurs in the vertebrates: a light locally causes a decrease in synaptic activity. In any case the important fact is that photoreceptors preserve the information of light intensity converted into neural activity.

### Bipolar Cells

The bipolar cells connect photoreceptors and ganglion cells, either directly or indirectly. They receive the synaptic input from either rods or cones (direct path) or also from intermediate horizontal cells (indirect path). As the photoreceptors, the bipolar cells communicate to other cells by graded potentials (continuous changes in electrical potential). The horizontal cells can introduce lateral inhibition and give rise to the center-surround inhibition. This process will be described in detail in the next section about ganglion cells because they precisely realize a similar process.

### Ganglion cells

The first experiments using micro-electrodes in the retinal cells were conducted on ganglion cells. The reason is they are the first cells in the visual system to produce spike discharges, so that they could be easily recorded from outside the cell. Kuffler and Barlow (1953) then recorded the cell's activity while different images where displayed to the animal's retina. The experiments showed the firing rate of the ganglion cell was highest for a spot of light at a specific location. Surprisingly, if the size of the spot was either increased of decreased, the firing rate diminished. This phenomenon is illustrated by Figure A.9. Thus a spatial selectivity of the ganglion cell around a certain spatial position happened. This antagonism between a selected center and its surrounding ring is called a center-surround mechanism and is shared by many type of cells in the visual system as we will see later.

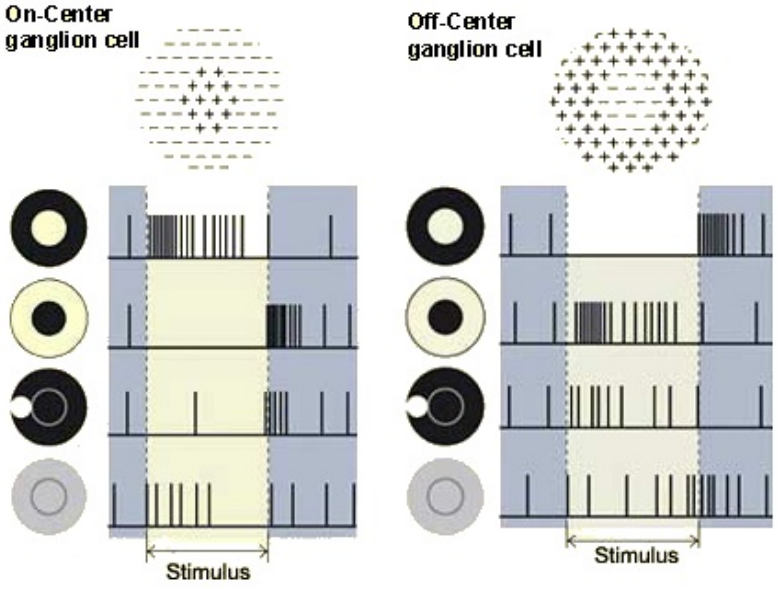



Figure A.9: Response of the two types of ganglion cells. Left: an on-center/off-surround cell. Right: an off-center/on-surround cell.

The ganglion cells have been later categorized into two different types: the on-center cells and the off-center ganglion cells. On-center cells respond to light at their receptive field center by spike discharges, while off-center cells respond to absence of light in their receptive center by spike discharges. On-center cells then have an excitatory response to light at their center. In contrast off-center cells respond by inhibition to light at their center. As highlighted by experiments, the area surrounding the central region always

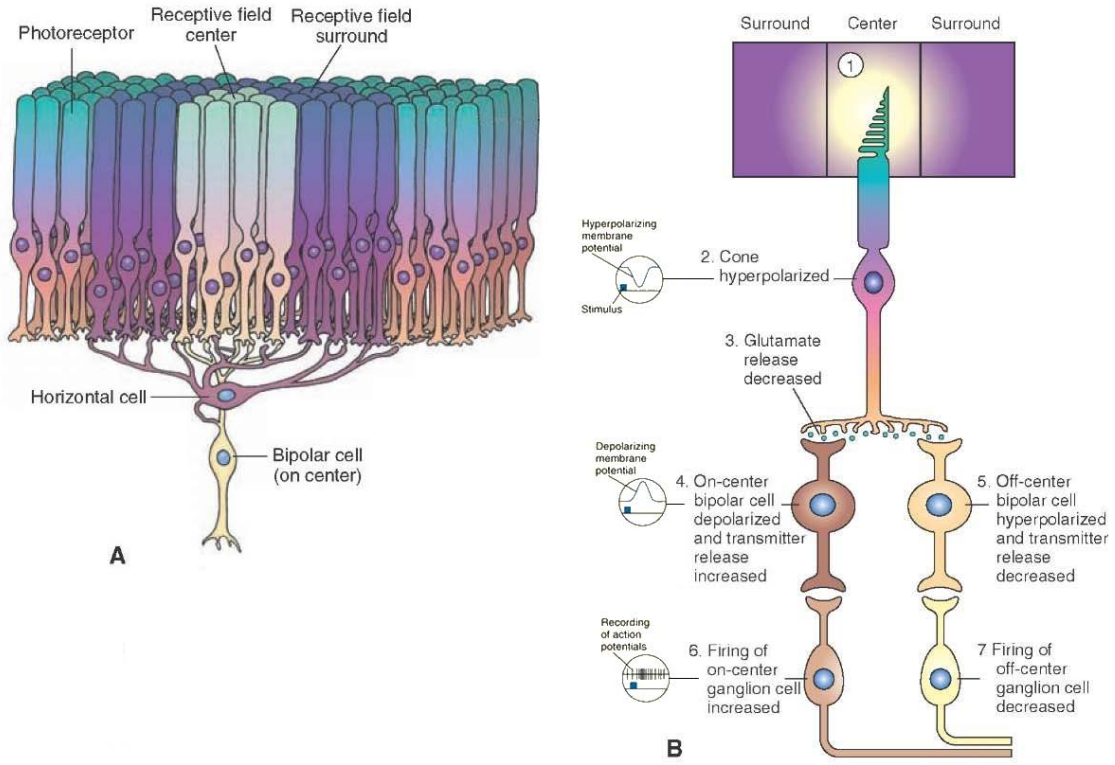



Figure A.10: (A) Receptive fields of photoreceptors and their connections. The receptive field center provides a direct input from the photoreceptors to the bipolar cell, and the receptive field surround provides indirect input from the photoreceptor to the bipolar cells via horizontal cells. (B) Responses of retinal bipolar and ganglion cells to darkness and illumination in the receptive field center. Changes in the electrical activity of the photoreceptor and on-center and off-center bipolar and ganglion cells when the photoreceptor receptive field center is illuminated. Modified from [108]

has the opposite characteristic. Then the on-center cells have off-surrounds, and are also called on-center-off-surround cells and inversely for off-center cells.

The connections of photoreceptors (with different receptive fields) to a single bipolar cell are illustrated in Figure A.10 (A). The Figure A.10 (B) describes a whole process of illumination in the receptive field center of a bipolar cell to the response of an on-center, off-surround ganglion cell. The light emission on the center produces an hyperpolarization of membrane potential of cones located in receptive field center. This provokes a release of neurotransmitter to the synapse between receptors and bipolar cells, which in turn provokes a depolarization of the membrane potential of the on-center bipolar cell. The transmitter release increase in turn provokes an increase of firing rate of an on-center ganglion cell. The opposite process occurs for off-center bipolar and ganglion cells connected to similar receptors in the receptive field center. Please also note that these bipolar and ganglion cells elicit opposite responses when the light is received at the receptive field surround and not at the center.

### Vision Neural Pathways

The axons from the ganglion cells leave the retina and the eye through the optic nerve. The set of nerves forms the optic chiasm and then separates in two pathways on each side

of the brain. The nerves from the nasal side of the fovea from each eye (see Figure A.11) are mapped to the opposite side of the brain. The nerves from opposite side (temporal) follow the same side. Thus the mapping from external visual fields to the visual cortex is crossed: the left half of the visual field (in purple in Figure A.11) goes to the right half of the brain, and conversely for the right visual field. Before the neurally encoded visual information is conveyed into the primary visual cortex, the optic chiasm is separated on each side. A small pathway goes to the superior colliculus, a nucleus mainly in charge of elementary localization and eye movement control. The larger amount of optic nerves goes to the lateral geniculate nucleus (LGN) of the thalamus.

**LGN** The ganglion cells -which process the information by center/surround mechanisms- finally synapse in the LGN having similar center/surround receptive fields. These paired nuclei however process the signal at a larger scale with a stronger inhibitory surround. Then, while bipolar and ganglion cells cover the two-dimensional surface of the retina and receive input from nearby cells, each LGN has a three-dimensional structure of cells which receives input from both eyes. However, each LGN cell responds to stimulation from one eye separately, these are monocular cells (the binocular cells can be found at a higher stage in the visual cortex and will be presented later).

The LGN structure is made of layered 2D-sheets of neurons folded as illustrated in Figure A.12. Six layers compose it.

The two lower ventral layers are the magnocellular layers. They receive the projection from magnocellular (M) ganglion cells. M ganglion cells receive themselves input from both rods and cones and are especially sensitive to black and white rather than colour. Then they project to LGN magnocellular cells in magnocellular layers, also not very selective to colour, but quite sensitive to contrast and with a fast response to retinal stimulation.

The upper four dorsal layers of each LGN contain smaller parvocellular cells receiving projection from parvocellular (P) ganglion cells. P cells receive input exclusively from cones, and so act as the reverse of M cells because they are more sensitive to colour than to black and white. The connected LGN parvocellular cells are also very selective to color but less sensitive to contrast and show a sustained response to retinal stimulation changes and then have a weak temporal resolution. These differences are summarized in Table A.1.

| | Parvocellulars | Magnocellular |
|---|---|---|
| Color sensitivity | High | Low |
| Contrast sensitivity | Low | High |
| Spatial resolution | High | Low |
| Temporal resolution | Slow | Fast |
| Receptive field size | Small | Large |

Table A.1: Main physiological difference between parvocellulars and magnocellular LGN cells. Reproduced from [101]

.

First, each layer contains cells getting signals from only one unique eye. The two magnocellular and four parvocellular layers receive input alternatively from the left or the right eye, bottom and top layers being connected to the opposite side eye. Then layers 1, 4 and 6 project to the opposite eye, while layers 2, 3 and 5 project to the eye on the same side of the head.

Second, each layer is arranged spatially like the retina that covers the back of the eye.

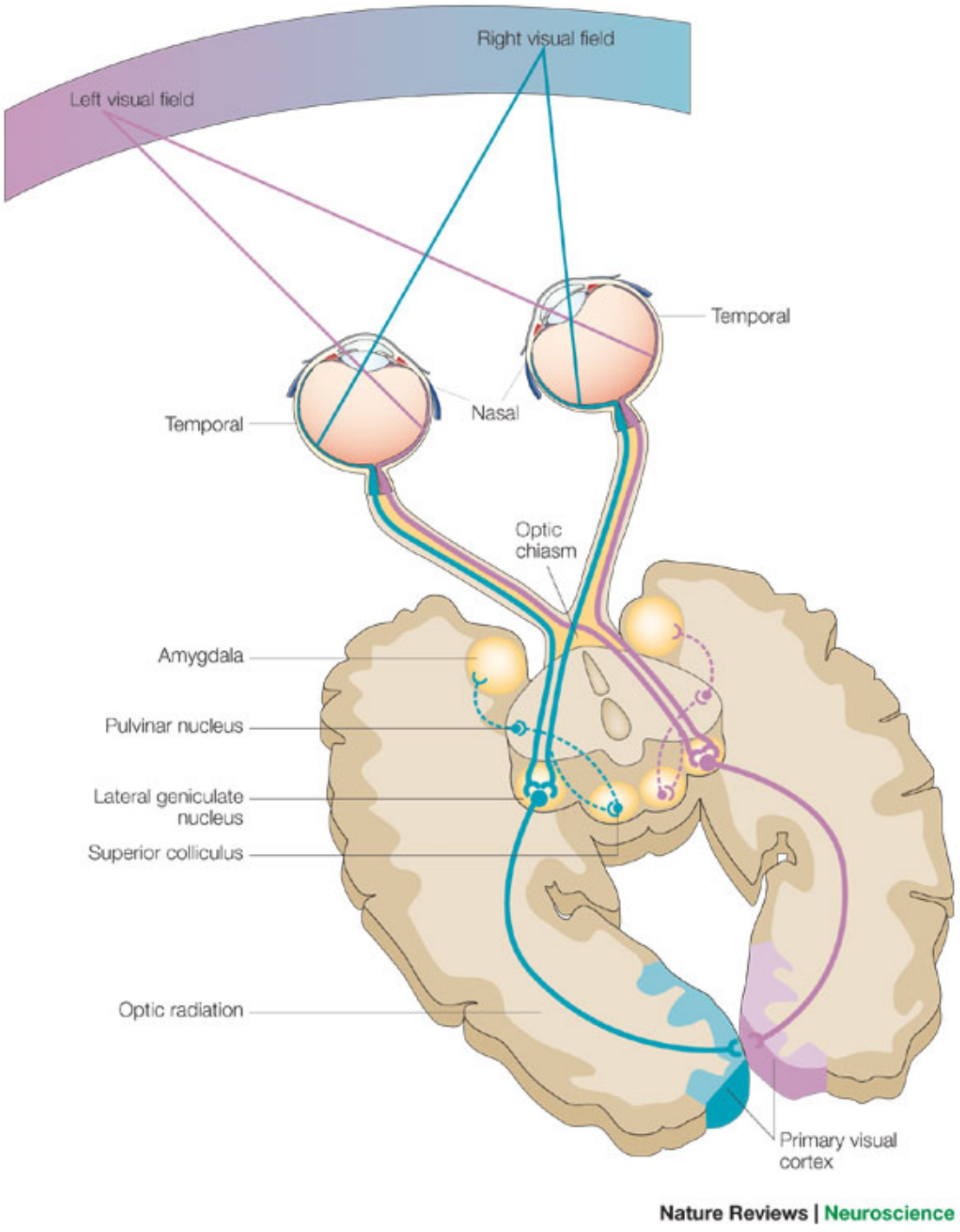



Figure A.11: The primary visual pathway i.e the retina-geniculate-striate system. Reproduced with permission from [45].

