Tout est dans le regard : reconnaissance visuelle du comportement humain en vue subjective

Francis Martinez

UNIVERSITÉ PIERRE ET MARIE CURIE - PARIS 6

ÉCOLE DOCTORALE ED SMAER

T H È S E

présentée par

Francis MARTINEZ

pour l'obtention du grade de

Docteur ès Sciences

de l'Université Pierre et Marie Curie - Paris 6

Mention : Informatique et Image

Tout est dans le regard : reconnaissance visuelle du comportement humain en vue subjective

soutenue publiquement le 9 juillet 2013

Composition du jury

Pr. Liming CHEN	-	LIRIS, Ecole Centrale de Lyon	Rapporteur
Dr. Jean-Marc ODOBEZ	-	IDIAP-EPFL, Martigny, Suisse	Rapporteur
DR. François BRÉMOND	-	INRIA Sophia-Antipolis	Examinateur
Pr. Alice CAPLIER	-	GIPSA-lab, Grenoble INP	Examinatrice
Pr. Matthieu CORD	-	LIP6, UPMC, Paris	Examinateur
Pr. Bruno GAS	-	ISIR, UPMC, Paris	Examinateur
Pr. Edwige PISSALOUX	-	ISIR, UPMC Paris	Directrice de thèse
Dr. Andrea CARBONE	-	CHArt LUTIN, Univ. Paris 8	Co-encadrant

Résumé

Dans ce manuscrit, nous nous intéressons à l'analyse visuelle du comportement humain à partir de l'information du regard. A l'inverse des caméras statiques et externes, nous adoptons un point de vue subjectif, ce qui permet de placer le contexte d'étude au centre de l'être humain et de ses interactions avec l'environnement. Pour atteindre cet objectif, nous avons développé un eye-tracker porté, ainsi que des outils d'analyse associés, en particulier la reconnaissance d'attention dans le cadre d'interactions sociales et la reconnaissance d'activités subjectives.

Dans la première partie de cette thèse, nous présentons un eye-tracker binoculaire tête porté à partir duquel nous estimons le regard du sujet. Contrairement à la plupart des systèmes basés sur l'éclairage infrarouge, notre approche fonctionne en éclairage visible. Pour cela, nous nous inspirons des méthodes basées apparence qui, au lieu, d'extraire des caractéristiques géométriques (par exemple, la pupille), exploitent l'image de l'œil dans sa globalité et elles permettent donc de prendre en compte toutes les caractéristiques de l'œil. Pour apprendre la relation entre les caractéristiques d'apparence et les coordonnées du point de regard dans l'image de la caméra scène, deux modèles de régression sont comparés : le Support Vector Regression et le Relevance Vector Regression.

Nous proposons, ensuite, une nouvelle méthode de reconnaissance d'attention en vue subjective. Le regard subjectif est obtenu à l'aide de notre eye-tracker, tandis que le regard d'autrui est construit à partir de l'estimation de l'orientation de la tête par régression à noyaux multiples localisés. En combinant ces deux types de regard, nous calculons alors des scores d'attention qui permettent d'identifier des motifs attentionnels dyadiques tels que le regard mutuel, mais aussi des motifs d'ordre supérieur émanant de la nature triadique de notre expérience.

Notre outil final d'analyse concerne la reconnaissance d'activités basée sur le regard et l'égo-mouvement. Ces mouvements sont quantifiés en fonction de leur direction et de leur amplitude et encodés sous forme de symboles. Des caractéristiques statistiques sont alors extraites via un codage multi-échelle et un partitionnement temporel. Pour la classification et la segmentation d'activités, nous décrivons une approche par apprentissage contextuel en intégrant des scores de prédiction d'un voisinage à longue portée. Une étude détaillée permet également de comprendre quelles caractéristiques jouent un rôle prédominant dans la représentation d'une activité.

Mots-clés : suivi du regard ⋆ estimation de la pose de la tête ⋆ vue subjective ⋆ reconnaissance d'attention et d'activités ⋆ modèle d'apparence ⋆ mouvements oculaires ⋆ régression ⋆ classification ⋆ apprentissage contextuel ⋆ égo-mouvement

Abstract

In this thesis, we focus on understanding human behavior from gaze information. In contrast to static and external camera viewpoint, we adopt a first-person point of view that allows carrying studies centered on humans and their interaction with the environment. To fulfill this goal, we developed a head-mounted eye-tracker and analysis tools for attention recognition during social interactions and for gaze-based first-person activity recognition.

In the first part of the thesis, we present a head-mounted binocular eye-tracker from which we infer the subject's gaze. Contrary to infrared systems, our approach works under visible light. Instead of extracting geometric features (e.g. pupil), we propose to use an eye appearance model in order to capture all available eye features. To learn the mapping between eye appearance and point of regard, two regression models are compared: Support Vector Regression and Relevance Vector Regression.

Then, we propose a novel approach for attention recognition from first-person vision. The first-person gaze is obtained using our eye-tracker, while the third-person gaze is computed from head pose estimation based on localized multiple kernel regression. Knowing the first- and third-person gaze direction, scores are computed which permit to assign dyadic attention patterns such as mutual gaze, and at the same time, higher-order patterns due to the triadic nature of the experiment.

Our final analysis tool involves activity recognition from first-person gaze and ego-motion. These motions are quantized according to their direction and their amplitude, and are encoded into a sequence of symbols. Statistical features are then extracted via multi-scale and temporal representation. For joint classification and segmentation of activities, we describe a contextual learning approach built upon confidence values from long-range neighborhood. Additionally, an in-depth study allows highlighting which features are relevant to each activity.

Keywords: gaze tracking ⋆ head pose estimation ⋆ first-person vision ⋆ attention and activity recognition ⋆ appearance model ⋆ eye movements ⋆ regression ⋆ classification ⋆ contextual learning ⋆ ego-motion

Remerciements

Tout d'abord, je tiens à remercier ma directrice de thèse, Edwige Pissaloux, pour son soutien, ses conseils et ses encouragements tout au long de ces trois années de thèse. Un grand merci aussi à mon co-encadrant, Andrea Carbone, pour son aide et les discussions techniques que nous avons pu partager.

Je suis également reconnaissant envers le directeur de l'ISIR, Philippe Bidaud, qui m'a accepté au sein de son laboratoire afin je puisse réaliser mes travaux de thèse.

Je tiens à exprimer toute ma gratitude à M. Liming Chen et M. Jean-Marc Odobez pour avoir accepté d'être les rapporteurs de ma thèse et pour avoir consacré une partie de leur temps à la relecture de mon manuscrit. Leurs commentaires et leurs conseils ont été très enrichissants. Je remercie également chaleureusement M. François Brémond, Mme Alice Caplier, M. Matthieu Cord et M. Bruno Gas d'avoir accepté de participer à mon jury. Je leur en suis très reconnaissant.

Je souhaite aussi remercier M. Thierry Baccino, Professeur à l'Université Paris 8 et directeur scientifique du LUTIN, pour m'avoir gentillement prêté le système eye-tracker SMI et m'avoir permis d'effectuer les expériences au sein de son laboratoire. Merci aux participants du LUTIN.

Mes remerciements vont également à mes collègues de bureau: Didier, Zefeng, Xiao. ainsi qu'aux étudiants qui ont accepté de participer à mes expériences. Merci aussi aux personnels de l'administration de l'ISIR: Michèle, Ludovic, Adela et Sylvie, ainsi qu'à M. Stéphane Régnier et Mme Marie Aubin de l'école doctorale. Je remercie également Mme Viviane Pasqui pour m'avoir donné l'accès à la salle Mouv' afin de réaliser mes expériences.

Je remercie aussi la Commission Européenne au travers du projet européen FP7 AsTeRICS, ainsi que le CNRS via la mission DEFISENS, pour leur soutien financier accordé à mes travaux de recherche et à mes déplacements à l'étranger.

Un immense merci à mes amis que j'ai eu la chance d'avoir à mes côtés: Alex, Magic Bram's, Nico, Béné, Sergio et Ludo pour les moments inoubliables que nous avons passé ensemble. Restez comme vous êtes !

Enfin, je tiens à exprimer toute ma reconnaissance envers mes parents et ma soeur pour leur soutien et leur affection sans lesquels l'accomplissement de cette thèse aurait été rendue plus difficile.

Table des matières

Acronymes

GPR	*Gaussian Process Regression*
HMM	*Hidden Markov Model*
HOG	*Histogram of Oriented Gradients*
KPLS	*Kernel Partial Least Squares*
KRR	*Kernel Ridge Regression*
LMKL	*Localized Multiple Kernel Learning*
LRR	*Linear Ridge Regression*
MAAE	*Mean Absolute Angular Error*
mAP	*mean Average Precision*
MSE	*Mean Squared Error*
PCR	*Pupil Corneal Reflection*
PoR	*Point of Regard*
SVM	*Support Vector Machine*
SVR	*Support Vector Regression*
RVR	*Relevance Vector Regression*

CHAPITRE 1

Introduction

"Le véritable voyage de découverte ne consiste pas
à chercher de nouveaux paysages,
mais à avoir de nouveaux yeux."

– Marcel Proust –

Sommaire

L'ANALYSE et la compréhension du comportement humain à partir de signaux visuels est une faculté dont dispose l'être humain. En effet, nous autres êtres humains pouvons aisément reconnaître les activités effectuées par une personne. Par exemple, en seul coup d'oeil, nous pouvons remarquer si une personne téléphone, mange, intéragit avec d'autres personnes etc. Par ailleurs, nous arrivons à anticiper certaines actions : par exemple, si une personne fait tomber ses clés, il est fort probable qu'elles les ramassent. Ceci étant dit, cette capacité d'analyse, aussi simple soit-elle, est en réalité très complexe et est le fruit de notre expérience, de notre vécu que nous assimilons au fil des années.

Partant de ce constat, l'homme, dans sa quête à concevoir des machines à l'image de lui-même, s'est intéressé à doter les machines de moyens d'acquisition et d'interprétation de données dans le but d'en faire des *systèmes intelligents* et pouvant réaliser des tâches à sa place.

De plus, en parallèle, l'évolution de notre société fait que de nouvelles technologies deviennent accessibles à l'être humain. Effectivement, ces derniers temps, nous entrons de plus en plus dans une ère dite *wearable and pervasive*, c'est-à-dire où les systèmes sont portables et omniprésents dans notre environnement. Ces technologies permettent alors d'ouvrir de nouvelles perspectives et nécessitent souvent le développement de nouveaux outils afin de traiter ces données et de proposer des fonctionnalités novatrices aussi bien pour étudier le comportement humain que pour assister une personne dans l'accomplissement de certaines tâches.

Dans ce manuscrit, nous proposons des approches pour la reconnaissance visuelle et automatique du comportement humain à partir de l'information du regard. Avant d'aborder le sujet plus en détails, nous apportons quelques réfléxions qui laissent penser que cette thématique reste de nos jours un enjeu important. Ensuite, nous présentons les questions sur lesquelles nous nous sommes penchés et auxquelles nous avons tenté d'apporter des solutions. Puis, à la fin de ce chapitre, nous énumérons succinctement les principales contributions de ce manuscrit.

1.1 Motivations

Trois thématiques majeures viennent nous conforter dans nos choix scientifiques (cf. figure 1.1) :

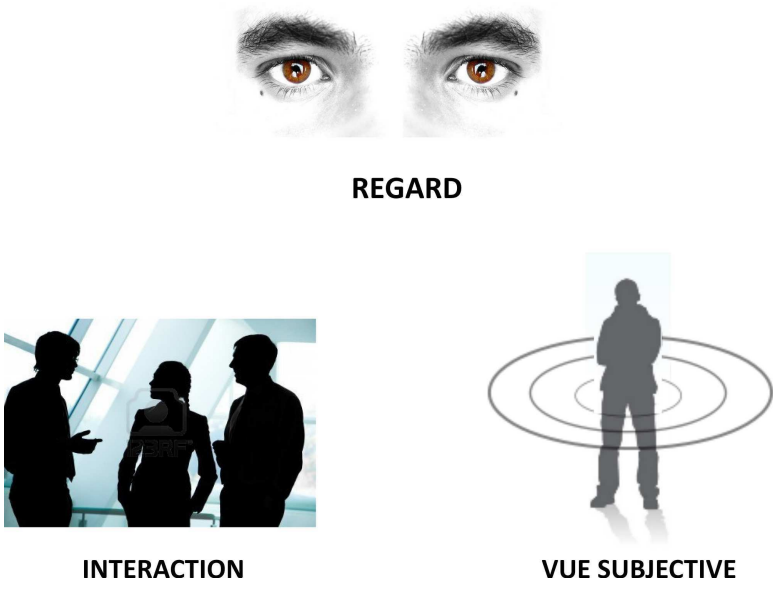


Figure 1.1 – Thématiques majeures : *Regard*, *Interaction* et *Vue subjective*.

Le regard - La direction des yeux vers un objet, plus connu sous le nom de regard, possède des informations riches : il est possible de comprendre et d'étudier les comportements humains, que ce soit leurs attitudes, leurs intérêts ou leurs intentions. En effet, le regard précède ces différents comportements humains : une personne regarde, en général, toujours à un endroit précis avant d'agir ou d'interagir.

Sur le plan personnel, le regard permet à un individu d'explorer son environnement en identifiant des points d'intérêt et fournit des informations sur la perception visuelle de scènes. Dans un contexte social, le regard agit comme un support de communication non-verbale et il constitue donc un outil indispensable pour étudier les interactions sociales (Argyle, 1969). En effet, de nombreux psychologues ont démontré que le regard permet de réguler les conversations entre les êtres humains, grâce à des mécanismes tels que le regard mutuel (Argyle *et al.*, 1973). La durée du regard, la fréquence des coups d'oeil, les trajectoires de fixation, la dilatation de la pupille ou encore les clignements des yeux sont autant de signaux importants qui permettent de décrypter la communication non-verbale.

Depuis des décennies, nombreux sont les chercheurs qui ont proposé des outils de suivi et d'analyse automatique du regard. Récemment, l'intérêt autour du regard a connu un essor considérable comme en témoigne le récent état-de-l'art (Hansen et Ji, 2010) et divers domaines font appel à l'information du regard : l'interaction homme-machine (IHM), les études psychologiques (par exemple, l'autisme), les technologies d'assistance, l'aide à la conduite automobile, l'analyse marketing et bien d'autres encore. Certains exemples d'application sont donnés dans la figure 1.2.

(a) (b)

(c) (d)

Figure 1.2 – Diverses applications à partir du suivi du regard : (a) Assistance à l'aide à la conduite (*source : Tobii*), (b) Analyse marketing (*source : SMI*) (c) Interaction avec un smartphone (*source : Samsung*), (d) Simulateur de vol (*source : Polhemus*).

Interaction - L'interaction est l'action réciproque qui s'exercent entre deux ou plusieurs systèmes physiques. Les processus d'interaction sont intéressants à étudier, en ce sens où ils permettent d'en savoir plus sur le comportement des systèmes concernés. Dans le cas de l'être humain, celui-ci interagit naturellement avec son environnement et ses interactions peuvent être du type personnelle (interaction homme-objet) ou interpersonnelle (interaction homme-homme). Cependant, ces interactions restent complexes et difficiles à analyser de part leur diversité (interaction multimodale, avec différents objets, avec un être humain), mais aussi de part leur nombre (interaction avec plusieurs personnes).

Vue subjective - Un point de vue est un aspect sous lequel une personne se place pour examiner une chose. Pour être plus précis, il existe différents points de vue qui traduisent chacun une manière d'interpréter les choses : le point de vue *objectif* et le point de vue *subjectif*. Nous retrouvons d'ailleurs ces concepts dans divers domaines allant de la littérature aux jeux vidéos, en passant par la peinture ou la cinématographie.

Ces derniers temps, la vue subjective (ou perspective subjective[1]) a connu un intérêt

[1]En anglais, nous retrouvons souvent les termes *first-person vision* ou *egocentric vision*.

considérable en partie dû à la miniaturisation des caméras[2] et dans le même temps, à leur utilisation de plus en plus fréquente par des particuliers[3] (par exemple, pour enregistrer des souvenirs vidéo). Elle se traduit par des données acquises à partir d'une caméra dont est équipée une personne. Cette caméra capture ce que "voit" la personne.

En vision par ordinateur, elle possède certains avantages en comparaison avec la vue objective utilisée en grande majorité jusqu'à présent :

- ▷ *Nature des données* : Les données acquises permettent de traiter différents aspects du comportement humain du fait de la vue subjective. En particulier, il est possible d'étudier les comportements d'une personne impliquée dans une interaction personnelle ou interpersonnelle depuis son propre point de vue.
- ▷ *Mobilité* : Ce point joue un rôle important dans la capacité à pouvoir explorer un espace plus important. En effet, dans le cas de la vue objective, la caméra est statique et elle est donc contrainte à capturer une zone constante de la scène.
- ▷ *Occultation et résolution* : La personne se place et se déplace, en général, de manière à choisir le meilleur angle de vue pour percevoir distinctement les zones d'intérêt. Ainsi, en principe, les occultations sont moins fréquentes, et la personne a tendance à se rapprocher de la zone d'intérêt.

1.2 Objectifs et méthodologie

Dans ce manuscrit, nous nous intéressons à l'analyse automatique du regard pour la compréhension du comportement humain en vue subjective. En particulier, nous nous concentrons sur trois principales thématiques : l'estimation du regard, l'interaction sociale et la reconnaissance d'activités. En effet, tout en exploitant l'information du regard, nous souhaitons aborder des questions telles que : (1) où une personne regarde-t-elle dans l'environnement ?, (2) qui regarde qui lorsque plusieurs personnes sont impliquées ?, (3) de quelle nature sont ces signaux attentionnels ?, (4) quelle activité est effectuée par une personne ?

Afin de répondre à ces questions, ce manuscrit propose des solutions qui ont été évaluées expérimentalement. L'élément central de ces contributions est caractérisé par l'utilisation d'un système eye-tracker porté. Un tel système permet à la fois, d'enregistrer une séquence vidéo des yeux d'une personne et d'obtenir une perception visuelle (partielle) de son environnement. Ainsi, il est possible de traiter deux types de regard : le regard *interne* obtenu via la séquence vidéo des yeux (*regard du soi*) et le regard *externe* issu de la présence d'une personne dans l'environnement (*regard d'autrui*). Le regard interne permet d'analyser les interactions personnelles, tandis que le regard externe, lui, permet d'élargir le champ d'applications à l'étude des interactions interpersonnelles.

[2] comme en témoigne le tout récent projet *Google Glass* (http://www.google.com/glass)
[3] Les caméras *GoPro* (http://gopro.com) en sont une parfaite illustration.

1.3 Contributions et organisation du manuscrit

Une vue d'ensemble des principales contributions est illustrée dans la figure 1.3. Ici, nous décrivons brièvement chacun des chapitres de ce manuscrit.

CHAPITRE 2 - **Méthodes de suivi du regard et applications**

Un bref état-de-l'art sur le suivi du regard est présenté et donne un récapitulatif des méthodes existantes qui exploitent les yeux ou, plus approximativement, la pose de la tête. Nous en profitons également pour aborder les applications potentielles telles que l'interaction sociale et l'analyse en vision subjective.

CHAPITRE 3 - **Suivi du regard à partir d'un eye-tracker porté**

Nous décrivons une nouvelle méthode de suivi du regard pour un eye-tracker binoculaire porté (Martinez *et al.*, 2012). Nous proposons :

- un modèle d'apparence basé sur des caractéristiques locales du gradient,
- une estimation du regard par régression non-linéaire, en particulier via une régression à vecteurs de relevance.

CHAPITRE 4 - **Reconnaissance d'attention en vue subjective**

Dans ce chapitre, nous reprenons les travaux du chapitre 3 et nous étendons le champ d'application à la reconnaissance d'attention en vue subjective (Martinez *et al.*, 2013). Nous proposons :

- une estimation, par *régression à noyaux multiples localisés*, de la pose de la tête des personnes externes,
- une modélisation des signaux sociaux, tels que le regard mutuel et partagé,
- un traitement en vue subjective à partir d'un eye-tracker porté basé sur un modèle d'apparence.

CHAPITRE 5 - **Reconnaissance d'activités basée sur le regard**

Nous présentons une nouvelle approche pour effectuer la reconnaissance d'activités en vue subjective et à partir du regard. Nous proposons :

- de nouvelles caractéristiques (multi-échelle et temporelle),
- une segmentation-classification d'activités par apprentissage contextuel.

CHAPITRE 6 - **Conclusion**

Nous concluons le manuscrit avec un résumé des principales contributions effectuées et abordons les directions possibles pour de futurs travaux.

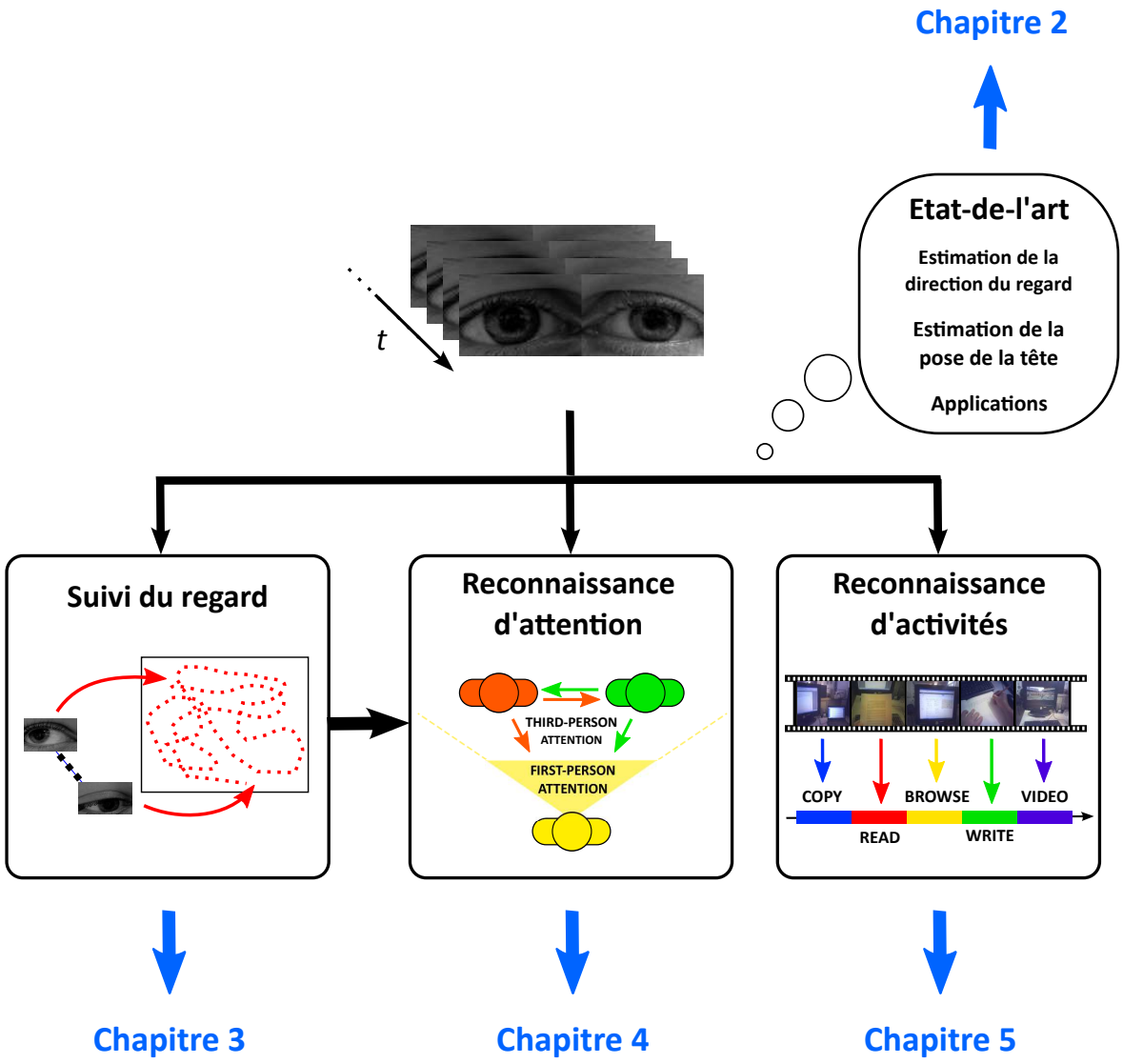


Figure 1.3 – Aperçu illustré des principales contributions de ce manuscrit. Le chapitre 3 s'intéresse à l'estimation du regard, notamment à l'extraction de caractéristiques des images des yeux, et à la relation entre ces caractéristiques et le point de regard. Dans le chapitre 4, le regard subjectif et objectif sont combinés dans le but de reconnaître des motifs attentionnels entre différentes personnes. Dans le chapitre 5, le regard est utilisé comme un moyen pour reconnaître les activités effectuées par une personne.

CHAPITRE 2

Méthodes de suivi du regard et applications

"Seul un regard peut créer l'univers."

– Christian Morgenstern –
dans *Borne kilométrique*

Sommaire

Dans ce chapitre, nous décrivons les principaux composants permettant d'estimer la direction du regard. Etant donné qu'il existe de nombreuses méthodes dans la littérature scientifique, nous présenterons ici uniquement celles qui sont pertinentes pour notre objet de recherche, à savoir l'estimation et l'analyse du regard.

Nous proposons donc de structurer cet état de l'art en trois parties :

1. l'estimation du regard basé sur l'information des yeux (cf. § 2.3),
2. l'estimation de la pose de la tête (cf. § 2.4),
3. quelques applications faisant appel à la connaissance de la direction du regard (cf. § 2.5) : *interaction sociale* et *activités en vue subjective*.

Mais tout d'abord, avant d'entrer dans le vif du sujet, nous introduisons quelques notions relatives à la vision humaine, en particulier la motricité oculaire, qui sont essentielles pour

comprendre le fonctionnement du regard (cf. § 2.2).

2.1 Introduction

D'un point de vue physiologique, l'être humain dispose de divers sens. Malgré l'absence de consensus au sein de la communauté scientifique, il est souvent admis que la perception sensorielle humaine peut se restreindre au cinq sens suivants : la vue, l'audition, le toucher, l'odorat et le goût.
Parmi ces sens physiologiques, la perception visuelle, dont l'oeil est l'organe récepteur, est souvent désignée comme l'un des sens le plus important et le plus complexe. L'oeil permet d'obtenir une représentation visuelle du monde extérieur grâce, entre autres, à la perception des formes, des couleurs, des contrastes et de la géométrie. Il a suscité et suscite toujours un grand intérêt de la part des scientifiques qui essaient de percer certains de ses mystères. En neurosciences cognitives, les chercheurs s'intéressent, par exemple, à la manière dont le cerveau traite les impulsions électriques issues des signaux lumineux captés par la rétine et qui donnent lieu à une représentation imagée de l'environnement. En médecine, les scientifiques essaient, quant à eux, de trouver des traitements pour les pathologies oculaires (thérapie génique) ou encore des moyens pour substituer des parties de l'oeil (oeil bionique, rétine artificielle).
De plus, l'oeil est aussi un organe mobile rendant possible les mouvements oculaires par le biais des muscles oculomoteurs. En agissant sur l'oeil, les trois paires de muscles oculomoteurs permettent d'orienter le regard de l'être humain. En moyenne, la mobilité des muscles oculomoteurs est de l'ordre de 30° à 40° de part et d'autre de la position moyenne où le regard est situé de face. Cependant, en général, les mouvements de la tête prennent spontanément le relai de celui des muscles oculomoteurs en deçà de ces limites physiologiques (typiquement pour des mouvements excédant 15° à 20°). Il existent aussi d'autres muscles tels que les muscles des paupières qui permettent au système lacrymal de protégér l'oeil en l'hydratant ou en supprimant les impuretés.

2.2 Vision humaine

2.2.1 Structure de l'oeil

La figure 2.1 représente la structure interne et externe de l'oeil humain. L'oeil est un globe oculaire de rayon 12 mm. D'un point de vue externe, l'oeil est composé de trois parties : la *pupille* (la partie noire), l'*iris* (la partie colorée) et la *sclère* (la partie blanche). La pupille permet d'absober les rayons lumineux et de les transmettre vers la rétine. L'iris contrôle la quantité de rayons lumineux en modifiant le diamètre et la taille de la pupille. La sclère, quant à elle, permet de contenir la pression interne de l'œil et de le protéger contre les

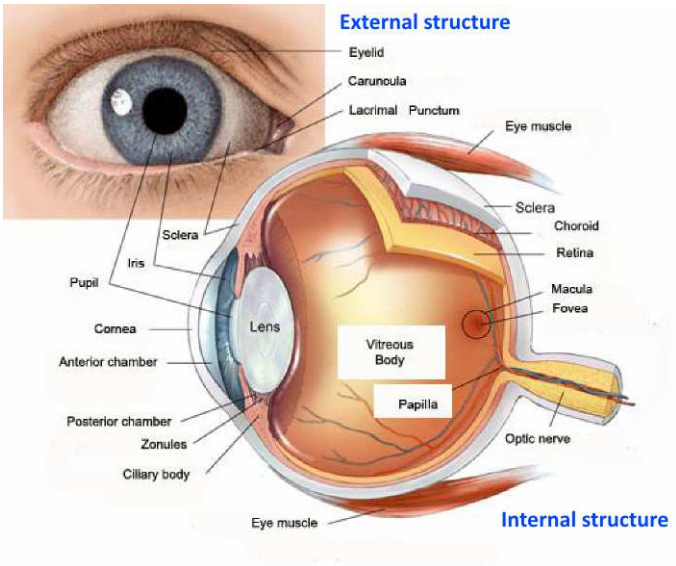


Figure 2.1 – Structure interne et externe de l'oeil humain. (*source : Dental Articles*)

agressions mécaniques. La partie sombre séparant la sclère de l'iris se nomme le *limbe* et elle intervient dans la résistance de l'oeil à la pression intra-oculaire.

2.2.2 Mouvements oculaires et fonctions visuelles associées

Dans une journée ordinaire, l'être humain est soumis à une multitude d'information visuelle. Les mouvements oculaires permettent alors de sélectionner les informations relevantes en fonction : (1) des intentions de la personne et (2) des zones saillantes de l'environnement. Pour pouvoir traiter ces informations, les mouvements oculaires possèdent deux fonctions principales qui permettent de répondre au besoin du système visuel. Nous décrivons uniquement les mouvements oculaires des deux yeux[1].

La première fonction, connue sous le terme anglais de *gaze-holding* (Carpenter, 1988), consiste à amener ou garder l'objet d'intérêt sur la partie centrale de la rétine, la fovéa, qui permet une vision de haute-qualité. Cette fonction fait appel aux mouvements oculaires suivants :

- ▷ **Fixation** - Les fixations apparaissent quand le regard est immobile et peuvent durer entre 100 et 150ms (Salvucci et Goldberg, 2000). Néanmoins, des micro-mouvements apparaissent également durant les fixations : dérives, micro-tremblements et micro-saccades, et servent à éviter que des zones détaillées de notre perception disparaissent (ce phénomème est connu sous le nom d'*effet Troxler* ou d'*effacement Troxler*). La

[1]par opposition, aux mouvements oculaires d'un seul oeil appelés aussi *ductions*

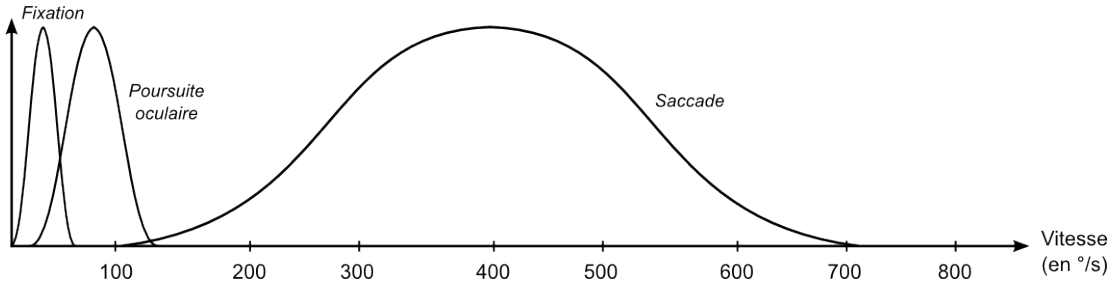


Figure 2.2 – Zones cinétiques approximatives des mouvements oculaires.

vitesse oculaire des fixations est, en général, inférieure à 20°/s (Salvucci et Goldberg, 2000).

- ▷ **Poursuite** : La poursuite oculaire (en anglais, *smooth pursuit*) est le mouvement continu des yeux sur une cible en déplacement régulier. Généralement, la poursuite oculaire s'effectue avec une vitesse comprise entre 1 et 30°/s.
- ▷ **Saccades** : Les saccades sont des mouvements brefs et très rapides, et peuvent atteindre une vitesse de 900°/s. Elles permettent d'explorer le champ visuel et dirigent la fovéa sur l'objet ou la région d'intérêt (entre 20 et 200ms suivant l'amplitude des saccades) afin de pouvoir effectuer une analyse plus détaillée à cet endroit. Une saccade a lieu entre deux fixations. Notons aussi que durant les fixations, les informations visuelles sont mémorisées, ce qui n'est pas le cas durant les saccades.
- ▷ **Vergence** : Il s'agit de mouvements oculaires symétriques au cours desquels les yeux se déplacent en sens inverse l'un de l'autre. Avec l'accomodation (déformation du cristallin), ils contribuent à améliorer la perception d'objets situés à des profondeurs de champ différentes.

La figure 2.2 schématise les zones cinétiques approximatives pour les saccades, les poursuites et les fixations.

La seconde fonction, appelée *gaze-shifting* (Carpenter, 1988), concerne la stabilisation rétinienne de l'image lorsque la personne bouge, ce qui induit les réflexes suivants :

- ▷ **Réflexe vestibulo-oculaire** : Il permet de compenser les mouvements de la tête afin de maintenir fixe la direction du regard grâce au système vestibulaire (responsable de la perception de la position angulaire de la tête et de son accélération).
- ▷ **Réflexe optocinétique** : Ce réflexe tend à faire suivre au regard le parcours d'un objet en mouvement quand la tête reste stationnaire ; il permet de compenser le déplacement de l'environnement visuel.

2.3 Estimation de la direction du regard

Afin de capturer les mouvements oculaires, plusieurs techniques ont été proposées :

Magnéto-oculographie (en anglais, *scleral search-coil*) - Les mouvements oculaires sont déduits des variations dans un champ magnétique. Ces variations sont dues à l'induction magnétique par trois bobines (*coils*) incorporées à une lentille spéciale posée sur la sclère de l'œil du sujet. Il s'agit d'une technique hautement invasive et elle requiert l'administration d'un anesthésique local au sujet.

Electro-oculographie (EOG) - Cette technique mesure les potentiels électriques continus cornéen-rétinien grâce à des électrodes placées autour des yeux. En fonction de la rotation de l'oeil qui agit comme un dipôle électrique : une différence de potentiel variable entre les électrodes est générée et permet de mesurer l'amplitude des mouvements oculaires.

Vidéo-oculographie (VOG) - Cette technique fait appel à un système d'acquisition d'images de l'oeil. Ces images sont, ensuite, traitées pour pouvoir extraire des caractéristiques oculaires. Il s'agit de la technique la plus utilisée de nos jours. Contrairement aux techniques précédentes, elle permet aussi d'obtenir une information *absolue* du mouvement oculaire.

Dans la suite du manuscrit, nous limiterons notre étude à l'estimation *visuelle* de la direction du regard, c'est-à-dire via des techniques de vision par ordinateur. Par ailleurs, dans la littérature scientifique, différentes configurations et différentes méthodes algorithmiques ont été proposées pour estimer le regard à partir de systèmes de vision. Nous invitons le lecteur à se référer à l'état de l'art (Hansen et Ji, 2010) et à l'édition spéciale (Ji *et al.*, 2005) pour un aperçu plus complet et plus détaillé des modèles existants.

Lorsqu'il s'agit de concevoir un système de suivi du regard, il existe 4 composants à définir : la *configuration* (cf. § 2.3.1), l'*éclairage* (cf. § 2.3.2), le *modèle de l'oeil* (cf. § 2.3.3) et le *modèle du regard* (cf. § 2.3.4). La figure 2.3 offre une vue générale des différents choix possibles.

2.3.1 Configurations

D'une manière générale, il existe deux types de configurations pour la mise en place d'un oculomètre (ou *eye-tracker*) :

- **Eye-tracker déporté** : Dans cette configuration, l'eye-tracker est extérieur à l'utilisateur et le ou les caméras sont orientées vers le visage de l'utilisateur de manière à capturer ses yeux. Ainsi, cette configuration est non intrusive, mais la position et les mouvements de la tête de l'utilisateur sont contraints. En effet, un oeil, au minimum,

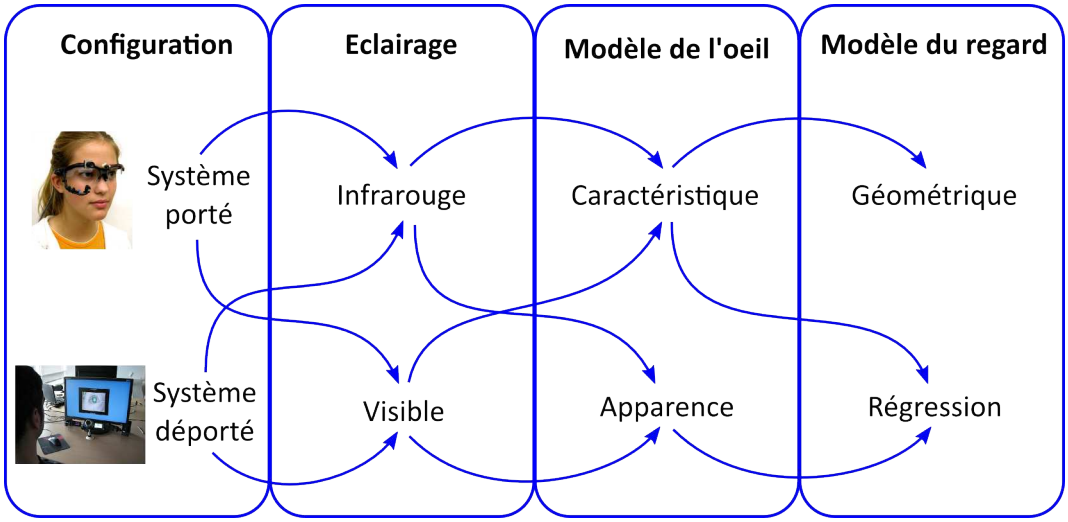


Figure 2.3 – Aperçu général des approches de suivi du regard.

doit rester dans le champ de vue de la caméra afin d'estimer le regard (la présence des deux yeux est, cependant, recommandée afin d'avoir une précision suffisante). De plus, le système peut comprendre plusieurs caméras afin d'obtenir un système de stéréovision, utile pour faciliter le suivi et l'estimation de la pose de la tête.

- **Eye-tracker porté** : La seconde configuration, quant à elle, consiste à porter l'eye-tracker sur la tête. L'avantage de cette configuration est que le système est fixe par rapport à la tête de la personne et l'oeil reste en permanence dans le champ de vue de la caméra. Ainsi, le regard de l'utilisateur peut être estimé à tout instant lorsque celui-ci se déplace dans l'environnement.

En principe, les eye-trackers déportés effectuent une calibration par rapport à un écran d'ordinateur dans le but de permettre une interaction avec ce dernier. Quant aux eye-trackers portés, ils sont généralement équipés d'une caméra additionnelle, appelée *caméra scène*, qui permet de capturer une information visuelle de l'environnement "perçu" par la personne. Toutefois, il est aussi possible d'avoir un système déporté avec une calibration dans l'environnement (Hennessey et Lawrence, 2008 ; Funes et Odobez, 2012) ou encore un eye-traker porté qui permet l'interaction via un écran d'ordinateur (San Agustin *et al.*, 2010).

2.3.2 Éclairage

Au départ, les eye-trackers ont été conçus en se basant sur l'éclairage infrarouge pour faciliter l'extraction visuelle de caractéristiques de l'oeil. Fort de son succès, de nos jours, la grande majorité, voire la totalité, des eye-trackers commerciaux, que ce soit déportés ou

portés, fonctionnent en lumière infrarouge.
Cependant, ces derniers temps, de nombreuses méthodes ont été proposées pour l'eye-tracking en éclairage visible (Noris *et al.*, 2011; Tsukada *et al.*, 2011). En effet, il existe plusieurs raisons qui peuvent expliquer cet intérêt croissant. Premièrement, pour des applications nécessitant l'utilisation d'un eye-tracker pour une longue durée, il est préférable d'éviter l'utilisation d'un éclairage infrarouge pour des raisons de sécurité médicale (Mulvey *et al.*, 2008). En effet, un éclairage prolongé en infrarouge peut entraîner une surchauffe au niveau de la rétine et provoquer des dommages irréversibles. De plus, l'éclairage infrarouge peut poser des problèmes lorsqu'un eye-tracker porté est utilisé en environnement extérieur, essentiellement à cause des rayons du soleil. Enfin, s'affranchir d'un éclairage actif permet aussi d'éviter de placer des sources lumineuses, ce qui simplifie la conception de l'eye-tracker.

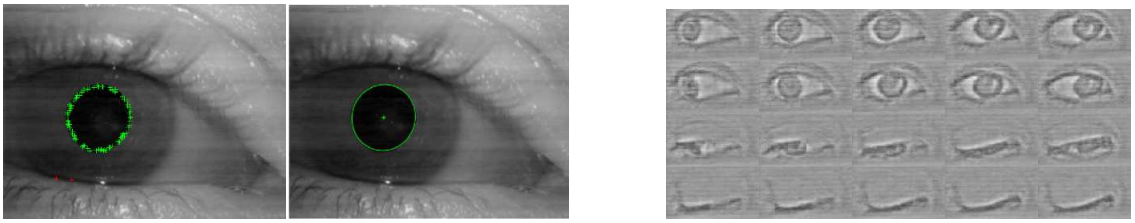
2.3.3 Modèles de l'oeil

Pour pouvoir estimer la direction du regard, il est tout d'abord indispensable de définir un modèle de l'oeil et d'extraire les caractéristiques associées. En fonction du type d'éclairage, le choix de ce modèle est différent. En éclairage infrarouge, les caractéristiques extraites sont la pupille et le(s) glint(s) (cf. § 2.3.3.1), alors qu'en éclairage visible, il s'agit plutôt soit de localiser l'iris (cf. § 2.3.3.1) ou soit d'avoir recours à l'apparence globale de l'oeil (cf. § 2.3.3.2).

2.3.3.1 Méthodes orientées *caractéristique*

Suivi de la pupille et glint(s) - En général, nombreux sont les systèmes qui font appel à l'éclairage infrarouge afin d'obtenir une image contrastée et donc faciliter le suivi de la pupille. Dans la littérature scientifique, il existe de nombreux algorithmes de suivi de la pupille et beaucoup d'heuristiques ont été proposées, notamment en appliquant des seuillages. Par exemple, Stiefelhagen *et al.* (1996) décrivent un algorithme par seuillage itératif permettant ainsi de considérer les changements d'illumination et les variations inter-sujet. D'autres (Ebisawa et Satoh, 1993) suggèrent d'utiliser deux sources d'illumination infrarouge et décrivent une méthode basée sur la diférence d'images pour détecter la pupille.

Suite aux travaux de Babcock et Pelz (2004) où les données sont traitées *a posteriori*, Li *et al.* (2005) ont proposé l'algorithme *Starburst* (cf. figure 2.4(a)) pour les systèmes portés qui bénéficient d'une image de plus haute résolution. Cet algorithme est certainement le plus connu et constitue souvent un point de départ pour développer de nouvelles approches. A partir d'une position de départ au sein de la pupille, des rayons sont radialement propagés jusqu'à leur intersection avec le contour de la pupille. Les points d'intersection sont, ensuite, transmis à un algorithme d'ajustement d'ellipse qui repose sur la méthode RANSAC (Fischler et Bolles, 1981). Plus récemment, certains se sont également intéressés à robustifier le suivi dans le cas où la caméra est fortement désaxée (Swirski *et al.*, 2012).



(a) (b)

Figure 2.4 – **Modèle de l'oeil** : (a) Suivi de la pupille par détection de points de contour et ajustement d'ellipse (Li *et al.*, 2005), (b) Modèle d'apparence par Retinex modifié (Noris *et al.*, 2011).

En plus du suivi de la pupille, le glint[2] est également localisé, car il est utile pour la modélisation du regard. Cependant, dans certains cas, en particulier pour les systèmes déportés, il est souvent nécessaire d'avoir des sources lumineuses supplémentaires. Il est alors possible de détecter d'autres reflets issus de ces sources lumineuses qui permettent de contribuer à l'estimation du regard, notamment en compensant les mouvements de la tête (cf. § 2.3.4.2).

Suivi de l'iris - Le suivi de l'iris concerne les méthodes en éclairage visible. Pour les systèmes portés, son principe est similaire à celui appliqué en éclairage infrarouge. Ainsi, en reprenant les mêmes idées que l'algorithme *Starburst* (Li *et al.*, 2005), Li et Parkhurst (2006) ont développé une méthode de suivi de l'iris en éclairage visible. Ryan *et al.* (2008) ont, ensuite, proposé des améliorations pour cet algorithme, notamment à l'aide d'un mécanisme alternant entre le suivi de la pupille ou de l'iris en fonction des points détectés. Ryan *et al.* (2010) restreignent la zone de recherche localement autour des points de contour de l'iris détectés à l'instant précédent. Cette méthode a l'avantage de pouvoir limiter le nombre de points parasites pour l'ajustement d'ellipse. Cependant, si les mouvements de l'iris entre deux instants sont larges, il est possible que l'algorithme ne soit plus capable de suivre l'iris et il est donc nécessaire de réinitialiser l'algorithme de suivi.

Tsukada *et al.* (2011) partent du principe que, dans les méthodes précédentes, le modèle d'ellipse utilise 5 paramètres : la position, l'angle de rotation et les axes mineur et majeur ; or ceux-ci peuvent être influencés par des points non issus de l'iris (paupières, cils, réflexions parasites). C'est pourquoi ils proposent de se baser uniquement sur la position de l'ellipse et introduisent un modèle 3D de l'oeil construit à partir d'une base de données de formes d'iris. Ils ont ensuite proposé une méthode afin de construire le modèle 3D de l'oeil automatiquement (Tsukada et Kanade, 2012). Plus récemment, Pires *et al.* (2013) ont proposé une méthode afin de corriger la déformation projective causée par la caméra, ce qui permet, ensuite, de chercher un cercle plutôt qu'une ellipse dans l'image de l'oeil.

