Estimation du regard dans un environnement contrôlé

Adel Lablack



Estimation du regard dans un environnement contrôlé

THÈSE

présentée et soutenue publiquement le 03 Février 2010

pour l'obtention du

Doctorat de l'Université des Sciences et Technologies de Lille

(spécialité informatique)

par

Adel Lablack

Composition du jury

Président :	Christophe Kolski, (Professeur)	Université de Valenciennes
Rapporteurs :	Claude Chrisment, (Professeur)	Université de Toulouse III
	Alan Hanjalic, (Professeur)	Delft University of Technology
Examinateurs :	El Mustapha Mouaddib, (Professeur)	Université de Picardie
	Gia Toan Nguyen, (DR)	INRIA Rhône-Alpes
Directeur de thèse :	Chabane Djeraba, (Professeur)	Université de Lille I

UNIVERSITÉ DES SCIENCES ET TECHNOLOGIES DE LILLE
Laboratoire d'Informatique Fondamentale de Lille — UPRESA 8022
U.F.R. d'I.E.E.A. – Bât. M3 – 59655 VILLENEUVE D'ASCQ CEDEX
Tél. : +33 (0)3 28 77 85 41 – Télécopie : +33 (0)3 28 77 85 37 – email : direction@lifl.fr

Dédicaces

A mes très chers parents

A mes grands parents

A mon frère

A toute ma famille

A tous mes amis

Je dédie ce mémoire

Adel

Remerciements

Tout d'abord, je remercie vivement le professeur Chabane Djeraba de m'avoir accueilli au sein de son équipe pour réaliser ma thèse. Je le remercie pour m'avoir guidé, conseillé et soutenu tout au long de la thèse.

Je suis très reconnaissant envers les professeurs Dan Simovici, Slimane Larabi et Mark Zhang pour leurs discussions, leurs idées et leurs disponibilités. Mes remerciements vont également à Nacim Ihaddadene pour sa collaboration. Je le remercie pour son aide et son soutien pendant ces trois années.

Je tiens à remercier les professeurs Claude Chrisment et Alan Hanjalic pour avoir accepté d'être les rapporteurs de mon mémoire de thèse, et les professeurs El Mustapha Mouaddib et Gia Toan Nguyen pour avoir accepté d'être examinateurs.

Je remercie vivement Christophe Kolski, professeur à l'Université de Valenciennes, d'être le Président de mon jury de thèse. Je remercie aussi l'ensemble des membres du LIFL, doctorants, chercheurs et personnels administratifs qui ont contribué de prés ou de loin à ma thèse.

Je tiens à saluer l'ensemble de l'équipe Fox-Miire dans laquelle j'ai réalisé ma thèse. Plus particulièrement, je remercie Thierry pour les nombreuses discussions, Fred, Jean et Marius pour l'aide apportée durant la rédaction de ma thèse, ainsi que les autres doctorants et Post-doctorants Yassine, Samir et Tarek.

Enfin, je remercie aussi l'ensemble des membres du projet européen MIAUCE pour leur collaboration. Je tiens aussi à saluer Rabah Saâdane et l'ensemble des joueurs de l'équipe nationale pour la qualification au mondial.

Résumé

L'objectif principal de mon travail de thèse est l'extraction de la direction du regard (attention visuelle) d'une personne à partir de la vidéo. Cette analyse est effectuée dans un environnement composé d'une scène cible et d'une zone d'observation. La scène cible est une région d'intérêt définie pour être analysée (e.g. un écran plasma large, une image projetée sur un mur, une affiche publicitaire, un linéaire dans un magasin, ou la vitrine d'un magasin). La zone surveillée quant à elle est l'emplacement d'où les personnes regardent la scène cible (e.g. la rue, un couloir ou bien les allées d'un supermarché). Les connaissances qui sont extraites sont alors utilisées pour comprendre le comportement visuel de personnes ainsi que pour la réorganisation de la scène cible.

Pour atteindre cet objectif, nous proposons une approche basée sur l'estimation de l'orientation de la tête et la projection du champ visuel pour localiser la région d'intérêt. Nous avons utilisé une méthode d'estimation de l'orientation de la tête basée sur l'apparence globale et sur un modèle cylindrique, et une méthode de projection géométrique pour extraire les régions d'intérêts basée sur les données physiologiques de la vision humaine. L'analyse du comportement visuel des personnes a été effectuée à l'aide d'un ensemble de métriques. Les méthodes proposées ont été validées sur des données vidéo et images.

Mots clés : Vision par ordinateur, extraction d'information, direction du regard, orientation de la tête, régions d'intérêts.