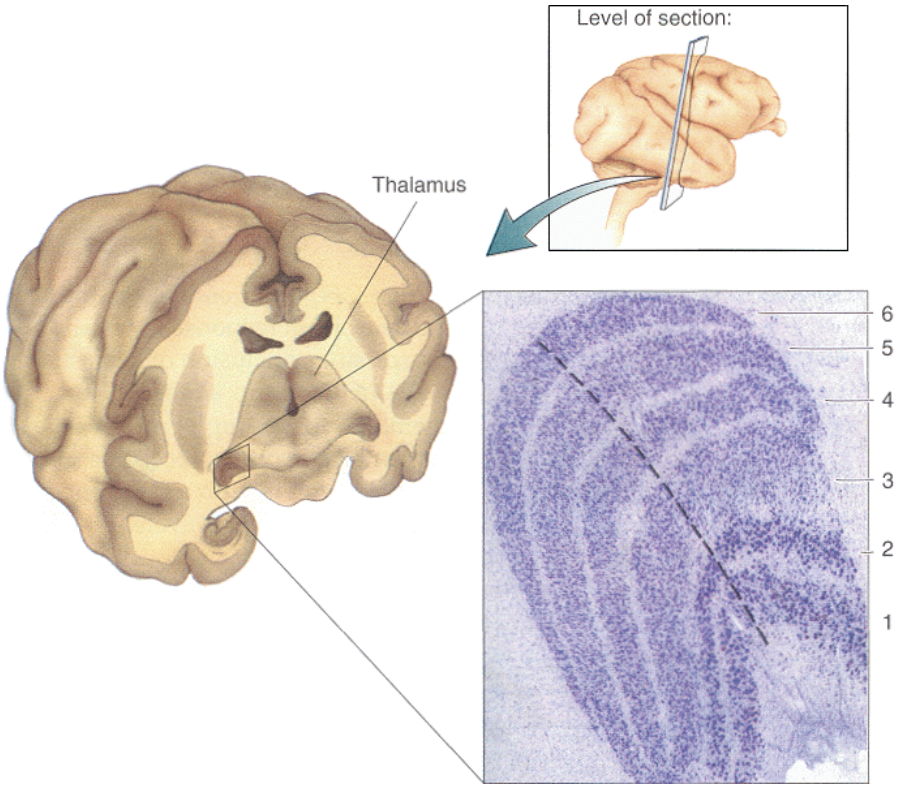



Figure A.12: View of a left lateral geniculate nucleus located in the most lateral inferior region of the thalamus. Its six layers project to different visual fields of each eye. The ipsilateral eye projecting to layers 2,3 and 5 and the contralateral eye projecting to layers 1,4 and 6. The two ventral layers, layers 1 and 2, contain the magnocellular cells while the four dorsal layers, layers 3 through 6, contain smaller parvocellular cells. Reproduced from [35].

This **retinotopic mapping** preserves the relative locations of cells from retina to LGN: the nearby regions on the retina mapped to nearby regions in the LGN.

In agreement with the visual field cross in the optic chiasm - each side of the brain receives input from only one half of the visual field- the left LGN then receives input from the right visual field only, either from the left eye (in layers 2, 3 and 5) or from the right eye (in layers 1,4 and 6), and conversely for the right LGN. This hemifield projection continues through the optic radiations (see Figure A.11) from LGN axons to the striate cortex.

### A.1.3 Visual Cortex

The visual cortex is the part of the cerebral cortex in charge of the visual information processing. It spreads in the back of the brain in the occipital lobe up to a part of the parietal and temporal lobe, which are divided into two cerebral hemispheres nearly symmetrical. The left hemisphere receives its input from the right visual field, and the right visual cortex from the left visual field.

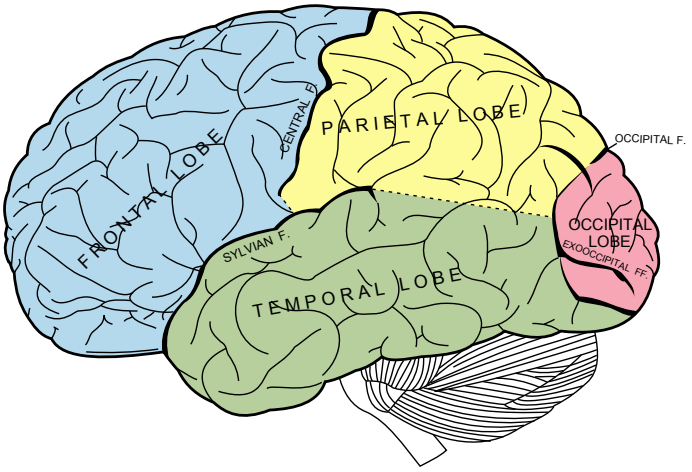



Figure A.13: Lateral view (left side) of the human brain. The frontal, parietal, occipital and temporal lobes are highlighted in blue, yellow, pink and green respectively. From [43].

The visual cortex covers the primary visual cortex (also known as striate cortex or V1) and extrastriate visual cortical areas such as V2, V3, V4 and V5. These anatomical distinctions are also physiological, as we will see in the next sections. There exist in the brain similar localized cortical areas that process other sensory modalities such as audition, taste, touch and smell. But up to now, the localization of functions is an open subject. Studies increasingly bear out however this localization of function, especially those concerning the first functions of the vision that lie in the occipital, parietal and temporal lobes.

#### The Visual Cortex: Occipital, Parietal and Temporal Lobes

The occipital lobe receives the neural information from the pair of LGN into the striate cortex. This area "outputs" to other different parts of the visual cortex, either to extrastriate cortical areas or areas in the parietal and temporal lobes.

The parietal and temporal lobes are often defined as being the termination of two primary pathways realizing two complementary functions on the visual information: the "where" and the "what" pathways respectively.

- The dorsal pathway starts from the striate cortex V1, goes through V2, then to the dorsomedial and visual area MT (V5) and ends up in the posterior parietal cortex. This "where pathway" is associated with motion, representation of object locations and control of the eyes and arms, such as visual saccades, grasping or reaching.

- The ventral pathway also finds its origin in V1, passes through V2, then through V4 and to the inferior temporal cortex. The ventral stream, also called the "what pathway" is believed to process the shape information to represent and identify objects.

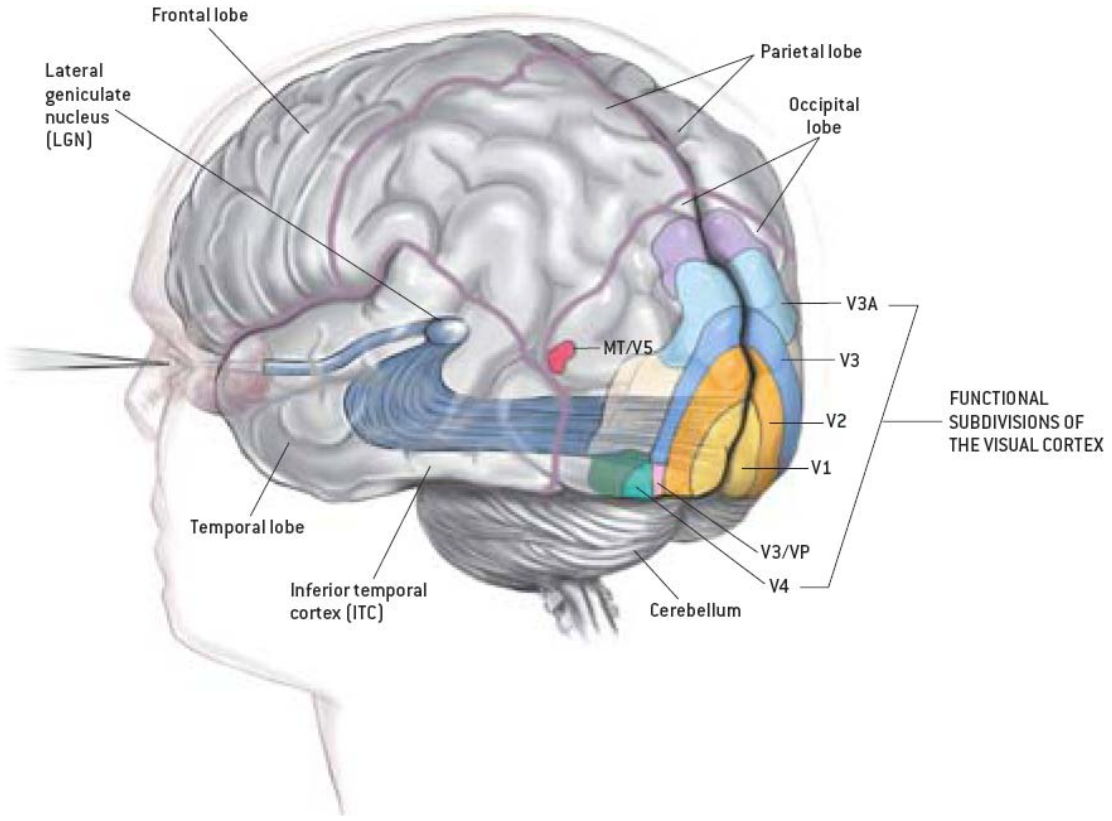



Figure A.14: The human visual pathway starts from the eyes and extends through LGN before ascending to different regions of the visual cortex. Reproduced with permission from Therese Winslow and [77].

This "big picture" of cortical functions described by a dichotomic dorsal/ventral pathways originates from physiologists Ungerleider and Mishkin [134, 90] who the first studied the impact of lesion in cortical areas on vision. The evidence comes from experiments where monkeys were required to do an object discrimination task or a landmark discrimination task following a central or a dorsal lesion. Object discrimination task is a very difficult task for monkeys with their inferior temporal cortex removed, while a landmark discrimination is difficult for monkeys with a parietal lesion.

This dichotomic view might sound like a simplistic approach of vision, but remains up to now contentious among vision scientists.

This distinction between two visual pathways clearly exists as well in humans since persons having damage in their temporal cortex suffer from visual agnosia (deficit in visually identifying some objects or faces). Thus the ventral pathway converging to the temporal lobe supports the hypothesis of a "what" system. Patients presenting damage in the parietal lobe suffer from the unilateral neglect syndrome: they show inability to grab objects in the half of the visual field opposite to brain damage, thus supporting the possibility of a damaged "where" system.

These study cases albeit characterizing the main functions of specific area in the brain, do not suffice to understand the neural processing underlying such complex abilities of object identification and localization. Vision scientists and physiologists henceforth observe the neurons though individual cells recordings to discover the neural relationship to specific stimuli, and more specifically their anatomical tracing from one area to another

and their physiological response or action potential to stimuli within their receptive field.

### The Primary Visual Cortex

The first cortical stage of visual processing is the striate cortex and is located in the occipital lobe at the back of the brain (see Figure A.14 for global position). It is the largest cortical area with 200 millions cells (while the LGN and retinal ganglion cells are respectively 1.5 million and 1 million).

**Location** As other cortical areas, it consists of highly convoluted sheet of cortical neurons, not visible in the exterior of the brain. The positions of V1 through V3 on the internal face of the occipital lobe are illustrated in Figure A.15 (left). A coronal slice illustrated in Figure A.15 (right) shows the transition between areas. The transition between V1 and other areas can be clearly delimited by a cellular frontier. Other areas are delimited by similar techniques but after V3, the delimitations do not meet consensus yet.


Figure A.15: Internal face of the occipital lobe (left) illustrating the location of V1 and several prestriate areas (V2 and V3). Right: The positions of V1 and the transition between V1, V2 and V3 are shown in an horizontal slice. Modified from [138] and [147].

Each striate cortex lobe receives the majority of ascending projections from the LGN, on the same side of the brain, so the visual input like the LGN is crossed. The right visual fields projects to the left striate cortex in the left hemisphere, the left visual field to the right striate cortex.

**The Retinotopic Map** Also like the LGN, the projection from the retina to the striate cortex is topographic or retinotopic: nearby retina's regions mapped to adjacent regions in the striate cortex. The qualitative spatial relations are respected, but the quantitative ones are not: the central area of the visual field, mainly integrated by the high-density cone fovea, is correspondingly projected to many more neurons in the striate cortex than the periphery visual field. This spatial distortion of central areas in regard to peripheral areas is called cortical magnification. It can be illustrated by an autoradiograph of the

retinotopic map in V1 of a macaque monkey (Figure A.16). A monkey was trained to fixate the center of the wheel-shaped pattern mixing black and white rectangles. The resulting cortical image from one hemisphere shows the five radial and the three semicircular lines mapped in a topographic manner from the half right pattern. The iso-excentricity circles project to vertical lines, while isopolar projections cross them orthogonally. This illustrates that the central -foveal- area of the visual field occupies a much larger area of cortex than the peripheral regions. It can be explained by the dense presence of receptor cells in the fovea that project to the upstream cortical cells.