[2]aussi appelé reflet *cornéen ou première image de Purkinje-Sanson*

Cette correction réduit le nombre requis de paramétres de 5 à 3, ce qui a pour avantage de diminuer l'influence de sur-ajustement et d'accélèrer le traitement.

Pour les systèmes déportés, l'iris peut être localisé à l'aide d'un modèle déformable de l'oeil (Colombo et A.D. Bimbo, 1999), d'un modèle à deux cercles plans (Wang et Sung, 2001), de la courbure isophote (Valenti et Gevers, 2012) ou de l'information du gradient (Timm et Barth, 2011). Pour le suivi de l'iris, Hansen et Pece (2005) décrivent une méthode probabiliste par filtrage particulaire avec un modèle du profile du gradient inclus dans la vraisemblance des observations.

2.3.3.2 Méthodes orientées *apparence*

Les approches orientées caractéristique requièrent l'extraction de caractéristiques géométriques de l'oeil, or cette opération peut s'avérer être compliquée et sujette à des erreurs (points parasites, non-localisation des caractéristiques, etc). Les approches orientées apparence, quant à elles, traitent directement le contenu de l'image de l'oeil et s'appliquent, en particulier, en éclairage visible.

Intensité des pixels - Il s'agit du modèle d'apparence le plus couramment employé. Le descripteur d'apparence est construit en extrayant, ligne par ligne, les valeurs d'intensité des pixels de l'image pour former le vecteur d'entrée du modéle d'estimation du regard (Baluja et Pomerleau, 1994). Des pré-traitements sont parfois également appliqués avant la vectorisation. Xu *et al.* (1998) procèdent à une normalisation d'histogramme. Noris *et al.* (2011) suggèrent de rajouter une étape supplémentaire en intégrant un modèle *Retinex* non-linéaire (Choi *et al.*, 2007) afin de prendre en compte les changements d'illumination (cf. figure 2.4(b)). Lu *et al.* (2011b) proposent de réduire la dimensionnalité "à la main" : l'image est divisée en 3×5 régions et la moyenne de ces régions est prise pour construire le descripteur d'apparence de taille égale à 15. Une étude par Analyse en Composantes Principales (ACP) leur a permis de montrer que ce modèle permet de retenir suffisamment d'information pour estimer le regard.

Fusion de caractéristiques - D'autres méthodes visent à combiner différentes caractéristiques pour bénéficier de différentes sources d'information. Par exemple, Williams *et al.* (2006) proposent de combiner l'intensité des pixels et la carte d'énergie des contours. Cependant, la combinaison de ces deux caractéristiques apporte une amélioration négligeable pour l'estimation du regard en comparaison avec les résultats obtenus uniquement à partir de l'intensité de l'image. Zhang *et al.* (2012) fusionnent : l'intensité, la couleur, les réponses à des filtres de Gabor et de Haar, et le spatiogramme de l'image. Ensuite, une étape de sélection de caractéristiques, via *minimum Redundancy Maximum Relevance* (mRMR, (Peng *et al.*, 2005)), est effectuée avant d'appliquer le modèle d'estimation du regard.

2.3.4 Modèles du regard

Estimer le regard revient à modéliser la relation $\mathbf{\Phi}$ entre la donnée m-dimensionnelle extraite de l'image de l'oeil et le "point de regard" (ou direction du regard), c'est-à-dire la position, dans l'espace de dimension d, pointéee par les yeux (Young et Sheena, 1975) :

$$\mathbf{\Phi} : \mathbb{R}^m \rightarrow \mathbb{R}^d \tag{2.1}$$

Cette relation peut être obtenue par une *fonction de régression* (cf. § 2.3.4.1) ou *géométriquement* (cf. § 2.3.4.2). La dimension d de l'espace de sortie est, en général, égale à 2 (pour un écran). Cependant, il est également possible d'estimer le regard dans l'espace 3D.

Le choix du modèle du regard dépend de la méthode utilisée pour extraire les caractéristiques de l'oeil. Ainsi, on pourra remarquer que les méthodes orientées caractéristique peuvent aussi bien faire appel à un modèle géométrique qu'à un modèle de régression, alors que les méthodes orientées apparence utilisent nécessairement un modèle de régression.

2.3.4.1 Fonctions de régression

Ces approches consistent à apprendre une relation entre l'oeil et une surface 2D. L'avantage de ces méthodes est qu'elles ne nécessitent pas de calibration géométrique (cf. § 2.3.4.2). En revanche, l'estimation du regard est restreint à un plan. En général, il s'agit d'un écran pour les eye-trackers déportés et de l'image de la caméra scène pour les eye-trackers portés.

Notons également que la taille du vecteur d'entrée est, ici, un facteur déterminant pour le choix du modèle du regard. Ainsi, les méthodes orientées caractéristique optent le plus souvent pour une approche polynomiale. En revanche, les méthodes orientées apparence doivent plutôt faire face une non-linéarité plus prononcée, car la taille du descripteur d'entrée a une plus grande dimensionnalité.

Pupille-glint ou iris - A l'origine, l'estimation du regard se faisait uniquement en suivant la pupille et en effectuant une calibration avec une surface plane. L'inconvénient de cette méthode est qu'elle nécessite d'avoir un déplacement rigide entre l'oeil et la caméra (ce qui n'est pas forcément le cas pour des systèmes déportés). C'est pourquoi la méthode pupille-reflexion cornéenne (*Pupil Corneal Reflection*, PCR (Merchant *et al.*, 1974)) a été proposée et permet de compenser des mouvements faibles de la tête. Elle consiste à calculer la position relative du centre de la pupille et du reflet brillant d'une source lumineuse sur l'oeil (le *reflet cornéen*). Pour estimer le regard à partir de ce vecteur, ils proposent d'employer une régression linéaire, puis polynomiale pour prendre en compte la non-linéarité.

Bien plus tard, Morimoto *et al.* (2000) suggèrent d'apprendre une régression polynomiale pour chaque dimension de l'espace du regard (avec, dans son cas, l'écran d'ordinateur comme référentiel) :

$$\begin{aligned} x_s &= a_0 + a_1 x_p + a_2 y_p + a_3 x_p y_p + a_4 x_p^2 + a_5 y_p^2 \\ y_s &= a_6 + a_7 x_p + a_8 y_p + a_9 x_p y_p + a_{10} x_p^2 + a_{11} y_p^2 \end{aligned} \tag{2.2}$$

avec (x_p, y_p) et (x_s, y_s), respectivement, les coordonnées du vecteur pupille-glint et du point de regard sur l'écran, et $(a_0, \ldots, a_{11})$ étant les paramètres des régresseurs.

Ji et Zhu (2002) emploient un réseau de neurones avec en entrée : les paramètres de la pupille (centre, orientation, rapport entre l'axe majeur et mineur), le déplacement pupille-glint et les coordonnées du glint. Cette méthode permet de compenser des mouvements faibles de la tête tout en gardant une précision de 5°. Zhu *et al.* (2006) suggèrent d'utiliser un *Support Vector Regression* (SVR, (Drucker *et al.*, 1996)) pour apprendre la relation entre le centre de la pupille et le glint, et les coordonnées de l'écran.

Notons que des approches similaires sont appliquées dans le cas où le modèle de l'oeil repose sur l'iris (Ryan *et al.*, 2008).

Apparence globale - De nombreuses approches ont recours à une transformation non-linéaire pour l'estimation du regard et requièrent, en général, un grand nombre d'échantillons d'apprentissage. Baluja et Pomerleau (1994) utilisent un réseau de neurones de type perceptron multi-couche pour apprendre la fonction de mise en correspondance entre les images des yeux et les coordonnées du regard sur l'écran avec 2000 échantillons. Xu *et al.* (1998) proposent une approche similaire mais avec 1000 échantillons de plus. Williams *et al.* (2006) présentent une nouvelle méthode de régression appelée S^3GP (pour *Sparse Semi-Supervised Gaussian Process*) qui repose sur une extension des Processus Gaussien au cas semi-supervisé et éparse, et estiment le regard à l'aide de 16 échantillons d'apprentissage partiellement annotés parmi 80 échantillons disponibles. Pour leur eye-tracker porté, Noris *et al.* (2011) effectuent l'estimation du regard à l'aide d'un SVR en utilisant 200 échantillons.

Variété de faible dimension - Ces approches assument que l'apparence de l'oeil, bien qu'ayant une dimensionnalité très élevée, appartient à un espace de dimension plus faible. Ainsi, ces méthodes modélisent la non-linéarité comme l'union d'espaces réduits et linéaires construits à partir d'une variété de faible dimension. Tan *et al.* (2002) utilisent une méthode de réduction de dimensionnalité par *Locally Linear Embedding* (LLE, (Roweis et Saul, 2000)) pour apprendre la variété de l'apparence de l'oeil à partir de 252 échantillons éparses. L'idée consiste à exprimer le descripteur d'apparence de l'oeil $\mathbf{e}_i$ comme une combinaison linéaire des apparences voisines $\mathcal{N}_i$ et de construire son image, c'est-à-dire les coordonnées du regard, $\mathbf{g}_i$ dans le nouvel espace en respectant cette relation. Pour un nouvel échantillon $\hat{\mathbf{e}}$, les poids $\mathbf{w}$ de la régression sont obtenus en minimisant l'erreur de reconstruction :

$$E(\mathbf{w}) = ||\hat{\mathbf{e}} - \sum_i w_i \mathbf{e}_i||_2 \;\; s.t. \;\; \mathbb{1}^\top \mathbf{w} = 1 \tag{2.3}$$

où $\mathbb{1}$ désigne la fonction indicateur. Ensuite, les coordonnées du regard sont estimées par interpolation locale en appliquant la formule suivante :

$$\hat{\mathbf{g}} = \sum_i w_i \mathbf{g}_i \tag{2.4}$$

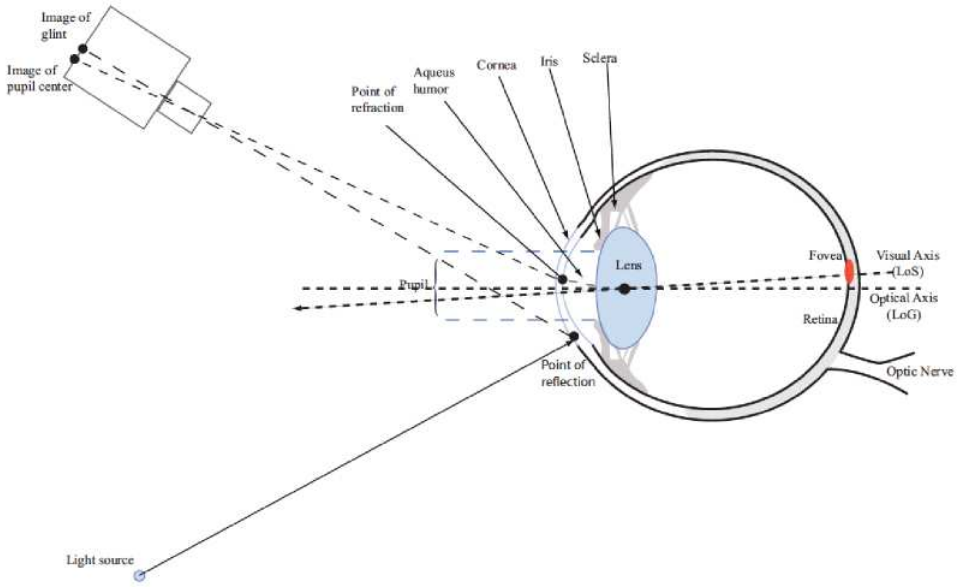


Figure 2.5 – Modèle géométrique : oeil, source lumineuse et caméra (Hansen et Ji, 2010).

avec $\mathbf{g}_i$ et $\hat{\mathbf{g}}$ les coordonnées, respectivement, d'apprentissage et de test.
Dans le même esprit (mais sans avoir recours à un apprentissage de la variété), Lu *et al.* (2011b) présentent une méthode par régression linéaire adaptative (*Adaptive Linear Regression*, ALR) où les poids du régresseur sont appris à l'aide de la minimisation sous contrainte suivante :

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} ||\mathbf{w}||_1 \quad s.t. \quad || E\mathbf{w} - \hat{\mathbf{e}}||_2 < \varepsilon \tag{2.5}$$

où $\hat{\mathbf{e}}$ est le descripteur d'apparence de test et E correspond à la matrice dont les colonnes sont les descripteurs d'apparence d'apprentissage. En minimisant $||\mathbf{w}||_1$, la méthode cherche, ainsi, une solution éparse. Ensuite, une interpolation locale équivalente à (2.4) permet alors d'obtenir les coordonnées du regard. De plus, ils proposent une approche pour déterminer ε automatiquement.

2.3.4.2 Méthodes géométriques

Les méthodes géométriques ont recours à un modèle géométrique de l'oeil (cf. figure 2.5). L'estimation du point de regard revient alors à trouver l'intersection entre la direction du regard et l'objet regardé. Cette direction est soit modélisée par l'*axe optique* (ou ligne de regard) qui passe par le centre de la pupille, de la cornée et du globe oculaire, soit par l'*axe visuel* (ou ligne de vue) qui relie la fovéa et le centre du globe oculaire. Cette dernière modélisation est celle qui s'avère être la véritable direction du regard. De plus, la connaissance du centre du globe oculaire ou de la cornée donne une information utile sur la

position de la tête dans l'espace 3D, ce qui explique l'enjeu de ces méthodes géométriques. Ainsi, pour obtenir une bonne précision, la plupart de ces méthodes géométriques requièrent une calibration géométrique *complète* (en anglais, *fully calibrated setup*), étudiée en détails dans (Guestrin et Eizenman, 2006), qui combine :

- ▷ **Calibration *interne*** : estimation des paramètres d'un modèle de l'oeil humain
- ▷ **Calibration *externe*** : calibration des caméras (via une mire) et connaissance des positions relatives des caméras, des sources lumineuses et de l'écran.

A cause de la complexité de l'anatomie de l'oeil, des hypothèses sont assez fréquemment formulées (cornée sphérique ou encore coïncidence de l'axe optique et visuel) ou alors ces méthodes nécessitent d'avoir des sources lumineuses supplémentaires pour bénéficier d'invariants projectifs.

Pupille-glint(s) - Il existe peu de méthodes (Ohno et Mukawa, 2004 ; Guestrin et Eizenman, 2006 ; Villanueva *et al.*, 2006) qui emploient une seule caméra avec une source lumineuse. Ces auteurs ont montré que cette configuration nécessite d'avoir la tête immobile ou de connaître la distance oeil-écran pour estimer le regard. En effet, la technique *Pupil Corneal Reflection* (PCR) dépend de la position de la pupille relativement au reflet cornéen, or ce vecteur varie également lors d'un mouvement de la tête. De plus, la non-sphéricité de l'oeil influe sur ce vecteur.

Les autres méthodes, quant à elles, proposent diverses solutions, notamment en se basant sur plusieurs caméras ou plusieurs sources lumineuses pour localiser des caractéristiques 3D de l'oeil et compenser les mouvements de la tête (Shih et Liu, 2004; Zhu et Ji, 2007). Par exemple, pour un système avec une caméra et deux sources lumineuses, Morimoto *et al.* (2002) effectuent différentes simulations pour évaluer leur modèle sphérique de l'oeil. Hennessey *et al.* (2006) proposent d'affiner le modèle de l'oeil en localisant le centre de la cornée par triangulation et estiment le point de regard comme l'intersection entre l'axe optique et l'écran d'ordinateur.

L'inconvénient de ces méthodes est qu'elles requièrent une calibration géométrique complète, ce qui est souvent laborieux à mettre en place, engendre des imprécisions et rend le système peu flexible (déplacement de caméra et de sources lumineuses).

Géométrie projective - Contrairement aux approches précédentes, Yoo *et al.* (2002) présentent une méthode connue sous le nom de *cross-ratio* et qui permet de pallier les limitations du PCR (cf. § 2.3.4.1). En projetant un motif lumineux rectangulaire sur la surface de la cornée, le point de regard peut être estimé par des propriétés d'invariance en géométrie projective, ce qui permet d'éviter d'effectuer une calibration géométrique complète. Un cinquième glint permet de prendre en compte la non-linéarité des déplacements des glints. Coutinho et Morimoto (2010) améliorent cette méthode avec un modèle plus précis de l'oeil et la rend également plus robuste aux mouvements de la tête (Coutinho et

Morimoto, 2012). Néanmoins, leur méthode reste sensible aux mouvements en profondeur et requièrent l'apprentissage de l'angle entre l'axe optique et l'axe visuel.

Dans ce même esprit, Hansen *et al.* (2010) présentent une méthode par normalisation d'homographie. En comparaison avec la méthode cross-ratio, elle a l'avantage de modéliser implictement l'offset entre l'axe optique et l'axe visuel, et ne requière que 4 glints contre 5.

Ces méthodes ont l'avantage de ne pas avoir recours à une calibration géométrique complète. En revanche, en pratique, elles peuvent être confrontées au problème de localisation des glints qui devient d'autant plus difficile lorsque la tête bouge ou que leur nombre augmente.

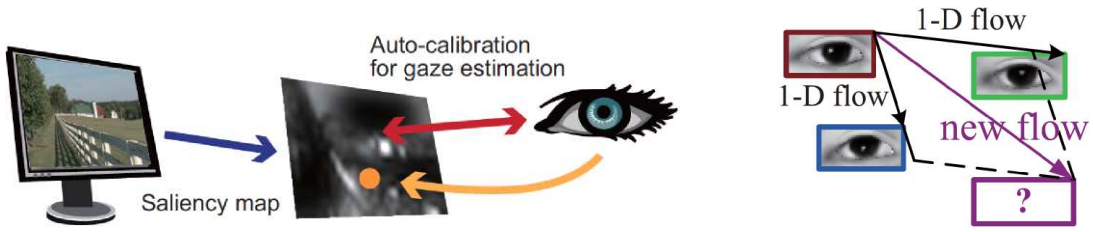
Point de regard 3D - Les approches citées précédemment s'intéressent à l'estimation du regard sur une surface plane dans le cas d'un système déporté. D'autres techniques visent à obtenir le point de regard 3D et s'appliquent essentiellement pour des systèmes portés (excepté (Hennessey et Lawrence, 2008)). Munn et Pelz (2008) procèdent en 4 étapes : (1) estimation du point de regard 2D dans l'image de la caméra scène, (2) estimation de l'ego-mouvement à partir du suivi de caractéristiques pour des images clés, (3) calcul de la position et de l'orientation de la tête du sujet, et (4) détermination du point de regard 3D dans l'environnement. Takemura *et al.* (2010) emploient une approche similaire. mais ils estiment le mouvement de la caméra scène par SLAM visuel (Davison *et al.*, 2007). Pour un système porté multi-caméras (2 caméras oeil et 2 caméras scène), Pirri *et al.* (2011) proposent une méthode exploitant les fondements de la géométrie d'images multiples (Hartley et Zisserman, 2006). Tout d'abord, les axes pupillaires 3D de l'oeil gauche et droit sont calculés. Ensuite, le point de regard 3D est obtenu par intersection de ces deux axes optiques. Bernet *et al.* (2011) s'intéressent à l'utilisation de la matrice fondamentale hybride (Puig *et al.*, 2008) pour estimer le point de regard à partir d'un eye-tracker binoculaire porté.

2.3.5 Calibration individuelle

Une grande majorité des eye-trackers nécessitent une procédure de calibration individuelle qui vise, à partir des échantillons d'apprentissage, à estimer les paramètres de la relation entre le modèle de l'oeil et le point de regard (en anglais, *gaze mapping calibration*). Par exemple, pour les systèmes déportés en éclairage infrarouge, cette calibration est utile pour estimer l'angle entre l'axe optique et l'axe visuel, alors que pour les eye-trackers portés, elle permet de déterminer les coefficients de régression.

Néanmoins, d'autres méthodes ont été présentées pour limiter ou éviter une calibration qui peut prendre du temps et être fastidieuse dans certains cas, tout en essayant de préserver une bonne précision :

Sans calibration - Yamazoe *et al.* (2008) emploient un modèle géométrique simple de l'oeil et ajustent le modèle à l'apparence de l'oeil. Pour un système porté monoculaire,



(a) (b)

Figure 2.6 – **Calibration** : (a) Saillance visuelle (Sugano *et al.*, 2010), (b) Synthèse d'images (Lu *et al.*, 2012).

Kohlbecher *et al.* (2008) proposent de procéder à la reconstruction 3D de la pupille par stéreovision. Noris *et al.* (2008) collectent des données de 33 sujets et apprennent un modèle d'apparence qui permet d'estimer le regard pour un nouveau sujet sans procédure de calibration et avec une précision moyenne de 2.34°. Pour approximer les angles entre l'axe optique et visuel de chaque oeil, Model *et al.* (2009) proposent une "calibration" automatique en minimisant la distance entre les intersections des deux axes visuels avec un écran lorsque le sujet le regarde naturellement (par exemple, en regardant une vidéo).

Calibration *adaptée* - Fehringer *et al.* (2012) étudient la possibilité d'adapter, par translation dans l'espace du regard, la calibration d'un sujet à un autre à partir d'un eye-tracker monoculaire en éclairage infrarouge.

Calibration *assistée* - Sugano *et al.* (2008) utilisent les mouvements naturels de la souris pour, incrémentalement, collecter des échantillons et effectuer la calibration en ligne.

Saillance visuelle - Ces approches partent du principe que l'estimation du regard (*top-down*) et la saillance visuelle (*bottom-up*) sont fortement liées. L'objectif est de construire un estimateur du regard à partir de la connaissance de la saillance visuelle acquise via des images ou des vidéos (cf. figure 2.6(a)). Sugano *et al.* (2010) décrivent une méthode probabiliste dont la relation oeil-saillance est apprise par une régression à processus Gaussien (GPR, (Rasmussen et Williams, 2006)). Chen et Ji (2011) étendent le modèle avec un modèle probabiliste 3D de l'oeil et prennent en compte les mouvements de la tête.

Synthèse d'images - Plus récemment, Lu *et al.* (2012) estiment le regard à partir de l'apparence des yeux soumis à des mouvements incontrôlés de la tête. Pour éviter une calibration répétitive, des images synthétiques des yeux sont générées par flot 1D pour différentes orientations de la tête (cf. figure 2.6(b)). Ces images sont construites à partir de 4 images acquises sous une orientation frontale de la tête. Une précision moyenne de 2.24° est obtenue.

2.4 Estimation de la pose de la tête

L'orientation de la tête est intrinsèquement liée la détermination de la direction du regard (Emery, 2000; Langton *et al.*, 2004) et l'utilité de son estimation varie selon le type d'application :

- **Focus d'attention visuel** : Dans certains cas où la résolution de l'image ne permet pas d'estimer la direction du regard à partir des yeux (par exemple, si la caméra est située trop loin de la personne), il est possible d'obtenir une estimation approximative du regard en s'intéressant à l'orientation de la tête. Etant donné qu'une personne peut regarder différentes cibles pour une même orientation, il est souvent indispensable de modéliser le focus d'attention visuel en tenant compte de l'orientation de la tête, mais aussi du contexte, c'est-à-dire en prédéfinissant les différentes cibles potentielles dans l'environnement (Stiefelhagen *et al.*, 1999).

- **Compensation des mouvements de la tête** : Pour certaines applications, il est impératif d'obtenir une estimation de la pose de la tête afin de pouvoir compenser les mouvements associés pour déterminer la direction du regard (Lu *et al.*, 2011a; Funes et Odobez, 2012).
  Dans d'autres cas, compenser les mouvements de la tête permet également d'améliorer la reconnaissance de visages (Blanz *et al.*, 2005) ou d'expressions faciales (Kumano *et al.*, 2009).

Dans la littérature scientifique, nous pouvons constater qu'il existe un grand nombre d'approches traitant de ce sujet. Les méthodes peuvent être classées en fonction du type de données qu'elles traitent, par exemple des images 2D ou des données de profondeur (Microsoft Kinect, Asus Xtion). Ici, nous présentons brièvement les différentes approches à partir d'images 2D et nous invitons le lecteur à se référer à (Murphy-Chutorian, 2009) pour plus de détails. Pour les méthodes exploitant les données de profondeur, un bref aperçu est proposé dans (Fanelli *et al.*, 2012).

Au sein des approches basées image 2D, nous distinguons les méthodes qui reposent sur (1) la *localisation de points caractéristiques faciaux* ou (2) l'*apparence globale du visage*. La première classe s'intéresse essentiellement à une estimation *indirecte* de la pose de la tête (la pose est déduite à partir des positions des caractéristiques faciales), alors que la seconde effectue cette estimation de manière *directe*.

2.4.1 Terminologie et challenges

2.4.1.1 Terminologie

Tout d'abord, nous revenons sur quelques termes couramment employés lorsque l'on aborde l'estimation de la pose de la tête, notamment pour nommer les mouvements d'orientation

(a) (b)

Figure 2.7 – (a) Angles de rotation de la tête (Murphy-Chutorian, 2009), (b) Variations de pose (deux premières lignes) et d'illumination (deux dernières lignes) (Little *et al.*, 2005).

de la tête. Il existe différentes manières de représenter ces mouvements de rotation dans l'espace. En vision par ordinateur, l'orientation de la tête est le plus souvent exprimée relativement à un référentiel, plus particulièremment par rapport au repère de la caméra. Ainsi, mécaniquement, il est possible d'assimiler la tête à un objet à 3 degrés de liberté (cf. figure 2.7(a)) :

- ▷ **Yaw** (en français, *lacet*) : Il s'agit du mouvement de rotation horizontal de la tête autour de son axe vertical (mouvement gauche-droite de la tête).
- ▷ **Pitch** (en français, *tanguage*) : Ce mouvement s'effectue selon l'axe transversal de la tête (mouvement avant-arrière du cou).
- ▷ **Roll** (en français, *roulis*) : La tête effectue un mouvement de type roll lorsque celle-ci effectue une rotation autour de son axe longitudinal (mouvement gauche-droite du cou).

2.4.1.2 Challenges

En pratique, l'estimation de la pose de la tête (cf. figure 2.7(a)) est rendue difficile à cause de certains facteurs :

- ▷ **Orientation de la tête** : L'apparence de la tête varie fortement en fonction de l'angle sous lequel la personne est observée (cf. figure 2.7(a)).
- ▷ **Variations intra-sujet** : Ces variations ont lieu au sein d'un même sujet et influent sur la géométrie (expressions faciales) ou l'apparence (cheveux, barbe, vieillesse) faciale.
- ▷ **Variations inter-sujet** : Chaque personne est unique et dispose, ainsi, de ses propres caractéristiques, géométriques ou d'apparence, faciales qui permettent de la différencier des autres.

- ▷ **Changement d'illumination** : Suivant la géométrie des sources lumineuses (position et orientation), les réflexions lumineuses peuvent avoir une influence, locale ou globale, sur l'apparence du visage (cf. figure 2.7(b)).
- ▷ **Occultations** : Les occultations peuvent survenir lorsque certaines parties du visage sont cachées par un objet (lunettes, écharpe, chapeau) ou par auto-occultation (rotation du visage, cheveux).

Enfin, il existe également des déformations propres à la caméra qui peuvent avoir un impact sur le rendu de l'image, telles que les distorsions optiques (par exemple, la distorsion radiale peut être plus ou moins importante selon l'objectif choisi), la résolution de l'image, ou encore le choix de la technologie de caméra (analogique ou numérique).

2.4.2 Méthodes orientées *caractéristiques*

Ces méthodes s'inspirent essentiellement de la manière dont la perception humaine capture les orientations de la tête, notamment par le biais de *caractéristiques faciales*.

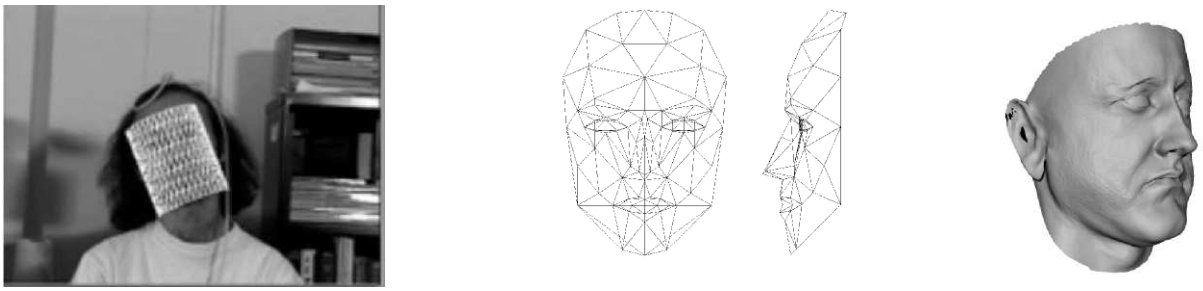
Parmi ces méthodes de localisation de caractéristiques, nous distinguons les approches par *détection de caractéristiques faciales* ou par *alignement d'un modèle*. La première approche vise à détecter des points caractéristiques faciaux tels que les coins des yeux, du nez ou encore de la bouche et elles font, en général, appel à des détecteurs et des heuristiques. La seconde stratégie, que nous décrivons plus en détails en § 2.4.2.1, consiste à aligner un modèle de forme à l'image pour localiser les points caractéristiques du visage.

L'estimation de la pose se fait, ensuite, en estimant une relation entre les positions des caractéristiques faciales et l'orientation de la tête. Cette relation peut être obtenue par *apprentissage* ou *géométriquement* (cf. § 2.4.2.2).

2.4.2.1 Alignement d'un modèle

Ces méthodes font appel à un modèle de forme du visage $\mathbf{s}$. Deux représentations sont possibles (cf. figure 2.8) : les modèles *rigides* et *non-rigides*. Parmi les modèles rigides, nous pouvons avoir un modèle en forme de plan (Black et Yacoob, 1995), de cylindre (Cascia *et al.*, 2000; Xiao *et al.*, 2003), d'ellipsoïde (Basu *et al.*, 2007; Morency *et al.*, 2008; Choi et Kim, 2008) ou de maillage (Vacchetti *et al.*, 2004). Les modèles non-rigides, quant à eux, utilisent un modèle déformable 3D du visage, tel que le modèle *Candide* (Rydfalk, 1987; Ahlberg, 2001), le *3D Morphable Model* (Blanz et Vetter, 1999) ou sa version plus détaillée (*Basel Face Model*, (Paysan *et al.*, 2009)), qui a l'avantage de pouvoir modéliser les expressions faciales. Il est également possible de créer son propre modèle soit manuellement soit en se basant sur un modèle existant (par décimation, par exemple).

Afin d'ajuster le modèle au visage cible, différents paramètres doivent être estimés :



(a) (b) (c)

Figure 2.8 – **Modèles de forme** : (a) modèle cylindre (Cascia *et al.*, 2000), (b) modèle *Candide-3* (Ahlberg, 2001), (c) *Basel Face Model* (Paysan *et al.*, 2009).

- ▷ **Paramètres externes** : ces paramètres permettent de modéliser le mouvement rigide d'un objet, notamment la roation, la translation et l'échelle.
- ▷ **Paramètres internes** : En plus des paramètres externes, les modèles non-rigides disposent d'un paramètre $\mathbf{p}_s$ qui permet la déformation du modèle. Nous le nommerons *paramètre de forme* par abus de language et son modèle de forme associé est $\mathbf{s}(\mathbf{p_s})$.

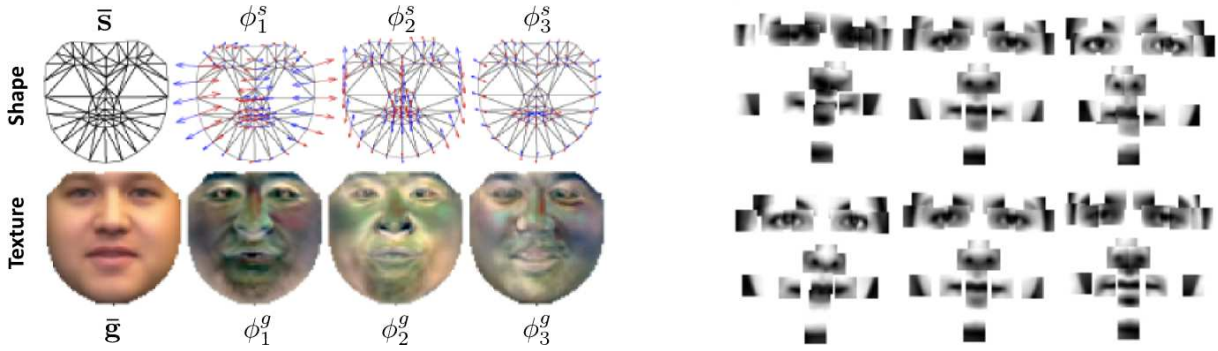
Active Appearance Models (AAMs) - Les modèles actifs d'apparences (Cootes *et al.*, 1998) sont des modèles paramétriques génératifs et sont les modèles les plus connus. Ils procèdent en deux étapes : la définition du modèle et l'ajustement (*fitting*) du modèle.

1) Définition du modèle : Initialement, le modèle est connu sous le nom de Modèle Actif de Forme (*Actif Shape Model*, ASM (Cootes *et al.*, 1995)). L'idée consiste à modéliser la forme du visage à l'aide de n points caractéristiques faciaux (*Point Distribution Model*, PDM (Cootes *et al.*, 2002)). Mathématiquement, la représentation de cette forme en dimension k peut se traduire par un vecteur $(k \times n)$-dimensionnel pour chaque image d'une base d'apprentissage. Avec une forme 2D ($k = 2$), nous obtenons le vecteur suivant :

$$\mathbf{s} = [x_1, y_1, \ldots, x_n, y_n]^\top \tag{2.6}$$

Ensuite, les différentes formes sont alignées par rapport à une référence à l'aide d'une analyse procustéenne (Stegmann et Gomez, 2002) qui permet la suppression des composantes de translation, d'échelle et de rotation. Les variations 2D de forme sont, ensuite, capturées en appliquant une Analyse en Composantes Principales (ACP) aux données précédemment normalisées et chaque exemple peut alors s'exprimer comme une *forme moyenne* $\bar{\mathbf{s}}$ plus une combinaison linéaire de n_s vecteurs propres ϕ^s ($n_s < k \times n$) (cf. figure 2.9(a)) :

$$\hat{\mathbf{s}} = \bar{\mathbf{s}} + \sum_{i=1}^{n_s} p_i^s \phi_i^s = \bar{\mathbf{s}} + \mathbf{\Phi}_s \mathbf{p}_s \tag{2.7}$$



(a) (b)

Figure 2.9 – **Alignement d'un modèle** : (a) *Active Appearance Model* (Gross *et al.*, 2006), (b) *Constrained Local Model* (Cristinacce et Cootes, 2006).

où $\mathbf{p}_s = \{p_i^s\}$ représente les paramètres de forme et $\mathbf{\Phi}_s = \{\phi_i^s\}$.
La technique fut, ensuite, étendue pour incorporer la connaissance de la texture (en plus de la forme) et donner lieu au Modèle Actif d'Apparence (*Active Appearance Model*, AAM (Cootes *et al.*, 1998)). L'alignement s'effectue par transformation affine $\mathbf{W}(\mathbf{x}; \mathbf{p}_s)$ de l'image vers la forme moyenne $\bar{\mathbf{s}}$ (pour $\mathbf{x} \in \bar{\mathbf{s}}$). Puis, après normalisation moyenne/écart-type, une ACP est appliquée au niveau de gris de l'image et nous obtenons une formulation similaire à celle de forme :

$$\hat{\mathbf{g}} = \bar{\mathbf{g}} + \mathbf{\Phi}_g \mathbf{p}_g \tag{2.8}$$

Ainsi, pour chaque exemple, un modèle de forme et de texture peut être synthétisé par les vecteurs $\mathbf{p}_s = \mathbf{\Phi}_s^\top(\mathbf{s} - \bar{\mathbf{s}})$ et $\mathbf{p}_g = \mathbf{\Phi}_g^\top(\mathbf{g} - \bar{\mathbf{g}})$ pour une forme test $\mathbf{s}$ et une texture test $\mathbf{g}$.

2) Ajustement du modèle : Il existe deux principales manières d'utiliser les modèles pour l'alignement à une nouvelle image (Matthews et Baker, 2004) : AAMs *combinés* ou *AAMs indépendants*. La première approche, qui est celle originalement proposée, capture la correlation entre les composantes $\mathbf{p}_s$ et $\mathbf{p}_g$ à l'aide d'une troisème ACP suivie d'une optimisation par descente de gradient. Quant aux AAMs indépendants, ils procèdent par optimisation en découplant les paramètres de forme et de texture :

$$\{\hat{\mathbf{p}}_s, \hat{\mathbf{p}}_g\} = \underset{\mathbf{p}_s, \mathbf{p}_g}{\operatorname{argmin}} ||\mathbf{I}(\mathbf{W}(\mathbf{x}; \mathbf{p}_s)) - \hat{\mathbf{g}}||_2^2 \tag{2.9}$$

Des extensions au modèles AAMs ont été proposées pour prendre en compte des orientations plus importantes de la tête (Cootes *et al.*, 2002; Gross *et al.*, 2006; Ramnath *et al.*, 2008).

Les principaux inconvénients des approches AAMs sont leur sensibilité aux changements d'illumination, leur biais vers le visage *moyen* (défini par la forme *moyenne*) et leur mauvaise généralisation à des visages différents de la base d'apprentissage (Gross *et al.*,

2005).

Autres modèles - De nombreux modèles s'inspirent également des modèles AAMs. Le *3D Morphable Model* (3DMM, (Blanz et Vetter, 1999; Storer *et al.*, 2009)) adopte une représentation de forme en dimension $k = 3$. Le *Pictorial Structure* (Felzenszwalb et Huttenlocher, 2005) introduit une méthode d'alignement de modèles locaux dont la forme est construite à l'aide de relations géométriques par paires. Le *Constrained Local Model* (CLM, (Cristinacce et Cootes, 2006)) utilise aussi une formulation basée sur le PDM. Mais contrairement aux modèles AAMs, il ne modélise pas l'apparence globale, mais plutôt l'apparence locale autour des points caractéristiques d'intérêt (cf. figure 2.9(b)) et combine ces prédictions indépendantes via une information a priori sur leur mouvement joint. Une procédure non-paramétrique régularisée d'alignement d'un modèle CLM est décrite dans (Saragih *et al.*, 2011).

2.4.2.2 Apprentissage vs. géométrie spatiale

Pour obtenir l'orientation de la tête, une relation entre les points caractéristiques faciaux localisés et la pose peut être obtenue par *apprentissage* ou *géométriquement*.

Les méthodes d'apprentissage reposent sur des algorithmes de classification ou de régression. Dans (Cootes *et al.*, 2002), une régression linéaire est utilisée pour estimer l'angle yaw à partir du modèle déformable AAM. Moon et Miller (2004) utilisent un *Support Vector Regression* (SVR, (Drucker *et al.*, 1996)) pour estimer la pose à partir de points caractéristiques du visage. Dans le même esprit, Ma *et al.* (2006) proposent de localiser 20 points faciaux et de calculer l'angle yaw et pitch de la tête par *Relevance Vector Regression* (RVR, (Tipping, 2001)). Les auteurs de (Martins et Batista, 2008) ont couplé un modèle AAM avec un algorithme *Pose from Orthography and Scaling with Iterations* (POSIT, (DeMenthon et Davis, 1995)) pour le suivi de l'orientation de la tête.

Quant aux méthodes géométriques, elles infèrent l'orientation de la tête à partir des positions relatives des points caractéristiques du visage tels que les yeux, la bouche et le nez. Par exemple, Gee et Cipolla (1994) proposent de modéliser l'orientation à l'aide de relations trigonométriques pour déterminer la direction du regard dans des tableaux de peinture. D'autres possibilités consistent à employer une régression linéaire ou polynomiale (Hu *et al.*, 2004), l'algorithme Espérance-Maximisation (Choi *et al.*, 1998) ou encore par géométrie projective (Wang et Sung, 2007). Après un alignement de type ASM, Zhou *et al.* (2003) procèdent à une inférence Bayesienne pour estimer les paramètres de pose.

2.4.3 Méthodes d'apparence globale

Les méthodes d'apparence globale s'intéressent à toute l'information de l'image du visage, à savoir les pixels contenus dans cette image, et elles ne cherchent donc pas à localiser des points caractéristiques faciaux. Cependant, des descripteurs sont extraits de l'image pour

rendre l'information de pose plus discriminante. Dans la suite, nous décrivons les méthodes permettant d'apprendre la relation entre ces descripteurs et la pose de la tête.

Appariement de gabarits (*Template matching*) - Les techniques par appariement de gabarits (Niyogi et Freeman, 1996; Beymer, 1994) comparent directement l'image d'un visage avec un ensemble d'exemples ; la pose est alors considérée égale à celle de l'exemple le plus proche. Ces approches sont, en général, sensibles aux changements d'identité, d'illumination et d'expression faciale étant donné que leur mise en correpondance s'effectue via une similarité par paire pour laquelle diverses fonctions de coût ont été employées : l'erreur quadratique moyenne (Niyogi et Freeman, 1996) ou la corrélation croisée normalisée (Beymer, 1994). En effet, une similarité dans l'espace d'image ne correspond pas forcément à une similarité dans l'espace de pose. De plus, ces méthodes n'ont pas recours à une étape d'apprentissage, c'est pourquoi elles requièrent beaucoup d'exemples pour obtenir une précision suffisante. Ainsi, estimer la pose revient à résoudre un problème de recherche le plus efficacement possible, c'est-à-dire en évitant de comparer le visage test à tous les exemples.

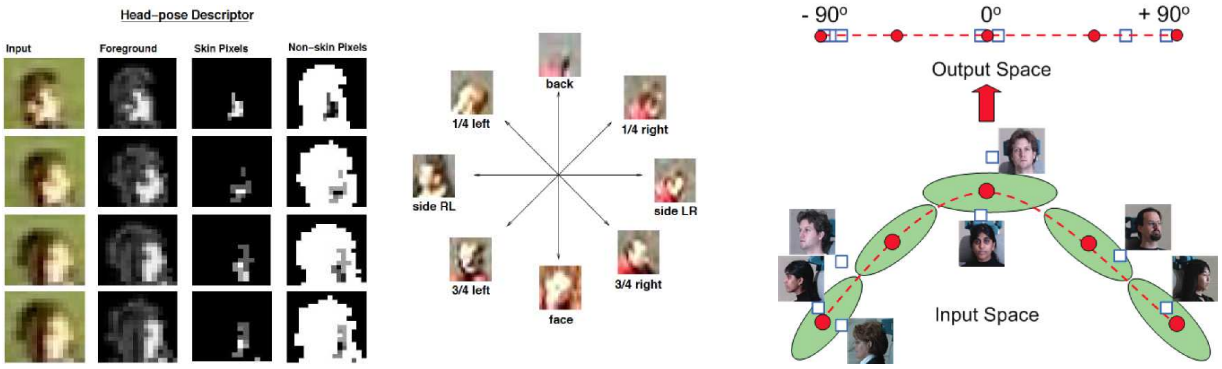
Classification - Ces méthodes apprennent une relation entre l'image d'entrée et l'espace *discret* de l'orientation de la tête. Cet espace discret est construit en regroupant des poses similaires en une classe et un classifieur est appris pour chacune des classes. Contrairement aux méthodes par appariement de gabarits, le visage test n'est pas comparé à tous les exemples, mais il est associé à la classe dont le score du classifieur est le plus élevé (dans certains cas, la pose finale peut également s'obtenir par une interpolation sur les scores des classifieurs, ce qui permet d'avoir une sortie continue plutôt que discrète). De plus, l'avantage de regrouper les poses en différentes classes permet de concentrer l'aprentissage sur les variations de pose et de limiter l'influence d'autres facteurs (cf. § 2.4.1.2).

Ces méthodes reposent, par exemple, sur un *Support Vector Machines* (SVM, (Vapnik, 1995)) multi-classe (Li *et al.*, 2001), une Analyse Discriminante Linéaire (*Linear Discriminant Analysis*, LDA) multi-classe ou encore sa version non-linéaire (KLDA, (Duda *et al.*, 2001)).

Parmi les approches SVM, Huang *et al.* (1998) proposent d'utiliser un noyau polynomial de degré 3 ou Gaussien pour apprendre 3 poses différentes. Plus récemment, dans des applications de vidéo-surveillance (Robertson, 2006; Orozco *et al.*, 2009), les SVM sont utilisés sur des images de faible résolution et l'espace de pose est divisé en 8 orientations (cf. figure 2.10(a)).

D'autres approches s'intéressent aux méthodes par *boosting*. Par exemple, en s'inspirant de (Viola et Jones, 2001), Jones et Viola (2003) proposent d'entraîner un détecteur multi-vue (frontal, profile, rotation) et un arbre de décision est, ensuite, appris pour déterminer la classe à laquelle un visage test appartient.

Régression - Les méthodes par régression apprennent une fonction entre l'image d'entrée et l'espace *continu* de la pose. En comparaison avec les techniques de type classification,



(a) (b)

Figure 2.10 – **Apparence globale** : (a) Exemple de descripteurs et de partitionnement de l'espace de la pose pour une méthode par *classification* (Robertson, 2006), (b) *Supervised Local Subspace Learning* (Huang *et al.*, 2011) : la méthode SL^2 apprend un ensemble de sous-espaces localisés paramétrés par un centre (cercle rouge) et un sous-espace tangent (ellipse verte).

elles n'ont pas besoin de partitionner les images en différentes classes, ce qui permet d'éviter les problèmes aux frontières des classes (appartenance à une classe plutôt qu'à une autre). Par ailleurs, une seule fonction est apprise à partir de toute la base d'apprentissage plutôt qu'un classifieur par classe, ce qui rend les méthodes par régression, en général, plus rapides dans la phase d'apprentissage, mais aussi de test. En revanche, elles ne profitent pas de l'avantage de la discrétisation de l'espace de pose (qui limite l'influence de facteurs externes) et elles doivent donc résoudre un problème de grande dimensionnalité.

De nombreuses méthodes utilisent les réseaux de neurones (Rae et Ritter, 1998; Stiefelhagen *et al.*, 2002), le *Support Vector Rgression* (Li *et al.*, 2004) ou le *Gaussian Process Regression* (Ranganathan et Yang, 2008). Plus récemment, l'algorithme *Kernel Partial Least Squares* (KPLS, (Rosipal et Trejo, 2001)) a été employé et modifié pour prendre en compte les erreurs d'alignement entre images (Al Haj *et al.*, 2012).

Une autre stratégie récente consiste à créer un modèle génératif pour l'estimation de la pose (Aghajanian et Prince, 2009). Cette méthode a été appliquée à des images acquises en environnements non contrôlés.