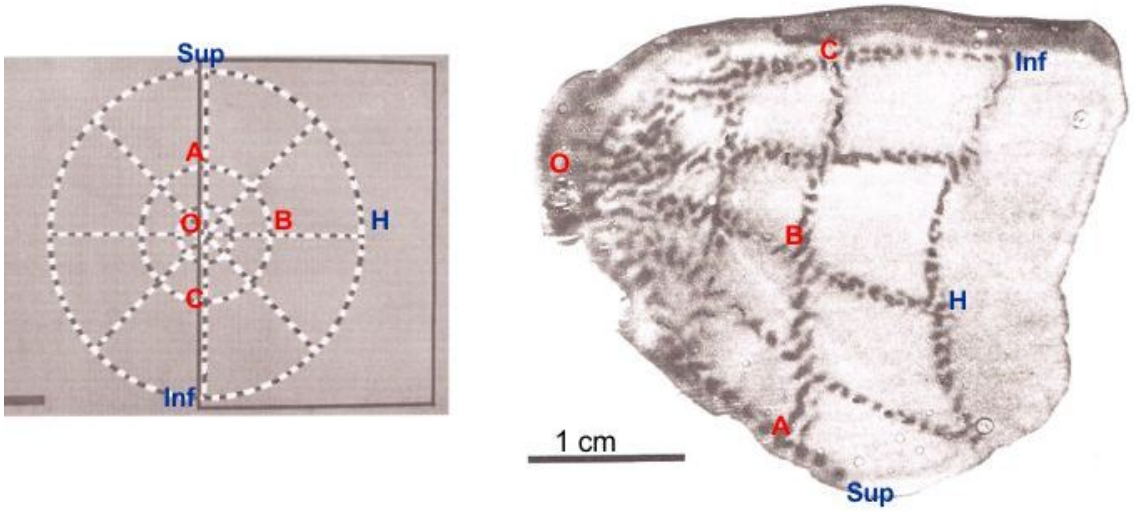



Figure A.16: An autoradiograph (Right) illustrates the retinotopic map in area V1 of a macaque monkey starring at the center of the pattern. From [129].

**Simple, Complex and Hyper Complex Cells** Hubert and Wiesel [52] discovered in 1959 the first three types of receptive fields in the striate cortex. They were using a single-cell recoding technique and trying to make a cell fire from a small spot that was mounted on a microscope slide in front of a projector. Sometimes while the slide was moved around, the cell responds by a strong burst of activity. They managed to stimulate a cortical cell by inadvertence without knowing why. They realized that the cell's output was not caused by the moving spot, but indeed was caused by the shadow of the slide's edge that was moving in a particular direction from a particular position. Thus they discovered that orientation and direction of edges were indeed processed by the first cortical area. They finally observed and distinguished three types of receptive fields: simple cells, complex cells and hypercomplex cells.

**Simple Cells** have responses that can be predicted from individual spots of light. They can respond by inhibition and excitation as cells seen previously, but seemed to aggregate and sum their responses to a set of small spots of light directed toward their receptive field. Hubel and Wiesel report that different subtypes of simple cells respond to lines or edges at specific retinal position and orientation. Hubert and Wiesel hypothesized that these edge and line detectors can be assembled from outputs of specifically aligned center/surround cells from the LGN. These cells could have receptive fields with centers aligned along the preferred orientation.

Recently it has been shown that a large variety of receptive field sizes exists among

the simple cells, small ones responding to fine spatial structure while large ones to coarse spatial structure.

**Complex Cells** differ from simple cells in the sense that their output is non-linear. They weakly respond to individual stationary spots, so their response to different orientation cannot be traced by observing their response to small spots in each retinal position. They also tend to be highly responsive to moving lines or edges with their receptive field. This motion sensitivity is also restricted to a direction of movement, as illustrated in Figure A.17. However complex cells are insensitive to the position of the stimuli within a certain limit. Thus small shifts of position of a bar will not impact the complex cell response rate. Finally, the complex cells have often larger receptive fields than simple cells. The conclusion of Hubert and Wiesel is that complex cells integrate the responses of different simple cells. Complex cells receive input from several simple cells having receptive field at the same orientation but different positions, as illustrated in Figure A.17. In practice, a few complex cells also receive input directly from the LGN.

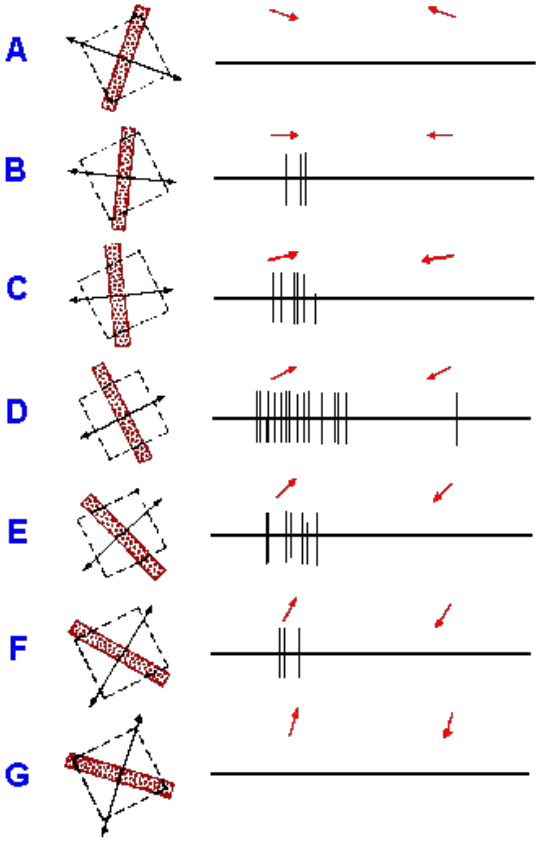



Figure A.17: Responses of a complex cell in right striate cortex (layer IVA) of Macaque to various orientations of a moving black bar. Receptive field in left eye is indicated by interrupted rectangles, approximately 3/8 x 3/8 degree in size. Duration of each record, 2 sec. Arrows indicate direction of stimulus motion. From [53].

**Hypercomplex cells** are a third type of cells located in the striate cortex. They are called hypercomplex cells because they are even more selective than complex cells. Nev-

ertheless, unlike complex cells, they are selective to position and to the termination of a line or an edge: extending in length an edge causes them to fire less than when the line is shorter. This "end-stopping" phenomenon might come from lateral inhibition and might modify the hypercomplex cells type as instead "end-stopped" simple or complex cells.

**Architecture** Hubel and Wiesel also discovered in 1968 [53] that the cells in V1 are arranged in an orderly fashion, so that neurons with similar properties lie nearby one another. They are grouped in hypercolumns, long thin columns of cortical tissue about 1 mm square by 1 mm of thickness, containing cells with the same receptive field location, but different orientation selectivities, direction selectivities and both (left and right-) eye dominance.

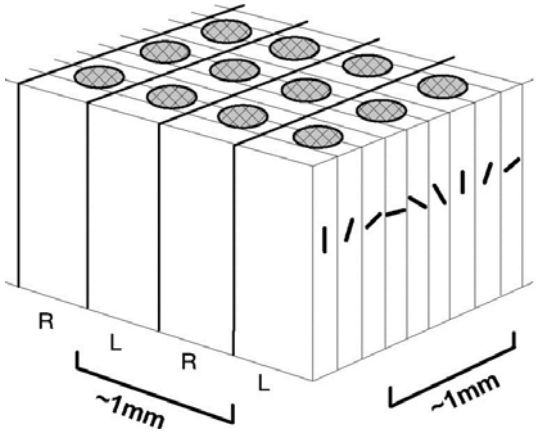



Figure A.18: The organization of an hypercolumn in the striate cortex. Modified from [132]

The fact that the processing of left and right retinal images is so close in each cortical hypercolumn indicates that disparity processing occurs at a very early stage along the visual pathway. This will be described later in this chapter.

The hypercolumn organization described by Hubel and Wiesel based on ocular dominance and orientation selectivity is still in dispute. Scientists converge on the subject of the precise shape of receptive fields in V1 and their specification, but questions remain on the functional significance of these cells: what are they doing in the HVS ?

On the basis of physiological evidence, De Valois and De Valois proposed in 1988 [31] that each hypercolumn processes an additional functional dimension. Orthogonal to the orientation dimension, a regular progression of size-scale values could exist from small to large. This has lead to a totally different macroscopic interpretation of the spatial processing by V1 that will be described in next section.

### The Psychophysical Channels

Psychophysics, the study of relations between people's conscious experience and properties of the physical world, measures people's performance in a perceptual task rather than record neural events. Thus psychophysicists methods instead rely on the study of sensitivity thresholds to variation of size and luminance of spots for example.

They have defined since 40 years the spatial frequency theory. It is based on an atomistic representation of an image as a set of primitive spatial atoms. The primitives are sinusoidal gratings characterized by their spatial frequency, orientation, amplitude

and phase. This image-based theory hypothesized that the visual system consist of many overlapping psychophysical channels that are selectively tuned to different ranges of spatial frequencies and orientations.

**Contrast Sensitivity Functions** Blakemore and Campbell [11] provide through experiments striking evidence for the existence of spatial frequency channels in the HVS. They measure the sensitivity of observers to specific spatial frequency gratings before and after adaptation. This sensitivity measurements results in **contrast sensitivity functions**. In practice, they measure the threshold at which a very low-contrast grating stops looking uniform and starts to look striped, at different spatial frequencies. The contrast sensitivity function is plotted in Figure A.19 (left) for an adult human. The shape of your own contrast sensitivity function to luminance gratings can be observed by looking at Figure A.19 (right). Then people are most sensitive to intermediate spatial frequencies of about

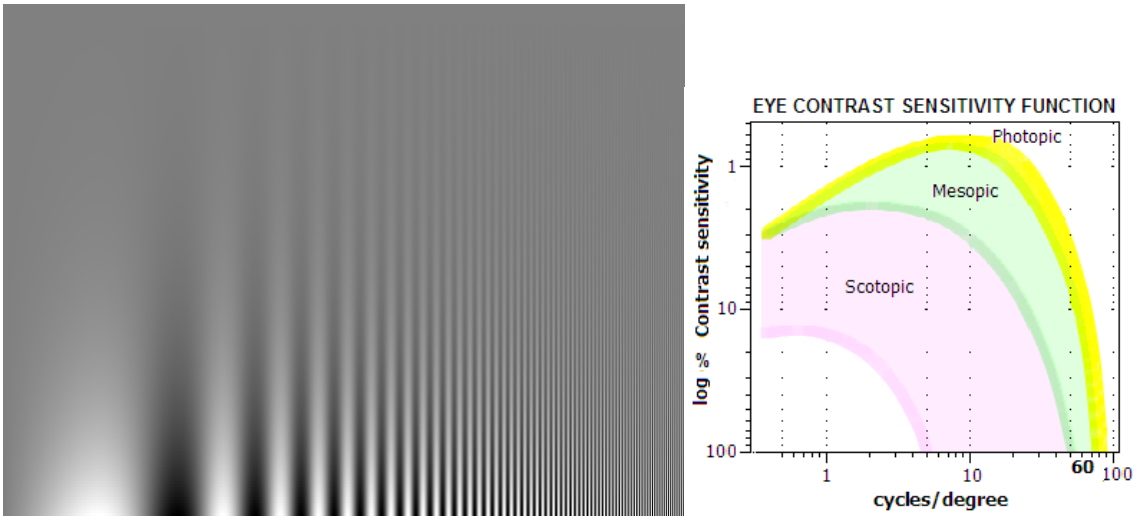



Figure A.19: Left: Contrast Sensitivity Functions as a function of spatial frequency (cycles/degree) vary with its illumination level. Right: Corresponding bell-shaped envelope can be demonstrated on this CSF chart for luminance gratings. Spatial frequency increases continuously from left to right, and contrast from top to bottom. From [16].

4 cycles per degree of visual angle. Interestingly, the CSF measured in low light conditions (scotopic) falls down, especially in the high frequencies. This can be explained by the poor spatial resolution of cones that operates in these conditions, and that are not present in the fovea of the retina, but are of greatest acuity in photopic conditions.

The selective adaptation of channels to particular spatial frequencies is proven by a second experiment where subjects adapt to a grating of a particular spatial frequency displayed for a few minutes. The sensitivity thresholds are remeasured, and show that the subject visual system adapts to and is less sensitive to a grating after prolonged viewing experience. This adaptation exists only near the particular spatial frequency and orientation of the adapting grating. The test is repeated at different spatial frequencies of adaptation and indicates that adaptation to the grating is not affected by test gratings of lower and higher frequency.

**Physiology** This nonlinear and selective adaptation to specific spatial frequency and orientation supports the idea of a psychophysical channel, where each separate channel

adapts to a degree reflecting its own selective sensitivity. Numerous other experiences have confirmed the spatial frequency theory as the dominant approach of spatial vision. However, the psychophysical channels are hypothetical processes inferred by psychophysiologists from behavioural measures and not biological processes of the nervous system. A second theory of **local spatial frequency analysis** has emerged from the function of cells discovered by Hubel and Wiesel. The striate cells, because they are spatially limited to a coverage of few degrees of visual angle, could realize a local, piecewise, spatial frequency analysis. The support for this analysis could be a Gabor function, patches of sinusoidal gratings fading-out with the distance, that would process local spatial frequency analysis. A two-dimensional Gabor function is illustrated in Figure A.20.

De Valois and De Valois [31] record the receptive field of striate cells and found evidence for multiple lobes of excitation and inhibition, as illustrated in Figure A.21. These recordings are very similar to profiles of Gabor functions and support their approach.

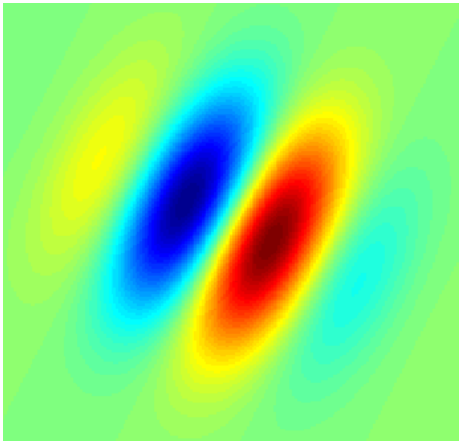


Figure A.20: Gabor filter-type receptive field for a simple cell. Blue regions indicate inhibition, red excitation. From [149].

De Valois also observed that the degree of tuning of cortical cells was continuous, some cells being sharply tuned and other broadly tuned. Simple cells are often more narrowly tuned than complex cells, and importantly the frequency and orientation tuning of these cells appear to be correlated. Then, by reconsidering the hypercolumn topology, De Valois and De Valois propose that hypercolumns define a two-dimensional space composed of spatial frequencies and orientations.

However, the model of the V1 receptive field by Gabor functions is controversial today because it does not conform to the anatomical layout of the visual system. It shorts-cuts the LGN and uses a 2D image as it is projected on the retina. Recently, a computational model of simple cells has been proposed [5], that implements the responses of model LGN cells with center-surround receptive fields. This model features cross orientation suppression, contrast invariant orientation tuning and response saturation, properties observed in real simple cells but not modelled by the Gabor function.

As we have seen, the spatial frequency theory is another interpretation of the properties of the V1 cells as filters rather than line and edge detectors. This theory is compatible with the line and edge detector proposed by Hubert and Wiesel. But even if the Gabor filters could be combined to form appropriate line and edge detectors, this combination could not appear at this stage but later along the visual pathway.

The nature of visual processing in higher level areas of cortex is less well understood than in area V1. We will focus in the next section on the processing of disparity and

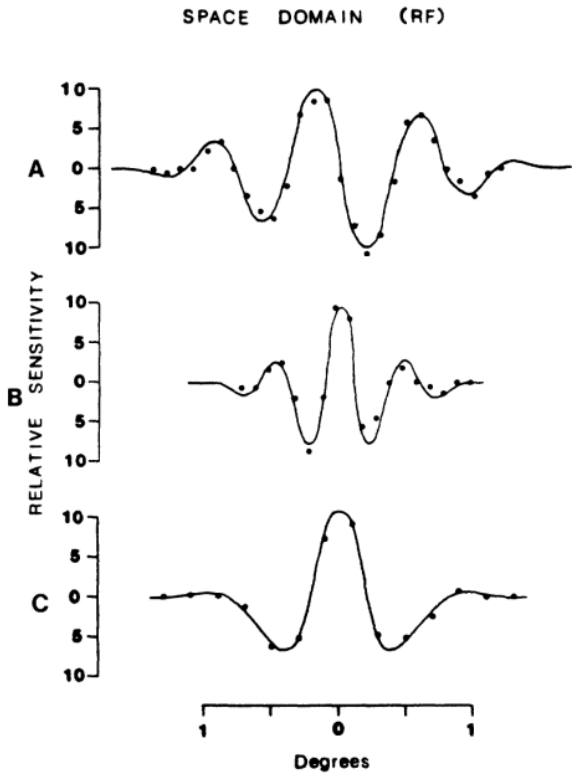



Figure A.21: Quantitative receptive field profile (in the space domain) of three simple cells. The receptive field profile was measured by recording the responses ($y$ axis) to a narrow and flickering black-white bar in different spatial positions ($x$ axis). The solid lines represent the profiles predicted by measuring the response to gratings of different spatial frequencies. From De Valois and De Valois [31] .

depth by the possible cortical areas. Before we could describe in detail what could be the processes that occur along the visual pathway to form the perception as a constructive act of spatial vision, we will introduce the computational modern approach to human vision.

## A.2 The Computational Approach to Vision

Modern theories of vision are considered within the foundation of the information processing theory. In this section, this paradigm in which modern theories of visual perception are cast will be described.

The information processing paradigm, as "a set of practices that define a scientific discipline at any period of time", theorizes about the nature of the human mind as a computational process. In this context, the nature of mental processes can be modelled by information processing events. As we have seen, Constructivists integrate this notion of information processing, while the Gibson theory of ecological optics supports the idea that visual perception is fully specified by optical information without mediating processes. Nevertheless, most vision scientists today propose their theory within the information processing framework.

### A.2.1 The Computer Metaphor

The invention of computers has strongly contributed to the development of the information processing theory.

First as a tool, the computer has served the implementation and the test of new theories of visual processing on "real" images. Second, within the information processing paradigm, the computer has served as an analogy for mental process. Then, mental processes such as visual perception are linked to the brain as programs are to the computer on which they run. In this view, brains are the "hardware" where minds are like programs or "software" for biological computation. This paradigm and its establishment through the development of computers has led this computer analogy to replace the theoretical analogies seen in classical theories of vision. Thus, the computer analogy is to the constructivism what the chemical analogy is to structuralist theory, what the field theoretical analogy is to structuralist theory or what the resonance analogy is to the information pickup theory of Gibson.

The information processing theorists proposed to decompose the description of an information systems in three levels of information:

- the Computational level, also called the conception level. This abstract stage defines the input information, the output ones and theorizes the process to pass from one to the other. It can be a mathematical statement, but not the mathematical function on how to achieve it. It is a functional definition or description of the inputs and outputs, and how inputs are formally related to the outputs.

- the Algorithmic level of description specify how a process is realized, by which operation. To construct an algorithm, a representation for the input and output information must be set. Processes will then transform the input representation into the output representation in a specified manner. Different algorithms are possible to realize a given computational description.

- the Implementation level specifies how an algorithm is implemented, embodied in a physical process within a physical system. Coming back to the analogy of brain to a biological computer, the same algorithm could be implemented on brains as in various kind of physical computers. As at the algorithmic level, there exist different ways to implement a dedicated algorithm in a physical manner.

These three levels of information have been theorized by the vision scientist David Marr in 1982 [83] in these terms and appear quite similar to the psychological approach of Palmer and Kimchi (1986) [102]. They theorized the assumptions underlying the information processing theories in cognitive psychology: the informational description of mental events as informational events, the recursive decomposition to describe a system into a hierarchy of components and the physical embodiment that bridges the gap between information, operations and the implementation on a real physical system.

The theoretical context given by these three levels of description of information processing systems helps tackling the interdisciplinary domain of vision. While the research at the computational level is done by the computer vision scientists, both computer vision and psychologists explore the decomposition of computational problems at the functional level. The first try to evaluate the matching of their algorithm in producing useful perception, while the second -psychologists- are more focused on evaluating how well the algorithm models human performance in experiments. Finally, physiologists try to understand how brains process the visual information at the implementation level while computer scientists evaluate their algorithm performance on electronic devices.

### A.2.2 The Four Stages of Visual Perception

The theoretical framework of "the four stages of visual perception" can be tackled in the light of the information processing approach. At the algorithmic level, this framework basically decomposes the visual perception into four stages, as illustrated in Figure A.22. Each stage describes the processes that operate on an input representation to compute an output representation. Up to now, it has constituted a solid framework to understand vision and its computational processes.

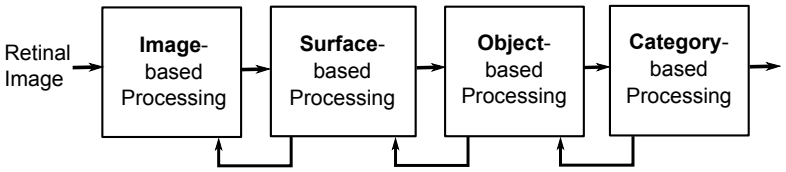



Figure A.22: The Four stages of the visual processing decompose the vision into four major stages beyond the retinal image itself: the image-based, surface-based, object-based and category-based processing.

1. The retinal representation is described in computational theories of vision as a homogeneous two-dimensional array of receptors, basically the 2D retina image of each eye.

2. The Image-Based stage includes among other mechanisms theoretical processes that are believed to be held in the striate cortex.Beyond the process of local edge detection, it includes their linking, the stereo matching between left and right eyes and the detection of other image-based features such as line terminations. At this image stage, the image local feature will be interpreted to form a global structure.

3. The Surface-Based stage deals with the extraction of intrinsic properties of visible surfaces from the features of the visual field represented in the image-based stage. The necessary perception of surface layout in the three dimensional environment was first intuitively proposed Gibson in 1950, while he was not an information processing theorist. He realized that before the perception of 3D object, the recovery of visible surface was a more primary task, but never suggest a dedicated representation. Marr [82] in 1977 and Barrow and Tennenbaum [6] proposed a surface-based representation and consequent algorithms to construct them from gray-scale images. Because it is halfway between the 2D structure of an image-based representation and the 3D structure of an object-based representation, Marr called it the 2.5D sketch. A flowchart illustrates (in Figure A.23) how this representation could be computed by a set of different processes extracting various information from the image-based representation.

4. The Object-based stage comes from behavioural tests showing that we have expectations about partly hidden surfaces of an object. There must be some form of three-dimensional inner representation in the visual system that allow us to infer the form of an object from a pair of retina images and its subsequent representations. The inclusion of information about unseen surfaces by inferences should then involve explicit representations of objects at the so named object-based stage.

5. The Category-based Stage states that at an higher cognitive-level, there must exist some processes of categorization of the perceived objects in the environment that

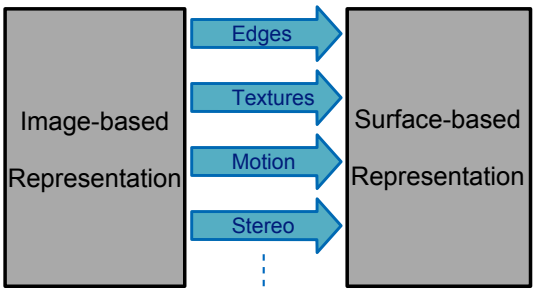



Figure A.23: The flowchart illustrates how a 2D image-based representation gives rise to a surface-based representation by a set of various parallel processes extracting surface orientation and depth information from stereo, motion, shading, edges, texture etc. It mainly originates from Marr and others [82].

would be necessary for human survival. Effortlessly and rapidly recovering some features of objects allow the perceiving organism to act appropriately in various situations. This implies that the ultimate goal of perception is to recover the functional properties of objects in their clutter environment.

## A.3 The Visual pathways

In the 1980s, a new hypothesis began to emerge to describe the overall architecture of early visual processing. It gives new insight on the relations between the anatomical structure, the physiological functions and the computational processes occurring in vision: the physiological pathway hypothesis.

It has been observed that the HVS is structured into separate pathways each processing different visual properties:

- form
- color
- motion
- depth or stereoscopic disparity

This proposal arose from several studies about higher visual areas. These suggest that distinct processing of different visual properties were involved in specific areas, like motion processing realized in MT area and color processing in V4 area (Zeki et al. in 1978 [6]) before projecting to the temporal and dorsal pathways. Later studies underline that this specialization had roots earlier in the visual system. Livingstone and Hubel follow the processing stream of earlier areas (retina and LGN) through pathways in higher cortical areas.