Projection sur une variété (*Manifold embedding*) - Ces approches consistent à projeter les données sur un espace de plus faible dimension via des techniques de réduction de dimensionnalité. Ensuite, l'orientation de la tête est estimée par classification (par exemple, à l'aide de la méthode des K plus proches voisins (KNN) (Fu et Huang, 2006)) ou par régression (par exemple, via une régression linéaire multivariée (Balasubramanian *et al.*, 2007)). Différentes réductions de dimensionnalité linéaires ont été proposées via ACP (Gong et Col-

lins, 1996) ou ACI (*Analyse en Composantes Indépendantes*), mais elles ne permettent pas de prendre en compte la non-linéarité engendrée par l'espace d'orientation de la tête. Des méthodes non-linéaires ont donc été proposées par *Isomap* (Tenenbaum *et al.*, 2000), *Local Linear Embedding* (LLE, (Roweis et Saul, 2000)) ou *Laplacian Eigenmaps* (LE, (Belkin et Niyogi, 2003)).

L'inconvénient des méthodes précédentes est qu'elles sont non-supervisées, c'est-à-dire qu'elles ne prennent pas en compte l'information de la pose. Ainsi, les variations de pose, mais également d'autres facteurs comme les variations inter-sujet, les changements d'illumination ou les expressions faciales, sont capturés dans un même espace, ce qui ne garantit pas d'avoir un espace suffisamment représentatif des variations de pose. Des méthodes par apprentissage supervisé ont donc été suggérées et permettent de partiellement résoudre ce problème. Avant l'étape de réduction de la dimensionnalité, un apprentissage de métriques par *Biased Manifold Embedding* (BME, (Balasubramanian *et al.*, 2007)) est proposé pour ne garder que les varations de pose, ce qui améliore les résultats en comparaison avec Isomap, LLE et LE. L'hypothèse sous-jacente est que des images de visages associées à des poses similaires ($P_i \approx P_j$) doivent être proches sur la variété et distantes en cas de poses différentes, cela indépendamment des variations intra-sujet. Ils procèdent en modifiant la topologie du voisinage par des contraintes par paires :

$$\hat{D}(i,j) = \mathcal{M}(D(i,j)) = \frac{\gamma \cdot P(i,j)}{\max\limits_{m,n} P(m,n) - P(i,j)} \cdot D(i,j) \qquad \forall i \neq j \tag{2.10}$$

avec $P(i,j) = |P_i - P_j|$ et γ une constante contrôlant la variation de la variété redéfinie. Sur ce même principe, BenAbdelkader (2010) propose une taxonomie de méthodes visant à incorporer l'information de la pose à différentes étapes de l'apprentissage de la variété, notamment pour la construction d'un graphe de voisinage.

Les méthodes précédentes reposent parfois sur une régression, et elles sont donc sujettes à des problèmes de sur-ajustement et de sensibilité à une base d'apprentissage non-uniforme. Pour partiellement résoudre ce problème, Huang *et al.* (2011) ont, récemment, proposé un modèle génératif appelé *Supervised Local Subspace Learning* (SL^2). Cette méthode apprend une mixture de modèles de sous-espaces tangents locaux (cf. figure 2.10(b)).

2.5 Application à la reconnaissance du comportement humain

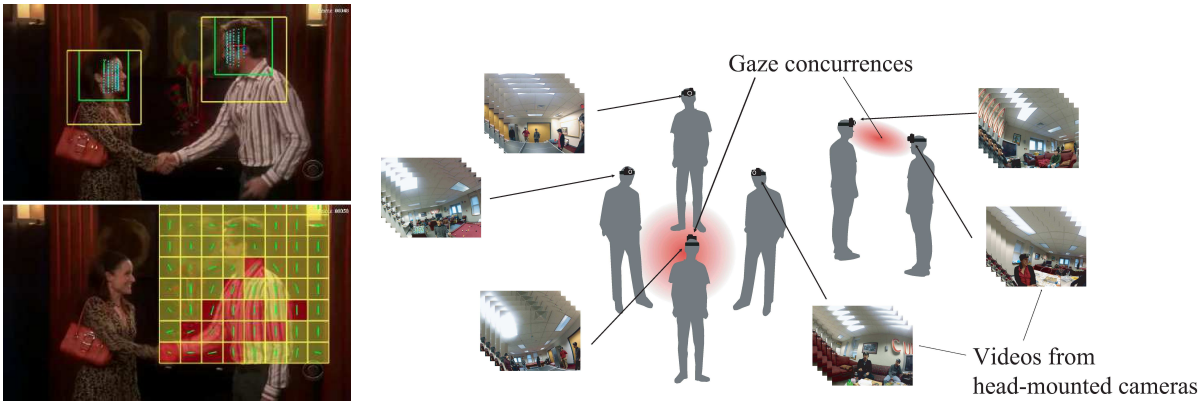
Il existe de nombreuses applications qui utilisent le regard comme signal d'entrée pour comprendre le comportement humain. Dans cette section, nous nous limiterons, en particulier, à la description d'applications dans un contexte, respectivement, social et individuel, à savoir l'*interaction sociale* et les *activités en vue subjective*.

2.5.1 Interaction sociale

Chez l'être humain, l'interaction sociale peut être vue comme une relation interhumaine par laquelle une action en provoque une autre en réponse, c'est-à-dire qu'elle implique un *échange* entre deux personnes. Cette interaction peut être soit *verbale*, soit *non-verbale* (via le regard, les expressions, les gestes ...). Dans la suite, nous nous intéressons uniquement à l'interaction sociale non-verbale basée sur le regard.

Vue externe - Stiefelhagen *et al.* (2002) et Ba et Odobez (2009) estiment les focus d'attention visuel dans des petits groupes de personnes (typiquement, lors de réunions) à l'aide de caméras placées en face des participants. Ils emploient, respectivement, un réseau de neurones et un suivi joint de la position et de la pose de la tête pour obtenir le focus d'attention visuel. Patron-Perez *et al.* (2010) proposent de reconnaître le type d'interaction entre deux personnes (se serrer la main, se taper dans la main, se serrer dans les bras et s'embrasser) dans des séries télévisées (cf. figure 2.11(a)). Leur descripteur est construit à partir de l'orientation de la tête pour délimiter une zone d'intérêt et d'une région autour de chaque personne ; la classification est effectuée à l'aide d'un SVM structuré (Joachims, 2006). Dans le même esprit, Marin-Jimenez *et al.* (2011) détectent si les personnes se regardent (*regard mutuel*). Leur approche de type *tracking-by-detection* repose sur le suivi du haut du corps et du visage par regroupement de détecteurs. L'orientation de la tête est, ensuite, obtenue via un GPR et un score est calculé pour savoir si les personnes se regardent mutuellement. Bazzani *et al.* (2011) introduisent une représentation 3D du champ visuel, ce qui permet de localiser la convergence de vues. Cristani *et al.* (2011) adoptent le concept de F-formation qui énumère toutes les configurations spatiales et angulaires des personnes afin de détecter et d'analyser les interactions sociales au sein des foules. Lan *et al.* (2012) présentent un modèle hiérarchique pour la reconnaissance d'activités multi-personne. Leur modèle repose sur des caractéristiques à différentes échelles, qui vont de la représentation d'actions de bas niveau jusqu'à celle d'évènements de haut niveau, en passant par la modélisation de rôles sociaux ; ils utilisent également un SVM structuré.

Vue subjective - Très récemment, des méthodes ont été proposées pour analyser l'interaction sociale en vue subjective, c'est-à-dire à partir d'une caméra montée sur la tête d'une personne. Fathi *et al.* (2012a) décrivent une chaîne de traitements qui permet de reconnaître des interactions sociales journalières en vue subjective. Tout d'abord, les visages sont localisés dans l'image et pour chacun de ses visages, le focus d'attention visuel est estimé. Ensuite, un champ aléatoire de Markov (MRF) est construit afin de prendre en compte l'interdépendance entre les orientations ; il permet de former un descripteur de rôles attentionnels (qui regarde qui, combien sont-ils à regarder une même personne ou un même endroit, existe-t-il un regard mutuel). En combinant le descripteur précédent à l'information du mouvement de la tête, ils apprennent un champ aléatoire conditionnel caché (HCRF, (Quattoni *et al.*, 2007)) pour intégrer une information temporelle et finalement



(a) (b)

Figure 2.11 – **Interaction sociale** : (a) Reconnaissance d'interaction dyadique à partir de l'orientation de la tête et de l'estimation du mouvement (Patron-Perez *et al.*, 2010), (b) Saillance sociale 3D construite à l'aide de plusieurs systèmes portés (Park *et al.*, 2012).

déterminer le type d'interaction sociale (monologue, dialogue, discussion), ceci suivant que la personne marche ou court. De plus, ils ont également construit un graphe de réseau social à partir du nombre de fois qu'un visage apparaît dans le champ de vue de la caméra.

Une approche complémentaire (Park *et al.*, 2012) a également été développée. Elle consiste à déterminer le point 3D d'intersection du regard de plusieurs personnes (en anglais, *3D gaze concurrence*). Ce point est un indicateur important de saillance sociale, car l'attention d'un groupe de personnes est centrée sur ce point spécifique (cf. figure 2.11(b)). De plus, il en existe plusieurs si différents groupes sont présents et ces multiples points saillants évoluent au cours du temps. Pour calculer la position de ce point attentionnel, Park *et al.* (2012) proposent d'équiper plusieurs personnes avec des caméras portées sur la tête. La direction du regard est approximée par une distribution en cône dont le sommet part du centre des yeux. Ensuite, ils reconstruisent la pose 3D de la caméra via l'estimation de structure à partir du mouvement (*Structure from Motion*, SfM) et modélisent la relation entre la pose de la caméra et la direction du regard. Finalement, un algorithme de recherche de modes multiples d'une distribution identifie les différents points attentionnels, ce qui permet de construire une carte 3D de saillance sociale.

2.5.2 Activités en vue subjective

Ces dernières années, le traitement et l'analyse en vue subjective ont connu un regain d'intérêt et incluent, en plus du suivi de la main, la reconnaissance d'objets ou encore la reconnaissance d'activités. Ces travaux sont importants, car ils ouvrent de nouvelles perspectives pour améliorer notre quotidien et peuvent être utilisés pour le suivi de patients malades ou dans le cadre des technologies assistives.

(a) (b)

Figure 2.12 – **Activités en vue subjective** : (a) Ego-mouvements lors d'activités sportives (Kitani *et al.*, 2011), (b) Reconnaissance objets-mains-actions-activités (Fathi *et al.*, 2011).

La reconnaissance d'activités peut être traitée de deux manières en considérant soit les mouvements du corps humain, soit le contexte (objet manipulé). Historiquement, la reconnaissance d'activités en vue subjective a connu ses débuts grâce à l'informatique "vestimentaire" et ubiquitaire (en anglais, *wearable and ubiquitous computing*). Parmi les nombreuses approches développées, Starner *et al.* (1998) ont été les premiers à s'intéresser à l'analyse en vue subjective. Ils décrivent une méthode pour reconnaître le language des signes américain à partir d'une caméra montée sur une casquette. Schiele *et al.* (1999) proposent un système de vision interactif et de réalité augmentée afin de retrouver des données médias basées sur les objets que l'utilisateur rencontre. A l'aide d'une caméra attachée à l'épaule, Mayol et Murray (2005) reconnaissent les gestes de la main en détectant les objets susceptibles d'être manipulés.

D'autres travaux utilisent des capteurs, tels que des centrales inertielles, positionnés sur le corps afin d'acquérir et traiter des données en perspective subjective et donnent de bons résultats pour la reconnaissance d'activité à long (Huynh *et al.*, 2008) ou court (Spriggs *et al.*, 2009) terme.

Plus récemment, Kitani *et al.* (2011) décrivent une méthode par apprentissage non-supervisé pour la détection d'égo-actions sportives en environnement extérieur (cf. figure 2.12(a)). Fathi *et al.* (2011) reconnaissent des égo-activités journalières en combinant la reconnaissance d'activités, d'actions et d'objets (cf. figure 2.12(b)).

Des travaux se sont aussi intéressés aux mouvements des yeux pour reconnaître et classer des actions. En effet, les mouvements oculaires sont une source d'information riche pour la reconnaissance d'activités. Doshi et Trivedi (2009) proposent d'utiliser la pose de la tête et des directions approximatives (devant, rétroviseurs intérieur/extérieur) du regard afin de prédire si l'automobiliste change de voie. A partir des mouvements oculaire, Courtemanche *et al.* (2011) décrivent une méthode pour identifier les interactions avec un écran à l'aide d'aires d'intérêt (*Area Of Interest*, AOI). L'inconvénient de ces méthodes est qu'elles nécessitent de définir des zones d'intérêt à l'avance, ce qui pour certaines applications n'est

pas envisageable.

A partir d'un système électro-oculographique, Bulling *et al.* (2011) proposent de catégoriser les mouvements oculaires en saccade, fixation et clignement. En se basant sur l'information spatiale et l'amplitude du signal, ces mouvements sont encodés en une séquence de symboles à partir de laquelle des statistiques sont extraites. Des statistiques relevantes sont, ensuite, sélectionnées via mRMR (Peng *et al.*, 2005) et permettent, finalement, d'entraîner un SVM linéaire pour reconnaître des activités en environnement intérieur. Ogaki *et al.* (2012) appliquent la méthode précédente (uniquement avec les saccades) en remplaçant l'EOG par la vidéo-oculographie et, en ajoutant une caméra scène, étendent la méthode pour intégrer les mouvements de la tête, ce qui améliore les résultats. Très récemment, Bulling *et al.* (2013) proposent une approche appelée *EyeContext* qui vise à reconnaître des signaux contextuels de haut niveau : social (interaction ou non), cognitif (concentré ou non), physique (actif ou non) et spatial (à l'intérieur ou non d'un bâtiment). Ils encodent les mouvements oculaires en une chaîne de caractères et emploient un *spectrum string kernel SVM* (Leslie *et al.*, 2002) pour la classification des signaux contextuels.

Dans un tout autre esprit, Fathi *et al.* (2012b) présentent une méthode pour identifier des activités journalières qui font appel à la coordination oeil-main et modélisent la relation spatio-temporelle entre le point de regard, les objets et l'activité effectuée. Ils arrivent également à prédire la séquence "optimale" des coordonnées du regard, ainsi que l'activité associée à partir d'une nouvelle séquence vidéo.

2.6 Conclusion

Dans ce chapitre, nous avons parcouru les différents points essentiels pour le suivi du regard. Tout d'abord, nous avons présenté les approches existantes pour estimer le point de regard à partir des yeux. Les eye-trackers déportés ont l'avantage d'être non-intrusifs et peuvent être très simples à mettre en place. Cependant, ils doivent compenser les mouvements de la tête et ces mouvements sont restreints à cause du champ de vue de la caméra. Les eye-trackers portés, quant à eux, offrent une plus grande mobilité et peuvent, via une caméra scène, capturer ce que "voit" le sujet.

Puis, nous nous sommes intéressés aux méthodes d'estimation de la pose de la tête comme indicateur plus général de la direction du regard. Elles ont recours soit à la localisation de points caractéristiques faciaux, soit à l'apparence globale du visage. La première approche est sensible à la détection erronée ou manquante de caractéristiques faciales et à des orientations importantes de la tête, alors que la seconde doit plutôt traiter le problème dans un espace de grande dimension.

Finalement, deux types d'application ont été abordés plus en détails : l'interaction sociale et les activités en vue subjective. Bien qu'étudiée depuis longtemps, l'analyse automatique de l'interaction sociale reste à ce jour un sujet important et difficile à traiter. L'analyse en vue subjective est, en revanche, une approche plus récente qui a pu se dé-

velopper grâce à la miniaturisation de capteurs et permet d'aborder un problème sous un autre angle de vue, centré sur le sujet.

Suite à cette étude, nous avons opté pour une approche de type eye-tracker porté afin d'offrir une certaine mobilité au sujet et de pouvoir connaître la position du regard dans l'environnement (chapitre 3). De plus, grâce à cette configuration, il est possible de traiter les données en vue subjective, ce qui ouvre de nouvelles perspectives intéressantes pour l'analyse automatique du comportement humain que nous verrons plus tard dans les chapitres 4 et 5.

CHAPITRE 3

Suivi du regard à partir d'un eye-tracker porté

"Chacun de nous ignore la couleur de l'iris de presque tous ses amis.
L'oeil est regard :
il n'est oeil que pour l'oculiste et pour le peintre."

– André Malraux –
dans *Les Voix du silence*

Sommaire

Dans ce chapitre, nous décrivons une nouvelle approche pour l'estimation du regard à partir d'un eye-tracker porté.

Nous proposons les deux principales contributions suivantes :

- ▷ **Caractéristiques locales** (cf. § 3.2.2) : Des caractéristiques issues du gradient de l'image sont extraites afin de constituer le modèle d'apparence de l'oeil.
- ▷ **Evaluation de modèles de régression** (cf. § 3.2.3) : Deux modèles de régression non-linéaire sont évalués dans la cadre de l'estimation du regard : Support Vector Regression (SVR) et Relevance Vector Regression (RVR).

Certains résultats de ce chapitre ont été publiés dans (Martinez *et al.*, 2012).

3.1 Introduction

Le suivi du regard permet de connaître les zones d'intérêt regardées par une personne. Il peut servir comme outil de communication pour l'interaction homme-machine, mais aussi d'étude comportementale dans le cadre des sciences cognitives. Ces derniers temps, les technologies d'eye-tracking ont fortement évolué à tel point qu'elles commencent à entrer dans le marché de masse, notamment via les nouvelles tablettes numériques.

Une autre voie, intéressante pour des applications futures, consiste à utiliser des systèmes portés par l'être humain et plus connu sous le nom de *mobile eye-tracking*. Différentes

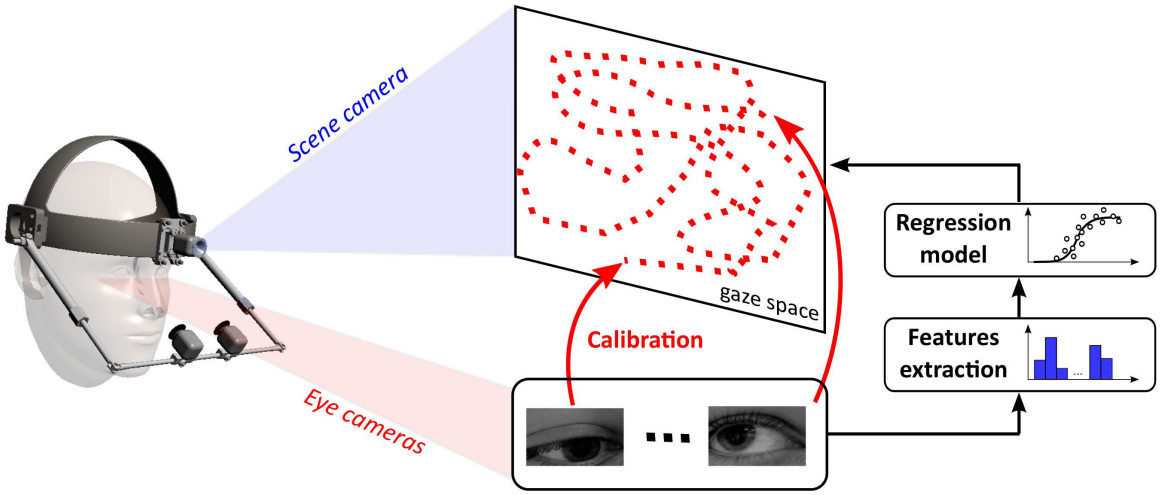


Figure 3.1 – Vue générale de l'estimation du regard.

approches ont été proposées et ont été vues plus en détails en § 2.3. Parmi elles, les systèmes en éclairage infrarouge donnent une meilleure précision de l'ordre de $0.5° - 1°$ d'angle visuelle, cependant leur utilisation de longue durée est peu recommandée pour des raisons de sécurité médicale. C'est pourquoi des recherches sont également menées pour le suivi du regard en éclairage visible.

Le but de ce chapitre est de poursuivre ces études sur le suivi du regard en éclairage visible et également de construire un système qui puisse servir pour une application future telle que celle décrite dans le chapitre 4. Nous présentons une méthode de suivi du regard basée sur un modèle d'apparence des yeux construit à partir de l'information du gradient de l'image.

3.2 Méthode proposée

Dans la figure 3.1, nous montrons la chaîne de traitement afin d'estimer le point de regard à partir des images des yeux. Elle est constituée de :

- ▷ *Phase de calibration* : Il s'agit de collecter des échantillons d'apprentissage pour lesquels nous disposons, à la fois, des images des yeux et des coordonnées des points de regard.
- ▷ *Définition du modèle d'apparence de l'oeil* : Des caractéristiques de gradient local sont extraites pour chaque image.
- ▷ *Construction du modèle de regard* : A partir des échantillons d'apprentissage, nous entrainons un modèle de régression pour estimer la relation entre les yeux et le point de regard.

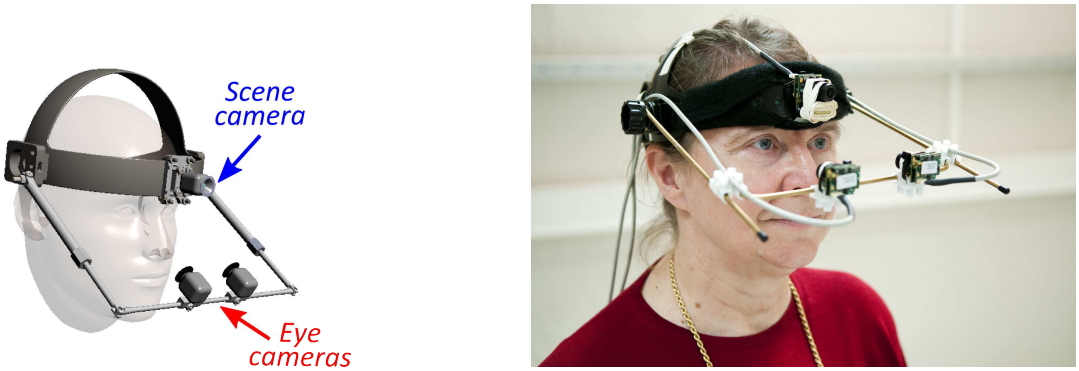


Figure 3.2 – Système eye-tracker binoculaire porté.

Finalement, pour un échantillon de test, nous appliquons le modèle de régression appris afin d'obtenir une estimation des coordonnées du point de regard.

3.2.1 Système eye-tracker porté

Tout d'abord, nous apportons quelques informations sur le système eye-tracker binoculaire porté (figure 3.2) qui a été réalisé. Il est composé de 3 caméras : 2 caméras permettent de capturer les images des yeux, tandis que la troisième caméra, appelée *caméra scène*, permet de capturer "ce que voit" le sujet. Le champ de vue horizontal de la caméra de scène est de 60°. Chacune des caméras délivre des images à la fréquence de 15 images par seconde et celles-ci ont une résolution de 640×576.

Une fois qu'une personne a mis le système sur la tête, la position et l'orientation des caméras des yeux sont ajustées de manière à ce que leurs champs de vue couvrent chacune un oeil. Les caméras sont légèrement placées en dessous des yeux et aussi légèrement inclinées vers le haut en diection des yeux de façon à limiter l'occultation due aux paupières (en particulier, la paupière supérieure). Ensuite, l'enregistrement des trois vidéos et les traitements des données *hors ligne* sont effectués sur un ordinateur portable.

3.2.2 Caractéristiques locales d'apparence

Les méthodes orientées apparence, aussi bien pour les systèmes déportés que portés, reposent, en grande majorité, sur un balayage ligne par ligne de l'intensité (normalisée) des pixels de l'image comme vecteur d'entrée de l'algorithme. Nous proposons, au contraire, d'utiliser l'information du gradient de l'image qui s'est montrée être une caractéristique pertinente dans les approches géométriques. Cependant, au lieu de l'utiliser sous forme de carte d'énergie de contours comme Williams *et al.* (2006), nous utilisons une approche

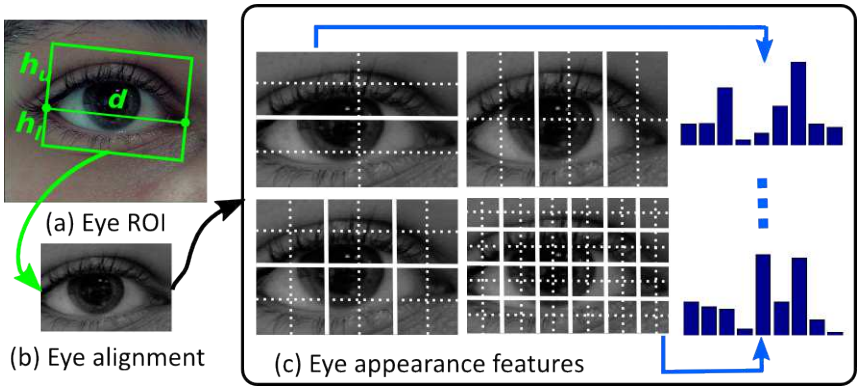


Figure 3.3 – Caractéristiques locales de l'oeil : (a) Région d'intérêt de l'oeil, (b) Alignement de l'oeil, (c) Extraction multi-échelle (1×2, 3×1, 3×2 et 6×4) des caractéristiques sous forme d'histogrammes de gradients orientés.

basée sur la construction d'histogrammes de gradients orientés[1] (HOG, Dalal et Triggs (2005)) à plusieurs échelles afin de capturer les variations oculaires. De plus, l'utilisation du descripteur HOG permet d'être robuste aux changements d'illumination.

Auparavant, une région d'intérêt de l'œil doit être définie afin d'extraire les caractéristiques. Dans la première image d'une vidéo capturée à l'aide d'une caméra oeil, les coins des deux yeux, l'œil gauche et droit, sont manuellement annotés. Ensuite, la région d'intérêt est définie de manière à ce qu'elle reste la même pour les images suivantes (cf. figure 3.3(a)) ; en effet, le système reste fixe par rapport à la tête quelques soient ses mouvements. Notons d la distance entre les coins de l'œil, la hauteur h de la région d'intérêt est divisée en $h_u = d/2$ et $h_l = d/4$ correspondant respectivement à la partie supérieure et inférieure de la région d'intérêt. Une fois la région d'intérêt définie, une rotation est appliquée à l'image de l'œil afin de l'aligner horizontalement par rapport aux coins de l'œil et l'image est, ensuite, redimensionnée à $w \times h = 120 \times 80$ (cf. figure 3.3(b)). L'amplitude et l'orientation du gradient sont calculées à l'aide de deux filtres dérivatifs 1D : $[-1\ 0\ 1]$ et $[-1\ 0\ 1]^\top$. Nous utilisons une orientation signée ($0°-360°$) afin de prendre en compte la différence de contraste entre les différentes parties de l'œil (iris, scléra, paupière). Des descripteurs HOG sont alors extraits avec différentes grilles de 1×2, 3×1, 2×3 et 4×6 blocs comme illustré sur la figure 3.3(c). Chaque bloc est composé de 2×2 cellules à l'intérieur desquelles un histogramme est construit avec $N_b = 9$ intervalles. Chaque bloc est, ensuite, normalisé via la norme L_2. Finalement, les descripteurs des deux yeux sont concaténés pour ne former qu'un vecteur de taille $M = 2520$. Les descripteurs HOG multi-échelle sont calculés rapidement à l'aide d'histogrammes intégraux et permettent d'obtenir un vecteur

[1] qui sont très similaires aux descripteurs SIFT Lowe (2004), mais ils utilisent une normalisation différente.

de faible dimension à partir de l'image haute résolution d'entrée : $M \ll w \times h$.

3.2.3 Modèle de régression

Nous formulons l'estimation du regard comme une approximation de fonction. Connaissant une base d'apprentissage $\mathcal{D} = \{(\mathbf{x}_i, t_i) \in \mathcal{X} \times [0,1] \,|i = 1, ..., N\}$ avec $\mathbf{x}_i$ le vecteur des caractéristiques d'entrée et t_i la valeur de sortie, le but est d'inférer une fonction $f : \mathcal{X} \to [0,1]$ entre le vecteur caractéristique de l'œil et la coordonnée du regard normalisée (par rapport à la largeur ou la hauteur de l'image de la caméra scène selon que le régresseur soit appris pour la coordonnée horizontale ou verticale du point de regard).

Nous étudions deux modèles de régression et, pour chacun de ces modèles, deux régresseurs 1D sont appris indépendamment, un pour chaque dimension de l'espace de sortie (coordonnée horizontale ou verticale). Ici, nous décrivons brièvement les modèles de régression et le lecteur est invité à se référer à, respectivement, (Drucker *et al.*, 1996; Smola et Schölkopf, 2004) et (Tipping, 2001; Tipping et Faul, 2003) pour plus de détails.

3.2.3.1 Régression à vecteurs de support

L'approche SVR (Smola et Schölkopf, 2004) consiste à apprendre une fonction de régression linéaire dans un espace de grande dimension où les données d'entrée ont été transformées à l'aide d'une fonction non-linéaire :

$$t_i = \sum_{j=1}^{N} w_j \phi_j(\mathbf{x}_i) + b \tag{3.1}$$

où ϕ_j caractérise la transformation de l'espace de représentation des données d'entrée vers un espace de grande dimension muni d'un produit scalaire, c'est-à-dire[2] $\phi_j(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}_j)$ et b est le biais.

Le principe de minimisation du risque structurel est appliqué permettant ainsi d'obtenir un compromis entre la complexité du modèle et le risque de sur-ajustement. Les paramètres optimaux du modèle à vecteurs de support sont estimés en minimisant une fonction d'erreur régularisante via une optimisation minimale séquentielle (SMO, (Platt, 1998)).

Pour des raisons de commodité, nous utilisons ν-SVR au lieu de ϵ-SVR afin de contrôler, à la fois, le nombre de vecteurs de support et l'erreur d'apprentissage par l'intermédiaire du paramètre ν.

[2]Nous adoptons une notation sous forme de fonction de base (*basis function*), similaire à (Tipping, 2001), afin de faciliter la comparaison entre les deux modèles de régression employés.

3.2.3.2 Régression à vecteurs de relevance

Formulation - Bien que similaire au SVR sur la forme fonctionnelle, RVR (Tipping, 2001), dont le modèle graphique est illustré par la figure 3.4(a), est une technique d'apprentissage Bayesienne qui permet de calculer les poids du régresseur à partir des données d'apprentissage :

$$t_i = \sum_{j=1}^{N} w_j \phi_j(\mathbf{x}_i) + w_0 + \epsilon_i = \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_i) + \epsilon_i \qquad \forall i \tag{3.2}$$

où $\mathbf{w} = [w_0, \ldots, w_N]^\top$ représente l'ensemble des poids, $\boldsymbol{\phi}(\mathbf{x}) = [1, \phi_1(\mathbf{x}), \ldots, \phi_N(\mathbf{x})]^\top$ incluant une valeur unitaire pour tenir compte du biais et le bruit d'observation est modélisé par $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Plus généralement, à partir de l'équation précédente, nous obtenons l'expression suivante :

$$\mathbf{t} = \boldsymbol{\Phi}\mathbf{w} + \epsilon \tag{3.3}$$

avec $\boldsymbol{\Phi} = [\boldsymbol{\phi}(\mathbf{x}_1), ..., \boldsymbol{\phi}(\mathbf{x}_N)]^\top$, connue sous le nom de *design matrix* :

$$\boldsymbol{\Phi} = \begin{bmatrix} 1 & k(\mathbf{x}_1, \mathbf{x}_1)) & \cdots & k(\mathbf{x}_1, \mathbf{x}_N)) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & k(\mathbf{x}_N, \mathbf{x}_1)) & \cdots & k(\mathbf{x}_N, \mathbf{x}_N)) \end{bmatrix} \tag{3.4}$$

Ainsi, en tenant compte de l'indépendance des distributions, la vraisemblance sur les données peut s'écrire[3] :

$$p(\mathbf{t}|\mathbf{w}, \sigma^2) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_i), \sigma^2) \tag{3.5}$$

Afin d'obtenir une solution éparse, la probabilité a priori sur les poids du régresseur est supposée être une distribution gaussienne de moyenne nulle et est controllée par un hyperparamètre indépendant pour chaque poids :

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=0}^{N} \mathcal{N}(w_i|0, \alpha_i^{-1}) \tag{3.6}$$

où $\boldsymbol{\alpha} = [\alpha_0, \ldots, \alpha_N]^T$ et permettent de contrôler la déviation des poids par rapport à la moyenne nulle et donc de privilégier une solution plus lisse.

Inférence - Le but est d'inférer les paramètres inconnus du modèle à partir des données d'apprentissage :

$$p(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{w}, \boldsymbol{\alpha}, \sigma^2)p(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2)}{p(\mathbf{t})} \tag{3.7}$$

[3] en omettant le conditionnement par rapport à $\mathbf{x}_1, \ldots, \mathbf{x}_N$

Cependant, il est impossible d'estimer l'intégrale $p(\mathbf{t})$ directement. Pour contourner ce problème, il est nécessaire de décomposer la distribution a posteriori de la façon suivante :

$$p(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2|\mathbf{t}) = p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \sigma^2)p(\boldsymbol{\alpha}, \sigma^2|\mathbf{t}) \tag{3.8}$$

Pour le premier terme de (8), en appliquant le théorème de Bayes et en utilisant la gaussianité des distributions, il est possible de montrer que la distribution a posteriori des poids $p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \sigma^2)$ est une distribution gaussienne de moyenne $\boldsymbol{\mu} = \sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t}$ et de covariance $\boldsymbol{\Sigma} = (\mathbf{A} + \sigma^{-2}\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$ avec $\mathbf{A} = diag(\alpha_0, ..., \alpha_N)$ d'après (Tipping, 2001).
Concernant le second terme de (8), les valeurs les plus probables (MP) des hyperparamètres, $\boldsymbol{\alpha}_{MP}$ et σ^2_{MP}, sont obtenues par la procédure d'optimisation rapide du maximum de vraisemblance (Tipping et Faul, 2003) qui consiste à maximiser (localement) la vraisemblance marginale $p(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2)$ par rapport aux hyperparamètres :

$$\mathcal{L}(\boldsymbol{\alpha}) = \log p(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2) = -\frac{1}{2}[N \log 2\pi + \log |\mathbf{C}| + \mathbf{t}^T\mathbf{C}^{-1}\mathbf{t}] \tag{3.9}$$

avec la covariance $\mathbf{C} = \sigma^2\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^T$.

Prédiction - Finalement, la prédiction y_* pour un vecteur test d'entrée $\mathbf{x}_*$ est donnée par la formule suivante (Tipping, 2001) :

$$p(t_*|\mathbf{t}, \boldsymbol{\alpha}_{MP}, \sigma^2_{MP}) \sim \mathcal{N}(t_*|y_*, \sigma^2_*) \tag{3.10}$$

avec :

$$y_* = \boldsymbol{\mu}^T\boldsymbol{\phi}(\mathbf{x}_*) \tag{3.11}$$

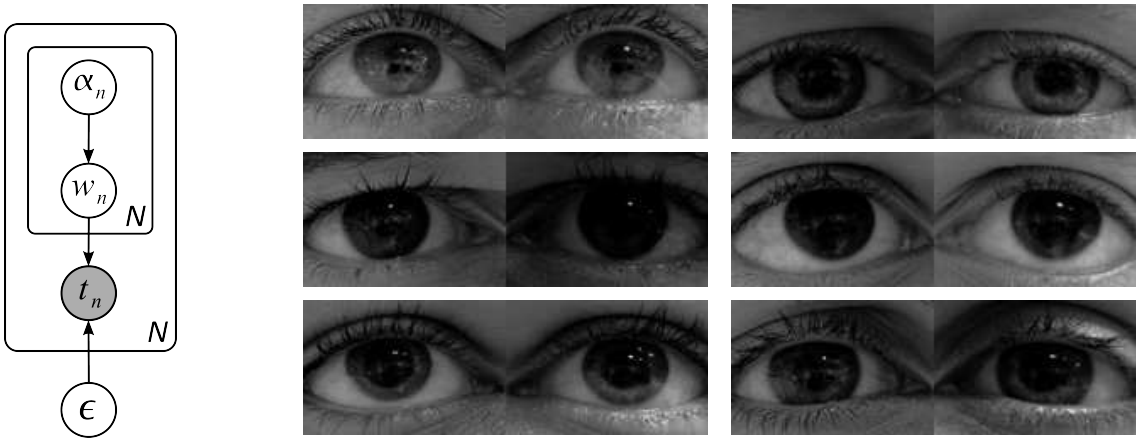
$$\sigma^2_* = \sigma^2_{MP} + \boldsymbol{\phi}(\mathbf{x}_*)^T\boldsymbol{\Sigma}\boldsymbol{\phi}(\mathbf{x}_*) \tag{3.12}$$

3.3 Résultats expérimentaux

3.3.1 Base de données

Afin d'évaluer la méthode proposée, nous avons utilisé le système précédemment décrit en section 3.2.1. Pour cela, nous avons créé une base de données. La base de données est constituée de 13 sujets (10 hommes et 3 femmes) ayant des yeux de différentes couleurs, allant des yeux clairs aux yeux foncés. Quelques exemples d'apparence des yeux sont montrés dans la figure 3.4(b).

Pour collecter les données, il est demandé au sujet de suivre une cible (déplacée par une autre personne) avec les yeux et de garder la tête immobile pour permettre à la cible de parcourir une grande partie du champ de vue de la caméra de scène. La distance de calibration, c'est-à-dire la distance entre les yeux et le plan où la cible est déplacée, est



(a) (b)

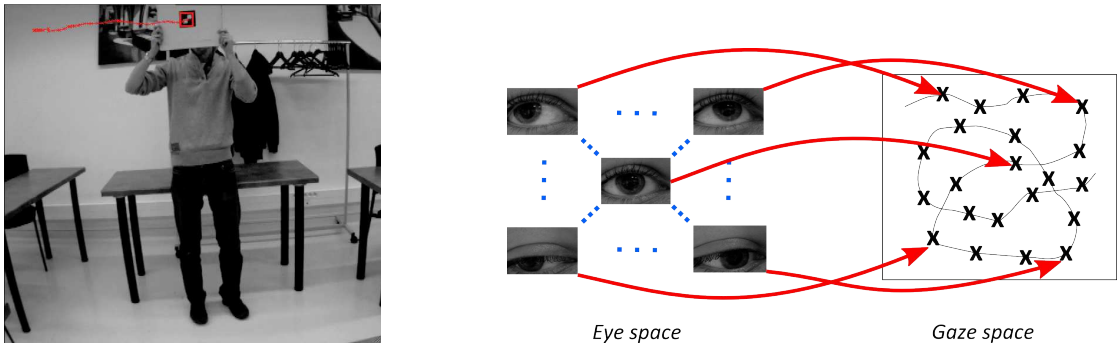
Figure 3.4 – (a) Modèle graphique du Relevance Vector Machine (Tipping, 2001). Le noeud grisé indique que la variable est observée. (b) Exemples d'images de la base de données pour 6 sujets différents.

d'environ 1m80. Afin de simplifier l'annotation, le centre de la cible (un motif 2×2 de carrés blancs et noirs) est automatiquement détecté et en cas d'échec, le suivi de la cible est manuellement réinitialisé (cf. figure 3.5(a)). Deux vidéos par sujet ont été enregistrées. La première vidéo sert à entraîner les régresseurs qui sont évalués sur la seconde vidéo. Ensuite, l'expérience est répétée en interchangeant les vidéos pour considérer différentes trajectoires du point de regard. L'erreur d'estimation du regard est calculée en prenant la moyenne des deux expériences. Pour chaque expérience, nous disposons, respectivement, de 200 et d'environ 400 échantillons d'apprentissage et de test par sujet.

3.3.2 Procédure de calibration et mesures de performance

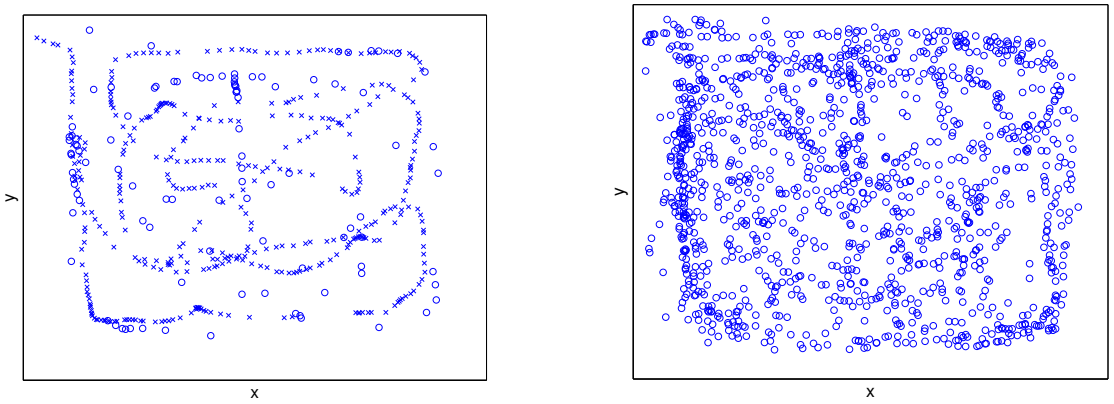
(1) Procédure de calibration - Pour chaque nouveau sujet, une calibration individuelle est effectuée, comme décrite précédemment (cf. § 3.3.1), afin de collecter des images qui vont permettre de modéliser les variations des yeux, propres au nouveau sujet (cf. figure 3.5(b)). Ensuite, le modèle de régression est construit à partir des données collectées. Sur la figure 3.6(a), nous montrons un exemple de distribution des points de regard pour un sujet, plus particulièrement avec $N = 100$ échantillons d'apprentissage et 400 échantillons de test.

(2) Mesures de performance - Les performances des deux modèles de régression ont, ensuite, été évaluées en terme d'erreur de prédiction, de robustesse au nombre réduit d'échantillons de calibration et de l'éparsité de la solution. Nous calculons l'erreur moyenne absolue angulaire (MAAE pour *Mean Angular Absolute Error* en anglais) ainsi que l'écart-type



(a) (b)

Figure 3.5 – Procédure de calibration individuel avec une trajectoire aléatoire de la cible couvrant le champ de vue de la caméra scène : (a) Exemple de calibration en pratique, (b) Relation entre l'apparence des yeux et les coordonnées du point de regard.



(a) (b)

Figure 3.6 – Distribution des points de regard : (a) pour un sujet avec $N = 100$ échantillons d'apprentissage et 400 échantillons de test (les ronds et les croix correspondent, respectivement, aux échantillons d'apprentissage et de test), (b) pour l'optimisation des paramètres avec 13 sujets et $N = 100$ échantillons par sujet.

entre la vérité terrain et l'estimation. Ces mesures de performance sont exprimées en degré d'angle visuel (°) et pour chacune des dimensions de l'espace de sortie, c'est-à-dire horizontalement et verticalement.

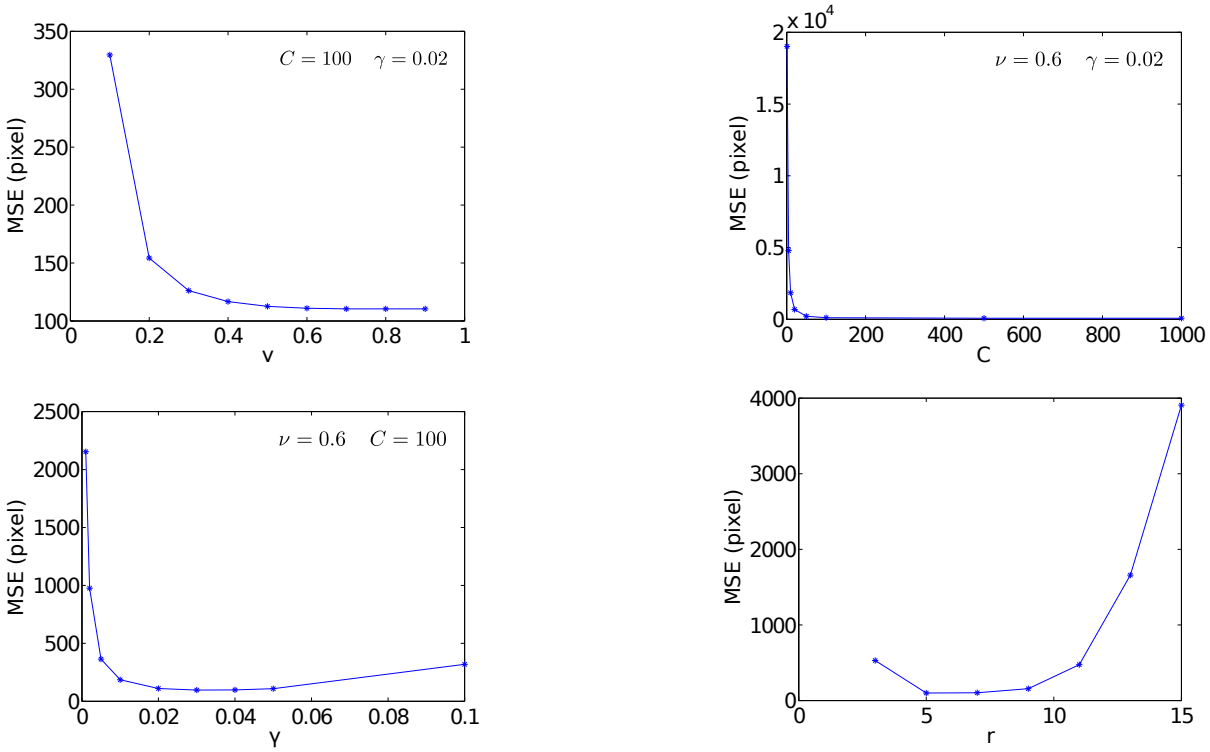


Figure 3.7 – Optimisation des pramètres (résultats unidimensionnels) avec l'erreur quadratique moyenne en ordonnée (exprimée en pixels) et le paramètre à optimiser en abscisse. Les trois premières courbes correspondent aux paramètres $\{\nu, C, \gamma\}$ du SVR, la dernière au paramètre r du RVR.

3.3.3 Optimisation des paramètres

Les paramètres des deux modèles de régression sont globalement optimisés pour éviter de les réestimer indépendamment pour chaque sujet et ils ne sont donc pas forcément optimaux pour chacun des sujets.
Dans les deux modèles de régression décrits précédemment, nous employons une fonction à base radiale (*Radial Basis Function*, RBF) comme noyau :

$$k(\mathbf{x}, \mathbf{x}_i) = \exp(-\gamma \|\mathbf{x} - \mathbf{x}_i\|_2^2) \tag{3.13}$$

avec γ le paramètre du noyau RBF. Nous avons conservé 100 échantillons d'apprentissage aléatoirement choisis par sujet, ce qui fait un total de 1300 échantillons. La distribution des points de regard est illustrée par la figure 3.6(b). Les paramètres ont été optimisés en effectuant une validation croisée en 10 groupes sur la base d'apprentissage.

Pour SVR, une grille de recherche sur les paramètres $\{\nu, C, \gamma\}$ a été établie. ν permet de contrôler la proportion de vecteurs de support et C représente le coût d'erreur de

classification (il permet d'obtenir un compromis entre l'erreur d'apprentissage et la largeur de la marge). Dans le cas de RVR, seul le paramètre de largeur de base $r = \gamma^{-1/2}$ a été optimisé à l'aide d'une grille unidimensionnelle (les hyperparamètres $\boldsymbol{\alpha}$ et σ^2 étant déterminés automatiquement, cf. § 3.2.3.2).

Les grilles de recherche sont les suivantes : $\nu \in \{0.1, ..., 0.9\}$, $C \in \{1, ..., 1000\}$, $\gamma \in \{0.001, ..., 0.1\}$ pour SVR et $r \in \{3, ..., 15\}$ pour RVR. Les résultats (unidimensionnels) sont présentés dans la figure 3.7. Les paramètres optimaux obtenus sont : $\nu = 0.6$, $C = 100$, $\gamma = 0.02$ pour SVR et $r = 7$ pour RVR. Avec le choix de ces paramètres, aucun des deux modèles de régression n'est avantagé puisque $\gamma \simeq \frac{1}{r^2}$.

3.3.4 Résultats

Dans cette section, nous proposons d'étudier le comportement des deux modèles de régression. Deux analyses expérimentales sont effectuées reposant, pour l'apprentissage, sur : un modèle *intra-sujet* avec N échantillons de calibration du sujet concerné, et un modèle avec connaissance *a priori* avec N échantillons de calibration du sujet concerné plus N échantillons par sujet pour les 12 autres sujets. Dans les deux cas, les paramètres des régresseurs employés sont ceux obtenus en § 3.3.3 et nous étudions l'influence du nombre d'échantillons de calibration N et l'éparsité de la solution obtenue. L'étude de l'influence du nombre d'échantillons permet de connaître le nombre minimum d'échantillons de calibration requis pour obtenir une précision suffisante. L'analyse de l'éparsité, quant à elle, nous renseigne sur la généralisation de l'algorithme, ainsi que sur le nombre effectif d'échantillons employés pour effectuer la régression. Elle s'obtient en calculant le ratio entre le nombre de vecteurs de support ou de relevance (suivant l'algorithme appliqué) sur le nombre total d'échantillons de calibration.

Il est important de souligner que tous les résultats ont été obtenus en sélectionnant N échantillons de calibration aléatoirement parmi les 400 échantillons disponibles et, par conséquent, ils ne correspondent pas forcément à une sélection optimale des échantillons. Cependant, cette méthode est adoptée pour des raisons de simplicité de mise en oeuvre. Une autre possibilité aurait été de sélectionner les échantillons selon une grille uniforme (ce qui est avantageux pour un modèle de régression) et de choisir le(s) plus proche(s) voisin(s) à un endroit donné de la grille. Néanmoins, cette méthode nécessite un échantillonnage assez dense de l'espace de sortie, ce qui n'est pas vraiment le cas avec les données obtenues en § 3.3.1. De plus, afin d'obtenir un échantillonnage suffisamment dense, la procédure de calibration serait plus longue. En moyenne, notre procédure de calibration prend environ 30s pour collecter 400 échantillons.

3.3.4.1 Résultats *intra-sujet*

(1) Influence du nombre d'échantillons de calibration N - La figure 3.8(a) présente les résultats sous forme de barres d'erreur et en fonction du nombre d'échantillons de

calibration pour une calibration individuelle. Nous représentons, respectivement, l'erreur selon la coordonnée horizontale x et verticale y en fonction de N. Nous pouvons remarquer que l'erreur de prédiction du regard diminue fortement en fonction du nombre d'échantillons de calibration et tend à se stabiliser autour de $N = 100$ avec une erreur moyenne de prédiction avoisinant 2°. Pour un nombre N faible, l'erreur reste cependant assez élevée. En ce qui concerne les modèles de régression, RVR donne de meilleurs résultats que SVR pour N significativement faible. Cela peut en partie s'expliquer à cause d'une meilleure généralisation due à une solution plus éparse que pour SVR. En revanche, les résultats apparaissent similaires et même en faveur de SVR lorsque $N > 100$. Au-delà de $N = 100$, le gain en terme de précision reste négligeable et doit être analysé parallèlement au nombre d'échantillons de calibration qui, lui, augmente considérablement. Il est également à noter que ces différences sont plus importantes selon la coordonnée x que selon la coordonnée y.

(2) Analyse de l'éparsité - La figure 3.8(b) indique que le nombre de vecteurs de support (SVs) augmente linéairement en fonction du nombre d'échantillons de calibration. En revanche, un nombre plus faible de vecteurs de relevance (RVs) sont requis pour l'estimation. Effectivement, SVR utilise la grande majorité des échantillons de calibration, alors que RVR ne garde que 20 à 60% (en fonction de N) de ces échantillons pour un résultat similaire ou supérieur. De plus, nous pouvons constater qu'il n'existe pas de réelle différence entre l'éparsité selon la cordonnée horizontale et la coordonnée verticale.

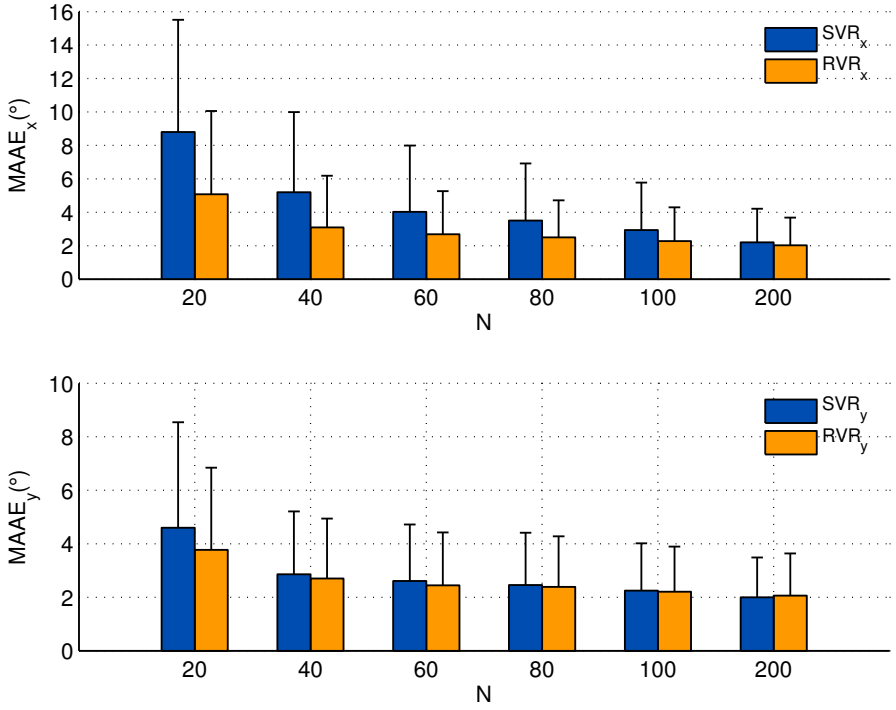
3.3.4.2 Résultats avec connaissance *a priori*

Nous avons également voulu voir quel pouvait être le gain en terme de prédiction du regard et du nombre d'échantillons de calibration, en incorporant la connaissance d'autres sujets, en plus du sujet impliqué.

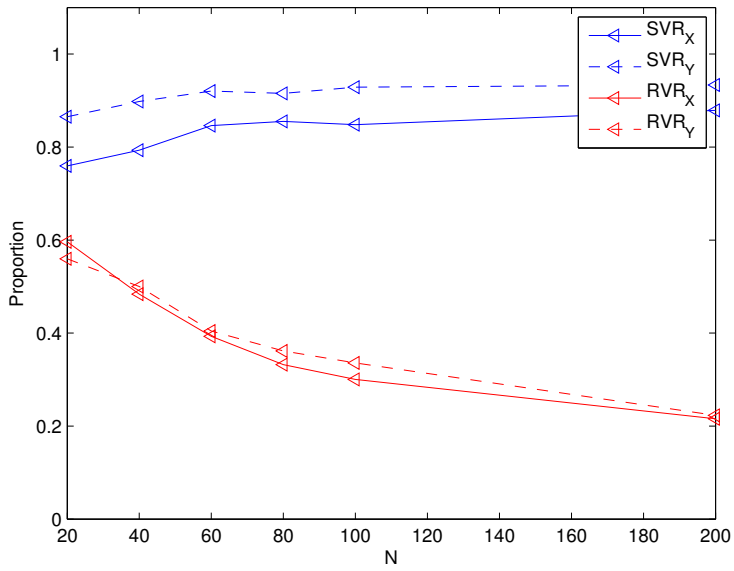
(1) Influence du nombre d'échantillons de calibration N - La figure 3.9(a) présente les résultats sous forme de barres d'erreur et en fonction du nombre d'échantillons de calibration pour une calibration avec *a priori*. Il est intéressant de voir que l'erreur de prédiction est faible (inférieure à 2° pour $N \geq 40$). Ces résultats permettent de mettre en évidence une certaine complémentarité entre les sujets lorsque peu d'échantillons sont disponibles dans certaines régions de l'espace de sortie. Ainsi, il est possible de réduire le nombre d'échantillons de calibration à collecter pour un nouveau sujet, si l'on dispose de données acquises avec d'autres sujets.

(2) Analyse de l'éparsité - Nous observons un comportement similaire avec les résultats *intra-sujet* (cf. § 3.3.4.1), mais avec une éparsité légèrement plus prononcée où 15 à 40% des échantillons sont utilisés (cf. figure 3.9(b)).

Pendant cette évaluation, nous avons également remarqué que du fait de l'ajustement

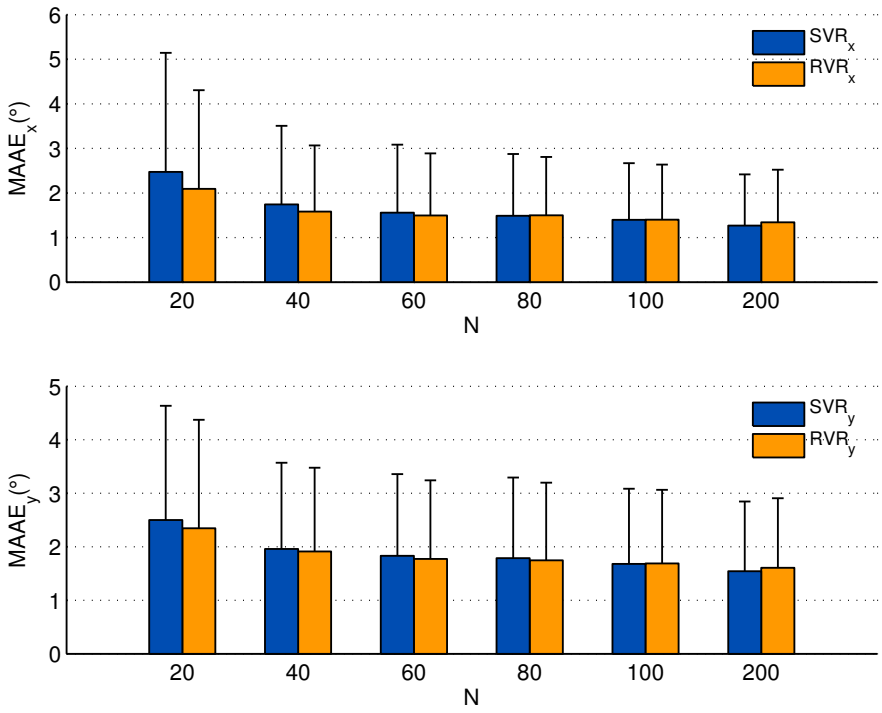


(a)

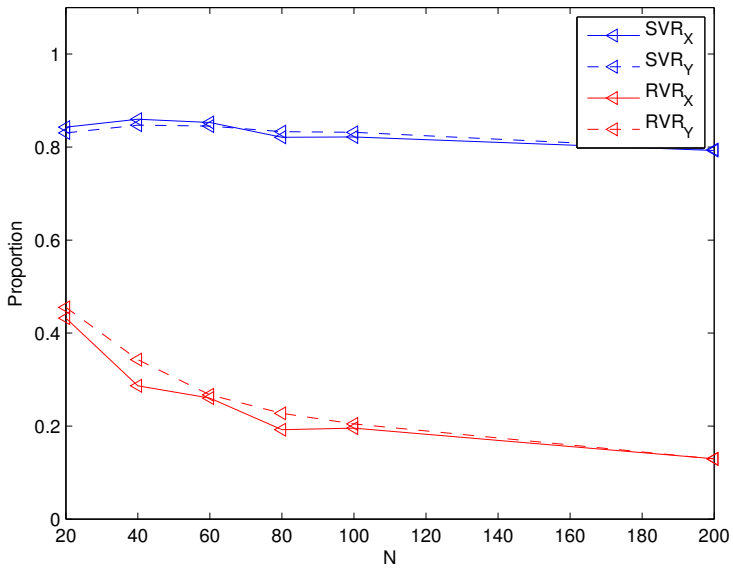


(b)

Figure 3.8 – Résultats *intra-sujet* : (a) Erreur moyenne de prédiction (en degré d'angle visuel), (b) Proportion d'échantillons de calibration retenue par le SVR et le RVR (obtenue en calculant le ratio entre le nombre de vecteurs de support ou de relevance sur le nombre total d'échantillons de calibration).



(a)



(b)

Figure 3.9 – Résultats avec connaissance *a priori* : (a) Erreur moyenne de pédiction et écart-type associé (en degré d'angle visuel), (b) Proportion d'échantillons de calibration retenue par le SVR et le RVR (obtenue en calculant le ratio entre le nombre de vecteurs de support ou de relevance sur le nombre total d'échantillons de calibration).

(position et orientation) manuel des caméras et des différences morphologiques des sujets, l'alignement des images entre divers sujets n'est pas optimal. En effet, nous avons aussi effectué une analyse en apprenant un modèle inter-sujet, c'est-à-dire avec 12 sujets et en le testant sur un autre sujet, et ceci pour les 13 sujets. Nous obtenons une erreur moyenne de 2,8° selon x et 5,3° selon y. En regardant plus en détails, l'erreur moyenne pour chacun des sujets, nous avons constaté que l'erreur selon y était très élevée pour 3 sujets en comparaison avec les autres sujets. Ce problème est dû à l'ajustement manuel des caméras qui fait que, pour un même point de regard, l'apparence entre des sujets est différente, en particulier l'ouverture de la paupière supérieure est plus ou moins importante. Si nous ne prenons pas en compte ces 3 sujets, nous obtenons une erreur moyenne de 2,7° et de 3,7° selon y. Nous pensons que là encore, le gain reste encore limité. Le modèle inter-sujet a également été évalué avec SVR et nous avons eu des résultats similaires.

Pour l'évaluation, nous pouvons affirmer que les bénéfices du modèle inter-sujet restent faibles pour certains sujets et ceci est principalement dû à la conception du système tête porté.

3.3.4.3 Résultats qualitatifs

Nous avons également enregistré deux séquences vidéo test afin d'illustrer comment la méthode proposée fonctionne : la première implique une seule personne dans le champ de vue de la caméra scène, alors que pour la seconde, deux personnes sont présentes. Pour cela, nous avons appliqué la même procédure qu'en § 3.3.4.1 : enregistrement d'une vidéo de calibration et de test, calibration *intra-sujet*, construction des modèles de régression RVR (pour x et y) et estimation du regard dans la vidéo de test. Notons également que le sujet de test n'était pas présent dans la base de données. Des résultats qualitatifs sont montrés dans la figure 3.10.

3.4 Conclusion

Nous avons proposé une méthode de suivi du regard en éclairage visible. A partir des images des yeux, des caractéristiques locales basées sur l'information du gradient à différentes échelles sont extraites et concaténées afin d'entraîner des régresseurs. Deux modèles de régression ont été évalués : la régression à vecteurs de support (SVR) et la régression à vecteurs de relevance (RVR). L'évaluation expérimentale montre que la seconde approche donne des résultats plus précis lorsque le nombre d'échantillons de calibration diminue. De plus, une méthode de calibration individuelle a aussi été proposée et permet de capturer les variations d'apparence des yeux avant de pouvoir apprendre ou tester un modèle de régression. La méthode que nous proposons donne des résultats satisfaisants ($\sim 2°$) avec environ 100 échantillons de calibration.

Nous avons également montré que si des données de calibration d'autres sujets sont

(a)

(b)

Figure 3.10 – Résultats du suivi du regard sur deux séquences vidéo test.

disponibles, alors il est possible de diminuer le nombre d'échantillons de calibration requis pour un nouveau sujet, tout en gardant une précision satisfaisante.

Finalement, nous avons aussi présenté quelques résultats qualitatifs à partir de deux séquences vidéo test comme étape préliminaire à notre prochaine contribution détaillée dans le chapitre 4.

CHAPITRE 4

Reconnaissance d'attention en vue subjective

"Nous ne sommes nous qu'aux yeux des autres
et c'est à partir du regard des autres
que nous nous assumons comme nous-mêmes."

– Jean-Paul Sartre –
dans *L'Être et le Néant*

Sommaire

Dans ce chapitre, nous présentons une nouvelle méthode de reconnaissance d'attention en vue subjective. Cette reconnaissance s'effectue en estimant le regard du sujet équipé d'un eye-tracker porté et celui des personnes présentes dans le champ de vue de la caméra scène.

Nous reprenons les travaux du chapitre 3 pour l'estimation du regard *subjectif* à l'aide d'un eye-tracker porté (cf. § 4.2) et nous proposons les deux principales contributions suivantes :

- ▷ **Estimation de la pose de la tête** (cf. § 4.3) : Lorsque la personne se trouve dans le champ de vue de la caméra scène, l'orientation de la tête permet d'obtenir une approximation de la direction du regard (cf. § 2.4). Le regard, que nous appellerons dorénavant *regard objectif*, est ensuite obtenu via un modèle cognitif.
- ▷ **Reconnaissance d'attention** (cf. § 4.4) : A partir du regard, l'attention de chaque individu vers une autre personne est modélisée. La combinaison de ces attentions permet alors d'identifier des motifs attentionnels tels que le *regard mutuel* et le *regard partagé*.

Les résultats de ces travaux ont été publiés dans (Martinez *et al.*, 2013).

4.1 Introduction

Les relations interhumaines permettent à des personnes de communiquer entre elles et constituent les bases de l'interaction sociale. Nous distinguons différents niveaux d'interaction suivant le nombre de personnes impliquées :

- ▷ **Interaction de face à face** : elle implique deux personnes qui communiquent entre elles.
- ▷ **Interaction de groupe** : elle se compose de 3 personnes au minimum et elle repose sur des mécanismes d'échanges plus complexes que pour l'interaction de face à face.
- ▷ **Interaction de foule** : l'interaction se traduit alors par la naissance de groupes abstraits ou temporaires qui participent souvent à des buts et des actions communes (auditoires, foules, rassemblements).

Lors d'interactions sociales, divers signaux sont échangés entre les différents interlocuteurs concernés. Parmi les signaux sociaux, le regard d'autrui joue un rôle important dans le cadre de l'interaction sociale. Il constitue un signal de communication riche et essentiel que nous décodons en fonction du contexte social dans lequel nous nous trouvons. Effectivement, la perception de divers regards déclenche des processus cognitifs distincts (Argyle *et al.*, 1973). Ainsi, un regard mutuel (ou *contact* par le regard) aura tendance à indiquer que l'intérêt d'autrui est porté vers nous-même. Par ailleurs, il constitue souvent le préambule fréquent aux interactions sociales et peut amener à susciter un regard partagé, c'est-à-dire que l'intérêt vers une même personne n'émane non plus d'une mais de plusieurs personnes.

De plus, le regard d'autrui est également un facteur de jugement de soi-même. En effet, nous pouvons, à ce titre, citer les travaux de C. H. Cooley (Cooley, 1902) qui affirme que la détermination du soi s'élabore par les relations sociales avec les autres, en particulier à travers le regard des autres. Ce concept de psychologie sociale est plus connu sous le nom de *Looking glass self*, c'est-à-dire l'image de soi réfléchie dans le regard porté par autrui sur soi.

Dans la suite, nous décrivons une méthode de reconnaissance d'attention en vue subjective. L'objectif est de pouvoir identifier différents motifs d'attention afin de savoir qui regarde qui et si regard mutuel et/ou partagé il y a. Nous nous intéressons à des interactions de groupe, en particulier des interactions triadiques. La chaîne de traitements proposée est illustrée par la figure 4.1.

4.2 Estimation du regard subjectif

Pour estimer le regard subjectif, nous utilisons la même approche que celle développée dans le chapitre 3, à savoir une méthode avec une représentation orientée apparence.

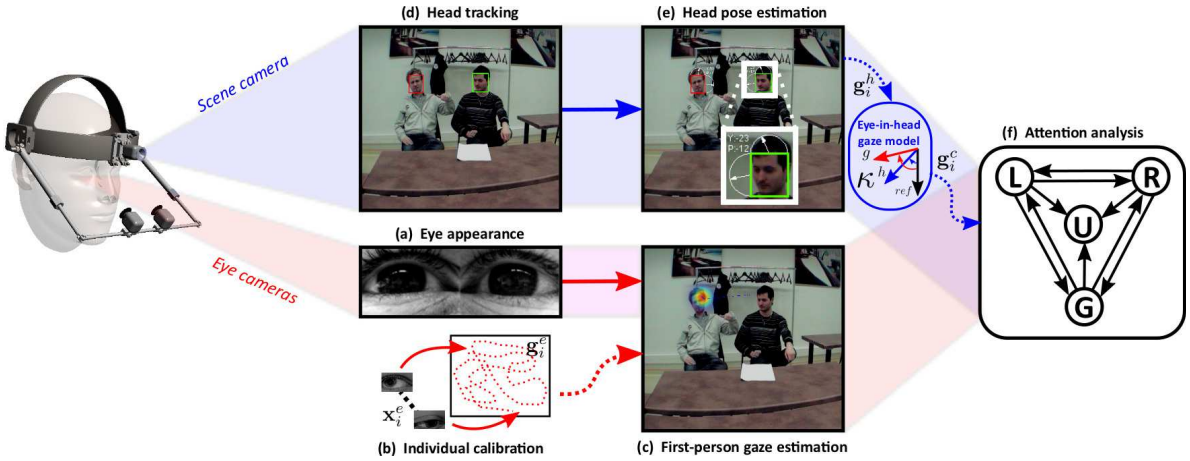


Figure 4.1 – Aperçu de la chaîne de traitements proposée : (a-c) Regard *subjectif*, (d-e) Suivi de la pose de la tête, (f) Reconnaissance et analyse d'attention.

Modèle d'apparence des yeux $\mathbf{x}^e$ - Tout d'abord, les caractéristiques, en particulier des HOG multi-niveaux, sont extraites afin de capturer les variations d'apparence associées au regard.

En supposant la position des coins des yeux connue[1], les images des yeux sont extraites et alignées. les images sont, ensuite, redimensionnées à la taille 120×80 et des caractéristiques HOG sont calculées sur des grilles de 1×2, 3×1, 3×2 et 6×4 blocs. Pour chaque bloc, des histogrammes de taille $N_b = 9$ sont construits au sein de cellules 2×2. Les caractéristiques sont extraites pour les deux yeux et concaténées pour former un vecteur final de taille égale à 2520.

Estimation du regard $\mathbf{g}^e$ - Nous construisons une base de données $\mathcal{D}_e = \{(\mathbf{x}_i^e, \mathbf{g}_i^e), i = 1 \dots N_g\}$ où $\mathbf{x}_i^e$ représente les caractéristiques d'apparence des yeux et $\mathbf{g}_i^e$ sont les coordonnées du regard. Nous entraînons un régresseur RVR (Tipping, 2001) différent pour chaque dimension de l'espace du regard, x et y :

$$g_i^e = \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i^e) \rangle + \epsilon_i \tag{4.1}$$

avec $\epsilon_i \sim \mathcal{N}(0, \sigma_e^2)$ le terme de bruit et $\boldsymbol{\phi}$ la matrice de fonctions de base construite à l'aide d'un noyau RBF :

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||_2^2}{r^2}\right) \tag{4.2}$$

où r est la largeur de base.

[1] nous identifions ces positions dans une seule image par oeil.

4.3 Estimation du regard objectif

4.3.1 Estimation continue localisée de la pose de la tête

Motivation - Nous considérons l'estimation de l'orientation de la tête comme une transformation des images du visage vers l'espace *continu* des poses de la tête. En particulier, nous nous intéressons aux méthodes par régression (cf. § 2.4.3). Traditionnellement, ces méthodes n'utilisent qu'un seul noyau pour effectuer la régression. Ces dernières années, dans le cadre des approches par classification, des méthodes dites *à noyaux multiples* ont fait leur apparition et consistent à combiner plusieurs noyaux suivant différentes manières. Nous invitons le lecteur à se référer à (Gönen et Alpaydin, 2011) pour plus de détails sur les différentes combinaisons possibles. Parmi celles-ci, l'approche *Multiple Kernel Learning* (MKL, (Bach *et al.*, 2004)) vise à construire une somme pondérée des valeurs de noyaux :

$$f(\mathbf{x}) = \sum_{m=1}^{P} \langle \mathbf{w}_m, \mathbf{\Phi}_m(\mathbf{x}) \rangle + b \tag{4.3}$$

dont les poids $\mathbf{w}_m$ de chaque noyau m sont optimisés lors de l'apprentissage. P désigne le nombre de noyaux et $\mathbf{\Phi}_m(\cdot)$ est la fonction noyau associée au $m^{\text{ième}}$ espace des caractéristiques.
Plus récemment, Gönen et Alpaydin (2008) ont présenté une méthode, appelée *Localized Multiple Kernel Learning* (LMKL) permettant de diviser l'espace d'entrée en plusieurs régions et d'apprendre des poids différents en fonction de la région sélectionnée (plutôt que d'avoir les mêmes poids pour l'ensemble des données d'entrée, comme c'est le cas pour MKL). Ainsi, il est possible de capturer les structures locales des données vu qu'à chaque donnée est associée un poids spécifique, d'ou le nom de *data-dependent combination*. Nous retrouvons cette même idée dans le cas des *Mixture-of-experts* (Jacobs *et al.*, 1991).

Les figures 4.2(a) et 4.2(b) illustrent le schéma de combinaison des poids pour, respectivement, MKL et LMKL[2].

Localized Multiple Kernel Regression - Pour estimer la pose de la tête, nous employons la methode *Localized Multiple Kernel Regression* (LMKR) proposée dans (Gönen et Alpaydin, 2010). Il s'agit de la version LMKL appliquée au problème de régression.
A partir des données d'apprentissage $\{\mathbf{x}_i^h, \mathbf{g}_i^h\}_{i=1}^N$ où $\mathbf{x}_i^h \in \mathbb{R}^d$ représente le descripteur des données et $\mathbf{g}_i^h \in \mathbb{R}^2$ les angles de rotations yaw α_i et pitch β_i, nous apprenons un régresseur (un pour chaque dimension de l'espace des poses) :

$$g_i^h = f_{\mathcal{R}}(\mathbf{x}_i^h) = \sum_{m=1}^{P} \eta_m(\mathbf{x}_i^h|\mathbf{V}) \langle \mathbf{w}_m, \mathbf{\Phi}_m(\mathbf{x}_i^h) \rangle + b \tag{4.4}$$

[2]Pour simplifier le schéma, le terme de biais b a été volontairement omis.

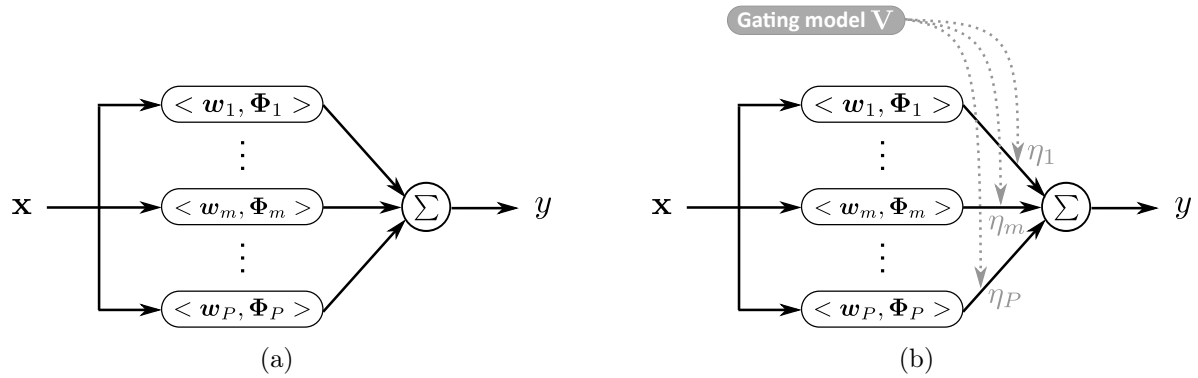


Figure 4.2 – Combinaison des poids : (a) *Multiple Kernel Learning* (Bach *et al.*, 2004), (b) *Localized Multiple Kernel Learning* (Gönen et Alpaydin, 2008).

avec $\eta_m(.|\mathbf{V})$ la fonction porte (*gating function*) correspondant au $m^{\text{ième}}$ espace des caractéristiques et $\mathbf{V}$ les paramètres de la fonction porte. A cause la nature non-convexe de l'optimisation jointe (4.4), une procédure d'optimisation alternée en deux étapes, similaire à (Rakotomamonjy *et al.*, 2008), est employée pour apprendre les coefficients de vecteurs de support et les paramètres de la fonction porte. Les coefficients de vecteurs de support sont appris avec un modèle de porte fixe, alors que le modèle de porte est mis à jour en calculant $\partial J(\mathbf{V})/\partial \mathbf{V}$ suivi d'une étape de descente de gradient ($J(\mathbf{V})$ étant la fonction objective de la formulation duale[3]). Pour le modèle de porte, nous utlisons une fonction softmax linéaire :

$$\eta_m(\mathbf{x}|\mathbf{V}) = \frac{\langle \boldsymbol{\nu}_m, \psi(\mathbf{x}) \rangle + \nu_m^0}{\sum_{h=1}^{P} \exp(\langle \boldsymbol{\nu}_h, \psi(\mathbf{x}) \rangle + \nu_h^0)} \tag{4.5}$$

où $\mathbf{V} = \{\boldsymbol{\nu}_m, \nu_m^0\}_{m=1}^P$ et $\psi(\mathbf{x}) = \mathbf{x}$ est une transformation linéaire permettant de diviser l'espace d'entrée et de sélectionner le régresseur associé. D'autres fonctions porte sont possibles telles que la fonction sigmoïde et il est également possible d'appliquer une transformation non-linéaire $\psi(\mathbf{x})$ auparavant.

Notons qu'une idée assez similaire a été proposée dans (Ho et Chellappa, 2012). Dans leur cas, des classifieurs, en particulier des SVM, sont appris pour diviser l'espace de pose de manière discrète. Ensuite, pour chaque classe, un SVR est appris afin d'obtenir une estimation continue et locale. Contrairement à (Ho et Chellappa, 2012), LMKL possède la particularité d'effectuer l'apprentissage conjointement.

[3]Son expression est disponible en annexe A.2.

4.3.2 De la pose de la tête au regard

Afin de pouvoir suivre le regard, nous intégrons deux étapes supplémentaires, en plus de l'estimation de la pose de la tête (cf. § 4.3.1) : l'une concerne le suivi du visage, l'autre permet d'obtenir une approximation du regard à partir de l'orientation de la tête.

Suivi probabiliste du visage - Le suivi de la pose de la tête reste une tâche difficile à accomplir à cause de la sensibilité au suivi et à l'estimation de la pose (cf. § 2.4.1.2). Nous proposons d'utiliser une approche de suivi en ligne.

L'approche proposée effectue un suivi individuel de personnes. Nous employons un algorithme de suivi de l'état-de-l'art, à savoir l'algorithme *Incremental Visual Tracking* (IVT, (Ross *et al.*, 2008)). De plus, afin de diminuer le risque de dérivation, nous avons modifié cet algorithme pour intégrer un détecteur de visage frontal combiné avec une détection de teinte chair dans le but d'affiner la boîte englobante. Ce détecteur est appelé toutes les k images (dans notre implémentation $k = 5$, c'est-à-dire qu'il est appelé 3 fois par seconde) et permet de mettre à jour le modèle dynamique et la vraisemblance.

Notons $\mathbf{s}_t$ la variable d'état cachée et $\mathbf{z}_{1:t} = \{\mathbf{z}_1, \ldots, \mathbf{z}_t\}$ l'ensemble des observations jusqu'à l'instant t. D'après le théorème de Bayes, l'inférence séquentielle pour le suivi peut être formulée par l'expression suivante :

$$p(\mathbf{s}_t|\mathbf{z}_{1:t}) \propto p(\mathbf{z}_t|\mathbf{s}_t) \int_{\mathbf{s}_t} p(\mathbf{s}_t|\mathbf{s}_{t-1}) p(\mathbf{s}_{t-1}|\mathbf{z}_{1:t-1}) d\mathbf{s}_{t-1} \tag{4.6}$$

Dans la suite, nous décrivons les différents termes permettant d'estimer la distribution de filtrage $p(\mathbf{s}_t|\mathbf{z}_{1:t})$.

Modèle d'état. L'état[4] est modélisé par un rectangle :

$$\mathbf{s} = \{x, y, s, a\} \tag{4.7}$$

où (x, y) désigne les coordonnées du centre de la région du visage, s le facteur d'échelle et a le rapport d'axes (*aspect ratio*). Notons que nous mettons à jour s et a après la détection du visage.

Modèle dynamique. Le modèle dynamique est défini par la mixture Gaussienne suivante :

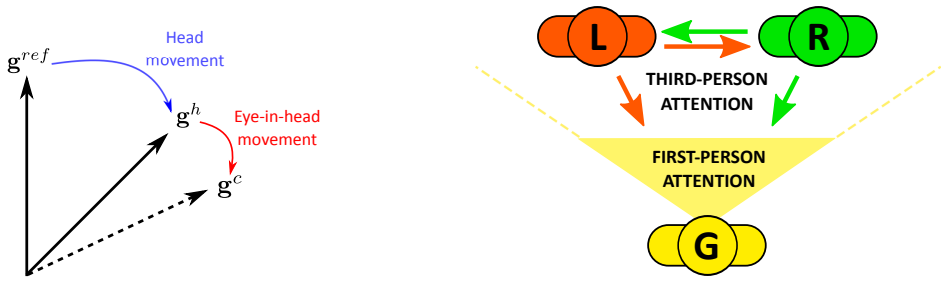
$$q(\mathbf{s}_t|\mathbf{s}_{t-1}, \mathbf{z}_t) = (1 - \gamma) \cdot p(\mathbf{s}_t|\mathbf{s}_{t-1}) + \gamma \cdot q_d(\mathbf{s}_t|\mathbf{z}_t) \tag{4.8}$$

avec :

$$\begin{aligned} p(\mathbf{s}_t|\mathbf{s}_{t-1}) &= \mathcal{N}(\mathbf{s}_t; \mathbf{s}_{t-1}, \mathbf{\Sigma}) \\ q_d(\mathbf{s}_t|\mathbf{z}_t) &= \mathcal{N}(\mathbf{s}_t; \mathbf{s}^d, \mathbf{\Sigma}^d) \end{aligned} \tag{4.9}$$

où le coefficient de mixture γ, $\mathbf{s}^d$ désigne la boîte englobante associée au visage détecté le plus proche et $\mathcal{N}(\mathbf{x}; \mu, \mathbf{\Sigma})$ est la fonction de densité de probabilité de la loi normale.

[4] Nous n'utilisons pas le modèle d'état complet de Ross *et al.* (2008).



(a) (b)

Figure 4.3 – (a) Couplage oeil-tête : le regard $\mathbf{g}^c$ est obtenue en appliquant un modèle linéairement proportionnel au déplacement de la tête par rapport à une référence $\mathbf{g}^{ref}$, (b) Motifs d'attention à reconnaître : attention *subjective* et *objective*.

Modèle d'observation. Le modèle repose sur une représentation d'un sous-espace linéaire (PCA) caractérisé par $(\boldsymbol{\mu}, \mathbf{B})$ où $\boldsymbol{\mu}$ est la moyenne du sous-espace dont les bases $\mathbf{B}$ sont calculées incrémentalement par SVD sur un ensemble d'images précédentes $\mathbf{I}_1 \dots \mathbf{I}_{t-1}$. En particulier, nous obtenons la vraisemblance suivante :

$$p(\mathbf{z}_t|\mathbf{s}_t) \propto e^{-||(\mathbf{I}_t-\boldsymbol{\mu})-\mathbf{B}\mathbf{B}^T(\mathbf{I}_t-\boldsymbol{\mu})||_2^2} \tag{4.10}$$

où $\mathbf{I}_t = \mathbf{I}(\mathbf{s}_t)$ est l'image du visage extraite de l'image $\mathbf{I}$ originale à l'aide de la boîte englobante définie par $\mathbf{s}_t$.

Couplage oeil-tête - L'estimation de la pose de la tête n'étant qu'une approximation de la direction du regard, il est nécessaire de tenir compte des mouvements oculaires. Cependant, dans notre cas, la faible résolution des images des yeux ne permet pas d'exploiter cette information. En revanche, nous proposons d'intégrer un modèle cognitif assumant un couplage linéaire entre les mouvements oculaires et les mouvements de la tête, similairement à (Ba et Odobez, 2009) :

$$\mathbf{g}^h - \mathbf{g}^{ref} = \kappa \cdot (\mathbf{g}^c - \mathbf{g}^{ref}) \tag{4.11}$$

avec les coefficients personnels $\kappa = [\kappa_\alpha, \kappa_\beta]$. $\mathbf{g}^{ref}$ and $\mathbf{g}^c$ sont, respectivement, le regard *référence* et le regard *compensé*. La figure 4.3(a) illustre le principe de ce modèle cognitif.

4.4 Reconnaissance d'attention

Nous présentons une méthode de reconnaissance d'attention dans des interactions en triade. Pour analyser les motifs d'attention, des scores sont calculés en se basant sur les prédictions des regards décrites en § 4.2 et § 4.3. Ensuite, pour chacune des personnes, nous associons un état attentionnel suivant : $S = \{$"G" = regarde le sujet équipé du système porté, "L" =

regarde la personne à gauche dans la caméra scène, "R" = regarde la personne à droite dans la caméra scène, "U" = regarde ailleurs}. La figure 4.3(b) permet de schématiser les différents motifs d'attention possibles.

4.4.1 Attention subjective

Ici, nous nous intéressons à modéliser l'attention du sujet équipé de l'eye-tracker porté, c'est-à-dire nous cherchons à savoir si cette personne regarde ou non une personne présente dans la caméra scène.
Nous supposons que si les coordonnées $\mathbf{g}^e$ du regard obtenu via l'eye-tracker sont contenus dans la région d'un des visages d'une personne présente dans la caméra scène, alors le sujet équipé du système eye-tracker regarde effectivement la personne considérée. Pour également prendre en compte une incertitude sur l'estimation du regard subjectif, nous proposons d'employer une distribution gaussienne centrée sur la région du visage. Le score d'attention subjective, défini à partir de la localisation d'un visage F_i dans la caméra scène et des coordonnées du regard subjectif, est alors donné par l'expression suivante :

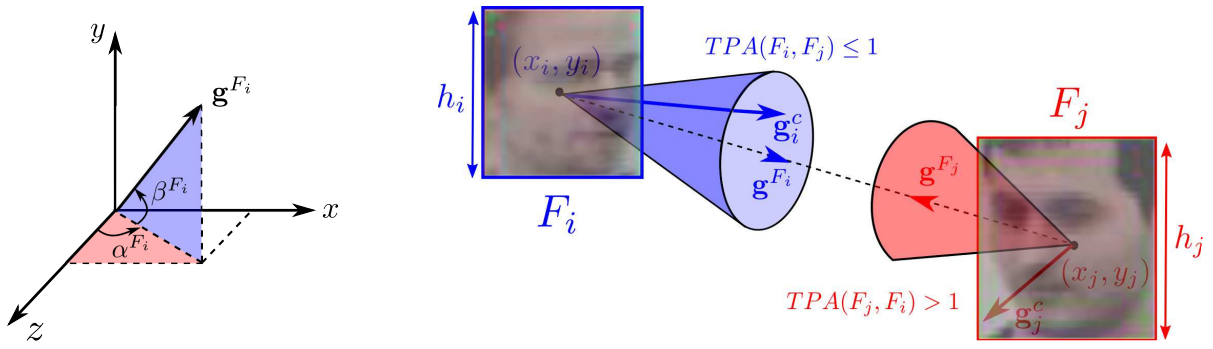
$$FPA(F_i) = \begin{cases} 1 \text{ si } \mathbf{g}^e \in \mathcal{O}_{F_i} \\ G_0 \exp\left(-\frac{||\mathbf{g}^e - L_{F_i}||_2^2}{2\sigma_g^2}\right) \text{ sinon} \end{cases} \tag{4.12}$$

où $\mathcal{O}_{F_i}$ représente l'ensemble des positions des pixels contenus dans la région du visage F_i, L_{F_i} est le centre de la boîte englobante du visage F_i, G_0 est un coefficient de normalisation et l'écart-type σ_g est ajusté en fonction de la précision du suivi du regard.

4.4.2 Attention objective

Dans le même esprit que pour l'attention subjective, nous proposons également de déterminer l'attention objective de personnes présentes dans la caméra scène, en particulier afin de savoir si ces personnes regardent une autre personne qui peut aussi bien être une personne présente dans la caméra scène que celle équipée de l'eye-tracker porté.

Score d'attention objective - L'idée consiste à prendre en compte la géométrie des personnes dans la camèra scène, en particulier le positionnement relatif des personnes obtenu via les localisations des visages, pour calculer un regard *référence* $\mathbf{g}^{F_i}$ auquel sera comparé le regard objectif issu de l'estimation de la pose du visage $\mathbf{g}^c = (\alpha^c, \beta^c)$ (cf. § 4.3).
Dans un premier temps, nous cherchons à déterminer le regard référence $\mathbf{g}^{F_i}$. Pour cela, nous adoptons une représentation angulaire $(\alpha^{F_i}, \beta^{F_i})$ comme indiquée sur la figure 4.4(a) et nous calculons les angles de rotation associés à partir des régions d'intérêt des visages détectés. Nous tenons également compte de la profondeur en exploitant l'information sur les hauteurs h_i, h_j des visages. Les angles de rotation sont alors obtenus à l'aide des expressions



(a) (b)

Figure 4.4 – (a) Passage du vecteur 3D du regard *référence* $\mathbf{g}^{F_i}$ à sa représentation angulaire $(\alpha^{F_i}, \beta^{F_i})$, (b) Modèle d'attention objective construit à partir du regard *référence* $\mathbf{g}^{F_i}$ reposant sur la localisation des visages et du regard obtenu via l'estimation de la pose de la tête $\mathbf{g}^c$ (dans cet exemple, F_i regarde F_j, alors que F_j regarde ailleurs).

suivantes :

$$\begin{aligned} \alpha^{F_i}(F_j) &= \tfrac{180}{\pi} \operatorname{atan2}(\Delta\bar{x}_{i,j}, \Delta\bar{z}_{i,j}) \\ \beta^{F_i}(F_j) &= \tfrac{180}{\pi} \operatorname{atan2}(\Delta\bar{y}_{i,j}, \sqrt{\Delta\bar{x}_{i,j}^2 + \Delta\bar{z}_{i,j}^2}) \end{aligned} \tag{4.13}$$

avec $\Delta\bar{x}_{i,j} = \frac{x_j - x_i}{\max(x_i, x_j)}$, $\Delta\bar{y}_{i,j} = \frac{y_j - y_i}{\max(y_i, y_j)}$ et $\Delta\bar{z}_{i,j} = \frac{h_j - h_i}{\max(h_i, h_j)}$ [5].
Ensuite, nous considérons un cône à base elliptique autour du regard référence afin de pouvoir prendre en compte l'incertitude sur les estimations des angles yaw et pitch (cf. figure 4.4(b)). En combinant l'information des deux regards via leur différence respective $\Delta\alpha_{i,j} = \alpha^c - \alpha^{F_i}(F_j)$ et $\Delta\beta_{i,j} = \beta^c - \beta^{F_i}(F_j)$, le score d'attention objective est alors obtenu par l'expression suivante :

$$TPA(F_i, F_j) = \sqrt{\mathbf{\Delta\Theta}_{i,j}^T \mathbf{Q} \mathbf{\Delta\Theta}_{i,j}} \quad \forall i \neq j \tag{4.14}$$

où $\mathbf{\Delta\Theta}_{i,j} = [\Delta\alpha_{i,j}, \Delta\beta_{i,j}]^T$ et $\mathbf{Q}$ est défini par :

$$\mathbf{Q} = \begin{bmatrix} 1/2\varphi_\alpha^2 & 0 \\ 0 & 1/2\varphi_\beta^2 \end{bmatrix} \tag{4.15}$$

φ_α et φ_β permettent de régler l'angle d'ouverture du cône dans la direction horizontale et verticale. Ainsi, $TPA(F_i, F_j) \geq 0$ et une personne P_i regarde une autre personne P_j si le score $TPA(F_i, F_j) \leq 1$.

[5] Une autre solution, plus précise, aurait consisté à calibrer le système comme c'est le cas pour (Fathi *et al.*, 2012a).

Filtrage temporel - Pour chaque personne présente dans le champ de vue de la caméra scène, nous employons un simple HMM pour lisser les observations d'attention o_t et estimer l'état attentionnel l_t en calculant la distribution a posteriori :

$$p(l_t|o_{1:t}) = \frac{p(o_t|l_t)p(l_t|l_{t-1})p(l_{t-1}|o_{1:t-1})}{\sum_{l'_t} p(o_t|l'_t)p(l'_t|l_{t-1})p(l_{t-1}|o_{1:t-1})} \tag{4.16}$$

où la probabilité a priori $p(l_0) = \pi_{l_0}$ est initialisée uniformément et les probabilités de transition $p(l_t = m|l_{t-1} = n) = A_{mn}$ ont des valeurs uniformes élevées pour $m = n$ et faibles pour $m \neq n$. La vraisemblance est calculée à l'aide des scores d'attention via l'expression suivante :

$$p(o_t|l_t) \propto \exp(-\lambda \cdot TPA^2) \tag{4.17}$$

où λ est un paramètre ajusté expérimentalement.