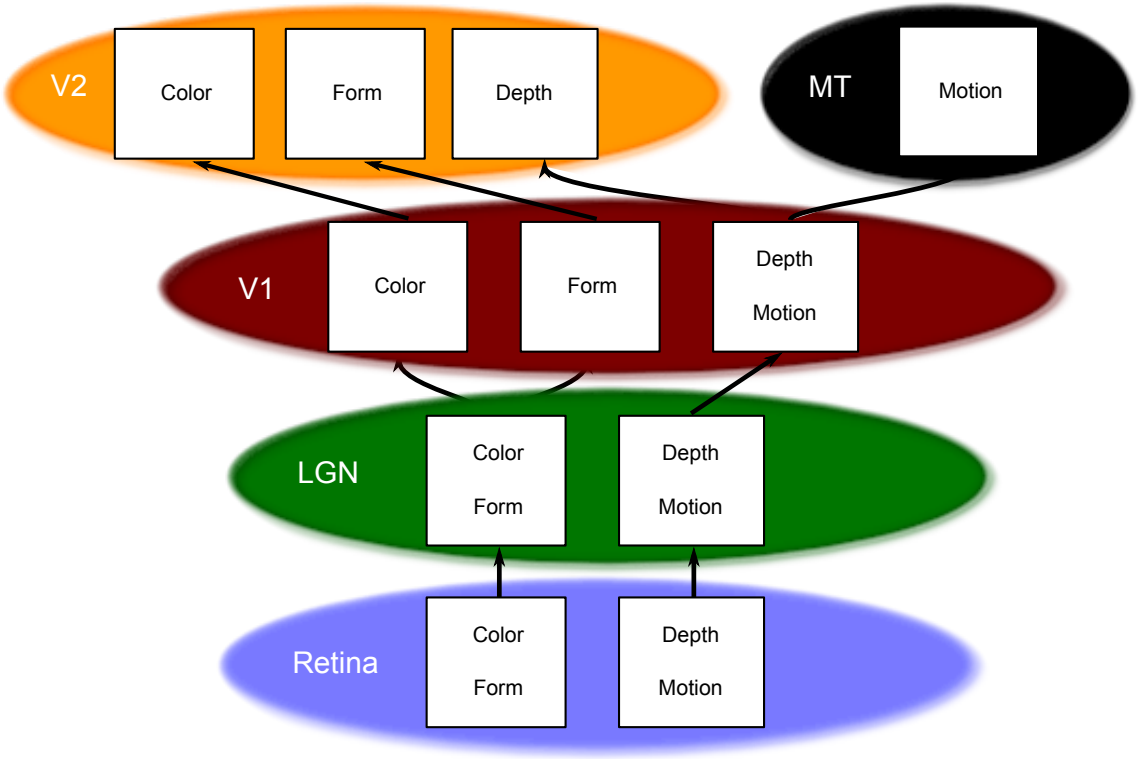



Figure A.24: Diagram of the four visual pathway hypothesis, where color, shape, motion and depth are supposed to be processed independently along the visual pathway and its areas.

They speculate on a structure of different functional visual pathways resting on a well-defined anatomical structure. The physiological functions along pathways are schematized

in Figure A.24, while the anatomical base of these functional pathways is represented in the Figure A.25 from Livingstone and Hubel [76].

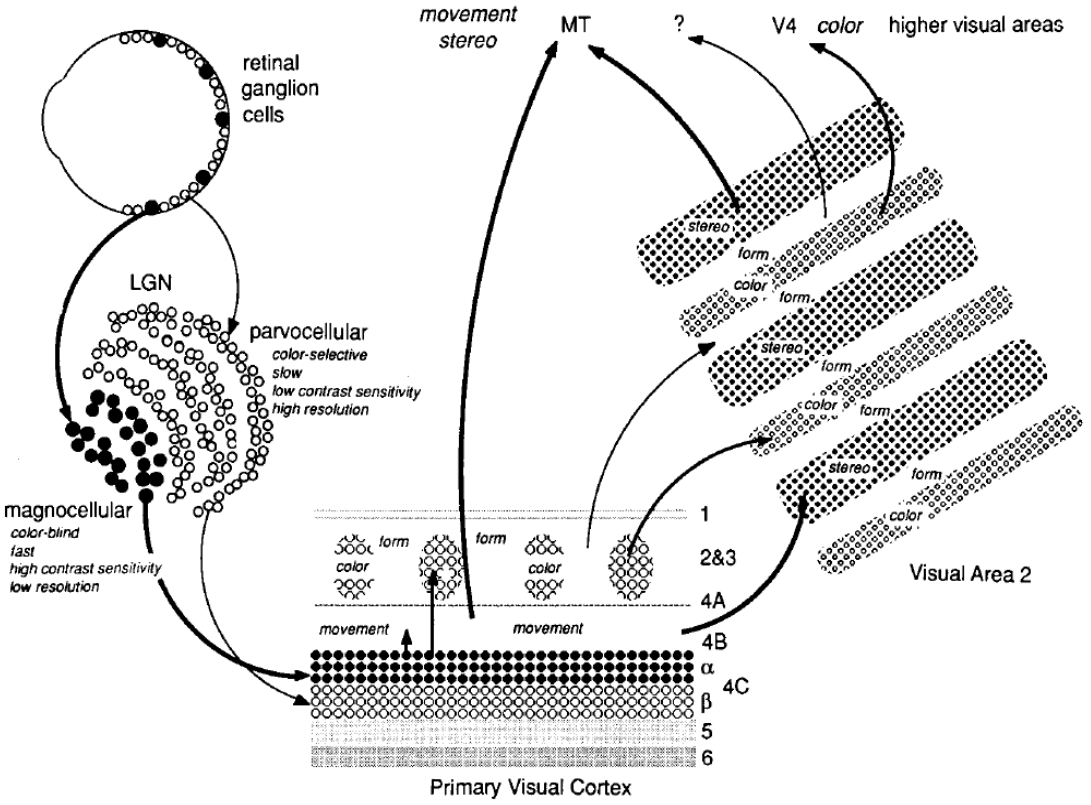



Figure A.25: The theory of functional pathways in the primate visual system by Livingstone and Hubel. They hypothesized that form, color, motion and stereo information is progressively separated along the visual pathway, from retina to the extrastriate visual cortex (V2 and beyond). From Livingstone and Hubel, 1988 [76].

Using single recordings, they observed that color, form, motion and stereo are processed in different subregions of V1 and V2, that project to distinct higher-level areas. They postulate that the M ganglion cells in the retina transmit motion and stereoscopic depth information and that the P ganglion cells transmit color and form information. As we have seen, the M and P cells are connected to the magno and parvocellular cells in the LGN, and as argued, they keep the functional splitting between motion and depth on one side, and color and form on the other. Livingstone and Hubel then proposed a further division in V1 after having observed that magno and parvocellulars cells from the LGN on to different sub-layers of the layer 4 of V1 (as illustrated in Figure A.25).

The V2 area also has a specific structure of thick and thin stripes interleaved with pale stripes. Livingstone and Hubel reported that the sub-layer 4B in V1 projects to the thick stripes, the blobs of layers 2 and 3 of V1 (illustrated by the dotted spots) project to the thin stripes in V2 and the interblobs in layers 2 and 3 of V1 project to the pale interstripe area of V2. The rest of the connections appears to be directly projected to V5 (or MT).

From these anatomical observations, Livingstone and Hubel [76] thus defined four potential functional pathways, traced in Table A.2.

This separation between pathways is importantly not complete, because there exist

| Pathway | retinal cells | LGN cells | V1 sub-layer | V1 2nd sub-layer | V2 stripe/MT | next area | next area |
|---|---|---|---|---|---|---|---|
| Stereo | M cells | LGN-magno | V1-4C$\alpha$ | V1-4B | V2-thick stripes | MT | IT... |
| Motion | M cells | LGN-magno | V1-4C$\alpha$ | V1-4B | MT | MST | ... |
| Color | P cells | LGN-parvo | V1-4C$\beta$ | V1-blobs | V2-thin stripes | V4 | ... |
| Form | P cells | LGN-parvo | V1-4C$\beta$ | V1-interb. | V2-pale stripes | V4 | IT... |

Table A.2: Hypothetical four functional pathways, defined by Livingstone and Hubel [76].

important crosstalks between these pathways. Even for a single pathway, the distinction between dedicated specific processes are far from clear, as will be described in the next section detailing the binocular neurons and associated areas.

There are strong claims however, that support the evidence of streams of processing. These are from the patients with brain damage first -as it helps in the past to understand the first visual functions- but also from psychophysical experiments.

Then, a study reports that patients having lost their ability to perceive motion didn't have difficulties perceiving color, depth or form [154], while in contrast some patients having lost color perception didn't report loss of motion or form or depth perception [28]. These facts support the idea of separate streams of processing, and seem difficult to explain differently.

Psychophysiologists Ramachandran and Gregory [110] realized experiments in which the stimuli present only differences along one dimension or one attribute. Randomly positioned green and red dots of equal luminance observed at little difference of position that would produce motion detection and then perception, appear thus to be static. But using black and white dots, the movement became detectable. Similar effects appear using equiluminant gratings of sinusoidally varying green and red patterns, where the perceived speed is much slower if not null. The color system then appears to be "motion-blind". A similar observation can be made between the separation of color and depth information: the perception of depth due to shading disappears with the same equiliminant light-to-dark to red-to-green transformation. Antagonist studies however show that the perception of motion [19] and depth [32] can be achieved with purely chromatic boundaries. The perceptual conclusions and uncertainties then appear to be like the physiological ones. The separation of visual information from different aspects into distinct pathways does not exist in its strong form. There exist connections and crosstalks among these pathways.

Also, up to now there is no clear understanding of the underlying processes that occur along these pathways. For example, from V1 through V2 up to V4 and IT, cells respond selectively to faces, hands and complex forms. But little is known on the way the local oriented spatial frequency filters of V1 process the information to detect face and hands in IT.

Thus the four visual pathway hypotheses can be seen as an integrative attempt to describe the basic separation that might occur in the first visual areas along the visual pathway. In the literature, it is now widely thought that the form and color pathways - the parvocellular pathway - project to the ventral "what" visual pathways. In contrast, the magnocellular pathway or depth and motion pathway is believed to be associated with the dorsal "where" pathway.

In the next section, we will try to give an insight into the processes of binocular depth perception occurring in vision and its supposed stereo pathway issued from these four visual pathway hypotheses.

### A.3.1 The stereo pathway

Binocular depth perception requires huge cognitive resources among the visual tasks we realize, and this precisely could explain why its processes start in the primary visual cortex. Because the two retinal images are obtained by two eyes (i.e binocular) whose viewpoint is shifted horizontally, they present differences called binocular disparities (see section 1.1). These differences must be overcome in order to be appropriately and precisely analysed. Indeed the HVS needs to be capable of registering a disparity smaller than the width of a cone photoreceptor. The operation of registration and integration of the binocular (also called "stereo") disparity between features from the left eye's image and the right eye's image is called stereo matching.

As we have seen, neurons respond to stimulation in their receptive fields by a tuning function for color, form, motion, and disparity. Recent studies have shown that neurons related to binocular depth perception respond more strongly to some binocular disparities than to others. This **disparity selectivity** can be described by a tuning function: the higher firing rate informs of the presence of a preferred binocular disparity.

In this section the hypothetical processes provided by the binocular vision (see chapter 1) will be mapped along the visual pathway and thereafter we review the functional areas processing the disparity and their hypothetical relationship to disparity scale and visual eccentricities.

#### Cortical sites

First -it would have been too simple- all regions of the visual cortex have neurons that respond to binocular disparity. Contrary to IT that is strongly believed to gather the information to process face and hands, no cortical site has been identify to be dedicated on binocular depth perception [103]. But as we will see hypotheses about specializations per visual areas have been pursued. Following the four visual pathway hypothesis, stereopsis is believed to be a unique element occurring along the magnocellular and the dorsal processing stream. However, other studies report that an important process occurs along the parvocellular and ventral stream, while others distinguish the coarse stereopsis dealt with by the dorsal pathway from the fine stereopsis [133, 117] believed to be processed by the ventral pathway.

A review of the recordings from binocular neurons of different visual areas of macaque monkeys trained to fixate a target (illustrated in Figure A.26(a)) on a random-dot figure (see section 1.1.1) with a varying disparity scale and eccentricity was realized by Cumming et al. [25]. As expected, the responses of binocular neurons can be described by Gabor functions with varying selectivity to disparity.

The plot of the disparity versus eccentricity of binocular neurons (see Figure A.27) indicates that there is no clear distinction between the range of disparity sensitivity and the location of a neuron along the dorsal (V5/MT) or ventral (V4) streams. But it seems clear that neurons from the V2 area process a higher range of disparity than neurons in V1, contributing to both the dorsal and ventral stream.

A conclusion is that the cortical site processing the binocular disparity are projections from binocular interactions initiated in V1. Also, there is no evidence of a coarser representation of depth beyond the V5/MT. Thus there exists distinct cortical sites in charge of the binocular depth processing. But in order to have a clear understanding of this process, one needs to better describe the format of the received signals by each area and which tasks they indeed support.

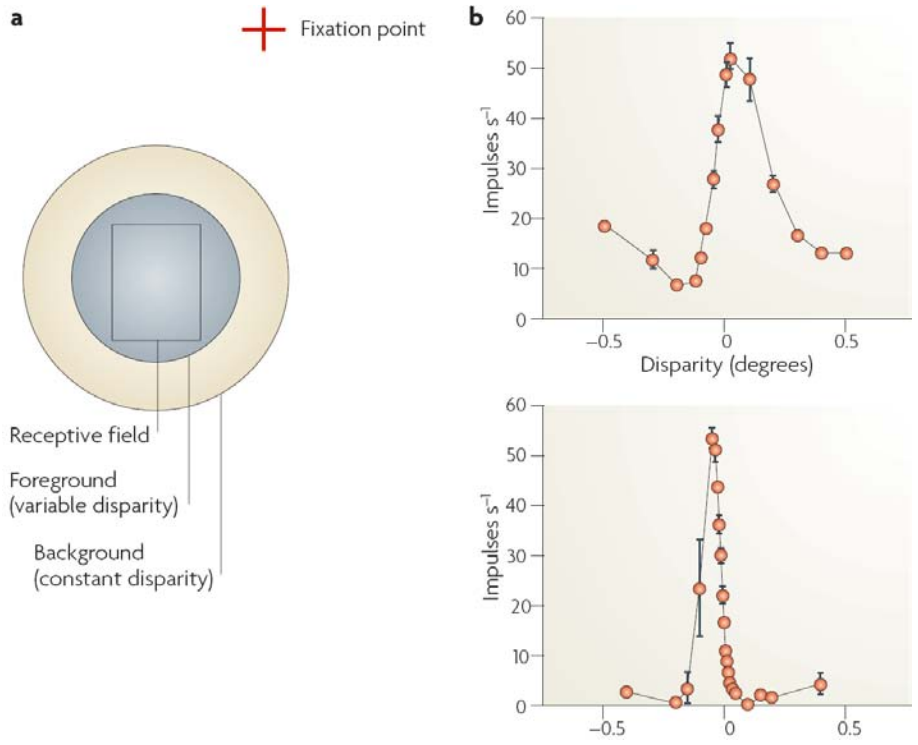



Figure A.26: Single-unit recording from binocular neurons, probed with random-dot figures (see section 1.1.1). (a): A target is fixed by a monkey, while different random-dot patterns, with different disparities are presented over the receptive field of a visual neuron. (b): Plot of the firing rate versus binocular disparity of binocular neurons from the V1 area. From [103].

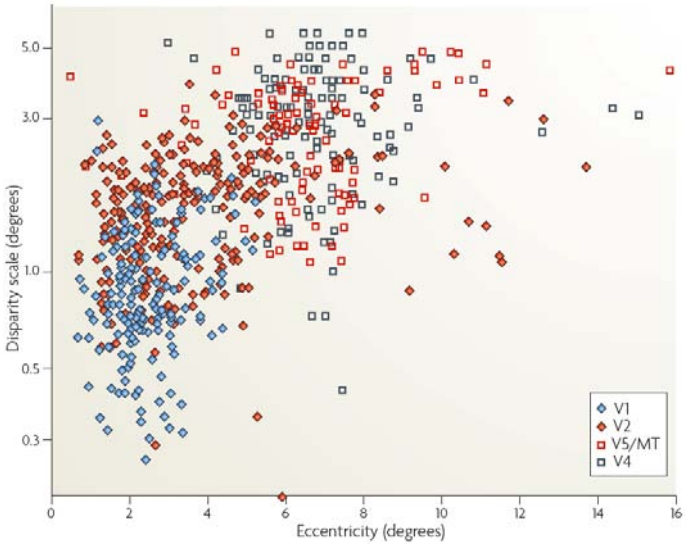



Figure A.27: Plot of the disparity scale (in degrees) -obtained from the frequency of the sine wave of the gabor functions recordings A.26- against visual eccentricity in the receptive field of different binocular neurons. From [103].

### Distinct tasks for distinct streams

Recent discussions on the contributions of the dorsal or ventral stream rejected the idea of a simple division of the coarse stereopsis in the dorsal stream and the fine stereopsis in

the ventral stream, but instead proposed a new hypothesis. A summary of these task-area associations from these numerous studies can be found in the Table A.3.

| | Cortical area | Anticorrelated response | Relative disparity |
|---|---|---|---|
| Early visual areas | V1 | Yes | None |
| | V2 | Yes, similar to V1 | Centre-surround |
| Ventral areas | V4 | Weaker than V1 | More than V2, centre-surround |
| | TEs | None | Sensitive to surface curvature |
| Dorsal areas | V5/MT | Similar to V1 | Surface slant and depth separation |
| | MST | Similar to V1 | Surface separation in depth |
| | CIP | Not determined | Surface slant |

Table A.3: Summary of the involvement of visual cortical areas in binocular depth perception, from [103]

The dorsal and ventral streams process different types of stereo computation. Each stream thus has a distinctive contribution to the realisation of the binocular depth extraction. The first stage of the visual pathway appears to process a direct computation of the binocular correlation between the left and right images. Because this early computation response to binocular anticorrelation by an inverted disparity response (the firing rates are inverted) [103], it is very plausible that this computation is a correlator.

The next V2 and the dorsal areas appear to rely on this "preprocessing" to compute a relative disparity (described in Section 1.1.1). The dorsal stream appears to compute the gradient of the extended surface and the segregation between surfaces. At a higher-level, the MST are selective for a link between optical flow computation and binocular depth which might contribute to the Motion parallax depth cues but also to the Kinetic Depth Effect (please refer to Section 1).

The tasks involved in the binocular computation along the ventral stream are much more complex. There might be there a full resolution of the stereo matching: neurons are specifically sensitive to relative depth between neighbouring features in the visual world. In the anterior inferotemporal cortex areas (TEs), neurons became sensitive to the shape and curvature of 3D surfaces.

### A.3.2 Conclusion

This section described the visual pathways through the four visual pathways hypothesis and detailed the pathways responsible for stereo depth perception and their connection to the ventral and dorsal stream.

Many studies advocated the existence of separated -but with cross-talk among them- visual pathways. Livingstone and Hubel speculate on the existence of four distinct pathways that appear to be wrong in detail on the basis of current scientific knowledge. But this integrative hypothesis precisely opened the way to physiological research and led to specific visual properties and functional areas of the human visual system.

Our understanding of the binocular system and associated visual areas has recently make significant progress. The pathways responsible for binocular depth perception are multiple and distinct, and operate in a multi-stage way. Dorsal and ventral streams perform different types of stereo computation, the ventral pathway contributes towards solving the multiple matching problems, while the dorsal pathway acts as a region-based cross-correlator [96].

## Conclusion

Light is a rich source of information about our spatial environment that the human visual system aims to exploit at best. The human vision manages to give a veridical perception of our environment in almost every cases, thanks to a learned and unconscious construction from bi-dimensional retinal images.

Following the retinal acquisition, the magnocellular and parvocellular ganglion neurons and LGN neurons are thought to be dedicated to fast-moving/disparity and high spatial/colour contrast stimuli respectively. These magnocellular and parvocellular are often associated with the dorsal and ventral cortical visual pathways respectively, simplified as the "where" and "what" pathways.

Livingstone and Hubel proposed that these two early magnocellular and parvocellular pathways initiate four functional pathways dedicated to color, form, binocular and motion processing. This perceptual separation of streams is supported by strong evidence, but does not suffice to explain the internal cortical processing.

The disparity or stereo function has been found to be processed in multiple stages along both the ventral and dorsal pathways. One hypothesis is that the dorsal and ventral streams are effectively adapted to different processing and consequently participate in different stages of the stereo computation.

Stereoscopy is however one of the multiple depth cues used to perceive the spatial structure of the scene. Questions are raised concerning the combination of multiple depth information sources in a global and coherent framework of depth perception. A long path remains towards the big picture.

# APPENDIX B

## Additional Subjective Results for Chapter 4

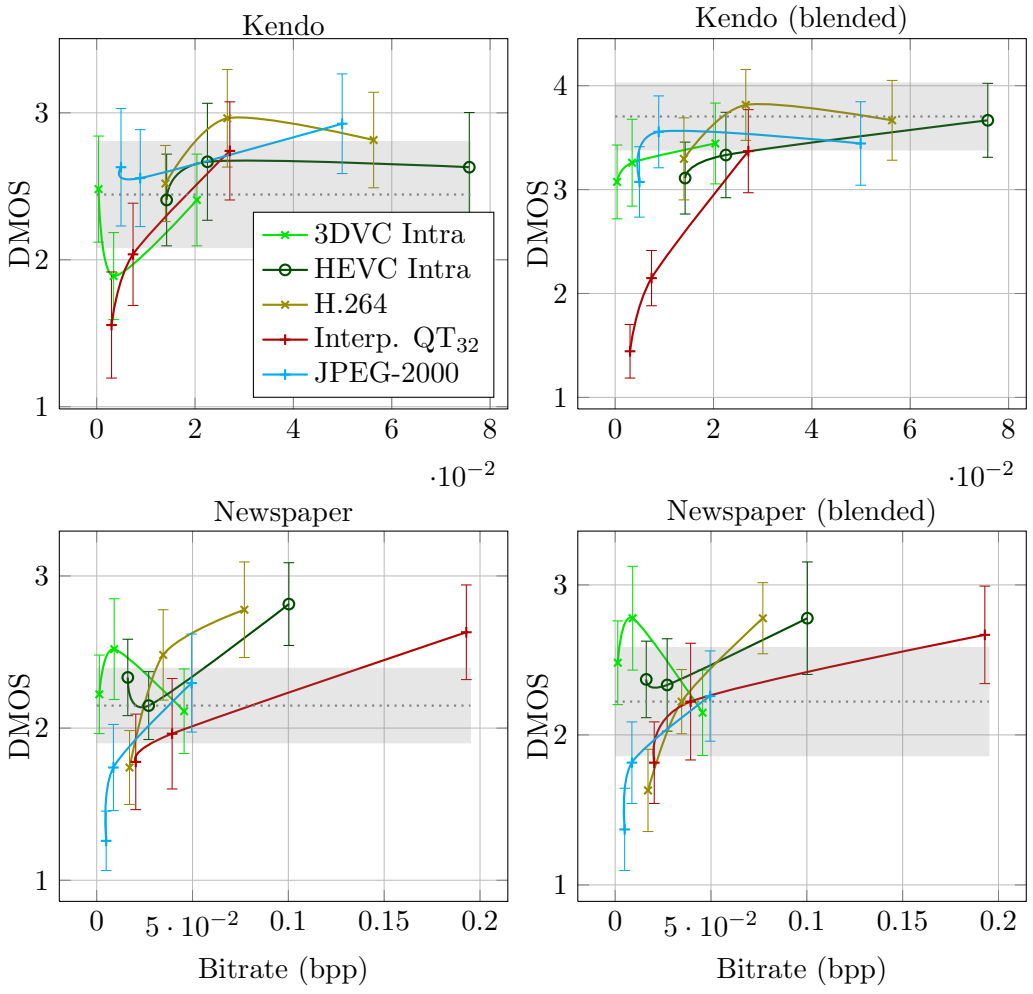



Figure B.1: Average DMOS reported by 27 observers on Class-C "Kendo" and "Newspaper" bullet-time synthesized videos for 3 classes of quality (high, middle and low quality) with 2 modes of rendering by the VSRS interpolation software (blending on/off to the right/left columns respectively) and for five depth map encoding methods.

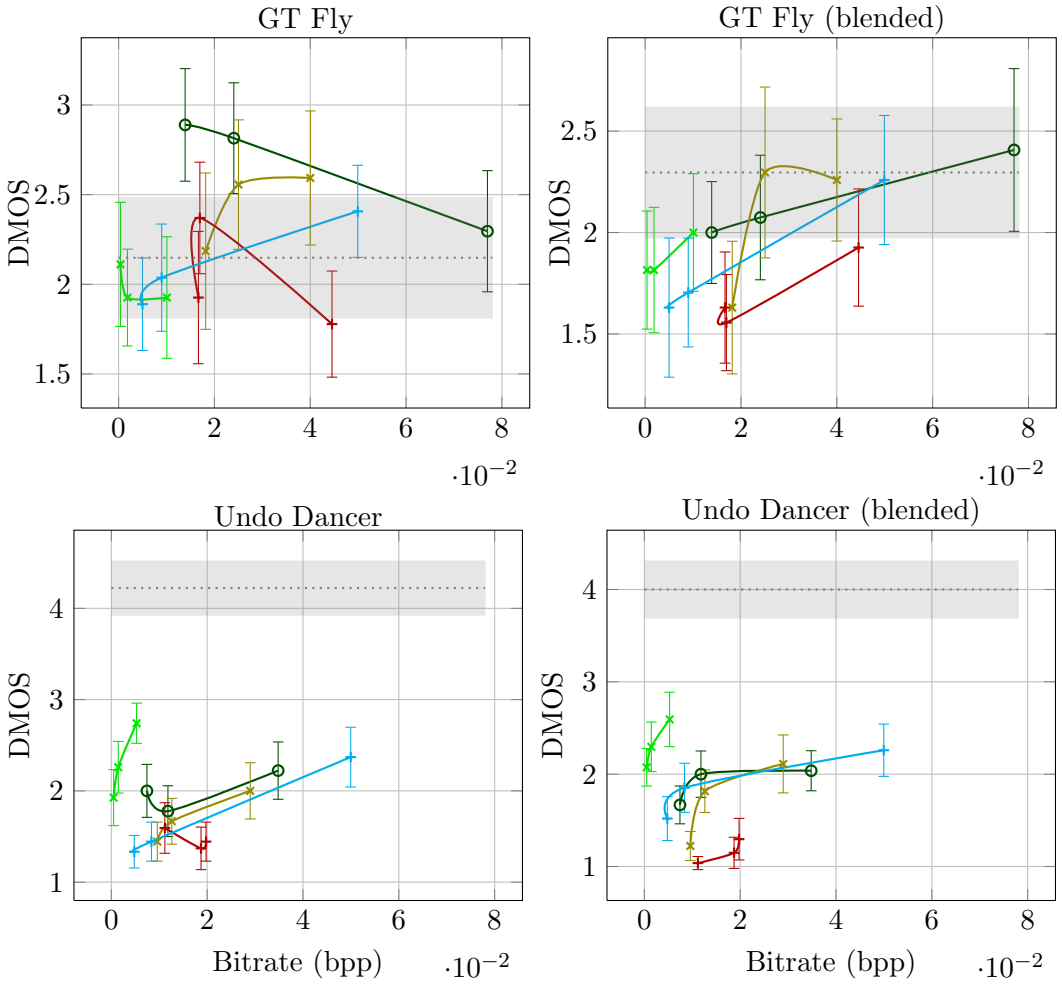



Figure B.2: Average DMOS reported by 27 observers on Class-A "Undo Dancer" and "GT Fly" bullet-time synthesized videos for 3 classes of quality (high, middle and low quality) with 2 modes of rendering by the VSRS interpolation software (blending on/off to the right/left columns respectively) and for five depth map encoding methods.