4.5 Résultats expérimentaux

Dans un premier temps, nous proposons d'évaluer les performances de la méthode d'estimation de la pose de la tête (cf. § 4.5.1). Ensuite, nous présentons des résultats obtenus pour la reconnaissance d'attention lors d'interactions en triade (cf. § 4.5.2).

4.5.1 Estimation de la pose de la tête

4.5.1.1 Bases de données

Nous avons évalué la méthode proposée sur deux bases de données publiquement disponibles :

- **CUbiC FacePix** (Little *et al.*, 2005) inclut 30 sujets avec des angles yaw couvrant $\pm 90°$ et avec un pas angulaire incrémental de $p_{inc} = 1°$. Au total, nous disposons de 5430 images. Etant donné que les images des visages sont alignées, nous avons sélectionné une région 98×98 centrée sur chacun des visages (comme c'est le cas pour (BenAbdelkader, 2010)).
- **PRIMA Pointing'04** (Gourier *et al.*, 2004) contient deux séries de 15 sujets pour lesquelles 91 images[6] sont capturées, ce qui fait un total de 2730 images. Pour l'angle yaw α, 13 poses discrètes sont disponibles couvrant $\pm 90°$ avec un pas $p_{inc} = 15°$. Pour l'angle pitch β, nous disposons de 7 poses discrètes $\{-60°, -30°, -15°, 0°, 15°, 30°, 60°\}$. Les régions des visages ont été manuellement déterminées en ajustant une boîte englobante autour de la région teinte chair.

Des exemples d'images de visages sous diférentes poses sont affichés dans la figure 4.5.

[6]Les angles pitch $\pm 90°$ sont ignorés, car ils sont sous-échantillonnées en comparaison avec les autres angles d'orientation.

(a) CUbiC FacePix (Little *et al.*, 2005)

(b) PRIMA Pointing'04 (Gourier *et al.*, 2004)

Figure 4.5 – Exemple d'images des bases de données avec : (a) $\alpha \in [-90°, +90°]$, (b) $\alpha \in [-90°, +90°]$ et $\beta \in [-60°, +60°]$

4.5.1.2 Comparaison avec d'autres méthodes

Nous avons comparé la méthode proposée avec d'autres méthodes de l'état-de-l'art.

Notons $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i) | i = 1, \ldots, n\}$ le corpus d'apprentissage dont les n échantillons d'apprentissage $\mathbf{x}_i \in \mathbb{R}^{d \times 1}$ sont des vecteurs d'entrée d-dimensionnel et $\mathbf{y}_i \in \mathbb{R}^{m \times 1}$ des vecteurs de sortie m-dimensionnel.

Linear Ridge Regression (LRR) - En régression linéaire, la fonction de prédiction possède une forme linéaire du type : $\mathbf{f}(\mathbf{x}) = \mathbf{W}^\top \mathbf{x}$, où $\mathbf{W} \in \mathbb{R}^{d \times m}$ est la matrice des poids à apprendre. En appliquant le principe inductif de minimisation du risque empirique régularisé, apprendre $\mathbf{W}$ revient à minimiser :

$$\min_{\mathbf{w}} \|\mathbf{Y} - \mathbf{W}^\top \mathbf{X}\|_F^2 + \lambda \|\mathbf{W}\|_F^2 \tag{4.18}$$

avec $\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_n]$ de taille $(d \times n)$ et $\mathbf{Y} = [\mathbf{y}_1 \ldots \mathbf{y}_n]$ de taille $(m \times n)$. Le paramètre λ permet de contrôler le degré de régularisation. Cette minimisation admet la solution

analytique suivante :

$$\mathbf{W} = (\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{Y}^\top \tag{4.19}$$

Kernel Ridge Regression (KRR) - Il s'agit du même principe que pour LRR mis à part que $\mathbf{x}$ subissent une transformation dans un espace non-linéaire $\boldsymbol{\phi}(\mathbf{x})$. Nous obtenons donc :

$$\mathbf{W} = (\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + \lambda\mathbf{I})^{-1}\boldsymbol{\Phi}\mathbf{Y}^\top \tag{4.20}$$

avec $\boldsymbol{\Phi} = [\boldsymbol{\phi}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}(\mathbf{x}_n)]$.

Gaussian Process Regression (GPR) - A partir de la base de données $\mathcal{D}$, nous souhaitons apprendre une fonction $f(\mathbf{x_i})$ transformant un vecteur d'entrée $\mathbf{x}_i$ en une sortie y_i en supposant un modèle d'observation bruité $y_i = f(\mathbf{x}_i) + \varepsilon_i$ avec $\varepsilon_i \sim \mathcal{N}(0, \sigma_n^2)$ (Rasmussen et Williams, 2006). En supposant que $f(\mathbf{x}_i)$ suive une distribution Gaussienne de moyenne nulle, il est possible d'affirmer que les sorties suivent une distribution Gaussienne $\mathbf{y} \sim \mathcal{N}(0, \mathbf{K} + \sigma_n^2 I)$ où $\mathbf{K}$ est la matrice de covariance dont la fonction de covariance est définie via le noyau[7] k suivant :

$$k(\mathbf{x}, \mathbf{x}') = \sigma_s^2 \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||_2^2}{2l^2}\right) \tag{4.21}$$

avec σ_f^2 la variance du signal et l la largeur de base. La distribution jointe de $\mathbf{y}$ et de $f(\mathbf{x}_*)$ pour un échantillon test $\mathbf{x}_*$ est alors donnée par la distribution Gaussienne multivariée suivante :

$$\begin{bmatrix} \mathbf{y} \\ f(\mathbf{x}_*) \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \mathbf{K} + \sigma_n^2\mathbf{I} & \mathbf{K}_*^\top \\ \mathbf{K}_* & k(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix}\right) \tag{4.22}$$

avec $\mathbf{K}_* = [k(\mathbf{x}_*, \mathbf{x}_1) \ldots k(\mathbf{x}_*, \mathbf{x}_n)]$. Finalement, d'après (Rasmussen et Williams, 2006), la distribution a posteriori $p(y_*|\mathbf{x}_*, \mathcal{D})$ suit une loi Normale $\mathcal{N}(\mu_l, \sigma_l^2)$ définie par la moyenne μ_l :

$$\mu_l = \mathbf{K}_* \left(\mathbf{K} + \sigma_n^2\mathbf{I}\right)^{-1} \mathbf{y} \tag{4.23}$$

et la variance σ_l^2 :

$$\sigma_l^2 = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{K}_* \left(\mathbf{K} + \sigma_n^2\mathbf{I}\right)^{-1} \mathbf{K}_*^\top \tag{4.24}$$

Les hyperparamètres du GPR à optimiser sont $\boldsymbol{\theta} = \left[\sigma_n^2, \sigma_f^2, l\right]$. Ils peuvent être obtenus automatiquement par maximisation de la log-vraisemblance marginale via une méthode d'optimisation telle que la méthode Quasi-Newton (Rasmussen et Williams, 2006).

Kernel Partial Least Squares Regression (KPLS) - La régression PLS suppose que les

[7] appelé *Squared Exponential covariance with isotropic distance measure*. Notons qu'il est possible d'utiliser d'autres fonctions de covariance : *Rational Quadratic*, *Neural Network* et *Matern*, ainsi que d'autres mesures de distance telles que *Automatic Relevance Determination*.

variables d'entrée et de sortie appartienent à un ensemble commun de variables cachées et elle approxime leur matrice correspondante en terme de projection sur p vecteurs latents :

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}^\top + \mathbf{E} \\ \mathbf{Y} &= \mathbf{U}\mathbf{Q}^\top + \mathbf{F} \end{aligned} \tag{4.25}$$

avec $\mathbf{T}$ et $\mathbf{U}$ des matrices de taille $(n \times p)$ dont les colonnes sont les vecteurs latents. $\mathbf{P}$ et $\mathbf{Q}$ sont appelées *loadings*. $\mathbf{E}$ et $\mathbf{F}$ sont les matrices résiduelles. Il existe différentes manières d'obtenir les décompositions dans (4.25). Nous employons la méthode la plus couramment utilisée, à savoir l'algorithme *Nonlinear Iterative Partial Least Squares* (NIPALS, (Rosipal et Trejo, 2001)) qui procède à p itérations des étapes suivantes :

1. Trouver une combinaison linéaire des colonnes de $\mathbf{X}$ et $\mathbf{Y}$ qui sont maximalement corrélées :
$$(\mathbf{w}, \mathbf{c}) = \underset{|\mathbf{r}|=|\mathbf{s}|=\mathbf{1}}{\operatorname{argmax}} \quad [cov(\mathbf{X}\mathbf{r}, \mathbf{Y}\mathbf{s})]^2 \tag{4.26}$$
2. Calculer les vecteurs latents $\mathbf{t} = \mathbf{X}\mathbf{w}$ et $\mathbf{u} = \mathbf{Y}\mathbf{c}$
3. Approximer (*deflate*) $\mathbf{X}$ et $\mathbf{Y}$ en se basant sur la direction $\mathbf{d} = \mathbf{t}/||\mathbf{t}||$:
$$\begin{aligned} \mathbf{X} &= \mathbf{X} - \mathbf{d}\mathbf{d}^\top\mathbf{X} \\ \mathbf{Y} &= \mathbf{Y} - \mathbf{d}\mathbf{d}^\top\mathbf{Y} \end{aligned} \tag{4.27}$$

En concaténant les p vecteurs $\mathbf{t}$ en $\mathbf{T}$ et $\mathbf{u}$ en $\mathbf{U}$, la prédiction $\mathbf{y}_*$ est obtenue via :

$$\mathbf{y}_* = \mathbf{x}_*\mathbf{X}^\top\mathbf{U}(\mathbf{T}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U})^{-1}\mathbf{T}^\top\mathbf{Y} \tag{4.28}$$

La régression Kernel PLS s'obtient similairement, mais en appliquant le "kernel trick" ($\mathbf{x} \rightarrow \phi(\mathbf{x})$). Il s'agit de résoudre (1) dans l'espace du noyau et l'étape (3) devient :

$$\mathbf{K} = \mathbf{K} - \mathbf{d}\mathbf{d}^\top\mathbf{K} - \mathbf{K}\mathbf{d}\mathbf{d}^\top + \mathbf{d}\mathbf{d}^\top\mathbf{K}\mathbf{d}\mathbf{d}^\top \tag{4.29}$$

avec $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^\top$. Pour la prédiction à partir d'un échantillon test $\mathbf{x}_*$, nous obtenons alors l'expression suivante :

$$\mathbf{y}_* = \mathbf{\Phi}_*\mathbf{\Phi}^\top\mathbf{U}(\mathbf{T}^\top\mathbf{K}\mathbf{U})^{-1}\mathbf{T}^\top\mathbf{Y} \tag{4.30}$$

Support Vector Regression (SVR) - Pour SVR, nous utilisons LMKR en fixant le nombre de noyaux à $P = 1$ (cf. § 4.3.1).

4.5.1.3 Procédure expérimentale

Extraction de caractéristiques - Les images des visages sont redimensionnées pour avoir une taille de 40×40. Des caractéristiques HOG (Dalal et Triggs, 2005) sont, ensuite, extraites avec $N_b = 8$ intervalles et en appliquant une grille 8×8 sur l'image. Aucune

méthode de réduction de dimensionnalité (PCA, ICA) n'est utilisée et nous obtenons donc un vecteur d'entrée de taille égale à 576.

Mesures de performance - Nous employons la moyenne de l'erreur angulaire absolue (MAAE pour *Mean Absolute Angular Error*) et l'écart-type (SD pour *Standard Deviation*) associé entre la valeur continue de l'orientation estimée et la valeur discrète de la vérité terrain.

Optimisation des paramètres - Lorsque les méthodes font appel à un noyau, nous utilisons un noyau Gaussien. Pour chaque méthode et chaque base de données, les paramètres (λ pour LRR ; λ et γ pour KRR ; σ_s, l et σ_n pour GPR ; γ et le nombre de facteurs latents pour KPLS ; C, γ et ε pour LMKR) ont été optimisés par validation croisée en $k = 5$ groupes sur l'ensemble d'apprentissage.

Pour la base de données FacePix, l'expérience a été répétée avec 10 essais aléatoires où deux tiers des sujets sont utilisés pour la phase d'apprentissage et le reste pour la phase de test. Pour l'apprentissage, nous avons pris un sous-ensemble des données en utilisant un pas incrémental de $p_{inc} = 10°$. Pour la phase de test, nous utilisons toutes les données (c'est-à-dire $p_{inc} = 1°$) des sujets restants. Nous avons fait attention à ce que les sujets de test n'apparaissent pas dans la phase d'apprentissage.

Concernant la base de données Pointing'04, la première série est employée comme corpus d'apprentissage, alors que la seconde série sert à tester les régresseurs appris.

Pour la méthode KPLS, nous avons utilisé l'algorithme NIPALS (Rosipal et Trejo, 2001) pour lequel nous avons obtenu le nombre optimal de facteurs latents $p = 40$.

Notons également que les algorithmes GPR et KPLS ont récemment été employés dans, respectivement, (Marin-Jimenez *et al.*, 2011) et (Al Haj *et al.*, 2012) pour l'estimation de la pose de la tête.

4.5.1.4 Résultats

Les tableaux 4.1 et 4.2 présentent les performances obtenues pour, respectivement, les bases de données FacePix et Pointing'04. Nous affichons également les résultats pour des orientations frontales ($|\alpha| \leq 45°$) et non-frontales ($|\alpha| > 45°$).
Nous remarquons que pour l'angle yaw, LMKR donne de meilleurs résultats en comparaison avec les méthodes de l'état-de-l'art. En particulier, nous pouvons constater qu'une amélioration significative est obtenue dans le cas où les visages sont de profil, c'est-à-dire pour $|\alpha| > 45°$, ce qui fournit une estimation moins sous-estimée pour cette région par rapport aux autres méthodes. Cependant, pour Pointing'04, nous remarquons que pour l'angle pitch, des résultats similaires sont obtenus avec KPLS, GPR et LMKR. Nous supposons que c'est en partie dû au faible nombre de poses (7 poses pour l'angle pitch contre 13 pour l'angle yaw) et cela indique que $P = 1$ noyau est suffisant étant donné la base de données.

TABLE 4.1 – Erreur moyenne (MAAE±SD) pour FacePix (avec $p_{inc} = 10°$)

		α	$\lvert\alpha\rvert \leq 45°$	$\lvert\alpha\rvert > 45°$
LRR		$9.57° \pm 7.55°$	$9.29° \pm 7.12°$	$10.40° \pm 8.38°$
KRR		$8.50° \pm 6.80°$	$8.20° \pm 6.49°$	$9.40° \pm 7.28°$
GPR		$8.12° \pm 6.68°$	$7.75° \pm 6.31°$	$9.23° \pm 7.20°$
KPLS		$7.99° \pm 6.78°$	$7.57° \pm 6.35°$	$9.26° \pm 7.36°$
LMKR	$P = 1$	$8.73° \pm 7.59°$	$8.03° \pm 6.71°$	$10.85° \pm 8.87°$
	$P = 2$	$\mathbf{6.79° \pm 5.86°}$	$\mathbf{6.81° \pm 5.48°}$	$\mathbf{6.72° \pm 6.39°}$

TABLE 4.2 – Erreur moyenne (MAAE±SD) pour Pointing'04

		α	$\lvert\alpha\rvert \leq 45°$	$\lvert\alpha\rvert > 45°$	β
LRR		$11.87° \pm 10.47°$	$11.21° \pm 9.89°$	$14.10° \pm 11.95°$	$10.97° \pm 8.84°$
KRR		$9.31° \pm 8.94°$	$8.59° \pm 8.41°$	$11.73° \pm 10.14°$	$8.24° \pm 7.40°$
GPR		$8.59° \pm 8.67°$	$7.81° \pm 8.07°$	$11.18° \pm 9.98°$	$7.48° \pm 6.99°$
KPLS		$8.53° \pm 8.88°$	$7.72° \pm 8.09°$	$11.40° \pm 10.66°$	$\mathbf{7.39° \pm 7.00°}$
LMKR	$P = 1$	$8.66° \pm 9.17°$	$7.77° \pm 8.32°$	$11.62° \pm 11.06°$	$7.57° \pm 7.27°$
	$P = 2$	$\mathbf{7.34° \pm 6.66°}$	$\mathbf{6.93° \pm 6.54°}$	$\mathbf{8.72° \pm 6.87°}$	$7.62° \pm 7.26°$

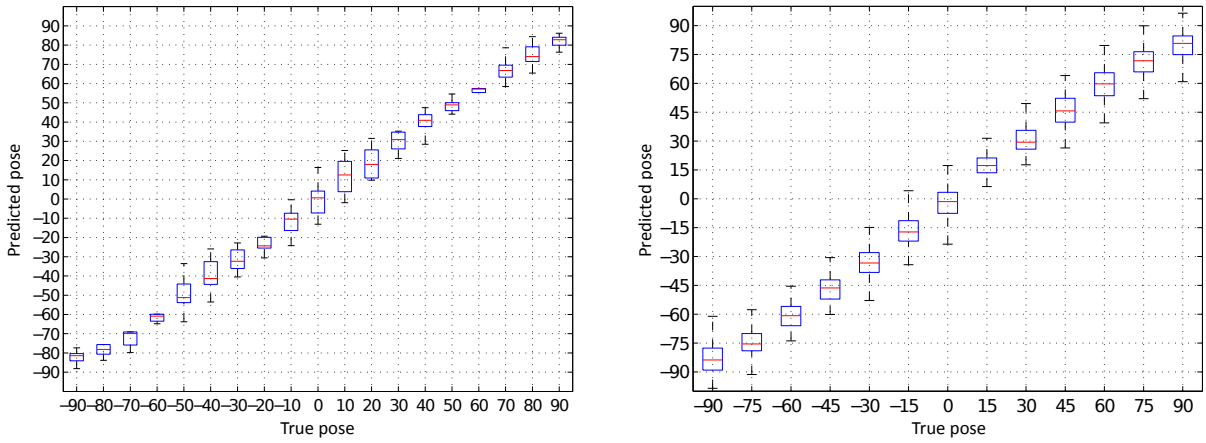
Les figures 4.6(a) et 4.6(b) montrent les diagrammes en boîte (*box plot*, en anglais) pour l'angle yaw α et les bases de données FacePix et Pointing'04.

Notons que les résultats obtenus pour Pointing'04 ne sont pas directement comparables avec (Al Haj *et al.*, 2012). Effectivement, ils utilisent 80% des données pour effectuer l'apprentissage, alors que dans notre cas, les régresseurs sont appris avec une seule série de données, c'est-à-dire 50% des données.

4.5.2 Evaluation de la reconnaissance d'attention

4.5.2.1 Base de données et procédure expérimentale

Calibration individuelle - Avant chaque expérience, nous procédons à une calibration afin de collecter des images des yeux du sujet équipé de l'eye-tracker porté et construire un modèle d'apparence de regard. Il est demandé au sujet de suivre une cible en gardant la tête immobile. Ensuite, le modèle de regard est appris à l'aide du paramètre de largeur de base $r = 7$ obtenu sur une base de données (cf. § 3.3.3) ne comprenant pas le sujet en question. Une étude préliminaire (cf. chapitre 3) a montré que l'erreur absolue moyenne



(a) FacePix (b) Pointing'04

Figure 4.6 – Diagrammes en boîtes pour l'angle yaw α prédit avec LMKR ($P = 2$)

est de l'ordre de 2° d'angle visuel avec 100 échantillons de calibration ; cette précision est suffisante pour l'application que nous visons.

Procédure d'évaluation - Pour évaluer notre détecteur d'attention, 4×3 sujets (4 vidéos sont enregistrées) sont impliqués dans des interactions en triade où l'un d'entre eux est équipé d'un eye-tracker porté alors que les autres sont assis en face lui. Chaque vidéo enregistrée contient entre 1500 et 6500 images acquises à une fréquence de 15 images/s, ce qui fait un total de ∼15 minutes de vidéo. Dans chaque image, nous avons attribué un état attentionnel (attention objectif) à chacune des personnes et les transitions (par exemple, lorsque la personne est en train de tourner la tête) sont supposées être dans un état "U". Ainsi, en connaissant où chaque personne regarde, il est possible de déterminer des motifs d'attention tels que :

1. qui regarde qui (LAP),
2. le *regard mutuel* (MG) entre tous les participants,
3. le *regard partagé* (SG), c'est-à-dire si deux personnes regardent une autre et même personne.

Détails d'implémentation - Les paramètres pour le suivi du visage cf § 4.3.2 sont les mêmes pour toutes les vidéos. Nous employons 700 particules[8] et le coefficient de mixture du modèle dynamique γ est fixé égal à 0.3. Pour IVT, le nombre de vecteurs propres est égal à 16 et le facteur d'oubli (*forgetting factor*) est égal à 1.

[8]Le nombre de particules est légèrement supérieur à celui de l'implémentation originale afin de mieux prendre en compte l'égo-mouvement.

TABLE 4.3 – Résultats pour la reconnaissance d'attention objective (LAP) et la détection du regard mutuel (MG) et du regard partagé (SG). Acc. correspond à la précision moyenne, F1 au score F1 et % au pourcentage d'apparition dans la base de données.

Exp.	**LAP**		**MG**			**SG**		
	Left person	Right person	Acc.	F1	%	Acc.	F1	%
A	0.89	0.75	0.82	0.85	59	0.87	0.90	71
B	0.75	0.76	0.75	0.79	58	0.84	0.87	63
C	0.82	0.75	0.87	0.88	57	0.77	0.78	44
D	0.91	0.89	0.93	0.95	67	0.91	0.93	68
Avg.	0.82		0.84	0.87	60	0.85	0.87	61

Concernant l'estimation de la pose de la tête, nous complétons la base de données Pointing'04 avec des images de visages (uniquement pour $\beta = 0°$) pivotées selon l'angle roll $\pm 15°$. Ainsi, nous obtenons au total 3380 visages pour les deux séries et nous entrainons LMKR avec $P = 2$ pour α et $P = 1$ pour β.

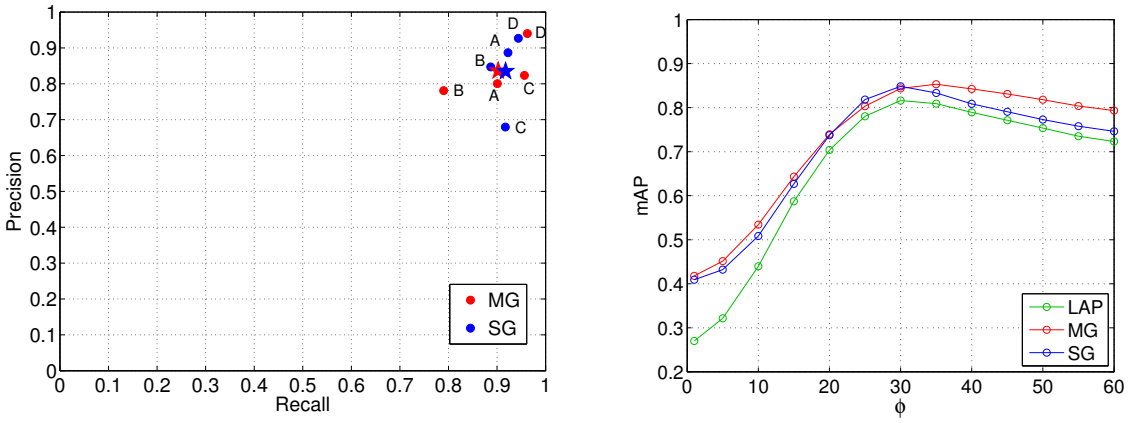
Pour toutes les expériences, le coefficient de couplage oeil-tête $\boldsymbol{\kappa} = [\kappa_\alpha, \kappa_\beta]$ est égal à $[0.6, 0.5]$. Connaissant la géométrie de l'interaction triadique, nous fixons $\mathbf{g}^{ref} = [0°, 0°]$, c'est-à-dire que le sujet équipé du système est choisi comme référence et il est supposé le rester pendant la durée de l'expérience.

En ce qui concerne le calcul du score d'attention subjective, nous avons fixé $\sigma_g = 30$ pixels en tenant compte de la précison de notre système porté. Pour l'attention objective (cf. § 4.4.2), les angles d'ouverture du cône elliptique sont $\varphi_a = \varphi_b = 30°$. Quant au score d'attention objective envers le sujet équipé de l'eye-tracker porté, nous fixons $\left[\alpha^F, \beta^F\right] = [0°, 0°]$ vu que celui-ci se trouve face aux autres participants.

4.5.2.2 Résultats

Etant donné que la section précédente a permis d'évaluer les performances de l'estimation de la pose de la tête pour le regard objectif, nous évaluons ici en particulier les performances de la reconnaissance d'attention objective (LAP) et de la détection du regard mutuel (MG) et partagé (SG). De plus, nous proposons aussi d'analyser des motifs d'attention à un niveau d'interaction plus général, à savoir celui du triade formé par les trois participants.

Evaluation individuelle et diadique - Nous employons, tout d'abord, une comparaison image-par-image (appelée aussi *Frame Recognition Rate*) avec la verité terrain. Le tableau 4.3 présente les résultats pour la reconnaissance d'attention objective (LAP) pour chaque expérience et chaque personne (à gauche ou à droite dans l'image de la caméra scène). Nous obtenons en moyenne un taux de bonne reconnaissance de 82%.



(a) (b)

Figure 4.7 – (a) Résultats de précision-rappel pour chaque expérience et chaque motif d'attention : regard mutuel (en rouge) et regard partagé (en bleu). Les étoiles ★ correspondent au résultat moyen par motif et pour toutes les expériences. (b) Influence de l'angle d'ouverture du cône. Nous considérons ici $\varphi = \varphi_\alpha = \varphi_\beta$.

Pour la détection du regard mutuel (MG) et du regard partagé (SG), nous affichons également, dans le même tableau, la F-mesure (ou *F1-score*) :

$$F = \frac{2 \cdot (precision \cdot rappel)}{precision + rappel} \tag{4.31}$$

où la précision et le rappel sont définis par :

$$precision = \frac{|\mathcal{R} \cap \mathcal{GT}|}{|\mathcal{R}|} \qquad rappel = \frac{|\mathcal{R} \cap \mathcal{GT}|}{|\mathcal{GT}|} \tag{4.32}$$

avec $\mathcal{R}$ et $\mathcal{GT}$ l'ensemble, respectivement, des état attentionnels reconnus et de la vérité terrain. Nous indiquons aussi le pourcentage d'apparition de ces motifs au sein de chaque expérience. Dans la figure 4.7(a), nous montrons également les résultats de précision-rappel pour la détection de MG et SG pour chaque séquence vidéo.

Nous remarquons que la méthode proposée donne, globalement, des résultats satisfaisants. Une précision moins bonne est obtenue pour l'expérience B. Cela est à mettre en parallèle avec un taux de reconnaissance plus faible pour LAP en comparaison avec les autres expériences. En effet, si nous regardons plus en détails la vidéo, nous constatons la présence de mouvements de la tête d'amplitude large selon l'angle roll, ce qui peut acroître la difficulté à estimer la pose de la tête. Des meilleurs résultats sont obtenus pour l'expérience D avec une précision moyenne de 90% pour LAP et un taux de détection supérieur

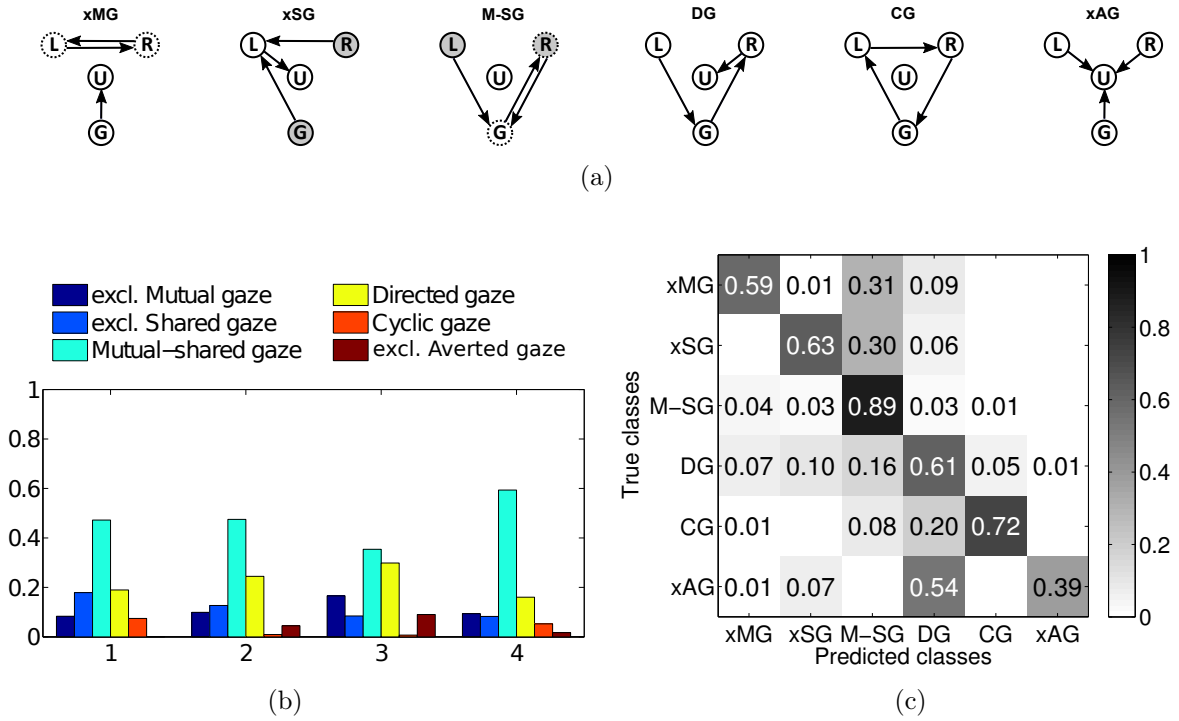


Figure 4.8 – Résultats en triade : (a) Exemples de motifs d'attention, (b) Statistiques d'apparition des motifs triadiques pour chaque expérience, (c) Matrice de confusion pour la reconnaissance d'attention en triade.

à 90% pour MG et SG. Dans certaines expériences telles que l'expérience A, le sujet avait tendance à très lègèrement orienter sa tête vers l'autre participant, ce qui impliquait des mouvements des yeux de plus large amplitude n'étant pas forcément bien appréhendés par le modèle de couplage oeil-tête qui est appliqué avec des paramètres inchangés pour toutes les expériences.

Dans la figure 4.7(b), nous affichons la précision moyenne en fonction de la valeur de l'angle d'ouverture du cône utilisé pour l'estimation du score d'attention objectif. Nous constatons que $\varphi = 30°$ constitue un choix approprié.

Evaluation triadique - Nous nous sommes aussi intéressés à des motifs d'attention d'ordre supérieur, notamment du fait de la nature triadique des expériences. Ces motifs sont essentiels pour comprendre certains aspects de l'interaction sociale à un instant donné (par opposition, à une séquence donnée). Dans notre cas, ces motifs correspondent à 6 catégories : regard mutuel exclusif (xMG), regard partagé exclusif (xSG), regard mutuel-partagé (M-SG), regard dirigé (DG), regard cyclique (CG) et regard dévié exclusif (xAG). Des exemples de motifs triadiques sont illustrés dans la figure 4.8(a). La figure 4.8(b) indique

les pourcentages d'apparition de ces motifs triadiques dans la base de données que nous avons construites. Nous constatons que le motif M-SG prédomine largement, ce qui est normal étant donné les conditions expérimentales. La figure 4.8(c) montre la matrice de confusion pour l'ensemble des expériences. Nous remarquons que le motif d'attention M-SG est, en général, bien détecté. Pour xAG, nous obtenons une confusion importante avec DG, ceci peut s'expliquer par une difficulté à détecter les transitions d'un état attentionnel vers un autre (tout comme pour xMG et xSG qui sont souvent confondus avec M-SG). En effet, ces confusions sont essentiellement dues à une estimation erronée de la pose de la tête d'une des deux personnes présentes dans la caméra scène, ce qui entraîne une mauvaise prédiction de l'état attentionnel associé. De plus, le motif xAG apparaît peu (moins de 5%) dans la base de données (cf. figure 4.8(b)). Une solution pour remédier à ce problème consisterait à améliorer les performances de suivi de la pose de la tête afin de mieux déterminer les états attentionnels.

Résultats qualitatifs - Les figures 4.9 et 4.10 illustrent des exemples de résultats pour : (a) expérience D avec $2000 \leq t \leq 3000$ et (b) expérience B avec $1000 \leq t \leq 2000$. Pour les images capturées par la caméra scène, nous affichons la boîte englobante de suivi du visage, l'estimation de la pose de la tête (indiquée par une flèche) et le point de regard subjectif (sous forme de zone saillante). En dessous des ces images, nous montrons les motifs reconnus et la vérité terrain pour les trois participants. Ensuite, nous affichons, dans l'ordre, les scores d'attention subjective, les estimations des poses des têtes, les scores d'attention objective pour la personne à gauche et à droite, et les motifs triadiques identifiés par la méthode proposée.

Dans la figure 4.9, les motifs d'attention sont reconnus de manière assez fiable même si les transitions posent parfois problème, mais celles-ci sont aussi difficiles à identifier lors de l'annotation des données. Pour la figure 4.10, la segmentation est plus fragmentée, ce qui est en partie dû à des variations angulaires plus importantes de la tête entraînant une estimation erronée de la pose de la tête.

En analysant les résultats, nous avons également constaté que les expressions apparaissant au niveau des yeux pouvait parfois affecter l'estimation du regard subjectif, notamment en influant sur la fermeture des paupières. Cela s'explique principalement à cause de la calibration qui est effectuée avec une expression neutre. Dans le cas où il s'agit uniquement d'évaluer la précision d'estimation d'un nouveau modèle de regard, ce type de calibration ne pose pas problème. En revanche, il est nécessaire de considérer les expressions des yeux en pratique, surtout quand des interactions sociales ont lieu. Cependant, il peut être compliqué de considérer ces expressions à cause de la difficulté à pouvoir faire la différence entre les variations des yeux traduisant un changement de regard ou émanant d'une expression.

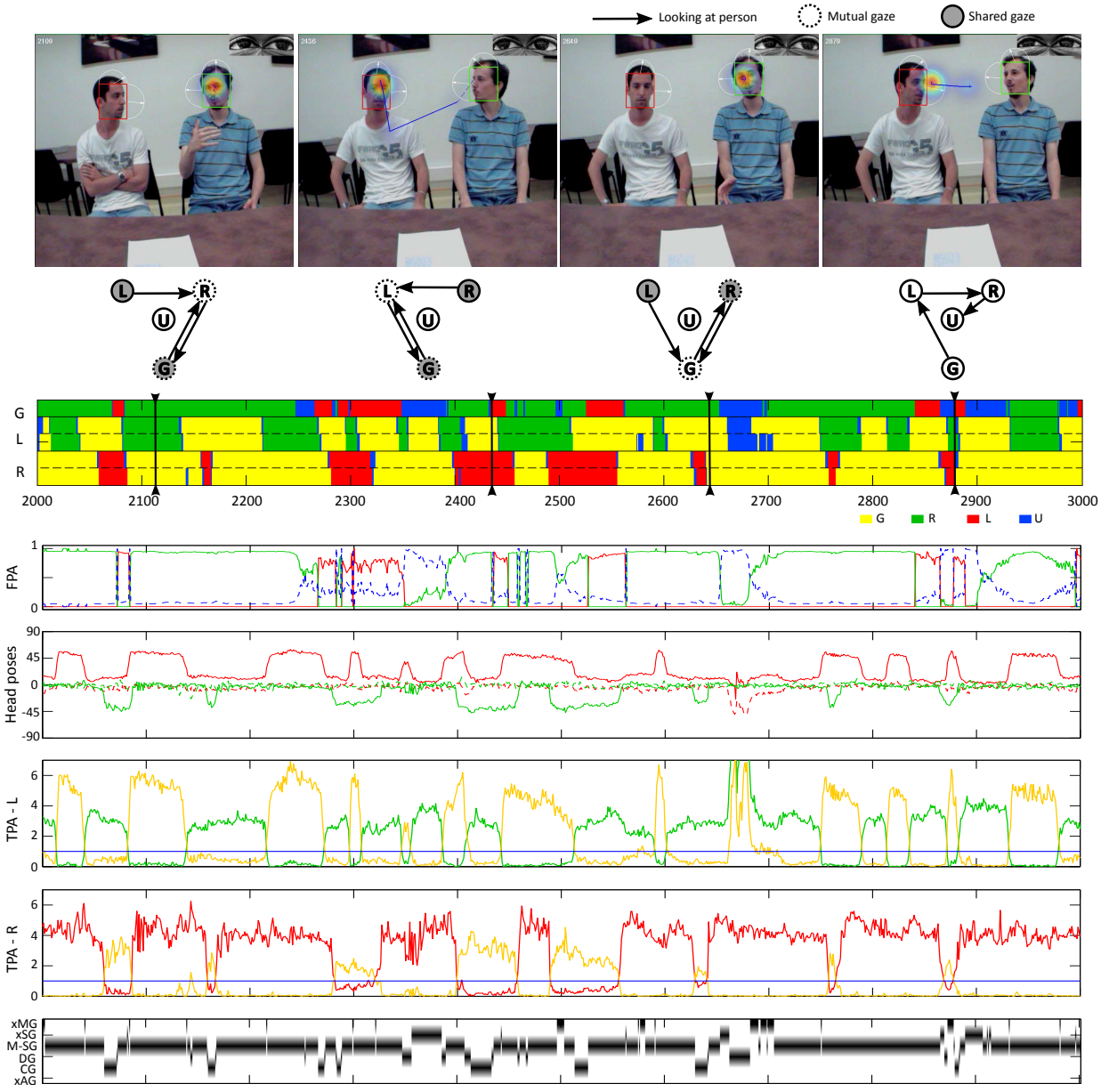


Figure 4.9 – Evaluation qualitative de l'expérience D : la vérité terrain et les motifs d'attention reconnus, les scores d'attention subjective, les estimations des poses des têtes (l'angle yaw α en trait continu et l'angle pitch β en trait discontinu), les scores d'attention objective de la personne à gauche et à droite (nous indiquons aussi le seuil en dessous duquel une personne regarde effectivement une autre personne) et les motifs triadiques reconnus.
La couleur rouge correspond à la personne à gauche, la couleur verte à la personne à droite, la couleur jaune à la personne équipée du système porté et la couleur bleu à ailleurs.

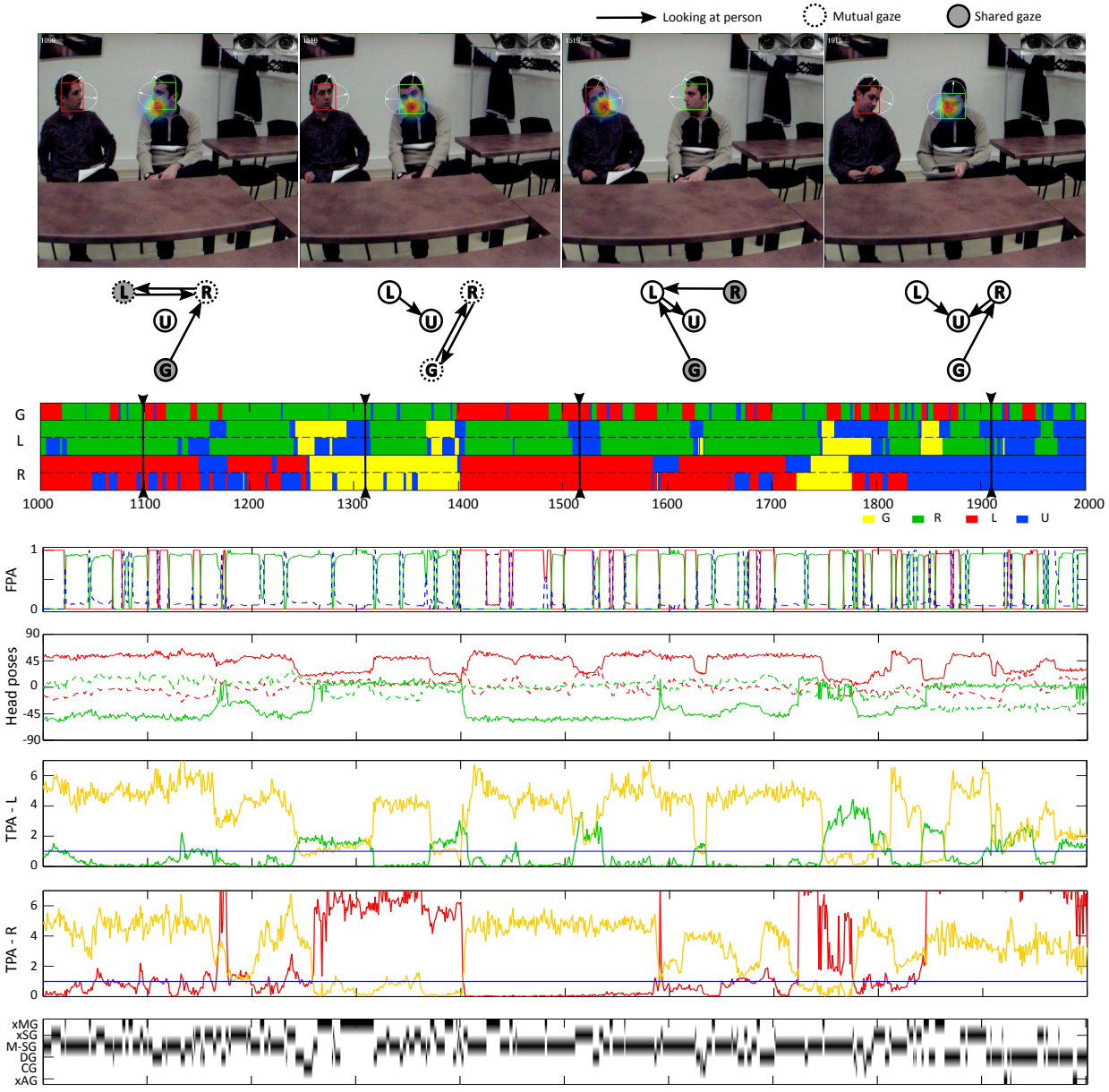


Figure 4.10 – Evaluation qualitative de l'expérience B : la vérité terrain et les motifs d'attention reconnus, les scores d'attention subjective, les estimations des poses des têtes (l'angle yaw α en trait continu et l'angle pitch β en trait discontinu), les scores d'attention objective de la personne à gauche et à droite (nous indiquons aussi le seuil en dessous duquel une personne regarde effectivement une autre personne) et les motifs triadiques reconnus.
La couleur rouge correspond à la personne à gauche, la couleur verte à la personne à droite, la couleur jaune à la personne équipée du système porté et la couleur bleu à ailleurs.

4.6 Conclusion

Nous avons proposé une méthode permettant d'identifier des motifs d'attention en vue subjective. Tout d'abord, le regard subjectif est calculé à l'aide d'un modèle d'apparence dont les paramètres de régression sont obtenus via une calibration individuelle. En parallèle, nous estimons le regard objectif en combinant le suivi du visage et l'estimation de l'orientation de la tête via une régression localisée.

L'évaluation de l'estimation de la pose de la tête a montré qu'une précision moyenne similaire ou supérieure à des méthodes de l'état-de-l'art était obtenue. De plus, des expériences ont permis d'évaluer l'approche dans le cadre d'interactions triadiques et ont donné des résultats encourageants pour la reconnaissance de motifs d'attention.

A travers cette étude, nous avons également démontré que, bien que l'estimation du regard subjectif soit moins précise que pour un système commercial en éclairage infrarouge, les eye-trackers portés reposant sur un modèle d'apparence sont toutefois appropriés pour l'analyse d'attention et qu'une haute précision n'est pas forcément nécessaire.

CHAPITRE 5

Reconnaissance d'activités basée sur le regard

"Que l'importance soit dans ton regard,
non dans la chose regardée !"

– André Gide –
dans *Les Nourritures terrestres*

Sommaire

Dans ce chapitre, nous présentons une nouvelle méthode de reconnaissance d'activités en vue subjective. Cette reconnaissance s'effectue par classification contextuelle de caractéristiques extraites à partir de données du regard encodées sous forme de symboles.

Nous proposons les deux principales contributions suivantes :

- ▷ **Caractéristiques hiérarchiques** (cf. § 5.2.2) : Deux types de caractéristiques sont présentés : un *codage multi-échelle* et un *partitionnement temporel*.
- ▷ **Apprentissage contextuel** (cf. § 5.2.3) : Nous proposons d'incorporer, dans l'apprentissage, de l'information contextuelle obtenue via les valeurs de confiance des classifieurs.

5.1 Introduction

La représentation et la reconnaissance d'activités humaines est un domaine important en vision par ordinateur où les activités sont souvent composées de multiples actions atomiques. L'objectif de la reconnaissance d'activités est d'identifier l'activité effectuée en cours à partir d'une séquence d'images. Il existe différents types d'activités (gestes, actions, interactions, activités de groupe) et chacune d'entre elles peut impliquer une ou plusieurs personnes. Pour une description plus détaillée des différentes approches pour la reconnaissance d'activités, le lecteur est invité à lire l'état-de-l'art (Aggarwal et Ryoo, 2011).