# List of Figures

# List of Tables

# Personal publications

## International Journal Paper

[1] Josselin Gautier and Olivier Le Meur. “A Time-Dependent Saliency Model Combining Center and Depth Biases for 2D and 3D Viewing Conditions”. In Cognitive Computation, Springer, Jan 1, 2012.

## International Conference Paper

[2] Josselin Gautier, Olivier Le Meur and Christine Guillemot. “Efficient depth map compression based on lossless edge coding and diffusion” in IEEE Picture Coding Symposium (PCS), 2012, p 81-84

[3] Olivier Le Meur, Josselin Gautier and Christine Guillemot. “Examplar-based inpainting based on local geometry”, 2011 18th IEEE International Conference on Image Processing (ICIP), p 3401-3404.

[4] Josselin Gautier, Olivier Le Meur and Christine Guillemot. “Depth-based image completion for view synthesis” in 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), IEEE, 2011,p 1-4.

## Domestic Conference Paper

[5] Josselin Gautier and Olivier Le Meur. “Un modÃ¨le de saillance dépendant du temps combinant les biais centré et de profondeur pour la visualisation en 2D et 3D” in Conference on Compression et REprésentation de Signaux Audiovisuels, Lille, 2012.

# Bibliography


[1] 3d4you. Online: http://www.3d4you.eu. 3

[2] 3dtv network of excellence. Online: http://www.3dtv-research.org. 3

[3] E.H. Adelson and J.R. Bergen. The plenoptic function and the elements of early vision. *Computational models of visual processing*, 1, 1991. 72

[4] A. Assee and N. Qian. Solving da vinci stereopsis with depth-edge-selective v2 cells. *Vision research*, 47(20):2585, 2007. 15, 16, 190

[5] G. Azzopardi and N. Petkov. A corf computational model of a simple cell that relies on lgn input outperforms the gabor function model. *Biological cybernetics*, pages 1–13, 2012. 173

[6] H.G. Barrow, J.M. Tenenbaum, SRI International. Artificial Intelligence Center. Computer Science, and Technology Division. *Recovering intrinsic scene characteristics from images*. Artificial Intelligence Center, SRI International, 1978. 176, 178

[7] Z. Belhachmi, D. Bucur, B. Burgeth, and J. Weickert. How to choose interpolation data in images. *SIAM Journal on Applied Mathematics*, 70(1):333–352, 2009. 105

[8] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424. ACM Press/Addison-Wesley Publishing Co., 2000. 128, 129, 130

[9] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher. Simultaneous structure and texture image inpainting. *Image Processing, IEEE Transactions on*, 12(8):882–889, 2003. 129

[10] M. Bindemann. Scene and screen center bias early eye movements in scene viewing. *Vision research*, 2010. xi, 35, 45

[11] C. Blakemore and FW Campbell. On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *The Journal of Physiology*, 203(1):237–260, 1969. 172

[12] D.E. Broadbent. Perception and communication. 1958. 153

[13] Bruce and J. Tsotsos. Saliency based on information maximization. *Advances in neural information processing systems*, 18:155, 2006. xii

[14] N.D.B. Bruce and J.K. Tsotsos. An attentional framework for stereo vision. In *Computer and Robot Vision, 2005. Proceedings. The 2nd Canadian Conference on*, pages 88–95. IEEE, 2005. xi, 34, 48

[15] J. Burge, C.C. Fowlkes, and M.S. Banks. Natural-scene statistics predict how the figure–ground cue of convexity affects human depth perception. *The Journal of Neuroscience*, 30(21):7269–7280, 2010. 29

[16] Published by Vladimir Sacek. Distribution of retinal photoreceptors, July 14. 2006. http://www.telescope-optics.net/eye_spectral_response.htm. 160, 172, 196, 197

[17] JC Carr, RK Beatson, JB Cherrie, TJ Mitchell, WR Fright, BC McCallum, and TR Evans. Reconstruction and representation of 3D objects with radial basis functions. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, page 76. ACM, 2001. 83

[18] A.D. Cate, M.A. Goodale, and S. Köhler. The role of apparent size in building-and object-specific regions of ventral visual cortex. *Brain research*, 1388:109–122, 2011. 23

[19] P. Cavanagh and S. Anstis. The contribution of color to motion in normal and color-deficient observers. *Vision research*, 31(12):2109–2148, 1991. 180

[20] C. Chamaret, J. C. Chevet, and O. Le Meur. Spatio-temporal combination of saliency maps and eye-tracking assessment of different strategies. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 1077–1080, 2010. 64

[21] X. Cheng, L. Sun, and S. Yang. Generation of layered depth images from multi-view video. In *IEEE International Conference on Image Processing, 2007. ICIP 2007*, volume 5, 2007. LDI. 75

[22] S. Cho, K. Yun, B. Bae, and Y. Hahm. Disparity-compensated coding using mac for stereoscopic video. In *Consumer Electronics, 2003. ICCE. 2003 IEEE International Conference on*, pages 170–171. IEEE, 2003. 86

[23] T. Colleu, S. Pateux, L. Morin, and C. Labit. A polygon soup representation for free viewpoint video. In *Proceedings of SPIE*, volume 7526, page 75260H, 2010. 79

[24] A. Criminisi, P. Pérez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212, 2004. xxiii, 128, 129, 130, 132, 133, 135, 136, 195

[25] BG Cumming and GC DeAngelis. The physiology of stereopsis. *Annual review of neuroscience*, 24(1):203–238, 2001. 181

[26] V. Cutsuridis. A cognitive model of saliency, attention, and picture scanning. *Cognitive Computation*, 1(4):292–299, 2009. 51

[27] J. E Cutting and P. M Vishton. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. *Perception of space and motion*, 5:69–117, 1995. 29, 34

[28] A. Damasio. *Principles of Behavioral Neurology*, chapter Disorders of complex visual processing: Agnosia, achromatopsia, Balint's syndrome and related difficulties of orientation and construction., pages 259–288. Philadelphia, PA: F.A. Davis, 1985. 180

[29] I. Daribo and B. Pesquet-Popescu. Depth-aided image inpainting for novel view synthesis. In *Multimedia Signal Processing (MMSP), 2010 IEEE International Workshop on*, pages 167–170. IEEE, 2010. 130, 136

[30] Ismael Daribo. *Codage et rendu de sequence video 3D; et applications Ã  la television tridimensionnelle (TV 3D) et a la television Ã  base de rendu de videos (FTV)*. PhD thesis, Telecom ParisTech, November 2009. xxiii, 70, 71, 120, 192, 194

[31] R.L. De Valois and K.K. De Valois. *Spatial vision*. Number 14. Oxford University Press, USA, 1988. 171, 173, 174, 197

[32] C.M.M. de Weert and KJ Sadza. New data concerning the contribution of colour differences to stereopsis. *Colour vision*, pages 553–562, 1983. 180

[33] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977. 51

[34] S. Di Zenzo. A note on the gradient of a multi-image. *Computer Vision, Graphics, and Image Processing*, 33(1):116–125, 1986. 132

[35] Nikos Drakos. Computer based learning unit, university of leeds, November 1999. http://fourier.eng.hmc.edu/e180/lectures/retina/node19.html. 165, 196

[36] A.A. Efros and T.K. Leung. Texture synthesis by non-parametric sampling. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1033–1038. Ieee, 1999. 128, 129

[37] L. Elazary and L. Itti. Interesting objects are visually salient. *Journal of Vision*, 8(3), 2008. 42

[38] C.C. Fowlkes, D.R. Martin, and J. Malik. Local figure–ground cues are valid for natural images. *Journal of Vision*, 7(8), 2007. 29

[39] M. Gervautz and W. Purgathofer. A simple method for color quantization: Octree quantization. In *Graphics Gems*, pages 287–293. Academic Press Professional, Inc., 1990. 82

[40] J.J. Gibson. The senses considered as perceptual systems. 1966. 152

[41] J.J. Gibson. The ecological approach to visual perception. *Boston: Houghtom Mifflin*, 1979. 152

[42] J.J. Gibson, G.A. Kaplan, H.N. Reynolds, and K. Wheeler. The change from visible to invisible. *Attention, Perception, & Psychophysics*, 5(2):113–116, 1969. 20

[43] H. Gray. *Anatomy of the human body*. Lea & Febiger, 1918. 166, 196

[44] A. Guzman-Arenas and A. Guzmaan. Computer recognition of three-dimensional objects in a visual scene. 1968. 26

[45] D.E. Hannula, D.J. Simons, and N.J. Cohen. Imaging implicit perception: promise and pitfalls. *Nature Reviews Neuroscience*, 6(3):247–255, 2005. 164, 196

[46] P. Harrison. A non-hierarchical procedure for re-synthesis of complex textures. In *Proc. Int. Conf. Central Europe Comp. Graphics, Visua. and Comp. Vision*, 2001. 128

[47] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*, volume 2. Cambridge Univ Press, 2000. 120

[48] Sebastian Bosse Heribert Brust Tobias Hinz Haricharan Lakshman Detlev Marpe Philipp Merkle Karsten Müller Hunn Rhee Gerhard Tech Martin Winken Thomas Wiegand Heiko Schwarz, Christian Bartnik. Description of 3d video technology proposal by fraunhofer hhi (hevc compatible, configuration a), November 2011. 88, 89, 90, 94, 193

[49] H. von Helmoltz. *Treatise on physiological optics*. New York: Dover publications, 1867/1925. 9, 153, 155

[50] T. Ho-Phuoc, N. Guyader, and A. Guerin-Dugue. A functional and statistical Bottom-Up saliency model to reveal the relative contributions of Low-Level visual guiding factors. *Cognitive Computation*, 2(4):344–359, 2010. xi, 37, 51, 55, 64

[51] H. Hoppe. View-dependent refinement of progressive meshes. In *Proc. Siggraph*, volume 97, pages 189–198, 1997. 77

[52] D.H. Hubel and T.N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574–591, 1959. 169

[53] D.H. Hubel and T.N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968. 170, 171, 196

[54] D.A. Huffman. Impossible objects as nonsense sentences. *Machine intelligence*, 6(1):295–323, 1971. 26

[55] H. Igehy and L. Pereira. Image replacement through texture synthesis. In *Image Processing, 1997. Proceedings., International Conference on*, volume 3, pages 186–189. IEEE, 1997. 128

[56] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259, 1998. xi, xii, 2, 34, 48

[57] S. Yea J. Heo, E. Son. Ce6.h region boundary chain coding for depth-map, July 2012. 102

[58] L. Jansen, S. Onat, and P. Konig. Influence of disparity on fixation and saccades in free viewing of natural scenes. *Journal of Vision*, 9(1), 2009. xi, 38, 39, 41, 42, 44, 47, 50

[59] V. Jantet. *Layered Depth Images for Multi-View Coding*. PhD thesis, IRISA / Université de Rennes 1 / ENS, 2012. 138, 139, 195

[60] V. Jantet, L. Morin, and C. Guillemot. Incremental-ldi for multi-view coding. 2008. LDI. 75

[61] T. Jost, N. Ouerhani, R. Wartburg, R. Muri, and H. Hugli. Assessing the contribution of color in visual attention. *Computer Vision and Image Understanding*, 100(1-2):107–123, 2005. 59

[62] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2106–2113. IEEE, 2009. xi, 36, 37, 51, 191

[63] B. Julesz. Foundations of cyclopean perception. 1971. 14, 189

[64] P. Kauff, M. Müller, F.Zilly, A.Smolic, and C.Varekamp. Deliverable d.2.1.2 : Requirements on post-production and formats conversion. Technical report, 3D4YOU consortium, 2008. public deliverable D1.1.2 available on the projects website: www.3d4you.eu. 74, 193

[65] S. Kircher and M. Garland. Progressive multiresolution meshes for deforming surfaces. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 191–200. ACM, 2005. 78

[66] K. Koffka. Principles of gestalt psychology. *Hartcourt, NY*, 1935. 152

[67] W. Köhler. *A sourcebook of Gestalt psychology*, chapter Physical Gestalten. New York: The Humanities Press, 1920-1950. 152

[68] M.S. Landy, L.T. Maloney, E.B. Johnston, and M. Young. Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision research*, 35(3):389–412, 1995. 29, 30

[69] O. Le Meur, T. Baccino, A. Roumy, et al. Prediction of the Inter-Observer visual congruency (IOVC) and application to image ranking. 2011. 44

[70] O. Le Meur and J. C Chevet. Relevance of a feed-forward model of visual attention for goal-oriented and free-viewing tasks. *Image Processing, IEEE Transactions on*, 19(11):2801–2813, 2010. 42, 59

[71] O. Le Meur and C. Guillemot. Super-resolution-based inpainting. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, 2012. 138

[72] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model bottom-up visual attentionnull. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(5):802–817, 2006. xii, 49

[73] M. Levoy and P. Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, page 42. ACM, 1996. light field. 72

[74] M. Levoy and T. Whitted. The use of points as a display primitive. *Tech. Report85-022, University of North Carolina at Chapel Hill*, 1985. Point based. 81

[75] D. Liu, X. Sun, F. Wu, S. Li, and Y.Q. Zhang. Image compression with edge-based inpainting. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(10):1273–1287, 2007. 129

[76] M. Livingstone and D. Hubel. Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science*, 240(4853):740, 1988. 2, 179, 180, 197, 199