Parmi les différentes techniques proposées, les approches syntaxiques visent à représenter ces actions atomiques à l'aide de symboles, similairement au concept de vocabulaire

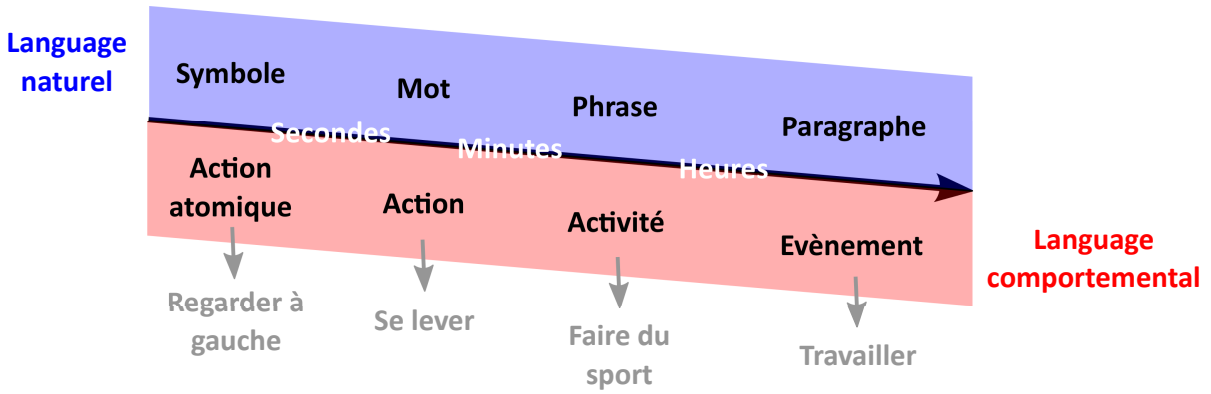


Figure 5.1 – Similarités possibles entre language naturel et comportemental à différentes échelles.

développé dans le cadre du *language naturel*. Le tableau 5.1 montre un aperçu de similarités possibles entre le language naturel et le language comportemental[1]. Notons que la définition de cette hiérarchie peut varier suivant le type d'activités et d'applications.

En pratique, de nombreuses méthodes quantifient les caractéristiques en classes discrètes appelées *mots* et l'ensemble de ces mots constitue un *sac de mots* (*Bag-of-Words*, BoW). Dans cette représentation, une activité est caractérisée par l'occurence de ces mots via un histogramme. L'ensemble des mots, aussi appelé *vocabulaire*, peut être soit prédéfini ou soit appris. Ensuite, une méthode d'apprentissage supervisé ou non-supervisé est appliquée pour classer les activités.

Dans la suite, nous nous intéressons, en particulier, à la reconnaissance d'activités à partir du regard et en vue subjective. Contrairement aux approches et applications standards, la personne n'est pas présente dans le champ de vue de la caméra. Ainsi, nous travaillons sur d'autres types d'information issus de la vision subjective.

Dans la suite, nous appellerons :

- ▷ **Alphabet $\mathcal{A}$** : Il s'agit d'un ensemble de symboles qui désigne une action atomique qui a lieu entre deux images d'une vidéo.
- ▷ **Vocabulaire** : Un vocabulaire (ou également *dictionnaire*) arbitraire désigne un vocabulaire constitué aussi bien de symboles que de mots.

[1]Cependant, en pratique, il arrive souvent que les mots "action" et "activité" soient interchangés, car il peut s'avérer difficile de déterminer la frontière en ces deux comportements. Ainsi, il peut arriver de rencontrer diverses expressions : *activités atomiques*, *activités composées* ou encore *macro-activités*.

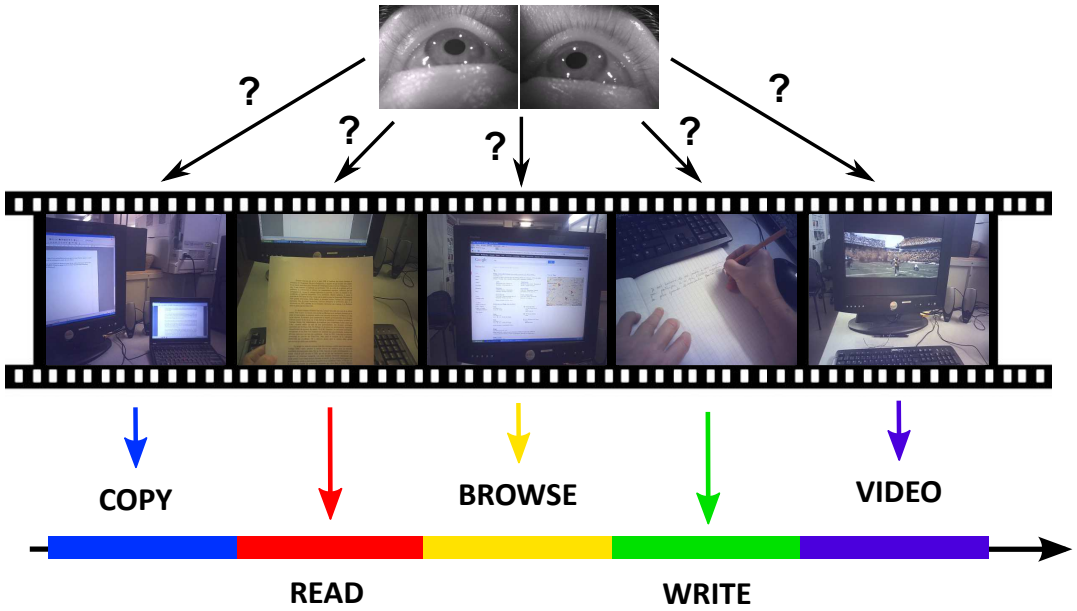


Figure 5.2 – Le principe de la méthode proposée consiste à reconnaître les différentes activités effectuées par un sujet en vue subjective.

5.2 Méthode proposée

En reprenant les travaux de (Bulling *et al.*, 2011) et (Kitani *et al.*, 2011), Ogaki *et al.* (2012) ont montré qu'en intégrant de l'information sur l'égo-mouvement en plus des mouvements oculaires, le taux de reconnaissance d'activités était meilleur ; en effet, ils obtiennent un gain de, respectivement, 10% et 17% par rapport aux mouvements oculaires et à l'égo-mouvement.

Néanmoins, un inconvénient majeur de l'approche (Bulling *et al.*, 2011), et donc de (Ogaki *et al.*, 2012), est que les performances sont en partie limitées par la simplicité de la mise en oeuvre, certes efficace, mais pas suffisante pour atteindre des taux de reconnaissance plus élevés. En effet, les caractéristiques statistiques sur les mouvements oculaires et l'égo-mouvement à un instant donné t sont obtenues au sein d'une fenêtre glissante de taille assez grande centrée en t, or au sein de cette fenêtre la distribution statistique peut fortement varier localement suivant le type d'activité effectuée. Par conséquent, cela peut engendrer une certaine ambiguïté lors de la phase d'apprentissage.

Nous proposons de reprendre ces travaux et de présenter de nouvelles contributions afin d'améliorer le taux de reconnaissance d'activités, notamment en y incorporant de l'information multi-échelle, temporelle et contextuelle pour mieux caractériser et apprendre les activités effectuées.

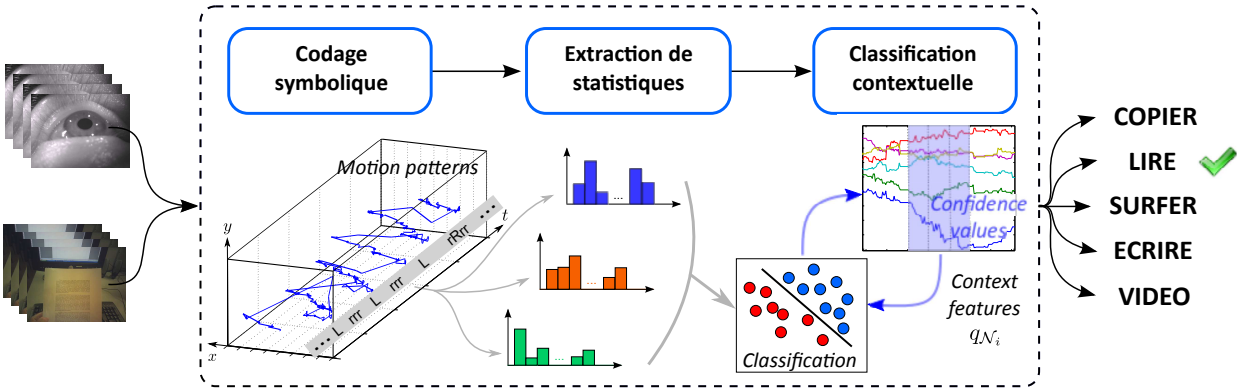


Figure 5.3 – Chaîne de traitement proposée.

A partir de données acquises en vue subjective, l'objectif est de reconnaître différentes activités, plus particulièrement des activités de bureau (cf. figure 5.2) :

- ▷ **copier** du texte d'un ordinateur à un autre
- ▷ **lire** un texte
- ▷ **surfer** sur internet
- ▷ **écrire** un texte
- ▷ regarder une **vidéo**

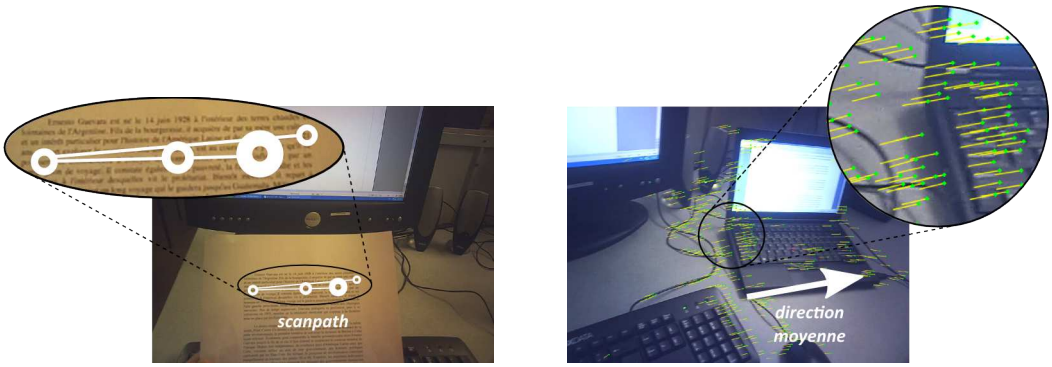
Plus spécifiquement, la chaîne de traitement (cf. figure 5.3) se compose de : (1) le codage symbolique de mouvements atomiques (cf. § 5.2.1), incluant l'extraction de données subjectives, (2) l'extraction de caractéristiques statistiques comprenant l'intégration d'information multi-échelle et temporelle (cf. § 5.2.2) et (3) la classification par apprentissage contextuel (cf. § 5.2.3).

5.2.1 Codage symbolique de mouvements atomiques

Pour pouvoir reconnaître les activités en vue subjective, nous traitons deux types de mouvements :

- ▷ **Mouvements oculaires** : Les mouvements oculaires sont capturés à l'aide d'un eye-tracker binoculaire porté fonctionnant en éclairage infrarouge[2]. Ce système renvoie les coordonnées g_x et g_y du point de regard (cf. figure 5.4(a)).
- ▷ **Ego-mouvement** : En plus du point de regard, l'eye-tracker acquiert des images de l'environnement par l'intermédiaire d'une caméra scène. A partir de ces images,

[2]Les raisons de ce choix, plutôt que l'utilisation du système eye-tracker porté proposé dans le chapitre 3, sont expliquées en § 5.3.1



(a) (b)

Figure 5.4 – (a) Mouvements oculaires (aussi appelés *scanpath*), (b) Flot optique épars.

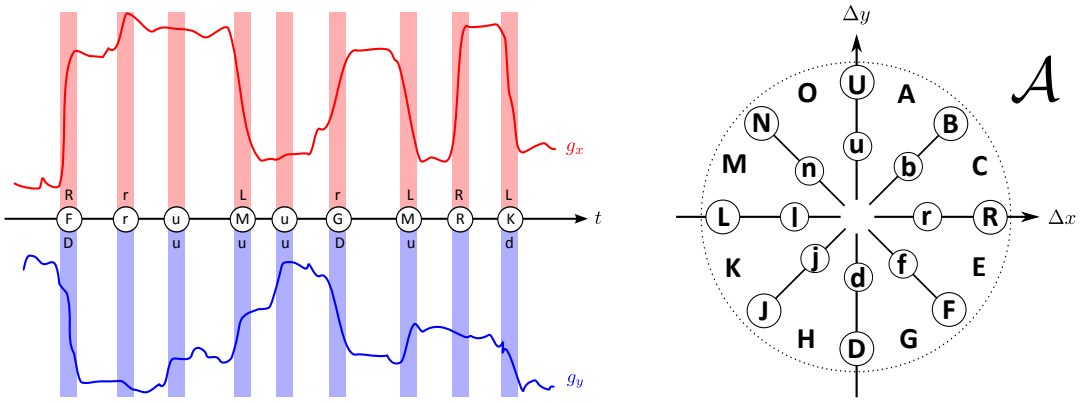
l'égo-mouvement est obtenu par estimation du flot optique épars. Le flot optique repose sur le suivi et la mise en correspondance de points d'intérêt entre des images consécutives. Nous utilisons l'implémentation pyramidale de Lucas-Kanade. Puis, les points, ne satisfaisant pas la contrainte d'homographie planaire, sont éliminés via l'algorithme RANSAC (Fischler et Bolles, 1981). Finalement, le flot moyen est calculé selon les deux directions, x et y (cf. figure 5.4(b)).

Les signaux issus de ces mouvements sont d'abord utilisés pour identifier différents types de mouvements atomiques, à savoir des mouvements de faible ou grande amplitude. Par exemple, dans le cas des mouvements oculaires, il s'agit de s'intéresser aux saccades oculaires. Ensuite, ces mouvements atomiques sont encodés en une séquence de symboles (cf. figure 5.5(a)).

Dans la suite, le même procédé d'extraction des caractéristiques est appliqué aux mouvements issus des yeux et de la tête, et seuls les paramètres sont ajustés différemment (cf. § 5.3.1 pour plus de détails sur les valeurs choisies des paramètres). Nous appliquons des étapes de traitement similaires à celles proposées dans (Bulling *et al.*, 2011).

Réduction du bruit - Tout d'abord, le signal est filtré à l'aide d'un filtre médian pour réduire l'influence du bruit. Dans le cas des yeux, ce bruit peut, par exemple, se traduire par une erreur de suivi de la pupille, alors que pour les mouvements de la tête, il peut s'agir de la mise en correspondance erronnée de points caractéristiques. Cette opération est effectuée tout en préservant la raideur et l'amplitude du signal pour la classification des mouvements atomiques.

Transformée en Ondelette Continue - Pour pouvoir détecter les mouvements saccadiques, nous appliquons une transformée en ondelette continue au signal $g(t)$ dont les



(a) (b)

Figure 5.5 – (a) Principe du codage symbolique construit à partir des signaux continus de mouvements horizontaux et verticaux, (b) Alphabet $\mathcal{A}$ pour le codage symbolique. Figures inspirées de (Bulling *et al.*, 2011).

coefficients d'ondelettes c_b^a sont obtenus via l'intégrale suivante :

$$c_b^a(g) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} g(t)\psi\left(\frac{t-b}{a}\right) dt \tag{5.1}$$

où b désigne le temps et a l'échelle. ψ représente l'ondelette mère de Haar dont l'expression est donnée par :

$$\psi(t) = \begin{cases} 1 & 0 \leq t < \frac{1}{2} \\ -1 & \frac{1}{2} \leq t < 1 \\ 0 & \text{sinon} \end{cases} \tag{5.2}$$

Seuillage - Le signal obtenu est quantifié en mouvement saccadique ($m_i \neq 0$) ou fixation ($m_i = 0$) par seuillage des coefficients d'ondelettes à différents niveaux d'amplitude. Deux niveaux d'amplitude sont, ici, considérés : une *faible* amplitude avec un seuil τ_1 et une *forte* amplitude avec un seuil τ_2. En fonction des intervalles de seuils, la discrétisation s'effectue alors de la façon suivante :

$$\forall i \; m_i = \begin{cases} 2 & c_i^a \geq \tau_2 \\ 1 & \tau_1 \leq c_i^a < \tau_2 \\ 0 & -\tau_1 < c_i^a < \tau_1 \\ -1 & -\tau_2 < c_i^a \leq -\tau_1 \\ -2 & -\tau_2 \leq c_i^a \end{cases} \tag{5.3}$$

avec i indiquant la position temporelle.

Quantification temporelle et spatiale - Le signal est, ensuite, quantifié temporellement pour obtenir un signal d'impulsions d'amplitudes différentes. Finalement, l'information des deux signaux, horizontal m^x et vertical m^y, est combinée afin de définir le symbole associé s. Plus spécifiquement, lorsqu'un mouvement atomique (caractérisé par une impulsion) est détecté dans l'un des deux signaux, le mouvement joint est représenté par un symbole issu d'un alphabet $\mathcal{A}$:

$$\begin{aligned} \Psi : \{-2,-1,0,1,2\}^2 \setminus (0,0) &\rightarrow \mathcal{A} \\ (m^x, m^y) &\rightarrow s \end{aligned} \tag{5.4}$$

Si parmi deux mouvements saccadiques joints, l'un possède une forte amplitude, alors le symbole résultant correspond à une lettre majuscule (ex. $(1,2) = (\text{"r"}, \text{"U"}) \rightarrow \text{"A"}$). En revanche, une lettre minuscule est attribuée dans le cas où aucune saccade n'a une forte amplitude (ex. $(-1,-1) = (\text{"l"}, \text{"d"}) \rightarrow \text{"j"}$). Dans notre cas, l'alphabet a une cardinalité égale à $|\mathcal{A}| = 24$ et il permet de coder les mouvements atomiques dans un espace joint direction-amplitude (cf. figure 5.5(b)). Finalement en répétant la procédure, pour chaque impulsion des deux signaux, nous générons une séquence de symboles $\mathbf{s}$.

5.2.2 Extraction de caractéristiques statistiques

A partir de la séquence de symboles obtenus par codage symbolique, des caractéristiques statistiques sont extraites au sein d'une fenêtre glissante. Dans un premier temps, nous décrivons la méthode proposée dans (Bulling *et al.*, 2011) et ses limitations. Ensuite, nous présentons deux modifications au schéma d'extraction des caractéristiques statistiques afin de mieux représenter une activité.

Statistiques n-grammes - Un *n-gramme* est une sous-séquence de n éléments consécutifs construite à partir d'une séquence donnée. La représentation n-gramme est fréquemment utilisée en raison de sa simplicité de mise en oeuvre et des bonnes performances qui peuvent en résulter.
Dans notre cas, le principe d'extraction de statistiques à partir de n-grammes est identique à celui décrit dans (Bulling *et al.*, 2011). Une fenêtre glissante (*sliding window*) de taille w parcourt une séquence donnée $\mathbf{s}$ et pour chaque instant t, elle permet de définir une sous-séquence :

$$\mathbf{s}_t = \left[s_{t-\frac{w}{2}}, \ldots, s_t, \ldots, s_{t+\frac{w}{2}} \right] \tag{5.5}$$

Puis, pour chacune des sous-séquences, un histogramme est construit à partir du nombre d'occurrences d'un n-gramme. Des statistiques du 1$^{\text{er}}$ ordre (moyenne, maximum, variance, différence entre maximum et minimum, et taille du vocabulaire) sont alors extraites à partir de l'histogramme afin de constituer le vecteur d'entrée d'un classifieur. Notons qu'en procédant ainsi, l'identité de l'élément n'est pas conservée[3].

[3]Derrière une telle représentation (tout comme pour la représentation *sac de mots*), nous retrouvons, en fait, l'idée d'isomorphisme discutée plus en détails dans (Edelman, 1999).

En procédant à l'extraction de statistiques, il est possible de partiellement contourner un problème majeur des modèles n-grammes, à savoir la définition d'un corpus d'apprentissage suffisamment représentatif de l'ensemble des n-grammes. En effet, certains n-grammes peuvent ne pas être présents au sein du corpus d'apprentissage, mais ils peuvent apparaître lors de la phase de test.

Dans notre cas, le corpus d'apprentissage est construit incrémentalement en fonction de l'apparition de nouveaux n-grammes présents dans la fenêtre glissante. Seule la similarité entre les distributions, en particulier des statistiques sur l'histogramme des n-grammes, servent à différencier les activités entre elles.

Malgré tout, le pouvoir discriminatif du n-gramme est limité par la taille n de ce dernier. En effet, pour un alphabet $\mathcal{A}$, le vocabulaire d'un n-gramme est de taille égale à $|\mathcal{A}|^n$ et donc sa dimensionalité croît exponentiellement en fonction de n. Ainsi, en pratique, un n de valeur faible est choisi, ce qui ne permet pas de prendre en compte une dépendance à longue portée.

Pour pouvoir remédier à ce problème, nous proposons deux schémas simples d'extraction de caractéristiques : un codage *multi-échelle* et un partitionnement *temporel*.

Codage multi-échelle - Coder les mouvements via l'alphabet $\mathcal{A}$ revient à considérer les mouvements fins des yeux et de la tête, or cela peut accroître l'effet de sur-ajustement lors de l'apprentissage. Il est donc nécessaire d'adopter une représentation plus générale qui peut être utile pour capturer une information plus globale.

Afin de partiellement résoudre ce problème, nous proposons d'employer un codage multi-échelle. En plus de l'alphabet $\mathcal{A}$ définit en § 5.2.1, nous considérons un sous-alphabet $\mathcal{A}^L \subset \mathcal{A}$ qui permet d'encoder des mouvements de plus large amplitude (cf. figure 5.6) :

$$\mathbf{S} = \{(\mathbf{s}_t, \mathbf{s}_t^L) \mid t = 1, \ldots, n\} \tag{5.6}$$

où $\forall t \ s_t \in \mathcal{A}, \ s_t^L \in \mathcal{A}^L$. L'idée est qu'une structure hiérarchique, constituée d'une trajectoire fine et d'une trajectoire simplifiée, peut permettre de capturer des informations à différents niveaux selon l'activité effectuée. Par exemple, une activité à court terme aura recours à des caractéristiques fines, alors qu'une activité de plus longue durée privilégira plutôt une information plus globale.

Partitionnement temporel - En plus du codage multi-échelle, nous proposons d'exploiter l'information temporelle (cf. figure 5.7). Nous divisons le contenu d'une fenêtre glissante $\mathbf{s}_t$ en différentes régions temporelles $\left[\mathbf{s}_t^1, \ldots, \mathbf{s}_t^P\right]$. L'objectif est de capturer des statistiques locales facilitant la similarité entre distributions via une comparaison par morceaux. En effet, un mélange de statistiques permet de mieux capturer la forme de la distribution de n-grammes.

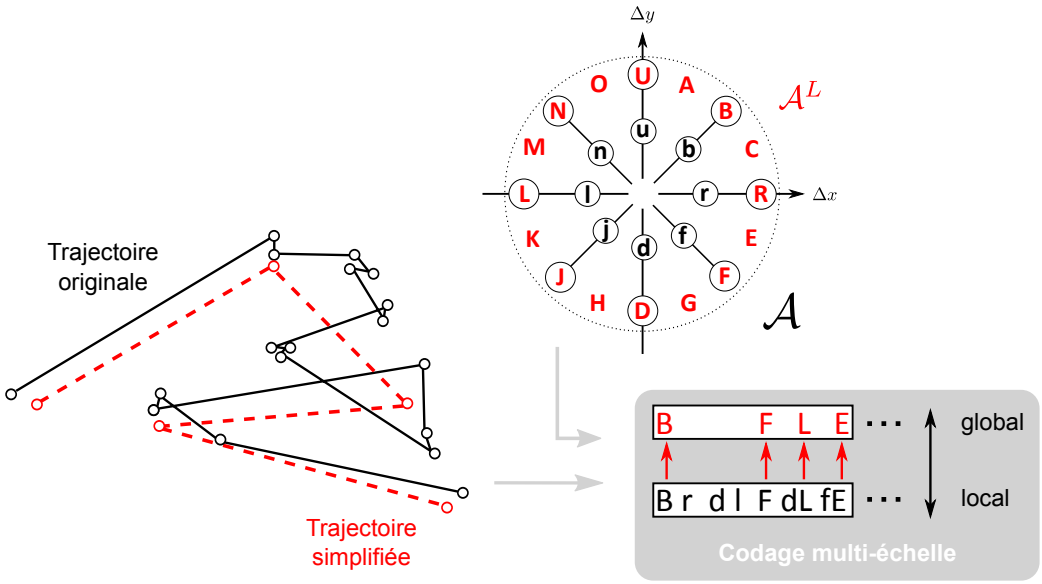


Figure 5.6 – Codage multi-échelle d'une trajectoire. Nous appliquons la procédure de codage symbolique pour une trajectoire simplifiée, en plus de la trajectoire originale.

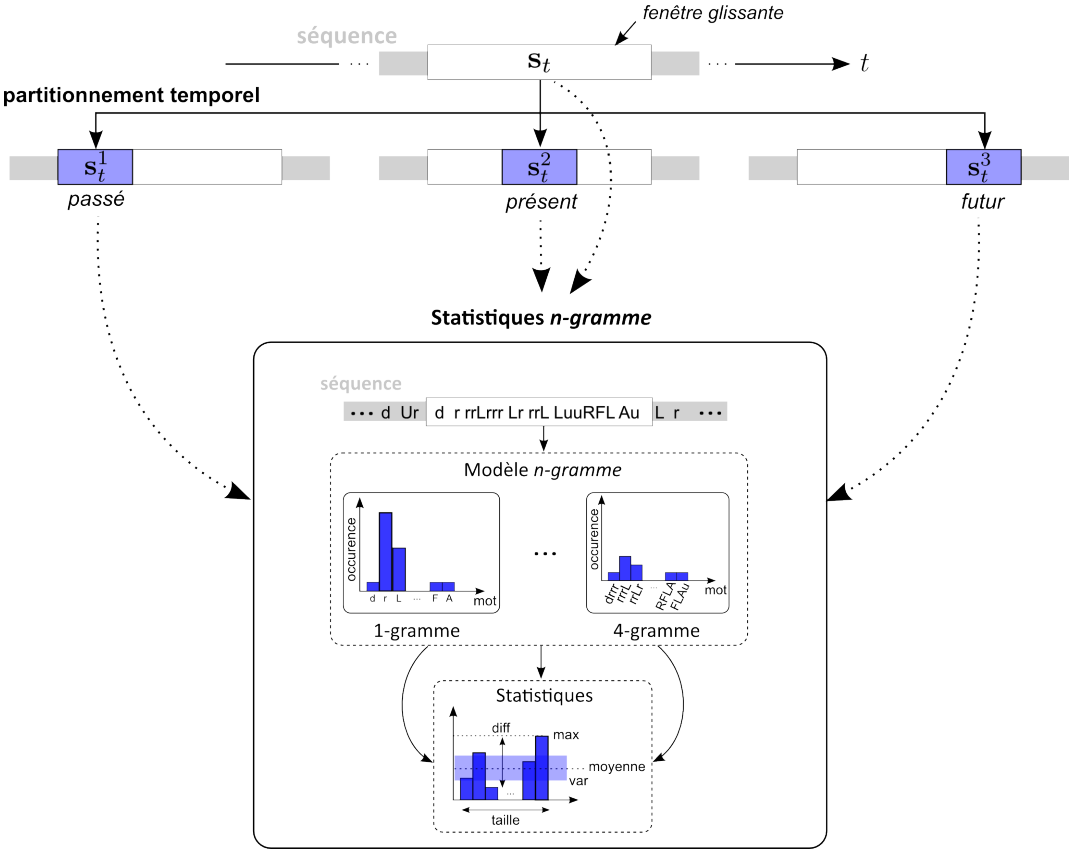


Figure 5.7 – Construction de statistiques temporelles (ici, $p = \{1, 3\}$, c'est-à-dire que nous considérons des statistiques extraites pour la fenêtre glissante principale et pour 3 de ses sous-fenêtres) à partir d'un modèle n-gramme à différents niveaux (ici, $n = 1, \ldots, 4$).

Nous pouvons, par exemple, supposer que certaines activités sont caractérisées par une information étalée dans le temps, alors que d'autres sont davantage concentrées sur l'instant présent.

5.2.3 Classification par apprentissage contextuel

Dans cette section, nous abordons le problème de classification des données. Tout d'abord, nous présentons le choix de la stratégie de classification employée. Puis, nous décrivons deux approches principales associées à cette stratégie choisie.

Supposons que nous disposons de n données d'apprentissage et de leur classes associées $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ avec $\mathbf{x}_i \in \mathbb{R}^d$ pour $i = 1, \ldots, n$. De plus, nous avons $y_i \in \{1, -1\}$ dans le cas binaire, alors que dans le cas multi-classe, $y_i \in \{1, \ldots, K\}$.

Pour apprendre un modèle de prédiction à partir des données d'apprentissage, une première stratégie, dite *générative*, consiste à optimiser, après décomposition en vraisemblance $p(\mathbf{X}|\mathbf{y})$ et probabilité a priori $p(\mathbf{y})$ via le théorème de Bayes, le *maximum a prosteriori* (MAP) :

$$\mathbf{y}^* = \underset{\mathbf{y}}{\operatorname{argmax}}\ p(\mathbf{y}|\mathbf{X}) = \underset{\mathbf{y}}{\operatorname{argmax}}\ p(\mathbf{X}|\mathbf{y})p(\mathbf{y}) \tag{5.7}$$

où $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$. Cependant, en pratique, il peut s'avérer difficile de modéliser la vraisemblance et la distribution a priori.

Une autre stratégie, dite *discriminante*, quant à elle, vise à étudier la probabilité a posteriori directement. Plus spécifiquement, elle s'intéresse à la distribution marginale :

$$p(y_i|\mathbf{X}) = \int_{\mathbf{y}_{-i}} p(y_i, \mathbf{y}_{-i}|\mathbf{X}) d\mathbf{y}_{-i} \tag{5.8}$$

où $\mathbf{y}_{-i}$ est égale à $\mathbf{y}$ privé de y_i. Une possibilité pour approximer cette distribution est d'avoir recours aux méthodes de classification.

Dans la suite, nous nous intéressons plus particulièrement à la seconde stratégie. Dans un premier temps, nous décrivons brièvement les méthodes traditionnelles de classification linéaire pour des grands ensembles de données : le SVM et la régression logistique. Ensuite, nous présentons la méthode d'apprentissage contextuel que nous proposons d'appliquer à notre problème de reconnaissance d'activités.

5.2.3.1 Classification linéaire de grands ensembles de données

Dans cette section, nous décrivons des méthodes de classification linéaire par apprentissage supervisé. L'objectif est d'apprendre une fonction de décision :

$$f(\mathbf{x}) \triangleq \bar{\mathbf{w}}^\top \phi(\bar{\mathbf{x}}) + b = \mathbf{w}^\top \phi(\mathbf{x}) \tag{5.9}$$

pour laquelle la décision est vraie si $f(\mathbf{x}) \geq 0$ et fausse sinon. Pour des raisons de clareté, nous supposons que $\mathbf{x} = \left[1, \bar{\mathbf{x}}^\top\right]^\top$ et $\mathbf{w} = \left[b, \bar{\mathbf{w}}^\top\right]^\top$ sont des vecteurs *augmentés* qui prennent en compte le terme de biais b.

Dans la suite, nous considérons uniquement la classification avec un noyau linéaire ($\phi(\mathbf{x}) = \mathbf{x}$), car elle donne des résultats compétitifs en comparaison avec des méthodes non-linéaires (Yuan *et al.*, 2012). De plus, il est plus facile d'interpréter le modèle de classification, car celui-ci repose sur la pondération directe des caractéristiques d'entrée.

Classification binaire - Pour générer une fonction de décision, la classification linéaire implique de résoudre le problème suivant de minimisation du risque :

$$\min_{\mathbf{w}} \; F(\mathbf{w}) \triangleq C \sum_{i=1}^{n} l(\mathbf{x}_i, y_i, \mathbf{w}) + R(\mathbf{w}) \tag{5.10}$$

avec $l(\mathbf{x}_i, y_i, \mathbf{w})$ une fonction de coût convexe non-négative, $R(\mathbf{w})$ la fonction de régularisation et C le coût de l'erreur de classification. Il existe deux types populaires de classification qui se différencient principalement par leur fonction de coût :

- ▷ **SVM** : Le but d'un classifieur SVM (Vapnik, 1995) est de trouver l'hyperplan séparateur qui, pour chaque classe, possède la distance la plus grande avec les échantillons d'apprentissage les plus proches. En lui conférant une marge suffisament large, le classifieur obtient alors une bonne généralisation, ce qui lui permet de mieux classer les échantillons de test. La classe associée au vecteur d'entrée $\mathbf{x}$ est attribuée en fonction du signe de la fonction de décision :

$$y = \text{sign}(\mathbf{w}^\top \mathbf{x}) \tag{5.11}$$

Deux fonctions de coût sont possibles et reposent sur la norme L_1 ou L_2 :

$$\begin{array}{rcl} l_{L_1}(\mathbf{x}_i, y_i, \mathbf{w}) & = & \left[1 - y_i \mathbf{w}^\top \mathbf{x}_i\right]_+ \\ l_{L_2}(\mathbf{x}_i, y_i, \mathbf{w}) & = & \left[1 - y_i \mathbf{w}^\top \mathbf{x}_i\right]_+^2 \end{array} \tag{5.12}$$

où $[z]_+ = \max\{0, z\}$ est la fonction coût charnière (*hinge loss*).

- ▷ **Régression logistique** : La régression logistique est un modèle probabiliste discriminant pour la classification dont la probabilité a posteriori pour un label y positif est :

$$p(y = +1|\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x}) \tag{5.13}$$

avec la fonction sigmoïde $\sigma(z) = (1 + e^{-z})^{-1}$. La fonction de coût est donnée par la log-vraisemblance négative d'une prédiction positive :

$$l_{LR}(\mathbf{x}_i, y_i, \mathbf{w}) = -\log \; p(y_i|\mathbf{x}_i, \mathbf{w}) = \log(1 + e^{-y_i \mathbf{w}^\top \mathbf{x}}) \tag{5.14}$$

En plus de la fonction de coût, il est nécessaire de définir le terme de régularisation[4] qui permet d'éviter l'effet de sur-ajustement soit via la norme L_1 (favorisant l'éparsité du modèle), soit via la norme L_2 :

$$\begin{aligned} R_{L_1}(\mathbf{w}) &= ||\mathbf{w}||_1 \\ R_{L_2}(\mathbf{w}) &= \frac{1}{2}||\mathbf{w}||_2^2 \end{aligned} \tag{5.15}$$

Pour un aperçu des techniques d'optimisation pour différentes combinaisons de fonction de coût et de régulariseur, on pourra se référer à (Yuan *et al.*, 2012).

Classification multi-classe - Concernant la régression logistique, cela revient à remplacer la distribution de Bernouilli par une distribution multinomiale.

Dans le cas du SVM, la formulation est moins triviale. Le problème de classification multi-classe est souvent résolu par la combinaison de différents classifieurs binaires telle que :

- **One-versus-rest** : K classifieurs binaires sont appris indépendemment. Chaque classifieur k est entraîné pour discriminer la classe k parmi toutes les autres classes et la donnée d'entrée $\mathbf{x}$ est assignée à la classe dont le score du classifieur est le plus élevé :

$$y^* = \underset{k=1,\ldots,K}{\operatorname{argmax}} \ \mathbf{w}_k^\top \mathbf{x} \tag{5.16}$$

- **One-versus-one** : $K(K-1)/2$ classifieurs binaires sont appris. Chaque classifieur k est construit entre 2 classes et la donnée d'entrée $\mathbf{x}$ est assignée à la classe qui récolte le plus de votes.

Notons également qu'il existe d'autres techniques qui considèrent toutes les données en même temps et proposent de définir un nouvelle fonction de coût (Weston et Watkins, 1999; Crammer et Singer, 2001).

En pratique, *one-versus-rest* est la technique la plus fréquemment utilisée, car elle est facile à implémenter et possède de meilleures propriétés de mise à l'échelle.

5.2.3.2 Modèle Auto-contexte

Motivation - L'inconvénient de l'approche précédente est qu'elle procède à une classification indépendante pour chaque $\mathbf{x}_t$ et elle ne prend donc pas en compte l'information contextuelle, en particulier l'information issue de son voisinage. Afin de résoudre ce problème, des techniques d'étiquetage structuré ont été proposées pour prédire les étiquettes $\mathbf{y}$ *jointement*. Une approche traditionnelle est d'employer un champ conditionnel aléatoire de Markov (CRF, (Lafferty *et al.*, 2001)) avec une fonction de potentiel d'ordre n

[4]Dans le cas de la régression logistique, cela revient à appliquer une distribution a priori sur $\mathbf{w}$: un prior Laplacien pour l_{L_1} et un prior Gaussien pour l_{L_2}.

(clique d'ordre n). Cependant, en pratique, les CRFs sont appris avec un voisinage fixe et un nombre limité de connexions pour rendre l'inférence tractable, ce qui ne permet pas récupérer des informations contextuelles d'ordre supérieur.

Nous proposons donc d'employer une procédure structurée d'étiquetage connue sous le nom d'apprentissage par *Auto-Contexte* (Tu et Bai, 2010).

Description du modèle - Considérons une séquence de données comme un graphe $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ dont les arêtes $\mathcal{E}$ définissent la topologie du graphe, c'est-à-dire le voisinage de chaque noeud. Pour l'étiquetage d'une séquence, dont les noeuds $\mathbf{v} = \{v_1, \ldots, v_n\} \in \mathcal{V}$ forment une chaîne, il est alors possible de définir le voisinage $\mathcal{N}_i$ de chaque noeud v_i comme les m noeuds précédents et suivants ce noeud :

$$\mathcal{N}_i = \{v_{t-\frac{m}{2}}, \ldots, v_{t-1}, v_{t+1}, \ldots, v_{t+\frac{m}{2}}\} \tag{5.17}$$

Dans un premier temps, un classifieur traditionnel est appris à partir des caractéristiques d'entrée $\mathbf{X}$. En évaluant ce modèle, nous obtenons alors des valeurs de confiance q_i associées à chaque noeud v_i :

$$\mathbf{q} = [q_1, \ldots, q_n]^\top \tag{5.18}$$

Ces valeurs[5] constituent les distributions a posteriori marginales (très approximatives) pour les données d'entrée $\mathbf{x}_i$.
Ensuite, un nouveau classifieur est appris en considérant ces valeurs de confiance, en plus des caractéristiques d'entrée :

$$q_i = f(\mathbf{x}_i, \mathbf{q}_{\mathcal{N}_i}; \theta) \tag{5.19}$$

avec θ le paramètre de la fonction f. $\mathbf{q}_{\mathcal{N}_i}$ sont les valeurs de confiance du voisinage $\mathcal{N}_i$ défini par (5.17) et elles permettent d'intégrer l'information contextuelle dans l'apprentissage. Plus généralement, pour tous les échantillons, nous obtenons la forme vectorisée :

$$\mathbf{q} = \mathbf{f}(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{q}; \theta) \tag{5.20}$$

avec $\mathbf{f}(\cdot) = [f(\mathbf{x}_1, \mathbf{q}_{\mathcal{N}_1}; \theta), \ldots, f(\mathbf{x}_n, \mathbf{q}_{\mathcal{N}_n}; \theta)]^\top$.
Le modèle Auto-Contexte est obtenu en itérant le procédé précédent plusieurs fois jusqu'à convergence :

$$\mathbf{q}^j = \mathbf{f}_j(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{q}^{j-1}; \theta_{j-1}) \tag{5.21}$$

En procédant ainsi, il a été montré que le modèle Auto-Contexte approxime $p(y_i|\mathbf{X})$ récursivement et asymptotiquement (Tu et Bai, 2010).

Les phases d'apprentissage et de test sont récapitulées, respectivement, via les algorithmes 5.1 et 5.2.

[5] Il peut s'agir aussi bien de scores de confiance que de probabilités.

Algorithm 5.1 Auto-Context model - *Learning*

Input: Training structures $\boldsymbol{\mathcal{G}} = \{\mathcal{G}_1, \ldots, \mathcal{G}_N\}$, number of iterations T
Output: Layered contextual prediction functions $f_1, \ldots, f_T$
  Initialize labeling confidences $\mathbf{q}^0$ to a uniform distribution
  for $j = 1, \ldots, T$ **do**
    Form contextually-augmented set $S_j = \{(\mathbf{x}_1, y_1, \mathbf{q}^{j-1}_{\mathcal{N}_1}), \ldots, (\mathbf{x}_n, y_n, \mathbf{q}^{j-1}_{\mathcal{N}_n})\}$ from $\boldsymbol{\mathcal{G}}$
    Train jth-layer contextual prediction function f_j on S_j
    Compute jth-layer labeling confidences $\mathbf{q}^j = \mathbf{f}_j(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{q}^{j-1}; \theta_j)$
  end for

Algorithm 5.2 Auto-Context model - *Testing*

Input: Testing structure $\mathcal{G}$, layered contextual prediction functions $f_1, \ldots, f_T$
Output: Testing labeling confidences $\mathbf{q}$ of $\mathcal{G}$
  Initialize labeling confidences $\mathbf{q}^0$ to a uniform distribution
  for $j = 1, \ldots, T$ **do**
    Form contextually-augmented set $S_j = \{(\mathbf{x}_1, y_1, \mathbf{q}^{j-1}_{\mathcal{N}_1}), \ldots, (\mathbf{x}_n, y_n, \mathbf{q}^{j-1}_{\mathcal{N}_n})\}$ from $\mathcal{G}$
    Compute jth-layer labeling confidences $\mathbf{q}^j = \mathbf{f}_j(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{q}^{j-1}; \theta_j)$
  end for

5.3 Résultats expérimentaux

5.3.1 Base de données et protocole expérimental

Dans cette section, nous présentons le système eye-tracker utilisé, la base de données que nous avons construite, ainsi que le protocole expérimental appliqué, afin d'évaluer la méthode proposée (cf. § 5.2).

Système eye-tracker porté - Pour enregistrer les données du regard et d'égo-mouvement, nous avons utilisé le système commercial *SMI eye tracking glasses*[6] (cf. figure 5.8) avec l'ordinateur portable *SMI-ETG* associé. L'eye-tracker dispose d'une caméra scène HD permettant l'acquisition d'images avec une résolution de 1280×960 à 24 fps. Son champ de vue, respectivement horizontal et vertical, est de 60° et 46°. Cependant, le système permet le suivi du regard dans un interval de 80° (horizontal) et de 60° (vertical). Les caméras oeil quant à elles, acquièrent les données à 30 fps. D'après les spécifications techniques, le système permet de suivre le regard avec une précision moyenne de 0.5° à n'importe quelle distance, au-delà de 40 cm. Pour chaque expérience et chaque sujet, nous avons effectué une calibration à 1 point. D'autres procédures de calibration proposées par le logiciel ont été, au préalable, testées : en pratique, une calibration à 3 points n'apportait pas d'amélioration significative et une calibration à 0 points, quant à elle, n'était pas satisfaisante.

[6] http ://www.eyetracking-glasses.com/

Figure 5.8 – Lunettes SMI.

Chiffres clés	
# classes	5 (+ 1 aléatoire)
# sujets	6 (4 ♂et 2 ♀)
# vidéos	12
# images	277 987

TABLE 5.1 – Données chiffrées.

Les principales raisons qui nous ont poussé à utiliser un système commercial plutôt que l'eye-tracker porté développé dans le chapitre 3 sont les suivantes. Premièrement, pour cette étude, nous avons besoin de capturer des mouvements oculaires fins, ce qui requiert de pouvoir estimer le regard avec une précision suffisante. Le système commercial utilisé possède une précision supérieure à celle du système que nous avons développé (0.5° contre 2°). Une autre raison réside dans la practicité du sytème commercial, notamment pour effectuer la calibration. Effectivement, la calibration s'effectue plus rapidement, car elle ne requiert qu'un point de calibration (contre 100 pour notre système). Les avantages cités précédemment sont essentiellement attribuable à l'utilisation d'un éclairage en infrarouge.

Base de données - Nous avons construit une base de données afin d'évaluer les performances des différentes solutions. Elle est constituée de 6 sujets et de 2 vidéos par sujet (cf. tableau 5.1). Pour chaque enregistrement, le sujet doit effectuer 5 activités prédéfinies à la suite, plus une activité aléatoire (par exemple, bouger la tête ou chanter) entre deux activités consécutives. Nous demandons aux sujets d'effectuer les mêmes activités que dans (Bulling *et al.*, 2011). Nous avons essayé de faire en sorte que chaque activité soit pratiquée durant environ 2min30, ce qui nous fait un total de plus de 3h de vidéos. Pour chacune des vidéos, nous disposons également des points de regard acquis à une fréquence différente égale à 30 fps.

Pour permettre d'effectuer les activités, différents matériels ont été rassemblés et les mêmes consignes ont été données à chacun des sujets :

- ▷ **Copier** : Il est demandé au sujet de copier, d'un ordinateur portable à un ordinateur fixe (équipé d'un logiciel de traitement de texte), un extrait du livre "L'Étranger" d'Albert Camus. Pour le second enregistrement, un deuxième extrait a été choisi.
- ▷ **Lire** : Le sujet doit lire deux extraits d'une courte biographie d'Ernesto Guevara.
- ▷ **Surfer** : Le sujet était libre de consulter différentes sites internet (*Fnac*, *Amazon* ...).
- ▷ **Ecrire** : Il est demandé au sujet d'écrire, sur un cahier, deux extraits (traduit en français) du discours *"I have a Dream"* de Martin Luther King, dictés à l'oral.
- ▷ **Vidéo** : La première vidéo d'une durée de 2min24 comprend la bande d'annonce de *"The Dark Knight Rises"* et une partie de la bande d'annonce de *"Indiana Jones and the Last Crusade"*. Concernant la seconde vidéo, la bande d'annonce de *"The Fifth*

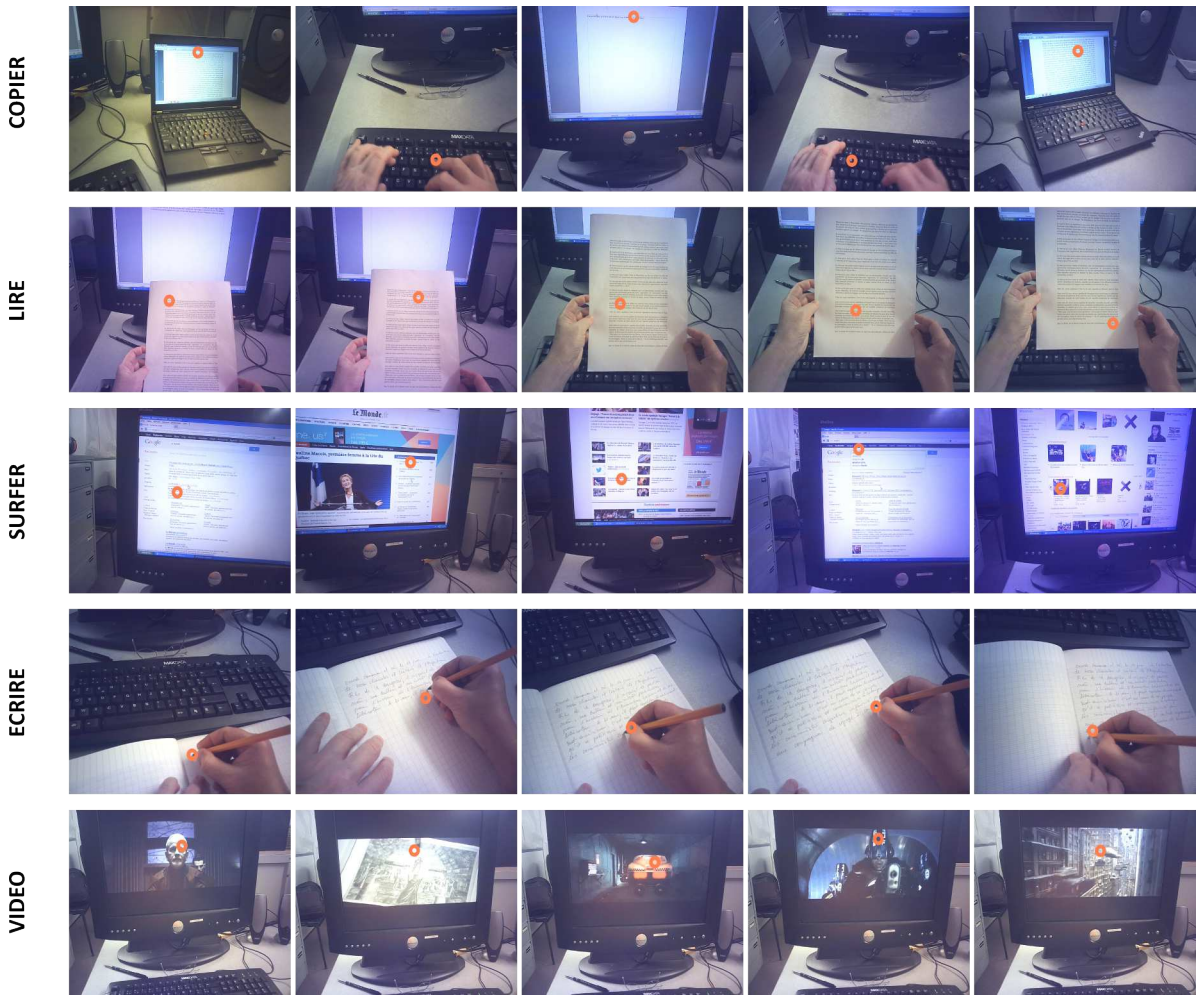


Figure 5.9 – Exemples d'images pour les cinq activités à reconnaître. Le curseur circulaire orange indique la point de regard dans le référentiel de l'image.