[77] N.K. Logothetis. Vision: a window on consciousness. *SCIENTIFIC AMERICAN-AMERICAN EDITION-*, 281:68–75, 1999. 167, 196

[78] K. Suzuki N. Fukushima Y. Mori M. Tanimoto, T. Fujii. Reference softwares for depth estimation and view synthesis, April 2008. 109

[79] M. Mainberger and J. Weickert. Edge-based image compression with homogeneous diffusion. In *Computer Analysis of Images and Patterns*, pages 476–483, 2009. xix, 102, 104, 105, 194

[80] A. Maki, P. Nordlund, and J. O Eklundh. A computational model of depth-based attention. In *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, volume 4, pages 734–739, 1996. xi, 34

[81] A. Maki, P. Nordlund, and J. O Eklundh. Attentional scene segmentation: integrating depth and motion. *Computer Vision and Image Understanding*, 78(3):351–373, 2000. xi, 34

[82] D. Marr. Representing visual information. 1977. 13, 176, 177, 197

[83] D. Marr et al. Vision: A computational investigation into the human representation and processing of visual information. *New York: Henry Holt and Co*, 1982. 131, 175

[84] C. M Masciocchi, S. Mihalas, D. Parkhurst, and E. Niebur. Everyone knows what is interesting: Salient locations which should be fixated. *Journal of vision*, 9(11), 2009. 42

[85] W. Matusik and H. Pfister. 3d tv: a scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes. In *ACM SIGGRAPH 2004 Papers*, pages 814–824. ACM, 2004. 71

[86] L. McMillan Jr. *An image-based approach to three-dimensional computer graphics.* PhD thesis, Citeseer, 1997. 122

[87] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Müller, T. Wiegand, et al. The effects of multiview depth video compression on multiview rendering. *Signal Processing: Image Communication*, 24(1-2):73–88, 2009. 99

[88] P Merkle, Yannick Morvan, Aljoscha Smolic, Dirk Farin, Karsten Muller, and Thomas Wiegand. The effects of multiview depth video compression on multiview rendering. *Singal Processing: Image Communication*, 24(1-2):73–88, 2009. xix

[89] W. Metzger. Gesetze des sehens 2nd edition (frankfurt am main: Waldemar kramer). 1953. 29

[90] M. Mishkin, L.G. Ungerleider, and K.A. Macko. Object vision and spatial vision: Two cortical pathways. *Trends in neurosciences*, 6:414–417, 1983. 167

[91] Y. Mori, N. Fukushima, T. Yendo, T. Fujii, and M. Tanimoto. View generation with 3D warping using depth information for FTV. *Signal Processing: Image Communication*, 24(1-2):65–72, 2009. 126

[92] Y. Morvan, P.H.N. de Witha, and D. Farina. Platelet-based coding of depth maps for the transmission of multiview images. In *Proceedings of SPIE, Stereoscopic Displays and Applications*, volume 6055, pages 93–100, 2006. 97, 98, 100, 102, 193

[93] K. Müller, A. Smolic, K. Dix, P. Merkle, P. Kauff, and T. Wiegand. View synthesis for advanced 3d video systems. *EURASIP Journal on Image and Video Processing*, 2008:1–11, 2008. 126

[94] S. Nagata. How to reinforce perception of depth in single two-dimensional pictures. *Spatial Displays and Spatial Instruments*, 1987. 29

[95] K. Nakayama and S. Shimojo. Da vinci stereopsis: depth and subjective occluding contours from unpaired image points. *Vision research*, 30(11):1811–1825, 1990. viii, 14, 131

[96] HK Nishihara. Prism: A practical real-time imaging stereo matcher. *Optical Engineering*, 23(5):536–545, 1984. 183

[97] K. J. Oh, S. Yea, and Y. S. Ho. Hole filling method using depth based in-painting for view synthesis in free viewpoint television and 3-d video. In *Picture Coding Symposium, 2009. PCS 2009*, pages 1–4, 2009. xxiii, 130

[98] A. Oliva. Gist of the scene. *Neurobiology of attention*, 696, 2005. 64

[99] N. Ouerhani and H. Hugli. Computing visual attention from scene depth. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 1, pages 375–378, 2000. xi, 34

[100] K. Ozkan and M.L. Braunstein. Background surface and horizon effects in the perception of relative size and distance. *Visual cognition*, 18(2):229–254, 2010. 22

[101] S. Palmer. *Vision: From photons to phenomenology*. Cambridge, MA: MIT Press, 2000. 22, 23, 27, 42, 52, 163, 190, 191, 199

[102] S.E. Palmer and R. Kimchi. The information processing approach to cognition. *Approaches to cognition: Contrasts and controversies*, pages 37–77, 1986. 175

[103] A.J. Parker. Binocular depth perception and the cerebral cortex. *Nature Reviews Neuroscience*, 8(5):379–391, 2007. 2, 181, 182, 183, 197, 199

[104] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107–123, 2002. 42, 43, 51

[105] F. Pedersini, A. Sarti, and S. Tubaro. Multi-camera systems. *Signal Processing Magazine, IEEE*, 16(3):55–65, 1999. 120

[106] A. Pentland. Shape information from shading: a theory about human perception. *Spatial vision*, 4(2-3):2–3, 1989. 25

[107] R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(8):2397–2416, Aug 2005. 41, 57

[108] The-Crankshaft Publishing. Receptive fields of photoreceptors and their connections. http://what-when-how.com/neuroscience/visual-system-sensory-system-part-2/. 162, 196

[109] F. T Qiu, T. Sugihara, and R. von der Heydt. Figure-ground mechanisms provide structure for selective attention. *Nature neuroscience*, 10(11):1492–1499, 2007. 64

[110] VS Ramachandran and RL Gregory. Does colour provide an input to human motion perception? *Nature*, 1978. 180

[111] A. Redert, M.O. de Beeck, C. Fehn, W. IJsselsteijn, M. Pollefeys, L. Van Gool, E. Ofek, I. Sexton, and P. Surman. Advanced three-dimensional television system technologies. In *3D Data Processing Visualization and Transmission, 2002. Proceedings. First International Symposium on*, pages 313–319. IEEE, 2002. 3

[112] Wikipedia Rhcastilhos. Eye, September 2012. http://en.wikipedia.org/wiki/File:Schematic_diagram_of_the_human_eye_en.svg. 157, 195

[113] E. Rubin. *Visuell wahrgenommene figuren*. Gyldendalske boghandel, 1921. 131

[114] E. Rubin. *Visuell wahrgenommene figuren: Studien in psychologischer analyse*. Gyldendalske boghandel, 1921. xii, 52

[115] A. Saxena, S.H. Chung, and A. Ng. Learning depth from single monocular images. *Advances in Neural Information Processing Systems*, 18:1161, 2006. 60

[116] A. Saxena, M. Sun, and A.Y. Ng. Make3d: Learning 3d scene structure from a single still image. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(5):824–840, 2009. 60

[117] P.H. Schiller. The effects of v4 and middle temporal (mt) area lesions on visual performance in the rhesus monkey. *Visual Neuroscience*, 10(04):717–746, 1993. 181

[118] H. Schwarz and K. Wegner. Iso/iec jtc1/sc29/wg11 mpeg2011/n12744 mpeg output document, test model under consideration for hevc based 3d video coding v3.0, April 2012. 100, 101, 194

[119] HA Sedgwick. Space perception in handbook of perception and human performance vol i. kr boff, l kaufman, jp thomas eds. *New York: Wiley*, 21:19–21, 1986. 22

[120] J. Shade, S. Gortler, L. He, and R. Szeliski. Layered depth images. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 231–242. ACM New York, NY, USA, 1998. LDI. 75

[121] H.R. Sheikh and A.C. Bovik. Image information and visual quality. *Image Processing, IEEE Transactions on*, 15(2):430–444, 2006. 140

[122] A. Smolic, K. Mueller, P. Merkle, P. Kauff, and T. Wiegand. An overview of available and emerging 3d video formats and depth enhanced stereo as efficient generic solution. In *Proc. PCS*, 2009. 3D rep. 76, 77, 193

[123] G. Sperling. The information available in brief visual presentations. *Psychological monographs: General and applied*, 74(11):1, 1960. 153

[124] J.M. Steger. Fusion of 3d laser scans and stereo images for disparity maps of natural scenes. *Publications of the Institute of Cognitive Science*, 14, 2010. 39

[125] K.A. Stevens. Surface perception from local analysis of texture and contour. 1980. 24

[126] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori. Reference softwares for depth estimation and view synthesis, April 2008. 125

[127] B. W Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14), 2007. xi, 35, 40, 45, 47, 51, 55, 191

[128] B. W Tatler, R. J Baddeley, and I. D Gilchrist. Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, 45(5):643–659, 2005. 42, 43, 51

[129] R.B. Tootell, E. Switkes, M.S. Silverman, and S.L. Hamilton. Functional anatomy of macaque striate cortex. ii. retinotopic organization. *The Journal of Neuroscience*, 8(5):1531–1568, 1988. 169, 196

[130] A. Torralba, A. Oliva, M. S Castelhano, and J. M Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological review*, 113(4):766–786, 2006. 42

[131] D. Tschumperlé. Fast anisotropic smoothing of multi-valued images using curvature-preserving pde's. *International Journal of Computer Vision*, 68(1):65–82, 2006. 132

[132] D.Y. Ts'o, M. Zarella, and G. Burkitt. Whither the hypercolumn? *The Journal of Physiology*, 587(12):2791–2805, 2009. 171, 197

[133] C.W. Tyler. A stereoscopic view of visual processing streams. *Vision research*, 30(11):1877–1895, 1990. 181

[134] LG Ungerleider and M. Mishkin. *Analysis of visual behaviour*, chapter 18: Two cortical visual systems, pages 549–586. Cambridge MA: MIT Press, 1982. 167

[135] R. VanRullen. Visual saliency and spike timing in the ventral visual pathway. *Journal of Physiology-Paris*, 97(2-3):365–377, 2003. 64

[136] B. T Vincent, R. Baddeley, A. Correani, T. Troscianko, and U. Leonards. Do we look at lights? using mixture modelling to distinguish between low-and high-level factors in natural image viewing. *Visual Cognition*, 17(6):856–879, 2009. xi, xii, 37, 38, 51, 191

[137] H. Wallach and DN O'connell. The kinetic depth effect. *Journal of Experimental Psychology; Journal of Experimental Psychology*, 45(4):205, 1953. 19

[138] B.A. Wandell and J. Winawer. Imaging retinotopic maps in the human brain. *Vision research*, 51(7):718–737, 2011. 168, 196

[139] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612, 2004. 140

[140] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *Signals, Systems and Computers, 2003. Conference Record of the Thirty-Seventh Asilomar Conference on*, volume 2, pages 1398–1402. IEEE, 2003. 140

[141] M. Waschbüsch, S. Würmlin, D. Cotting, and M. Gross. Point-sampled 3d video of real-world scenes. *Signal Processing: Image Communication*, 22(2):203–216, 2007. 81, 82, 193

[142] J. Weickert. Coherence-enhancing diffusion filtering. *International Journal of Computer Vision*, 31(2):111–127, 1999. xxiv, 133

[143] M. Wexler and N. Ouarti. Depth affects where we look. *Current Biology*, 18(23):1872–1876, 2008. 63

[144] Y. Wexler, E. Shechtman, and M. Irani. Space-time completion of video. *IEEE transactions on PAMI*, pages 463–476, 2007. 135

[145] C. Wheatstone. Contributions to the physiology of vision.–part the first. on some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical transactions of the Royal Society of London*, 128:371–394, 1838. 2

[146] C. Wheatstone. Contributions to the physiology of vision.–part the second. ons some remarkable, and hitherto unobserved, phenomena of binocular vision (continued). *Philosophical Transactions of the Royal Society of London*, 142:1–17, 1852. 2

[147] W. Wichmann and W. Müller-Forell. Anatomy of the visual system. *European journal of radiology*, 49(1):8–30, 2004. 168, 196

[148] Wikipedia. Neuron, September 2012. http://en.wikipedia.org/wiki/File:Complete_neuron_cell_diagram_en.svg. 158, 196

[149] Joe Pharos Wikipedia. Gabor filter, November 2006. http://en.wikipedia.org/wiki/File:Gabor_filter.png. 173, 197

[150] S.U. Yoon, E.K. Lee, S.Y. Kim, Y.S. Ho, K. Yun, S. Cho, and N. Hur. Coding of layered depth images representing multiple viewpoint video. In *Proc. of Picture Coding Symposium (PCS) SS3-2*, 2006. LDI. 75

[151] Y. Zhang, G. Jiang, M. Yu, and K. Chen. Stereoscopic visual attention model for 3D video. *Advances in Multimedia Modeling*, pages 314–324, 2010. xi, xii, 34

[152] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of vision*, 11(3), 2011. 36, 37, 51

[153] L. Zhaoping, N. Guyader, and A. Lewis. Relative contributions of 2D and 3D cues in a texture segmentation task, implications for the roles of striate and extrastriate cortex in attentional selection. *Journal of vision*, 9(11), 2009. 51

[154] J. Zihl, D. Von Cramon, and N. Mai. Selective disturbance of movement vision after bilateral brain damage. *Brain*, 106(2):313–340, 1983. 180

[155] C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. In *International Conference on Computer Graphics and Interactive Techniques*, pages 600–608. ACM New York, NY, USA, 2004. LDI. 74, 108, 126