Element" a été diffusée et dure 2min35.

Notons que pour certaines personnes, le point de regard n'apparait parfois pas dans l'image de la caméra scène. Cela s'explique par des mouvements des yeux d'amplitude bien plus importante que ceux de la tête et qui font que le point de regard se trouve alors en dehors du champ de vue de la caméra scène. En pratique, cela ne pose pas de problème, car le logiciel associé aux lunettes SMI permet d'extrapoler le regard dans une certaine limite au-delà du champ de vue de la caméra (cf. § 5.3.1) et nous disposons donc bien des coordonnées du point de regard.

Protocole expérimental - Pour évaluer l'approche proposée, nous adoptons un protocole expérimental similaire à (Ogaki *et al.*, 2012). Nous disposons de 2 séries de vidéos

acquises pour chaque sujet. La première série de vidéos (soit 6 vidéos) est utilisée pour l'apprentissage des classifieurs. Dans la phase de test, nous appliquons les classifieurs sur la seconde série de vidéos. La procédure est répétée en échangeant les deux séries de vidéos. Finalement, nous analysons les résultats en terme de performances moyennes sur ces deux expériences.

Détails d'implémentation - Identiquement à (Ogaki *et al.*, 2012), les tailles des fenêtres glissantes sont égales à $w^e = 15$s (soit $15 \times 30 + 1 = 451$ images par fenêtre) et $w^h = 30$s (soit $30 \times 24 + 1 = 721$ images par fenêtre). Ces tailles sont fixées en seconde afin de les rendre invariantes à la fréquence d'acquisition des données. Les autres paramètres ont été déterminés expérimentalement. Ils restent les mêmes pour tous les sujets et ils ne sont donc pas optimaux pour chacun des sujets. Nous avons appliqué le filtre médian avec une fenêtre de taille $w^m = 5$. Pour la transformée en ondelette continue, nous avons choisi une échelle $b = 5$ pour les deux types de mouvements. Pour classer les mouvements continus, les seuils, respectivement pour les mouvements oculaires et de la tête, sont fixés à ($\tau_1^e = 50$, $\tau_2^e = 200$) et à ($\tau_1^h = 5$, $\tau_2^h = 15$). Concernant les statistiques n-grammes, nous fixons $n = 1, \ldots, 4$, ce qui constitue un vecteur de taille égale à 20 pour chacune des modalités.

Pour la classification linéaire, nous utilisons un SVM linéaire implémenté via LIBLINEAR (Fan *et al.*, 2008), en particulier nous employons une optimisation directe dans la forme primale, car le nombre de caractéristiques est très inférieur au nombre d'instances ($d \ll n$). Nous avons choisi la fonction de coût l_{L_2} et un régulariseur basé sur la norme L_2 (cf. § 5.2.3.1). L'optimisation est effectuée à l'aide de la méthode *trust region Newton* (Lin *et al.*, 2008). En ce qui concerne les paramètres, nous avons utilisé les mêmes valeurs que (Bulling *et al.*, 2011), c'est-à-dire $C = 1$ et $\epsilon = 0,1$. Nous adoptons une stratégie multi-classe de type *one-versus-rest* (cf. § 5.2.3.1). Notons ici que le choix du SVM est justifié étant donné que le nombre de caractéristiques d'entrée est relativement faible et que cela est tout à fait compatible avec le modèle Auto-Contexte (Tu et Bai, 2010).

Caractéristiques contextuelles $\mathbf{q}_{\mathcal{N}_i}$ - Jusqu'à présent, nous n'avons pas défni les caractéristiques contextuelles $\mathbf{q}_{\mathcal{N}_i}$ de l'équation (5.19). En effet, en pratique, il n'est pas possible de considérer toutes les valeurs de confiance de chaque noeud v_i, autrement nous aurions un nombre excessif de caractéristiques (dans notre cas, $>$ 4k) et l'apprentissage serait extrêmement long. De plus, certaines études ont montré que des résultats similaires peuvent être obtenus avec un nombre faible de caractéristiques contextuelles (Tu et Bai, 2010). Il est donc nécessaire de capturer suffisamment d'information contextuelle tout en trouvant un bon compromis entre la précision de classification et le temps d'apprentissage.

La procédure proposée est décrite dans la figure 5.10. Les caractérstiques contextuelles sont construites à partir des valeurs de confiance de chaque classifieur $\mathbf{q}^k$. Pour chaque classe $k \in \{1, \ldots, K\}$, nous proposons d'utiliser les caractéristiques contextuelles $\mathbf{q}^k_{\mathcal{N}_i}$ suivantes :

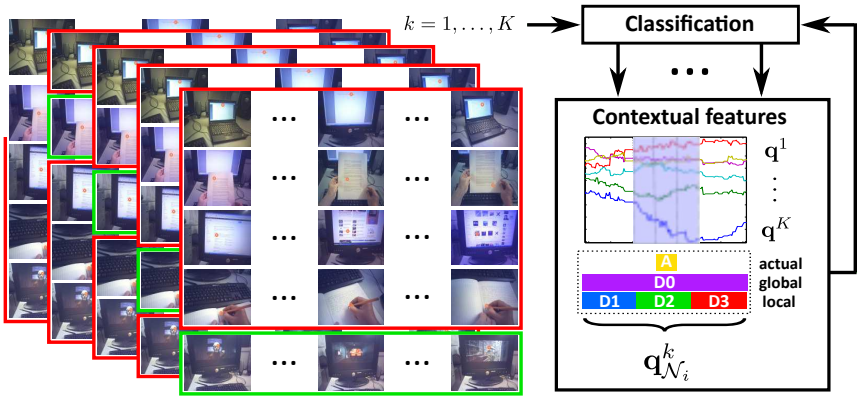


Figure 5.10 – Procédure de classification par apprentissage contextuel. Les caractéristiques contextuelles sont construites à partir des valeurs de confiance issues des classifieurs. Ces caractéristiques contextuelles sont ensuite utilisées, en plus des caractéristiques initiales $\mathbf{x}_i$, pour apprendre de nouveaux classifieurs. Cette procédure est répétée plusieurs fois jusqu'à convergence.

Caractéristiques contextuelles $\mathbf{q}^k_{\mathcal{N}_i}$		
valeur actuelle (A)	q^k_i	
différence relative globale (D0), locale (D2)	$q^k_i - \mu^k(i - \frac{w}{2}, i + \frac{w}{2})$	$q^k_i - \mu^k(i - \frac{w}{4}, i + \frac{w}{4})$
différence relative passée (D1), future (D3)	$q^k_i - \mu^k(i - \frac{w}{2}, i)$	$q^k_i - \mu^k(i, i + \frac{w}{2})$

avec :

$$\mu^k(i_1, i_2) = \frac{1}{i_2 - i_1 + 1} \sum_{i=i_1}^{i_2} q^k_i \tag{5.22}$$

Ainsi, nous proposons de définir :

$$\mathbf{q}_{\mathcal{N}_i} = \left[\mathbf{q}^1_{\mathcal{N}_i}, \ldots, \mathbf{q}^K_{\mathcal{N}_i}\right] \tag{5.23}$$

Notons qu'en définissant les cractéristiques contextuelles de cette façon, nous intégrons à la fois de l'information temporelle, mais également de l'information inter-classe.

5.3.2 Résultats

Pour évaluer les performances de classification, nous effectuons : (1) une analyse globale (cf. § 5.3.2.1), (2) une analyse par classe (cf. § 5.3.2.2) et (3) une analyse par sujet (cf. § 5.3.2.3).

Dans la suite, les résultats sont essentiellement comparés par rapport à la référence de base (Ogaki *et al.*, 2012). Elle repose sur l'extraction de statistiques sur des histogrammes construits à partir de $n = 1, \ldots, 4$-grammes pour les mouvements oculaires et l'égo-mouvement encodés avec l'alphabet $\mathcal{A}$ et sur l'utlisation d'un SVM linéaire pour la classification.

Mesures de performance - Nous évaluons les résultats en terme de moyenne de la précision moyenne (*mean Average Precision*, mAP).

$$mAP = \frac{1}{N_s} \sum_{j=1}^{N_s} \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbb{1}(\hat{y}_{i,j} = y_{i,j}) \tag{5.24}$$

avec $\mathbb{1}$ la fonction indicateur, n_j le nombre d'échantillons pour chaque sujet et N_s le nombre de sujets. Nous prenons la moyenne des mAP sur les deux expériences (correspondant aux deux séries de vidéos de la base de données) afin de constituer la performance globale d'une solution.

Nous adoptons les abbréviations suivantes : E pour désigner les mouvements oculaires, H pour désigner l'égo-mouvement, L pour l'utilisation de l'information multi-échelle, T pour l'ajout de l'information temporelle et AC pour l'utilisation de l'apprentissage contextuel.

5.3.2.1 Analyse globale

Influence du codage multi-échelle - Nous avons, tout d'abord, évalué si une information de plus haut niveau des mouvements oculaires et de la tête pouvait améliorer le taux de reconnaissance. Pour cela, nous avons extrait des statistiques n-grammes d'ordre 1 et 2 (au-delà, le gain est négligeable) construites à partir de l'alphabet $\mathcal{A}_L$, en plus des mouvements fins caractérisés par $\mathcal{A}$ ($n = 1, \ldots, 4$). Les résultats sont indiqués dans le tableau 5.2. Par exemple, E-H et $l = 2$ correspondent à l'ajout de 1- et 2-grammes pour les mouvements des yeux et de la tête encodés avec l'alphabet $\mathcal{A}_L$, en plus des caractéristiques de la référence de base. E-H et $l = \{2, 1\}$ correspondent à l'ajout de 1- et 2-grammes pour les mouvements des yeux et de 1-grammes pour ceux de la tête encodés avec l'alphabet $\mathcal{A}_L$. Comme nous pouvons le constater, le gain est supérieur à 4%, suivant que les mouvements proviennent des yeux ou de la tête uniquement et augmenter la taille du n-gramme n'apporte pas d'amélioration. De plus, ce gain augmente lorsque les caractéristiques de mouvements oculaires et d'égo-mouvement sont combinées. Effectivement, nous notons un gain significatif, plus de 7%, pour les tailles $l = \{2, 1\}$ en comparaison avec la référence de base.

Influence du partitionnement temporel - Similairement à l'analyse précédente, nous avons voulu connaître l'impact de l'information temporelle. Nous avons appliqué différents schémas de partionnements temporels p, en plus du codage multi-échelle E-H-L (cf. tableau 5.3). Par exemple, $p = \{1, 3\}$ revient à extraire les caractéristiques statistiques pour

TABLE 5.2 – Influence du codage multi-échelle. Référence désigne les résultats obtenus avec des statistiques calculées sur les 1- à 4-grammes pour les mouvements oculaires E et l'égo-mouvement H. $l = 2$ correspond à 1- et 2-grammes pour E et H, en plus de la référence. $l = \{2, 1\}$ correspond à 1- et 2-grammes pour E et à 1-grammes pour H, en plus de la référence.

Caract.	Référence	$l = 1$	$l = \{1, 2\}$	$l = \{2, 1\}$	$l = 2$
E	57,90%	62,30%	-	-	62,78%
H	40,96%	45,50%	-	-	45,31%
E-H	65,58%	71,94%	71,69%	72,67%	72,03%

TABLE 5.3 – Influence du partitionnement temporel. Nous considérons initialement un codage multi-échelle E-H-L de type $l = \{2, 1\}$. p indique comment la fenêtre glissante est divisée. Par exemple, $p = 1$ implique que les statistiques sont collectées au sein de la fenêtre glissante principale et $p = \{1, 3\}$ indique que nous considérons également des statistiques provenant de 3 sous-régions de la fenêtre glissante principale.

Caract.	$p = 1$	$p = \{1, 2\}$	$p = \{1, 3\}$	$p = \{1, 2, 3\}$
E-H-L $l = \{2, 1\}$	72,67%	75,89%	77,20%	77,45%
# caract.	55	165	220	330

la fenêtre initiale et pour 3 de ses sous-régions. Dans ce même tableau, nous indiquons également le nombre de caractéristiques. Ainsi, pour $p = 1$, nous avons un vecteur de dimension égale à 55 : 20 pour E, 20 pour H, 10 pour E-L (1- et 2-grammes) et 5 pour H-L (1-grammes). En observant les résultats, nous remarquons que, là aussi, l'information temporelle améliore les performances avec un gain de plus de 4% pour $p = \{1, 3\}$. Un schéma pyramidale $p = \{1, 2, 3\}$, quant à lui, améliore de façon négligeable le taux de reconnaissance par rapport $p = \{1, 3\}$. Afin de garder un compromis entre taux de reconnaissance et nombre de caractéristiques, nous adoptons un partitionnement temporel $p = \{1, 3\}$ pour la suite.

Dans la figure 5.11, nous affichons les performances en fonction du type de caractéristiques. Nous notons que la combinaison de l'information multi-échelle et temporelle permet d'améliorer les résultats de plus de 11% par rapport à la référence de base, alors que ce gain est d'environ 7% si ces informations sont considérées individuellement. En ne considérant que les mouvements des yeux ou de la tête séparément, nous avons un gain supérieur à, respectivement, 4% et 8%. Ces résultats confirment l'importance de combiner l'information

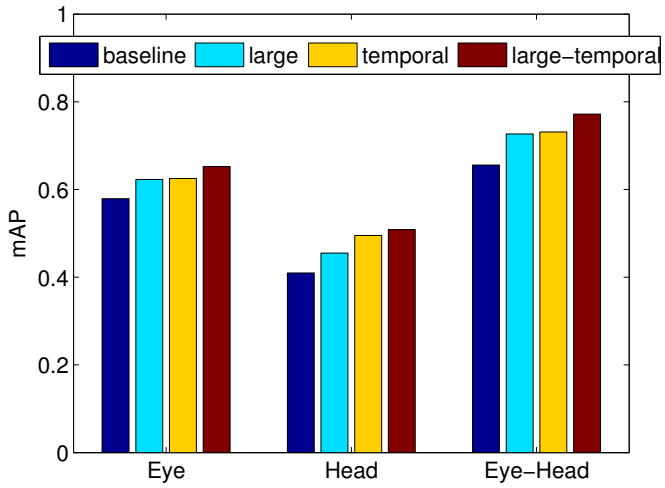


Figure 5.11 – Influence de l'information multi-échelle et temporelle pour les mouvements oculaires, l'égo-mouvement et la combinaison des deux mouvements.

multi-échelle et temporelle.

Modèle Auto-Contexte - Nous proposons d'évaluer les performances obtenues avec le modèle Auto-Contexte. Dans la figure 5.12(a), nous montrons les précisions moyennes obtenues lors des étapes d'apprentissage et de test, et pour deux solutions : l'une sans informations multi-échelle et temporelle (E-H), l'autre avec (E-H-L-T). Premièrement, dans les deux cas, nous constatons que le modèle Auto-Contexte améliore sensiblement les résultats par rapport à un classifieur traditionnel. Deuxièmement, la plus forte contribution est observée dès la $2^{\text{ème}}$ itération en comparaison avec les itérations suivantes.

Dans la figure 5.12(b), la proportion de caractéristiques contextuelles du modèle Auto-Contexte est affichée pour chaque activité et en fonction du nombre d'itérations (l'apprentissage est effectué avec l'information multi-échelle et temporelle). Ce ratio est obtenu à partir des valeurs absolues des poids des classifieurs, notamment en prenant la somme des poids contextuels sur celle de tous les poids (statistiques et contextuels) ; pour chacune des activités, nous prenons, ensuite, la moyenne sur les 6 sujets et les deux séries de vidéos. Nous indiquons également la moyenne sur toutes les activités par la courbe en trait continu. Après deux itérations, la proportion contextuelle avoisine les 30% pour chaque classe. Au-delà, parmi les différentes classes, nous remarquons que "ECRIRE" est celle qui a le moins recours à l'information contextuelle. Nous verrons plus tard que cela peut s'expliquer à cause de son taux de reconnaissance initialement très élevé sans information contextuelle (cf. § 5.3.2.2). En revanche, les activités "VIDEO" et "SURFER" sont celles qui profitent le plus du modèle Auto-Contexte, en particulier "VIDEO" qui atteint 50%, voire plus. Là

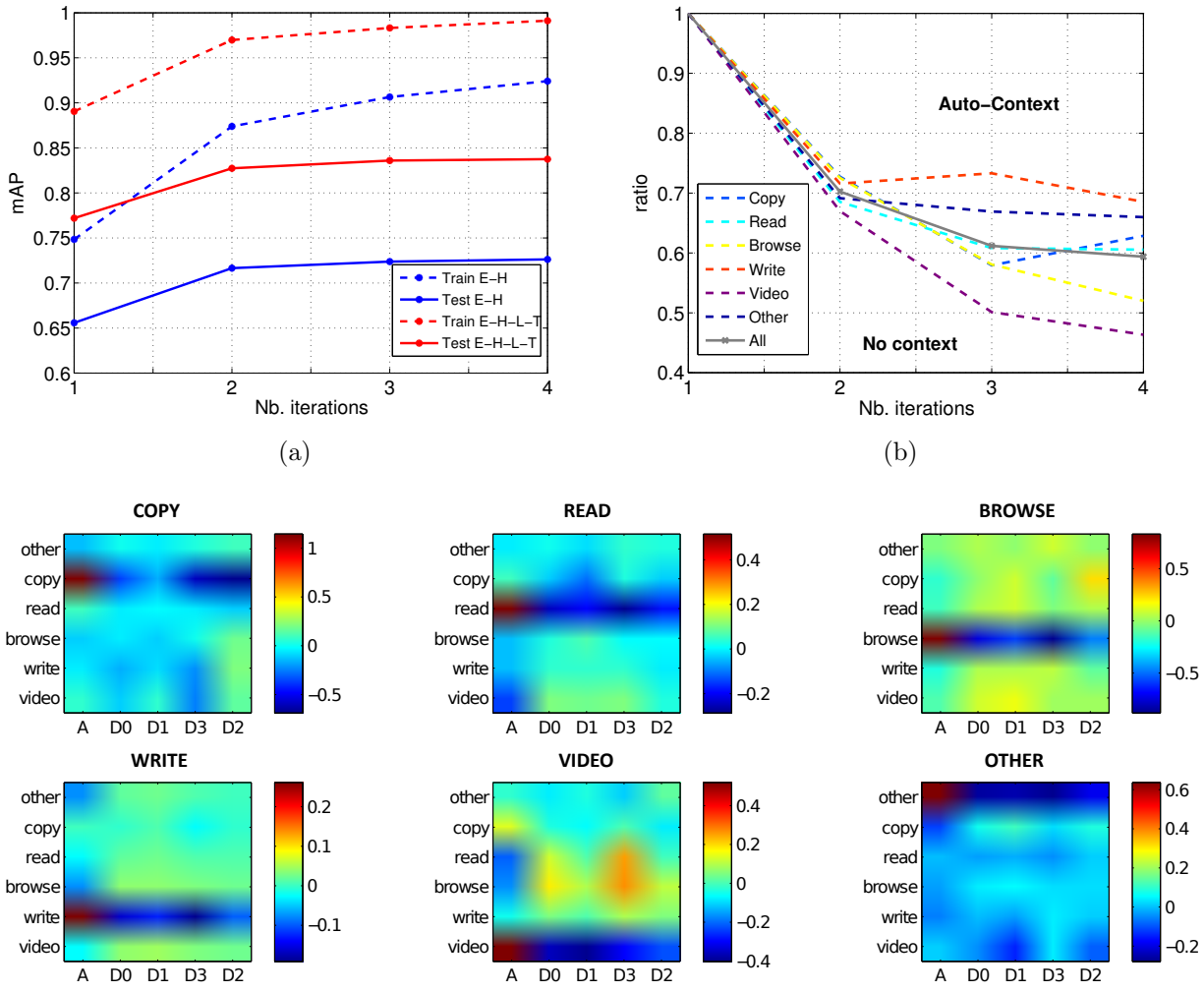


Figure 5.12 – Modèle Auto-Contexte : (a) Précisions moyennes pour les phases d'apprentissage et de test en fonction du nombre d'itérations, (b) Proportion de caractéristiques contextuelles pour chaque classe en fonction du nombre d'itérations (elle est obtenue en prenant le ratio entre la somme des valeurs absolues des poids contextuels et celle des valeurs absolues des poids de l'ensemble des caractéristiques), (c) Contribution des poids contextuels pour chaque activité à l'itération la plus bénéfique $T = 2$ (cf. § 5.3.1 pour la définition des caractéristiques contextuelles A et D0-D3).

aussi, nous verrons plus tard que leur taux de reconnaissance augmente considérablement grâce à l'information contextuelle (cf. § 5.3.2.2).

TABLE 5.4 – Synthèse des résultats

Caract.	Sans AC		Avec AC	
	SVM	**SVM Temp.**	**SVM**	**SVM Temp.**
E	57,90%	62,50%	71,23%	75,51%
H	40,96%	49,50%	50,54%	55,68%
E-L	62,30%	65,24%	73,72%	76,56%
H-L	45,50%	50,85%	52,36%	58,35%
E-H	65,58%	73,12%	72,38%	79,12%
E-H-L	72,67%	77,20%	78,53%	**83,59%**

E : Eye H : Head L : Codage multi-échelle ($l = \{2, 1\}$)
Temp : Caractéristiques temporelles ($p = \{1, 3\}$) AC : Auto-contexte ($T = 3$ iter.)

En regardant plus en détails les caractéristiques[7] associées aux poids contextuels à l'itération la plus bénéfique $T = 2$ (cf. figure 5.12(c)), nous observons que le poids dominant de chaque classe k est représenté essentiellement par la valeur actuelle q_i^k. En seconde position, nous retrouvons, en général, la différence relative future (D3) et celle-ci apporte plus d'information que la différence relative globale (D0) ou locale (D2).

Pour la suite, le modèle Auto-Contexte (AC) par défaut sera calculé avec $T = 3$ itérations, car au-delà le gain reste assez négligeable.

Synthèse - Le tableau 5.4 récapitule l'ensemble des principaux résultats en fonction des différents types de solution :

- ▷ **SVM** - Le classifieur est appris uniquement avec les statistiques n-grammes, similairement à (Ogaki *et al.*, 2012).
- ▷ **SVM + Temp.** - Le classifieur est appris avec les statistiques n-grammes temporelles.
- ▷ **SVM + AC** - Le classifieur est appris avec les statistiques n-grammes et en construisant un modèle Auto-Contexte.
- ▷ **SVM + Temp. + AC** - Le classifieur est appris avec les statistiques n-grammes temporelles et en construisant un modèle Auto-Contexte.

Nous affichons aussi les performances obtenues individuellement pour les mouvements oculaires et l'égo-mouvement. Notons que nous obtenons de bons résultats pour les mouvements oculaires seuls, en tenant compte uniquement de l'information contextuelle (71,23% contre 57,90%). Ces résultats dépassent même la référence de base (65,58%), ce qui confirme l'importance de la modélisation contextuelle.

[7] Leurs expressions ont été données en § 5.3.1.

Plus généralement, tous résultats confondus, nous obtenons un gain de 18% par rapport à la référence de base et de plus de 26% vis-à-vis des mouvements oculaires seuls.

5.3.2.2 Analyse par classe

Analyse des poids du classifieur - Nous proposons de regarder plus en détails l'influence de chaque caractéristique, notamment le codage multi-échelle et le partitionnement temporel. Pour cela, nous avons calculé la moyenne (sur les deux séries de vidéos) des poids des classifieurs et nous avons affiché les poids dominants positifs et négatifs dans la figure 5.13. Une description des caractéristiques analysées est indiquée dans le tableau 5.5.

D'un point de vue général, nous constatons que ces deux informations ont bien été intégrées, puisqu'elles interviennent de façon dominantes dans toutes les classes. Essayons de comprendre plus en détails quelles caractéristiques sont prépondérantes pour reconnaître les différentes activités :

- ▷ **Type de mouvements** : Les classes "COPIER" et "AUTRE" reposent essentiellement sur la connaissance des mouvements de la tête, ce qui est normal, car "COPIER" implique de passer d'un écran à un autre ou au clavier, et "AUTRE" consiste à effectuer des mouvements aléatoires de la tête. Pour les classes "LIRE" et "SURFER", le nombre de caractéristiques liées aux mouvements de la tête est plus faible (2×H sur 8). Cependant, pour "LIRE", l'une d'entre elles occupe un rôle dominant.
- ▷ **Partitionnement temporel** : "SURFER" et "VIDEO" sont très peu décrites par des caractéristiques temporelles (elles sont plutôt représentées par L0 et L2 qui privilégient une information centrée sur le présent), ce qui confirme le fait que ces activités ne sont pas vraiment répétitives. En revanche, les autres activités ont tendance à prendre en compte cette notion temporelle. "LIRE" et "ECRIRE" utilisent des caractéristiques du passé, alors que "COPIER" s'appuie pour une information future des mouvements de la tête.
- ▷ **Codage multi-échelle** et **taille des n-grammes** : "LIRE" et "ECRIRE" font appel à des statistiques n-grammes pour lesquelles la taille n est élevée, ceci est compréhensible du fait de la nature répétitive de ces activités. Étonnament, pour "ECRIRE", il s'agit des mouvements de la tête. "LIRE" repose sur une combinaison d'information locale (n3) et globale (N2) des mouvements oculaires. Les autres activités, quant à elles, utilisent des informations de plus courte durée.
- ▷ **Type de statistiques** : Nous remarquons que la taille de l'histogramme des n-grammes est une statistique qui revient fréquemment dans les poids dominants et semble occuper un rôle important pour reconnaître une activité.

Evaluation quantitative - Le tableau 5.6 montre les taux moyens de reconnaissance pour chaque activité. Le codage multi-echelle et le partionnement temporel ont permis de considérablement améliorer les résultats pour toutes les activités, exceptée la classe

TABLE 5.5 – Description des caractéristiques

Type de mouvements	E : Mouvements oculaires H : Ego-mouvement
Partitionnement temporel	L0 : Présent (global) L1 : Passé L2 : Présent (local) L3 : Futur
Codage multi-échelle	n : Codage fin N : Codage approximatif
N-gramme	taille du n-gramme ($n = 1, \ldots, 4$ ou $N = 1, 2$)
Type de statistiques	mean, maximum, variance, size, range

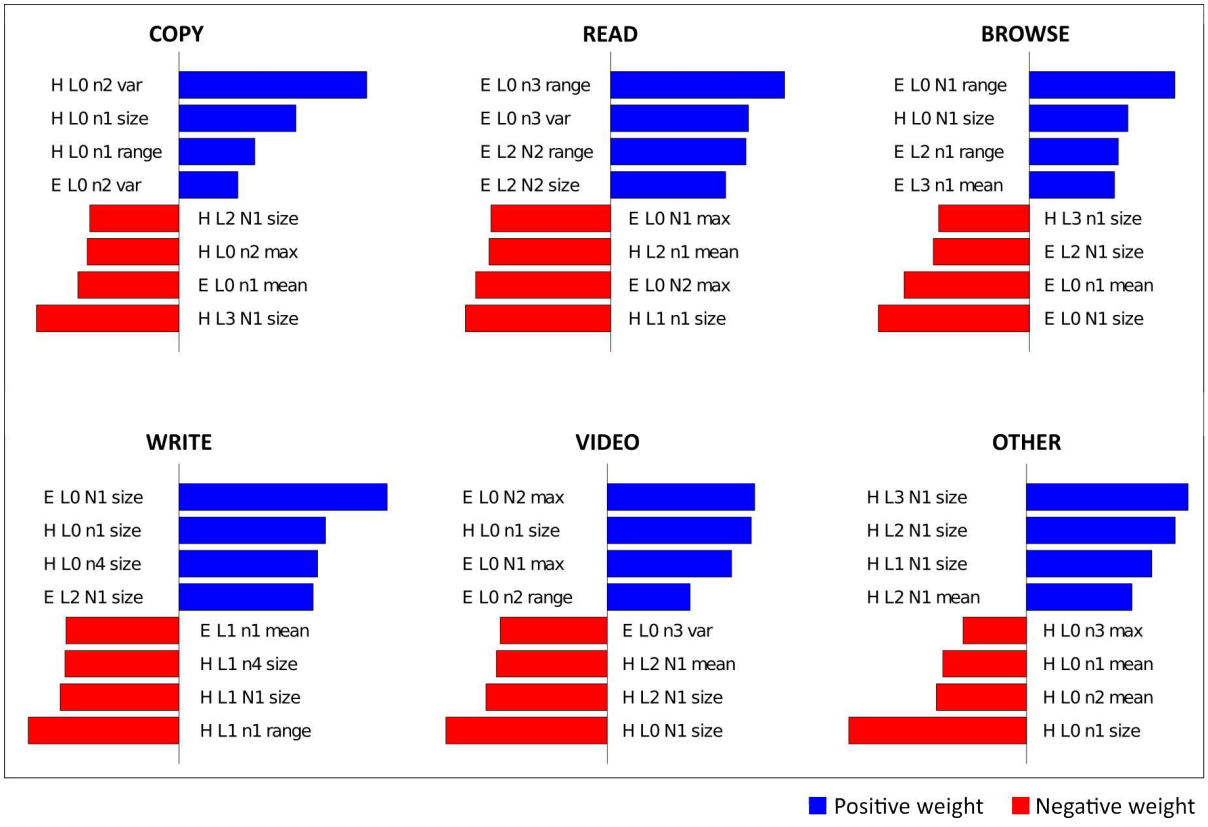


Figure 5.13 – Contribution relative des poids dominants positifs (en bleu) et négatifs (en rouge) pour chaque classe. Une description de chaque caractéristique est donnée dans le tableau 5.5.

TABLE 5.6 – Résultats inter-classe. Proposé 1 correspond à l'intégration du codage multi-échelle et du partitionnement temporel. Proposé 2 inclut l'apprentissage contextuelle en plus de Proposé 1.

Activités	**mAP**		
	Référence	Proposé 1	Proposé 2
Copier	0.7719	0.8903	0.9178
Lire	0.7111	0.8670	0.9360
Surfer	0.3076	0.4780	0.6192
Ecrire	0.8687	0.9093	0.9294
Video	0.6046	0.7222	0.8312
Autre	0.6166	0.7475	0.7593
Total	**0.6467**	**0.7691**	**0.8322**

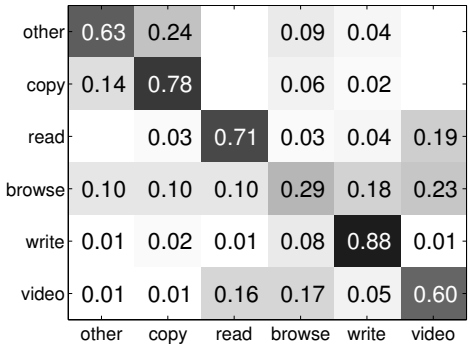
"ECRIRE" pour laquelle le gain est seulement de 2% (cependant, les performances sont, à la base, déjà très élevées). Les activités "LIRE", "SURFER" et "VIDEO" ont le plus bénéficié du modèle Auto-Contexte.

Nous avons aussi affiché les matrices de confusion pour différentes combinaisons de caractéristiques (cf. figure 5.14). Comme nous pouvons le remarquer, les activités structurées (avec mouvements répétitifs) - "COPIER", "LIRE", "ECRIRE" - sont celles qui sont les mieux reconnues. Dans notre évaluation, l'activité "SURFER" est la plus difficile à classer. Elle est souvent confondue avec la classe "VIDEO", ce qui paraît, en partie, normal vu qu'il s'agit d'activités qui nécessitent de regarder un écran d'ordinateur et elles impliquent des mouvements non-structurées (plutôt de faible amplitude) des yeux et de la tête. Nous remarquons que le codage multi-échelle est fortement utile pour minimiser le risque de confondre "LIRE" et "VIDEO", ainsi que "SURFER" et "ECRIRE". Le partitionnement temporel permet de diminuer la confusion entre la classe "SURFER" et les classes "COPIER" et "LIRE". Cela est compréhensible vu que l'activité "SURFER" est composée de phases de lecture sur écran et de nouvelle recherche via le clavier.

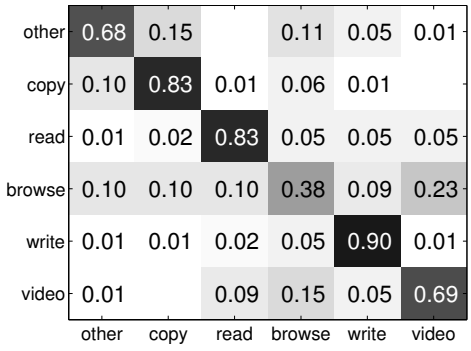
5.3.2.3 Analyse par sujet

Evaluation quantitative - La figure 5.15(a) montre le taux moyen de reconnaissance par sujet et pour différentes combinaisons de caractéristiques. Nous remarquons qu'en prenant en compte toutes les informations (multi-échelle, temporelle et contextuelle), ce taux atteint plus de 80% par sujet, ce qui n'est jamais le cas avec la référence de base. Sur cette même figure, nous confirmons aussi ce qui a été dit précédemment, à savoir que l'information multi-échelle et temporelle jouent un rôle important pour reconnaître une activité.

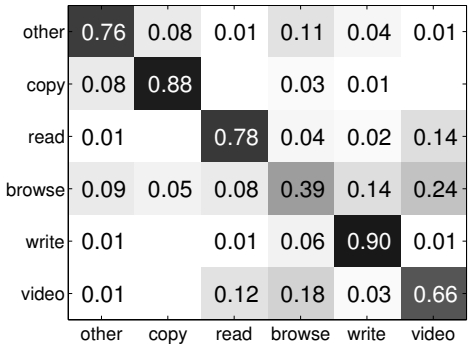
La figure 5.15(b) indique les gains obtenus entre la référence de base et la combinaison



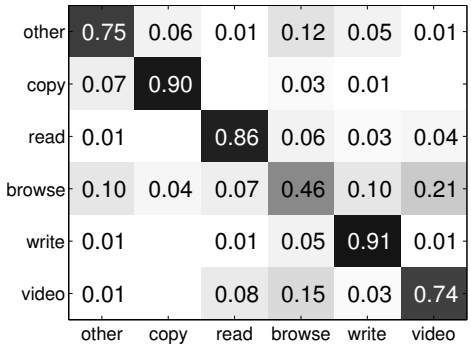
(a) E-H (Ogaki *et al.*, 2012)



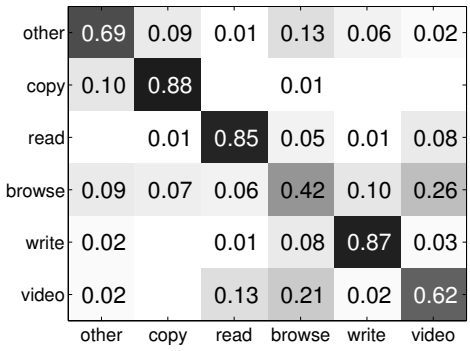
(b) E-H-L



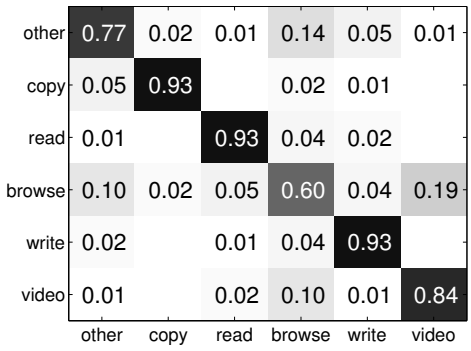
(c) E-H-T



(d) E-H-L-T

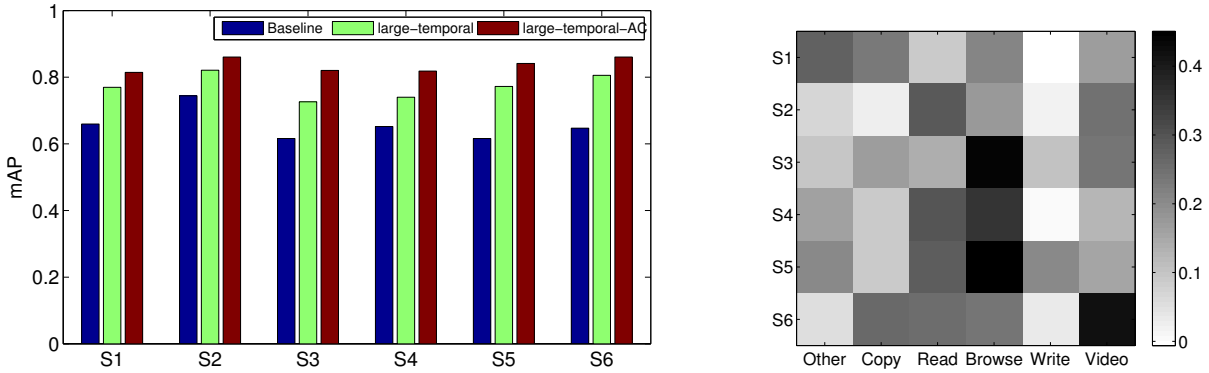


(e) E-H-AC



(f) E-H-L-T-AC

Figure 5.14 – Matrices de confusion pour la reconnaissance multi-classe (E : mouvements oculaires, H : égo-mouvement, L : information multi-échelle, T : information temporelle, AC : information contextuelle).



(a) (b)

Figure 5.15 – Analyse par sujet : (a) Précision moyenne pour chaque sujet et pour différentes combinaisons de caractéristiques, (b) Gain en classification, pour chaque activité et chaque sujet, entre la référence de base et la méthode proposée.

de toutes les informations. De plus, les résultats sont affichés pour chaque classe et chaque sujet. Les résultats indiquent que l'utlisation de l'information contextuelle dépend aussi bien de l'activité effectuée que du sujet qui la pratique.

Comme nous avons pu le constater précédemment, les activités "SURFER" et "ECRIRE" sont celles qui ont, respectivement, le plus et le moins bénéficié de la combinaison de toutes les caractéristiques. Nous remarquons aussi que les sujets 3 et 5 sont ceux qui ont connu les plus fortes améliorations pour chaque classe. Un gain de plus de 40% est obtenu par ces mêmes sujets pour l'activité "SURFER". Le sujet 6 obtient un gain élevé similaire pour l'activité "VIDEO". En revanche, les sujets 1 et 4 bénéficient d'une amélioration quasi nulle pour la classe "ECRIRE".

Evaluation qualitative - La figure 5.16 montre les résultats de la segmentation des activités pour chaque sujet et chaque vidéo. Nous remarquons, là aussi, que la classe "SURFER" est celle qui pose le plus de problème et est parfois confondue avec les classes "VIDEO", "AUTRE" et "LIRE". Sur cette figure, nous constatons que les transitions sont, en général, assez difficile à segmenter, notamment lorsqu'il s'agit de déterminer le début et la fin de l'activité). Dans notre cas, ces transitions ont lieu entre une activité prédéfinie et une activité aléatoire. Cette dernière, bien que de très courte durée (moins de 30s), est plutôt bien identifiée. Notons également que les activités dans les premières et les dernières secondes des vidéos sont, parfois, mal reconnues, ceci s'explique par le fait que la fenêtre glissante à une plus petite taille.

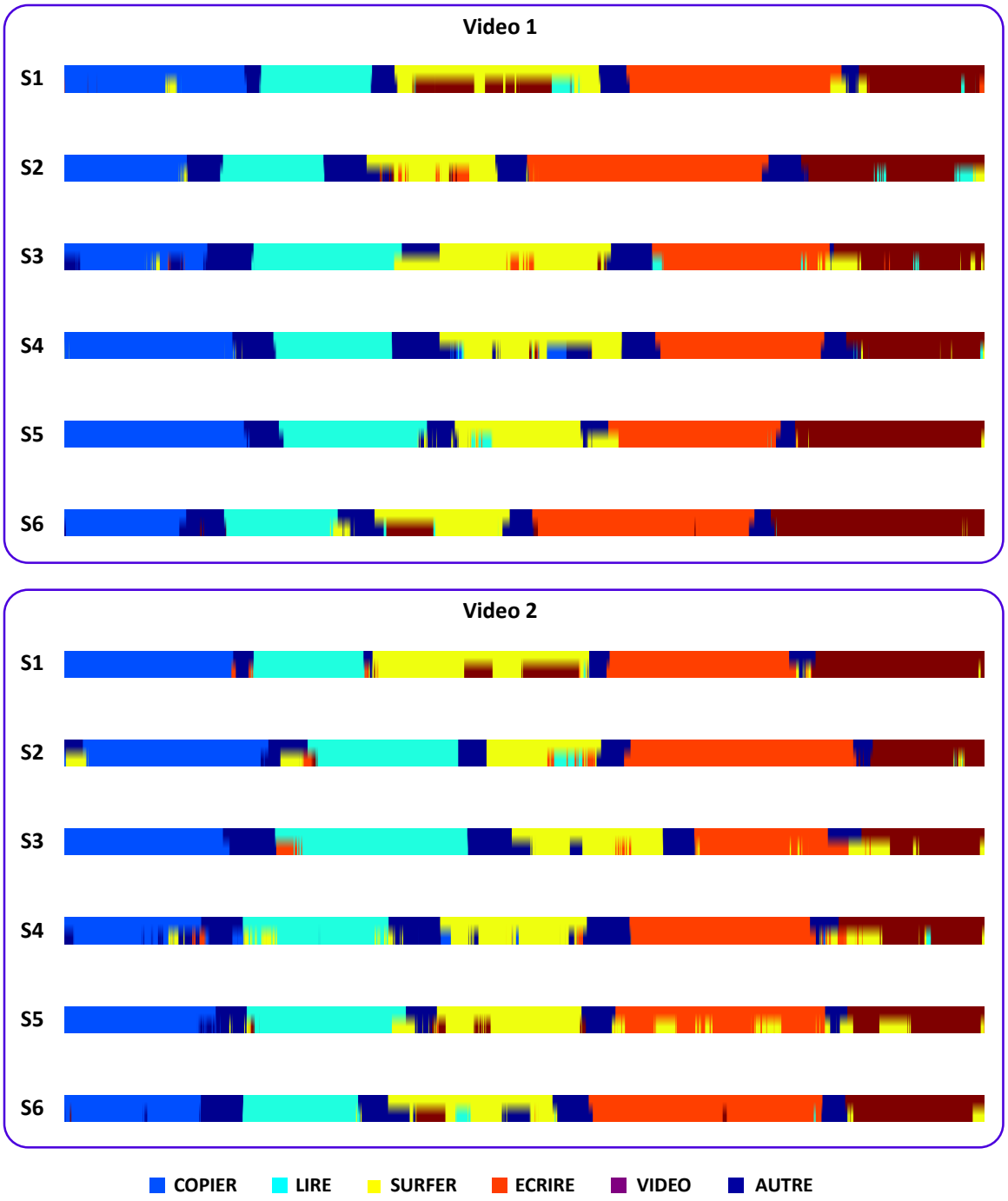


Figure 5.16 – Analyse inter-sujet : Evaluation qualitative pour E-H-L-T-AC. La première ligne de chaque séquence d'étiquetage correspond à la vérité terrain, la seconde à la prédiction obtenue. Les activités sont effectuées les unes après les autres : copier, lire, surfer, écrire et vidéo. Entre deux activités, la personne effectue une activité aléatoire de plus courte durée (mouvements aléatoires de la tête ou chanter).

5.4 Conclusion

Dans ce chapitre, nous nous sommes intéressés à la reconnaissance d'activités en vue subjective à partir des données du regard et de l'égo-mouvement.

La reconnaissance s'effectue en codant les données de mouvements continus en une séquence de symboles et en extrayant des statistiques n-grammes. Nous proposons des caractéristiques additionnelles afin de mieux représenter les activités. Dans un premier temps, nous présentons un codage multi-échelle afin de considérer une description plus globale des mouvements. Ensuite, nous appliquons un partitionnement temporel pour collecter les statistiques n-grammes. Finalement, nous procédons à un apprentissage contextuel via le modèle Auto-Contexte afin de segmenter les activités au sein des vidéos.

Une étude détaillée des résultats a permis de montrer que l'intégration de ces nouvelles caractéristiques améliore significativement le taux moyen de reconnaissance des activités.

CHAPITRE 6

Conclusion

"Le monde de la réalité a ses limites ;
le monde de l'imagination est sans frontières."

– Jean-Jacques Rousseau –

Sommaire

L'estimation et l'analyse du regard sont des outils essentiels pour la compréhension du comportement humain depuis un système de vision. Dans cette thèse, nous avons exploré trois principales pistes : (1) l'estimation du regard, (2) la reconnaissance d'attention et (3) la reconnaissance d'activités.

Ci-après, nous résumons les contributions et les conclusions présentées dans les chapitres précédents. En § 6.2, nous analysons les perpectives ouvertes par ces travaux.

6.1 Conclusions

Estimation du regard pour un eye-tracker porté - Dans le chapitre 3, nous avons présenté un système eye-tracker binoculaire porté pour estimer le regard. Ce système permet de savoir où une personne regarde dans l'environnement. Contrairement à une grande majorité des systèmes, nous employons un modèle d'apparence de l'oeil. En particulier, nous extrayons des descripteurs construits à partir de l'information du gradient de l'image. Afin d'apprendre la relation entre l'apparence des yeux et le point de regard 2D, nous avons considéré deux modèles de régression : le Support Vector Regression (SVR) et le Relevance Vector Regression (RVR). Expérimentalement, nous avons montré que ces deux modèles donnent des résultats similaires avec un léger avantage pour le RVR lorsque le nombre d'échantillons d'apprentissage est faible. Nous avons également proposé une manière simple et semi-automatique pour effectuer la calibration du système. Finalement, la méthode proposée a été appliquée sur des vidéos de test.

Reconnaissance d'attention - En reprenant les travaux du chapitre 3, nous avons proposé, dans le chapitre 4, une chaîne de traitement pour la reconnaissance d'attention en vue subjective. L'idée principale repose sur la combinaison de la connaissance du regard subjectif et du regard objectif. Le regard subjectif est obtenu à l'aide du système eye-tracker porté et de la méthode de suivi du regard développée dans le chapitre 3. Quant au regard objectif, il est calculé à partir du suivi de l'orientation de la tête et d'un modèle de couplage oeil-tête. Ensuite, nous proposons de calculer des scores d'attention à partir de ces regards et de la localisation des visages dans l'image de la caméra scène. En combinant ces scores, il est alors possible de reconnaître des motifs attentionnels dyadiques tels que le regard mutuel, mais aussi des motifs d'ordre supérieur issus de la nature triadique des

expériences. Pour l'évaluation expérimentale, nous avons procédé en deux étapes : l'évaluation de l'estimation de la pose de la tête qui a donné des résultats similaires ou supérieurs aux méthodes de l'état-de-l'art, et l'évaluation de la reconnaissance d'attention en donnant des résultats encourageants pour l'approche proposée. Ces travaux ont également permis de montrer la possibilité d'utiliser le système de suivi du regard présenté dans le chapitre 3 pour une application à la reconnaissance d'attention.

Reconnaissance d'activités - Dans le chapitre 5, nous nous sommes intéressés à la reconnaissance d'activités en vue subjective. Pour cela, nous avons repris les travaux (Bulling *et al.*, 2011; Ogaki *et al.*, 2012) sur le codage symbolique de mouvements oculaires et d'égo-mouvements. Nous avons proposé un codage multi-échelle et un partitionnement temporel pour capturer différentes propriétés d'une activité. En plus de ces caractéristiques, nous avons proposé une classification par apprentissage contextuel afin de prendre en compte des scores de prédiction d'un voisinage à longue portée. Une évaluation expérimentale sur une base de données collectées à l'aide d'un système eye-tracker porté commercial a permis de montrer l'importance des informations multi-échelle, temporelle et contextuelle, et que la méthode proposée donne des performances supérieures en comparaison avec l'état-de-l'art.

6.2 Limitations et directions futures

Il existe de nombreuses manières d'améliorer les approches proposées dans ce manuscrit. De plus, il est également possible de réutiliser certaines solutions proposées pour construire de nouvelles approches.

Dans cette section, nous abordons certaines limitations des solutions proposées, ainsi que quelques voies de recherche potentielles à explorer.

Calibration adaptée - Dans le chapitre 3, nous avons effectué une calibration "traditionnelle" en construisant un modèle de régression via un apprentissage supervisé. Il serait intéressant d'explorer d'autres types de calibration permettant d'alléger ou d'éviter la procédure de calibration. Dans la littérature, différentes méthodes ont été proposées et ont été mentionnées en § 2.3.5. Parmi celles-ci, la calibration *adaptée* est un moyen de bénéficier des données d'une autre personne afin de calibrer le système pour un nouveau sujet (Fehringer *et al.*, 2012). Cependant, elle a été uniquement explorée dans le cas d'un eye-tracker porté en éclairage infrarouge, ce qui sous-entend qu'elle a été développée pour un modèle orienté caractéristique. Une piste envisageable serait d'appliquer cette idée avec un modèle d'apparence. Néanmoins, la même méthode ne peut être employée en raison des modèles d'oeil différents. Une solution serait, par exemple, d'adapter le modèle de régression d'un sujet à un autre, notamment en contraignant le nouveau modèle à rester proche du modèle de référence. Cette idée est inspirée des approches par *domain adaptation* qui ont essen-

tiellement été explorées en classification.

Suivi joint de la position et de la pose de la tête - Dans le chapitre 3, nous avons présenté un suivi de la pose de la tête pour pouvoir calculer les scores d'attention objective. Ce suivi est effectué en deux étapes séquentielles : le suivi du visage et l'estimation de la pose. Il serait intéressant de robustifier ce suivi pour diminuer le bruit de la localisation du visage, mais aussi de l'estimation de la pose de la tête. Une solution serait d'essayer de voir la faisabilité d'utiliser le concept de régression localisée dans le cadre du suivi joint de la position et de l'orientation de la tête (Ba et Odobez, 2009). Des idées pourraient également être trouvées du côté du filtrage dynamique de la pose décrit dans (Fergie et Galata, 2012).

Alternative à l'apprentissage en cascade - Pour l'étiquetage structuré tel que l'apprentissage contextuel exploité dans le chapitre 5, nombreuses sont les méthodes qui apprennent des classifieurs en cascade. Il serait intéressant de construire des alternatives aux structures en cascade employées pour l'apprentissage, notamment pour réduire le temps d'apprentissage. Très récemment, une stratégie a été présentée dans (Li *et al.*, 2013). Ce modèle est, conceptuellement, très similaire au modèle Auto-Contexte que nous avons utilisé (cf. § 5.2.3.2). En revanche, les auteurs proposent d'apprendre une seule fonction de prédiction (au lieu de T classifieurs pour le modèle Auto-Contexte). Ce classifieur est appris avec les valeurs $\mathbf{q}$ répliquées et perturbées ($\bar{\mathbf{q}} = \mathbf{q} + \delta\mathbf{q}$) afin, lors de la phase de test, de pouvoir être également efficace aux étages supérieurs de la cascade. Dans le cadre de nos expériences, nous avons implémenté et testé ce modèle, mais sans réel succès.

Reconnaissance précoce d'activités (ou *early activity recognition*) - Dans le chapitre 5, nous nous sommes intéressés à la reconnaissance d'activités dont le but est d'identifier, pour chaque image, l'activité effectuée à partir d'informations passées et futures. Une autre approche, qui n'a pas encore été explorée dans le contexte du suivi du regard, consiste à reconnaître les activités en cours d'exécution dans un délai le plus court possible, comme c'est le cas pour (Ryoo, 2011). L'intérêt d'un tel concept réside, ensuite, dans la possibilité de choisir rapidement des actions adaptées en réponse aux activités identifiées.

Activités journalières - A plus long terme, il serait intéressant de construire un dispositif complet de reconnaissance d'activités journalières, notamment en combinant différentes solutions. Par exemple, nous pourrions imaginer de combiner la reconnaissance d'activités avec la reconnaisance d'attention : l'un serait le "déclencheur" de l'autre ou bien ils fonctionneraient en parallèle. Divers sous-systèmes pourraient alors être proposés.

Cependant, le problème du suivi du regard dans le cadre d'activités journalières n'est toujours pas résolu à ce jour. Bulling *et al.* (2013) ont néanmoins proposé une alternative, notamment en exploitant des signaux issus de l'électro-oculographie (EOG).

Biométrie dynamique via le regard - Une application qui serait également intéressante

à étudier est la biométrie dynamique basée sur l'information du regard (ou les mouvements oculaires). En effet, des études en sciences cognitives (Risko *et al.*, 2012) suggèrent que les différences individuelles de personnalité peuvent être capturées à l'aide des mouvements oculaires durant l'exploration visuelle d'une scène. Ainsi, des algorithmes de reconnaissance de l'identité personnelle : profil démographique (sexe, âge, culture ...) ou bien encore du type de personnalité pourraient être proposés en exploitant des caractéristiques extraites via des trajectoires oculaires ou des cartes de fixation.

Données multimodales - Finalement, dans ce manuscrit, nous avons seulement exploité l'information du regard. D'autres sources d'information peuvent être envisagées et peuvent venir en complément du regard pour améliorer ou créer de nouvelles applications. L'eye-tracker porté pourrait, par exemple, être combiné avec d'autres capteurs portables tels qu'un accéléromètre (pour, par exemple, capturer les mouvements des bras) ou encore un électroencéphalogramme (EEG pour mesurer l'activité cérébrale). En ce qui concerne la vision subjective, une information sur la localisation de la personne obtenue via un GPS peut également être utile, par exemple, pour reconnaître des lieux ou des objets rencontrés à des endroits où la personne est passée.

Enfin, dans ce manuscrit, nous nous sommes uniquement intéressés au traitement et à l'analyse en vue subjective. Une autre voie ensisageable et novatrice serait de construire un système multi-caméra, notamment en combinant la vue subjective avec la vue externe (via des caméras fixes). La vue subjective permettrait, par exemple, de connaître plus précisément les zones d'intérêt d'une personne ou les personnes rencontrées, tandis que la vue externe permettrait d'avoir une information plus gobale sur l'environnement. De nouvelles applications pourraient alors voir le jour, notamment dans le domaine des technologies assistives ou de la sécurité.

ANNEXE A

Localized Multiple Kernel Learning : Formulation primale et duale

En appliquant le principe de *Localized Multiple Kernel Learning* (Gönen et Alpaydin, 2008) pour la régression, la fonction de décision peut se réécrire :

$$f_{\mathcal{R}}(\mathbf{x}) = \sum_{m=1}^{P} \eta_m(\mathbf{x}|\mathbf{V}) \langle \mathbf{w}_m, \mathbf{\Phi}_m(\mathbf{x}) \rangle + b \tag{A.1}$$

Il est possible d'en déduire les formulations primale et duale.

A.1 Formulation primale

A partir de (A.1) et par régularisation des poids $\mathbf{w}_m$, la forme primale peut s'écrire :

$$\begin{array}{ll} \min & \frac{1}{2} \sum_{m=1}^{P} ||\mathbf{w}_m||_2^2 + C \sum_{i=1}^{n} \left(\xi_i^+ + \xi_i^-\right) \\ \text{w. r. t.} & \mathbf{w}_m, \xi^+, \xi^-, \mathbf{V}, b \\ \text{s. t.} & \epsilon + \xi_i^+ \geq y_i - f_{\mathcal{R}}(\mathbf{x}_i) \quad \forall i \\ & \epsilon + \xi_i^- \geq f_{\mathcal{R}}(\mathbf{x}_i) - y_i \quad \forall i \\ & \xi_i^+ \geq 0 \quad \forall i \\ & \xi_i^- \geq 0 \quad \forall i \end{array} \tag{A.2}$$

où ξ^+, ξ^- sont les variables ressort (*slack variables* en anglais). C et ϵ sont, respectivement, le coût d'erreur et la largeur du tube d'insensibilité.

A.2 Formulation duale

La formulation duale, quant à elle, s'écrit de la façon suivante :

$$\begin{array}{lll} \max & J(\mathbf{V}) = & \sum_{i=1}^{n} y_i(\alpha_i^+ - \alpha_i^-) - \epsilon \sum_{i=1}^{n}(\alpha_i^+ - \alpha_i^-) \\ & & -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-)k_\eta(\mathbf{x}_i, \mathbf{x}_j) \\ \text{w. r. t.} & & \alpha^+, \alpha^- \\ \text{s. t.} & & \sum_{i=1}^{n}(\alpha_i^+ - \alpha_i^-) = 0 \\ & & C \geq \alpha_i^+, \alpha_i^- \geq 0 \quad \forall i \end{array} \tag{A.3}$$

avec $k_\eta(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{P} \eta_m(\mathbf{x}_i|\mathbf{V})k(\mathbf{x}_i, \mathbf{x_i})\eta_m(\mathbf{x}_j|\mathbf{V})$.
La fonction de décision peut alors se réécrire sous la forme :

$$f_{\mathcal{R}}(\mathbf{x}) = \sum_{i=1}^{n}(\alpha_i^+ - \alpha_i^-)k_\eta(\mathbf{x}_i, \mathbf{x}) + b \tag{A.4}$$

Bibliographie

Aggarwal, J. et Ryoo, M. (2011). Human activity analysis : A review. *ACM Computing Surveys*, 43(16):16–33. 82.

Aghajanian, J. et Prince, S. (2009). Face pose estimation in uncontrolled environments. *In British Machine Vision Conference (BMVC)*, pp. 1–11. 32.

Ahlberg, J. (2001). Candide-3 – an updated parameterized face. *Report No. LiTH-ISY-R-2326*. 27 et 28.

Al Haj, M., Gonzàlez, J. et Davis, L. S. (2012). On partial least squares in head pose estimation : How to simultaneously deal with misalignment. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2602–2609. 32, 71 et 72.

Argyle, M. (1969). Social interaction. *Travistock Publications*. 3.

Argyle, M., Ingham, R., Alkema, F. et McCallin, M. (1973). The different functions of gaze. *In Semiotica*. 3 et 59.

Ba, S. O. et Odobez, J.-M. (2009). Recognizing visual focus of attention from head pose in natural meetings. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 39:16–33. 34, 64 et 116.

Babcock, J. S. et Pelz, J. B. (2004). Building a lightweight eyetracking headgear. *In ACM Symposium on Eye Tracking Research and Applications (ETRA)*, pp. 109–114. 16.

Bach, F. R., Lanckriet, G. R. G. et Jordan, M. I. (2004). Multiple kernel learning, conic duality, and the smo algorithm. *In International Conference on Machine Learning (ICML)*. 61 et 62.

Balasubramanian, V., Jieping, Y. et Panchanathan, S. (2007). Biased manifold embedding : A framework for person-independent head pose estimation. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–7. 32 et 33.

Baluja, S. et Pomerleau, D. (1994). Non-intrusive gaze tracking using artificial neural networks. *In Advances in Neural Information Processing Systems (NIPS)*. 18 et 20.

Basu, S., Essa, I. et Pentland, A. (2007). Motion regularization for model-based head tracking. *In International Conference on Pattern Recognition (ICPR)*, vol. 3, pp. 611–616. 27.

Bazzani, L., Cristani, M., Tosato, D., Farenzena, M., Paggetti, G., Menegaz, G. et Murino, V. (2011). Social interactions by visual focus of attention in a three-dimensional environment. *In Expert Systems*. 34.

Belkin, M. et Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396. 33.

BenAbdelkader, C. (2010). Robust head pose estimation using supervised manifold learning. *In European Conference on Computer Vision (ECCV)*, pp. 518–531. 33 et 67.

Bernet, S., Sturm, P., Cudel, C. et Basset, M. (2011). Study on the interest of hybrid fundamental matrix for head mounted eye tracker modeling. *In British Machine Vision Conference (BMVC)*, pp. 1–10. 23.

Beymer, D. (1994). Face recognition under varying pose. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 756–761. 31.

Black, M. J. et Yacoob, Y. (1995). Tracking and recognizing rigid facial motions using local parametric models of image motion. *In International Conference on Computer Vision (ICCV)*, pp. 374–381. 27.

Blanz, V., Grother, P., Phillips, P. J. et Vetter, T. (2005). Face recognition based on frontal views generated from non-frontal images. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 454–461. 25.

Blanz, V. et Vetter, T. (1999). A morphable model for the synthesis of 3d faces. *In SIGGRAPH*, pp. 187–194. 27 et 30.

Bulling, A., Ward, J., Gellersen, H. et Troster, G. (2011). Eye movement analysis for activity recognition using electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33:741–751. 37, 84, 86, 87, 88, 96, 98 et 115.

Bulling, A., Weichel, C. et Gellersen, H. (2013). Eyecontext : Recognition of high-level contextual cues from human visual behaviour. *In ACM SIGCHI International Conference on Human Factors in Computing Systems (CHI)*. 37 et 116.

Carpenter, R. H. S. (1988). *Movements of the Eyes*. Pion Ltd, 2nd edition. 12 et 13.

Cascia, M. L., Sclaroff, S. et Athitsos, V. (2000). Fast, reliable head tracking under varying illumination : An approach based on registration of texture-mapped 3d models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22:322–336. 27 et 28.

Chen, J. et Ji, Q. (2011). Probabilistic gaze estimation without active personal calibration. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 609–616. 24.

Choi, D. H., Jang, I. H. an Kim, M. H. et Kim, N. C. (2007). Color image enhancement based on single-scale retinex with a jnd-based nonlinear filter. *In IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 3948–395. 18.

Choi, K. N., Carcassoni, M. et Hancock, E. R. (1998). Estimating 3d facial pose using the em algorithm. *In British Machine Vision Conference (BMVC)*, pp. 1–10. 30.

Choi, S. et Kim, D. (2008). Robust head tracking using 3d ellipsoidal head model in particle filter. *Pattern Recognition (PR)*, 41:2901–2915. 27.

Colombo, C. et A.D. Bimbo, A. D. (1999). Real-time head tracking from the deformation of eye contours using a piecewise affine camera. *Pattern Recognition Letters (PRL)*, 20:721–730. 18.

Cooley, C. H. (1902). Human nature and the social order. *New York : Scribner's*. 59.

Cootes, T. F., Edwards, G. J. et Taylor, C. J. (1998). Active appearance models. *In European Conference on Computer Vision (ECCV)*, vol. 2, pp. 484–498. 28 et 29.

Cootes, T. F., Taylor, C. J., Cooper, D. H. et Graham, J. (1995). Active shape models - their training and application. *Computer Vision and Image Understanding (CVIU)*, 61:38–59. 28.

Cootes, T. F., Wheeler, G. V., Walker, K. N. et Taylor, C. J. (2002). View-based active appearance models. *Image and Vision Computing (IVC)*, 20:657–664. 28, 29 et 30.

Courtemanche, F., Admeur, E., Dufresne, A., Najjar, M. et Mpondo, F. (2011). Activity recognition using eye-gaze movements and traditional interactions. *Interacting with Computers*, 23:202–213. 36.

Coutinho, F. L. et Morimoto, C. H. (2010). A depth compensation method for cross-ratio based eye tracking. *In ACM Symposium on Eye Tracking Research and Applications (ETRA)*, pp. 137–140. 22.

Coutinho, F. L. et Morimoto, C. H. (2012). Augmenting the robustness of cross-ratio gaze tracking methods to head movement. *In ACM Symposium on Eye Tracking Research and Applications (ETRA)*, pp. 59–66. 22.

Crammer, K. et Singer, Y. (2001). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research (JMLR)*, 2:265–292. 93.

Cristani, M., Bazzani, L., Paggetti, G., Fossati, F., Tosato, D., Del Bue, A., Menegaz, G. et Murino, V. (2011). Social interaction discovery by statistical analysis of f-formations. *In British Machine Vision Conference (BMVC)*, pp. 1–12. 34.

Cristinacce, D. et Cootes, T. F. (2006). Feature detection and tracking with constrained local models. *In British Machine Vision Conference (BMVC)*, pp. 929–938. 29 et 30.

Dalal, N. et Triggs, B. (2005). Histograms of oriented gradients for human detection. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 886–893. 43 et 70.

Davison, A. J., Reid, I. D., Molton, N. et Stasse, O. (2007). Monoslam : Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29:1052–1067. 23.

DeMenthon, D. F. et Davis, L. S. (1995). Model-based object pose in 25 lines of code. *International Journal of Computer Vision (IJCV)*, 15:123–141. 30.

Doshi, A. et Trivedi, M. M. (2009). On the roles of eye gaze and head dynamics in predicting driver's intent to change lanes. *IEEE Transactions on Intelligent Transportation Systems (ITS)*, 10:453–462. 36.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J. et Vapnik, V. N. (1996). Support vector regression machines. *In Advances in Neural Information Processing Systems (NIPS)*, pp. 155–161. 20, 30 et 44.

Duda, R., Hart, P. et Stork, D. (2001). Pattern classification. *John Wiley & Sons*. 31.

Ebisawa, Y. et Satoh, S.-I. (1993). Effectiveness of pupil area detection technique using two light sources and image difference method. *IEEE International Conference on Engineering in Medicine and Biology Society (EMBS)*, pp. 1268–1269. 16.

Edelman, S. (1999). Representation and recognition in vision. *MIT Press*. 88.

Emery, N. J. (2000). The eyes have it : the neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews*. 25.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. et Lin, C.-J. (2008). Liblinear : A library for large linear classification. *Journal of Machine Learning Research (JMLR)*, 9:1871–1874. 98.

Fanelli, G., Gall, J. et van Gool, L. (2012). Real time 3d head pose estimation : Recent achievements and future challenges. *In International Symposium on Communications, Control and Signal Processing (ISCCSP)*. 25.

Fathi, A., Farhadi, A. et Rehg, J. M. (2011). Understanding egocentric activities. *In International Conference on Computer Vision (ICCV)*, pp. 407–414. 36.

Fathi, A., Hodgins, J. K. et Rehg, J. M. (2012a). Social interactions : A first-person perspective. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1226–1233. 34 et 66.

Fathi, A., Li, Y. et Rehg, J. M. (2012b). Learning to recognize daily actions using gaze. *In European Conference on Computer Vision (ECCV)*, vol. 1, pp. 314–327. 37.

Fehringer, B., Bulling, A. et Krüger, A. (2012). Analysing the potential of adapting head-mounted eye tracker calibration to a new user. *In ACM Symposium on Eye Tracking Research and Applications (ETRA)*, pp. 245–248. 24 et 115.

Felzenszwalb, P. et Huttenlocher, D. (2005). Pictorial structures for object recognition. *International Journal on Computer Vision (IJCV)*, 61:55–79. 30.

Fergie, M. et Galata, A. (2012). Dynamical pose filtering for mixtures of gaussian processes. *In British Machine Vision Conference (BMVC)*. 116.

Fischler, M. A. et Bolles, R. C. (1981). Random sample consensus : A paradigm for model fitting with applications to image analysis and automated cartography. *In Commun. ACM*, vol. 24, pp. 381–395. 16 et 86.

Fu, Y. et Huang, T. S. (2006). Graph embedded analysis for head pose estimation. *In International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 3–8. 32.

Funes, K. A. M. et Odobez, J.-M. (2012). Gaze estimation from multimodal kinect data. *In CVPR Workshop on Face and Gesture and Kinect demonstration competition*, pp. 25–30. 15 et 25.

Gee, A. et Cipolla, R. (1994). Determining the gaze of faces in images. *Image and Vision Computing (IVC)*, 12:639–647. 30.

Gönen, M. et Alpaydin, E. (2008). Localized multiple kernel learning. *In International Conference on Machine Learning (ICML)*, pp. 352–359. 61, 62 et 119.

Gönen, M. et Alpaydin, E. (2010). Localized multiple kernel regression. *In International Conference on Pattern Recognition (ICPR)*, pp. 1425–1428. 61.

Gönen, M. et Alpaydin, E. (2011). Multiple kernel learning algorithms. *Journal of Machine Learning Research (JMLR)*, 12:2211–2268. 61.

Gong, S. McKenna, S. et Collins, J. J. (1996). An investigation into face pose distributions. *In International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 265–270. 32.

Gourier, N., Hall, D. et Crowley, J. L. (2004). Estimating face orientation from robust detection of salient facial features. *In Proc. of Pointing 2004, ICPR, International Workshop on Visual Observation of Deictic Gestures*. 67 et 68.

Gross, R., Matthews, I. et Baker, S. (2005). Generic vs. person specific active appearance models. *Image and Vision Computing (IVC)*, 23:1080–1093. 29.

Gross, R., Matthews, I. et Baker, S. (2006). Active appearance models with occlusion. *Image and Vision Computing (IVC)*, 24:593–604. 29.

Guestrin, E. D. et Eizenman, M. (2006). General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transaction on Biomedical Engineering*, 53:1124–1133. 22.

Hansen, D. W. et Ji, Q. (2010). In the eye of the beholder : A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32:478–500. 3, 14 et 21.

Hansen, D. W. et Pece, A. E. C. (2005). Eye tracking in the wild. *Computer Vision and Image Understanding (CVIU)*, 98:155–181. 18.

Hansen, D. W., San Agustin, J. et Villanueva, A. (2010). Homography normalization for robust gaze estimation in uncalibrated setups. *In ACM Symposium on Eye Tracking Research and Applications (ETRA)*, pp. 13–20. 23.

Hartley, R. et Zisserman, A. (2006). Multiple view geometry in computer vision. *Cambridge University Press (2nd ed.)*. 23.

Hennessey, C. et Lawrence, P. (2008). 3-d point-of-gaze estimation in a volumetric display. *In ACM Symposium on Eye Tracking Research and Applications (ETRA)*, p. 59. 15 et 23.

Hennessey, C., Noureddin, B. et Lawrence, P. (2006). A single camera eye-gaze tracking system with free head motion. *In ACM Symposium on Eye Tracking Research and Applications (ETRA)*. 22.

Ho, H. T. et Chellappa, R. (2012). Automatic head pose estimation using randomly projected dense sift descriptors. *In International Conference on Image Processing (ICIP)*, pp. 1961–1964. 62.

Hu, Y., Chen, L., Zhou, Y. et Zhang, H. (2004). Estimating face pose by facial asymmetry and geometry. *In International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 651–656. 30.

Huang, D., Storer, M., De la Torre, F. et Bischof, H. (2011). Supervised local subspace learning for continuous head pose estimation. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2928. 32 et 33.

Huang, J., Shao, X. et Wechsler, H. (1998). Face pose discrimination using support vector machines. *In International Conference on Pattern Recognition (ICPR)*, pp. 154–156. 31.

Huynh, T., Fritz, M. et Schiele, B. (2008). Discovery of activity patterns using topic models. *In International Conference on Ubiquitous Computing (UbiComp)*, pp. 10–19. 36.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J. et Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computing*, 3(1):79–87. 61.

Ji, Q., H., W., Duchowski, A. T. et Flickner, M. (2005). Special issue : eye detection and tracking. *Computer Vision and Image Understanding (CVIU)*, 98. 14.

Ji, Q. et Zhu, Z. (2002). Eye and gaze tracking for interactive graphic display. *In International Symposium on Smart Graphics*, pp. 79–85. 20.

Joachims, T. (2006). Training linear svms in linear time. *In ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 217–226. 34.

Jones, M. J. et Viola, P. A. (2003). Fast multi-view face detection. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 31.

Kitani, K. M., Okabe, T., Sato, Y. et Sugimoto, A. (2011). Fast unsupervised ego-action learning for first-person sports videos. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3241–3248. 36 et 84.

Kohlbecher, S., Bardinst, S., Bartl, K., Schneider, E., Poitschke, T. et Ablassmeier, M. (2008). Calibration-free eye tracking by reconstruction of the pupil ellipse in 3d space. *In ACM Symposium on Eye Tracking Research and Applications (ETRA)*, pp. 135–138. 24.

Kumano, S., Otsuka, K., Yamato, J., Maeda, E. et Sato, Y. (2009). Pose-invariant facial expression recognition using variable-intensity templates. *International Journal of Computer Vision (IJCV)*, 83: 178–194. 25.

Lafferty, J. D., McCallum, A. et Pereira, F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. *In International Conference on Machine Learning (ICML)*, pp. 282–289. 93.

Lan, T., Sigal, L. et Mori, G. (2012). Social roles in hierarchical models for human activity recognition. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1354–1361. 34.

Langton, S. R. H., Honeyman, H. et Tessler, E. (2004). The influence of head contour and nose angle on the perception of eyegaze direction. *Perception and Psychophysics*, 66:752–771. 25.

Leslie, C., Eskin, E. et Noble, W. S. (2002). The spectrum kernel : A string kernel for svm protein classification. *In Pacific Symposium on Biocomputing*, pp. 566–575. 37.

Li, D. et Parkhurst, D. (2006). Open-source software for real-time visible-spectrum eye tracking. *In Proceedings of the COGAIN Conference*. 17.

Li, D., Winfield, D. et Parkhurst, D. (2005). Starburst : A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. *In IEEE Conference on Computer Vision and Pattern Recognition - Workshop (CVPR-W)*. 16 et 17.

Li, Q., Wang, J., Wipf, D. et Tu, Z. (2013). Fixed-point model for structured labeling. *In International Conference on Machine Learning (ICML)*. 116.

Li, S., Fu, Q., Gu, L., Scholkopf, B., Cheng, Y. et Zhang, H. (2001). Kernel machine based learning for multi-view face detection and pose estimation. *In International Conference on Computer Vision (ICCV)*, pp. 674–679. 31.

Li, Y., Gong, S., Sherrah, J. et Liddell, H. (2004). Support vector machine based multi-view face detection and recognition. *Image and Vision Computing (IVC)*, 22:752–771. 32.

Lin, C.-J., Weng, R. C. et Keerthi, S. S. (2008). Trust region newton method for large-scale logistic regression. *Journal of Machine Learning Research (JMLR)*, 9:627–650. 98.

Little, G., Krishna, S., Black, J. et Panchanathan, S. (2005). A methodology for evaluating robustness of face recognition algorithms with respect to changes in pose and illumination angle. *In International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2. 26, 67 et 68.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60:91–110. 43.

Lu, F., Okabe, T., Sugano, Y. et Sato, Y. (2011a). A head pose-free approach for appearance-based gaze estimation. *In British Machine Vision Conference (BMVC)*, pp. 1–11. 25.

Lu, F., Okabe, T., Sugano, Y. et Sato, Y. (2011b). Inferring human gaze from appearance via adaptive linear regression. *In International Conference on Computer Vision (ICCV)*, pp. 153–160. 18 et 21.

Lu, F., Sugano, Y., Okabe, T. et Sato, Y. (2012). Head pose-free appearance-based gaze sensing via eye image synthesis. *In International Conference on Pattern Recognition (ICPR)*, pp. 1008–1011. 24.

Ma, Y., Konishi, Y., Kinoshita, K., Lao, S. et Kawade, M. (2006). Sparse bayesian regression for head pose estimation. *In International Conference on Pattern Recognition (ICPR)*, pp. 507–510. 30.

Marin-Jimenez, M. J., Zisserman, A. et Ferrari, V. (2011). "here's looking at you, kid". detecting people looking at each other in videos. *In British Machine Vision Conference (BMVC)*, pp. 1–12. 34 et 71.

Martinez, F., Carbone, A. et Pissaloux, E. (2012). Gaze estimation using local features and non-linear regression. *In International Conference on Image Processing (ICIP)*, pp. 1961–1964. 6 et 40.

Martinez, F., Carbone, A. et Pissaloux, E. (2013). Combining first-person and third-person gaze for attention recognition. *In International Conference on Automatic Face and Gesture Recognition (FG)*. 6 et 58.

Martins, P. et Batista, J. (2008). Accurate single view model-based head pose estimation. *In International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–6. 30.

Matthews, I. et Baker, S. (2004). Active appearance models revisited. *International Journal of Computer Vision (IJCV)*, 60:135–164. 29.

Mayol, W. W. et Murray, D. W. (2005). Wearable hand activity recognition for event summarization. *In International Symposium on Wearable Computers (ISWC)*, pp. 122–129. 36.

Merchant, J., Morrissette, R. et Porterfield, J. (1974). Remote measurements of eye direction allowing subject motion over one cubic foot of space. *IEEE Transactions on Biomedical Engeneering*, 21:309–317. 19.

Model, D., Guestrin, E. D. et Eizenman, M. (2009). An automatic calibration procedure for remote eye-gaze tracking systems. *In IEEE International Conference on Engineering in Medicine and Biology Society (EMBS)*, pp. 4751–4754. 24.

Moon, H. et Miller, M. (2004). Estimating facial pose from a sparse representation. *In International Conference on Image Processing (ICIP)*, pp. 75–78. 30.

Morency, L.-P., Whitehill, J. et Movellan, J. (2008). Generalized adaptive view-based appearance model : Integrated framework for monocular head pose estimation. *In International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–8. 27.

Morimoto, C. H., Amir, A. et Flickner, M. (2002). Detecting eye position and gaze from a single camera and 2 light sources. *In International Conference on Pattern Recognition (ICPR)*, pp. 314–317. 22.

Morimoto, C. H., Koons, D., Amir, A. et Flickner, M. (2000). Pupil detection and tracking using multiple light sources. 18:331–335. 19.

Mulvey, F., Villanueva, A., Sliney, D., Lange, R., Cotmore, S. et Donegan, M. (2008). D5.4 exploration of safety issues in eyetracking. *In Communication by Gaze Inter action (COGAIN), IST-2003-511598 : Deliverable 5.4*. 16.

Munn, S. M. et Pelz, J. B. (2008). 3d point-of-regard, position and head orientation from a portable monocular video-based eye tracker. *In ACM Symposium on Eye Tracking Research and Applications (ETRA)*, pp. 181–188. 23.

Murphy-Chutorian, E. andTrived, M. (2009). Head pose estimation in computer vision : A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 31:607–626. 25 et 26.

Niyogi, S. et Freeman, W. T. (1996). Example-based head tracking. *In International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 374–378. 31.

Noris, B., Benmachiche, K. et Billard, A. (2008). Calibration-free eye gaze direction detection with gaussian processes. *In International Conference on Computer Vision Theory and Applications (VISAPP)*, pp. 611–616. 24.

Noris, B., Keller, J.-B. et Billard, A. (2011). A wearable gaze tracking system for children in unconstrained environments. *Computer Vision and Image Understanding (CVIU)*, 115:476–486. 16, 17, 18 et 20.

Ogaki, K., Kitani, K. M., Sugano, Y. et Sato, Y. (2012). Coupling eye-motion and ego-motion features for first-person activity recognition. *In CVPR Workshop on Egocentric Vision (ECV)*, pp. 1–7. 37, 84, 97, 98, 100, 104, 108 et 115.

Ohno, T. et Mukawa, N. (2004). A free-head, simple calibration, gaze tracking system that enables gaze-based interaction. *In ACM Symposium on Eye Tracking Research and Applications (ETRA)*, pp. 115–122. 22.

Orozco, J., Gong, S. et Xiang, T. (2009). Head pose classification in crowded scenes. *In British Machine Vision Conference (BMVC)*, pp. 1–11. 31.

Park, H. S., Jain, E. et Sheikh, Y. (2012). 3d social saliency from head-mounted cameras. *In Advances in Neural Information Processing Systems (NIPS)*, pp. 431–439. 35.

Patron-Perez, A., Marszalek, M. Zisserman, A. et Reid, I. D. (2010). High five : Recognising human interactions in tv shows. *In British Machine Vision Conference (BMVC)*, pp. 1–11. 34 et 35.

Paysan, P., Knothe, R., Amberg, B., Romdhani, S. et Vetter, T. (2009). A 3d face model for pose and illumination invariant face recognition. *In IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS))*, pp. 296–301. 27 et 28.

Peng, H., Long, F. et Ding, C. (2005). Feature selection based on mutual information : criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27:1226–1238. 18 et 37.

Pires, B. R., Devyver, M. S., Tsukada, A. et Kanade, T. (2013). Unwrapping the eye for visible-spectrum gaze tracking on wearable device. *In Workshop on the Applications of Computer Vision (WACV)*, pp. 369–376. 17.

Pirri, F., Pizzoli, M. et Rudi, A. (2011). A general method for the point of regard estimation in 3d space. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 921–928. 23.

Platt, J. C. (1998). Sequential minimal optimization : A fast algorithm for training support vector machines. *In Microsoft Research Tech. Report, MSR-TR-98-14*, pp. 1–21. 44.

Puig, L., Guerrero, J. et Sturm, P. (2008). Matching of omnidirectional and perspective images using the hybrid fundamental matrix. *In Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras (OMNIVIS)*. 23.

Quattoni, A., Wang, S., Morency, L.-P., Collins, M. et Darrell, T. (2007). Hidden-state conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29:1848–1852. 34.

Rae, R. et Ritter, H. (1998). Recognition of human head orientation based on artificial neural networks. *IEEE Transactions on Neural Networks (NN)*, 9:257–265. 32.

Rakotomamonjy, A., Bach, F. R., Canu, S. et Grandvalet, Y. (2008). Simplemkl. *Journal of Machine Learning Research (JMLR)*, 9:2491–2521. 62.

Ramnath, K., Koterba, S., Xiao, J., Hu, C., Matthews, I., Baker, S., Cohn, J. et Kanade, T. (2008). Multi-view aam fitting and construction. *International Journal of Computer Vision (IJCV)*, 76:183–204. 29.

Ranganathan, A. et Yang, M.-H. (2008). Online sparse matrix gaussian process regression and vision applications. *In European Conference on Computer Vision (ECCV)*, pp. 468–482. 32.

Rasmussen, E. et Williams, C. K. I. (2006). Gaussian processes for machine learning. *The MIT Press*. 24 et 69.

Risko, E. F., Anderson, N. C., Lanthier, S. et Kingstone, A. (2012). Curious eyes : Individual differences in personality predict eye movement behavior in scene-viewing. *Cognition*, 122:86–90. 117.

Robertson, N. M.and Reid, I. D. (2006). Estimating gaze direction from low-resolution faces in video. *In European Conference on Computer Vision (ECCV)*, pp. 402–415. 31 et 32.

Rosipal, R. et Trejo, L. J. (2001). Kernel partial least squares regression in reproducing kernel hilbert space. *Journal of Machine Learning Research (JMLR)*, 2:97–123. 32, 70 et 71.

Ross, D., Lim, J., Lin, R. et Yang, M.-H. (2008). Incremental learning for robust visual tracking. *International Journal of Computer Vision (IJCV)*, 77(1-3):125–141. 63.

Roweis, S. et Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326. 20 et 33.

Ryan, W. J., Duchowski, A. T. et Birchfield, S. T. (2008). Limbus/pupil switching for wearable eye tracking under variable lighting conditions. *In ACM Symposium on Eye Tracking Research and Applications (ETRA)*, pp. 61–64. 17 et 20.

Ryan, W. J., Duchowski, A. T., Vincent, E. A. et Battisto, D. (2010). Match-moving for area-based analysis of eye movements in natural tasks. *In ACM Symposium on Eye Tracking Research and Applications (ETRA)*, pp. 235–242. 17.

Rydfalk, M. (1987). Candide, a parameterized face. *Report No. LiTH-ISY-I-866*. 27.

Ryoo, M. S. (2011). Human activity prediction : Early recognition of ongoing activities from streaming videos. *In International Conference on Computer Vision (ICCV)*, pp. 1036–1043. 116.

Salvucci, D. D. et Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. *In ACM Symposium on Eye Tracking Research and Applications (ETRA)*, pp. 71–78. 12 et 13.

San Agustin, J., Skovsgaard, H. H. T., Møllenbach, E., Barret, M., Tall, M., Hansen, D. W. et Hansen, J. P. (2010). Evaluation of a low-cost open-source gaze tracker. *In ACM Symposium on Eye Tracking Research and Applications (ETRA)*, pp. 77–80. 15.

Saragih, J. M., Lucey, S. et Cohn, J. F. (2011). Deformable model fitting by regularized landmark mean–shift. *International Journal of Computer Vision (IJCV)*, 91:200–215. 30.

Schiele, B., Oliver, N., Jebara, T. et Pentland, A. (1999). An interactive computer vision system dypers : Dynamic personal enhanced reality system. *Computer Vision Systems*, pp. 51–65. 36.

Shih, S. et Liu, J. (2004). A novel approach to 3-d gaze tracking using stereo cameras. *IEEE Transactions on Systems, Man and Cybernetics (SMC)*, 34:234–245. 22.

Smola, A. J. et Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14:199–222. 44.

Spriggs, E. H., De la Torre, F. et Hebert, M. (2009). Temporal segmentation and activity classification from first-person sensing. *In CVPR Workshop on Egocentric Vision*, pp. 17–24. 36.

Starner, T., Weaver, J. et Pentland, A. (1998). Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20:1371–1375. 36.

Stegmann, M. B. et Gomez, D. D. (2002). A brief introduction to statistical shape analysis. *Technical University of Denmark, Lyngby*. 28.

Stiefelhagen, R., Finke, M., Yang, J. et Waibel, A. (1999). From gaze to focus of attention. *In VISUAL*, pp. 761–768. 25.

Stiefelhagen, R., Yang, J. et Waibel, A. (1996). A model-based gaze tracking system. *In IEEE International Joint Symposia on Intelligence and Systems*, pp. 304–310. 16.

Stiefelhagen, R., Yang, J. et Waibel, A. (2002). Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, 13:928–938. 32 et 34.

Storer, M., Urschler, M., et Bischof, H. (2009). 3d-mam : 3d morphable appearance model for efficient fine head pose estimation from still images. *In ICCV Workshop on Subspace Methods*, pp. 192–199. 30.

Sugano, Y., Matsushita, Y. et Sato, Y. (2010). Calibration-free gaze sensing using saliency maps. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 24.

Sugano, Y., Matsushita, Y., Sato, Y. et Koike, H. (2008). An incremental learning method for unconstrained gaze estimation. *In European Conference on Computer Vision (ECCV)*, pp. 656–667. 24.

Swirski, L., Bulling, A. et Dodgson, N. (2012). Robust real-time pupil tracking in highly off-axis images. *In ACM Symposium on Eye Tracking Research and Applications (ETRA)*, pp. 173–176. 16.

Takemura, K., Kohashi, Y., Suenaga, T., Takamatsu, J. et Ogasawara, T. (2010). Estimating 3d point-of-regard and visualizing gaze trajectories under natural head movements. *In ACM Symposium on Eye Tracking Research and Applications (ETRA)*, pp. 157–160. 23.

Tan, K.-H., Kriegman, D. J. et Ahuja, N. (2002). Appearance-based eye gaze estimation. *In Workshop on the Applications of Computer Vision (WACV)*, pp. 191–195. 20.

Tenenbaum, J., Silva, V. et Langford, J. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323. 33.

Timm, F. et Barth, E. (2011). Accurate eye centre localisation by means of gradients. *In International Conference on Computer Vision Theory and Applications (VISAPP)*, pp. 125–130. 18.

Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research (JMLR)*, 1:211–244. 30, 44, 45, 46, 47 et 60.

Tipping, M. E. et Faul, A. C. (2003). Fast marginal likelihood maximisation for sparse bayesian models. *In Workshop on AI & Statistics*. 44 et 46.

Tsukada, A. et Kanade, T. (2012). Automatic acquisition of a 3d eye model for a wearable first-person vision device. *In ACM Symposium on Eye Tracking Research and Applications (ETRA)*, pp. 213–216. 17.

Tsukada, A., Shino, M., Devyver, M. S. et Kanade, T. (2011). Illumination-free gaze estimation method for first-person vision wearable device. *In ICCV Workshop on Computer Vision in Vehicle Technology : From Earth to Mars*, pp. 2084–2091. 16 et 17.

Tu, Z. et Bai, X. (2010). Auto-context and its application to high-level vision tasks and 3d brain image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32:1744–1757. 94 et 98.

Vacchetti, L., Lepetit, V., et Fua, P. (2004). Stable real-time 3d tracking using online and offline information. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 26:1385–1391. 27.

Valenti, R. et Gevers, T. (2012). Accurate eye center location through invariant isocentric patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34:1785–1798. 18.

Vapnik, V. N. (1995). The nature of statistical learning theory. *Springer Verlag*. 31 et 92.

Villanueva, A., Cabeza, R. et Porta, S. (2006). Eye tracking : Pupil orientation geometrical modeling. *Image and Vision Computing (IVC)*, 24:663–679. 22.

Viola, P. A. et Jones, M. J. (2001). Rapid object detection using a boosted cascade of simple features. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 511–518. 31.

Wang, J. G. et Sung, E. (2001). Gaze determination via images of irises. *Image Vision and Computing (IVC)*, 19:891–911. 18.

Wang, J. G. et Sung, E. (2007). Em enhancement of 3d head pose estimated by point at infinity. *Image Vision and Computing (IVC)*, 25:1864–1874. 30.

Weston, J. et Watkins, C. (1999). Support vector machines for multi-class pattern recognition. *In European Symposium on Artificial Neural Networks (ESANN)*, pp. 219–224. 93.

Williams, O., Blake, A. et Cipolla, R. (2006). Sparse and semi-supervised visual mapping with the s3gp. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 230–237. 18, 20 et 42.

Xiao, J., Moriyama, T., Kanade, T. et Cohn, J. F. (2003). Robust full-motion recovery of head by dynamic templates and re-registration techniques. *International Journal of Imaging Systems and Technology (IST)*, 13:85–94. 27.

Xu, L. Q., D. Machin, D. et Sheppard, P. (1998). A novel approach to real-time non-intrusive gaze finding. *In British Machine Vision Conference (BMVC)*. 18 et 20.

Yamazoe, H., Utsumi, A., Yonezawa, T. et Abe, S. (2008). Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions. *In ACM Symposium on Eye Tracking Research and Applications (ETRA)*, pp. 245–250. 23.

Yoo, D. H., Lee, B. R. et Chung, M. J. (2002). Non-contact eye gaze tracking system by mapping of corneal reflections. *In International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 94 – 99. 22.

Young, L. et Sheena, D. (1975). Methods and designs : Survey of eye movement recording methods. *Behavior Research Methods and Instruments*, 7:397–429. 19.

Yuan, G.-X., Ho, C.-H. et Lin, C.-J. (2012). Recent advances of large-scale linear classification. *In Proceedings of the IEEE*, vol. 100, pp. 2584–2603. 92 et 93.

Zhang, Y., Bulling, A. et Gellersen, H. (2012). Towards pervasive eye tracking using low-level image features. *In ACM Symposium on Eye Tracking Research and Applications (ETRA)*, pp. 261–264. 18.

Zhou, Y., Gu, L. et Zhang, H. J. (2003). Bayesian tangent shape model : Estimating shape and pose parameters via bayesian inference. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 109–116. 30.

Zhu, Z. et Ji, Q. (2007). Novel eye gaze tracking techniques under natural head movement. *IEEE Transactions on Biomedical Engineering*, 54:2246–2260. 22.

Zhu, Z., Ji, Q. et Bennett, K. P. (2006). Nonlinear eye gaze mapping function estimation via support vector regression. *In International Conference on Pattern Recognition (ICPR)*, vol. 1, pp. 1132–1135. 20.