Abstract

The aim of this thesis is to analyze the behaviour of the people passing in front of a target scene. We consider an environment composed of a so-called target scene (a specific scene under analysis, such as a large plasma screen, a projected image, an advertising poster, a shop window, etc.) and a monitored area (place from which people look at the target scene, such as a street or shopping mall). Computer vision provides promising techniques enabling to obtain such information by analyzing videos captured by cameras monitoring this area. Such information is useful in order to simplify technologies that uses the output of the studies about a target scene.

In this thesis, we propose an approach that estimates the visual gaze of a person in a controlled environment. The visual gaze of a person is estimated from the head pose. It is followed by its projection on the target scene that allows to estimate the approximate location of interest. Finally, an analysis of the region of interest allows an accurate explanation of the human activity and interest.

Keywords : Computer Vision, Gaze Estimation, Head Pose Estimation, Region of interest extraction.

Table des matières

Table des figures

Liste des tableaux

List of Figures

List of Tables

Chapitre 1

Summary

Sommaire

1.1 Introduction

In applications where human activity is under observation from a static camera, the knowledge about where a person is looking provides observers with important information that offers an accurate explanation of the scene activity and human interest. Some of the most common examples are the applications used to capture user attention while driving, and to control devices for disabled people. This is done by extracting the visual gaze information.

We are presenting an application of computer vision techniques to obtain specific information about the behaviour of the subjects passing in front of a target scene. This is done by analyzing videos captured by cameras monitoring an area under surveillance. The target scene could be a large plasma screen, a projected image, an advertising poster or a shop window. An example of the type of information that can be obtained is the number of people passing in the area (possibly even making a stop), those who show some interest in the target scene (i.e. looking in its direction), and the specific locations of interest inside the target scene. The person detection counts the number of persons subsequently followed by the person tracking which determines who has stopped and who is still moving. The head pose estimation denotes whether they are looking at the target scene or not. Finally the projection of the visual field extracts the location of interest in the target scene. All these tasks are improved by taking in account any environmental information.

We will apply these techniques to an environment composed of a so-called target scene (a specific scene under analysis, such as a large plasma screen, a projected image, an advertising poster, a shop window, etc.) and a monitored area (place from which people look at the target scene, such as a street or shopping mall) as seen in Figure 1.1.

Our aim is to analyze the behaviour of the people passing in front of a target scene. We are proposing an approach that analyzes the videos captured from a camera that observes a person looking at the target scene. Using head pose estimation the visual field of the user is determined and its projection on the target scene allows creating a product targeting system. The analysis of the target scene provides the information necessary to reorganize the target scene. An overview of the approach is shown in Figure 1.2.

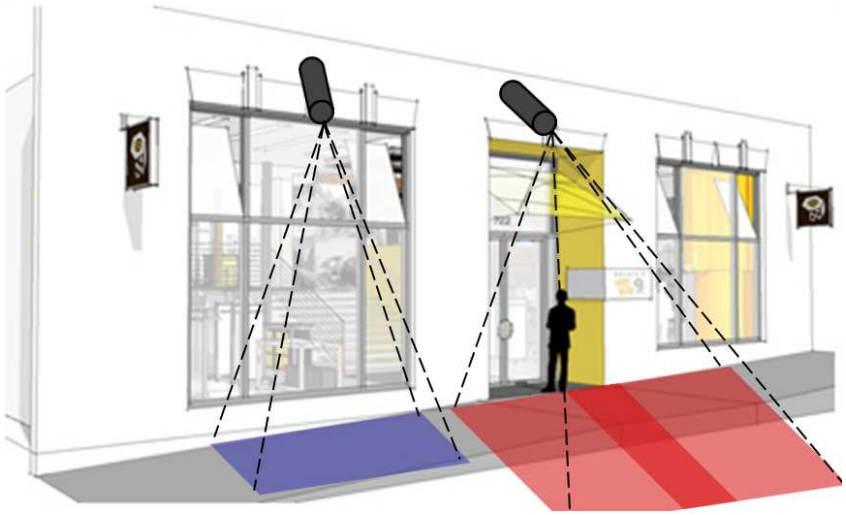


Figure 1.1: Example of an environment.

1.2 Content of the thesis

Firstly, we have outlined the history of the visual gaze estimation. We have then described the essential component of our approach which consists on head pose estimation. After that, we describe our proposed methodology for performing visual field projection (on the target scene) and extracting the regions of interest. Finally, we have described the method used to perform multi-person detection and tracking.

1.2.1 Visual gaze estimation

The analysis of gaze has been studied for over a century in several disciplines, including physiology, psychology, psychoanalysis, and cognitive sciences. The purpose is to analyze the gaze direction of persons watching a given scene, in order to obtain several kinds of information. With the recent development of low-cost gaze tracker devices, the possibility of taking advantage of the information conveyed in a subject's gaze has opened many opportunities for research, namely in image compression in which users' gaze is used to set variable compression ratios at different places in an image, in marketing for detecting products which

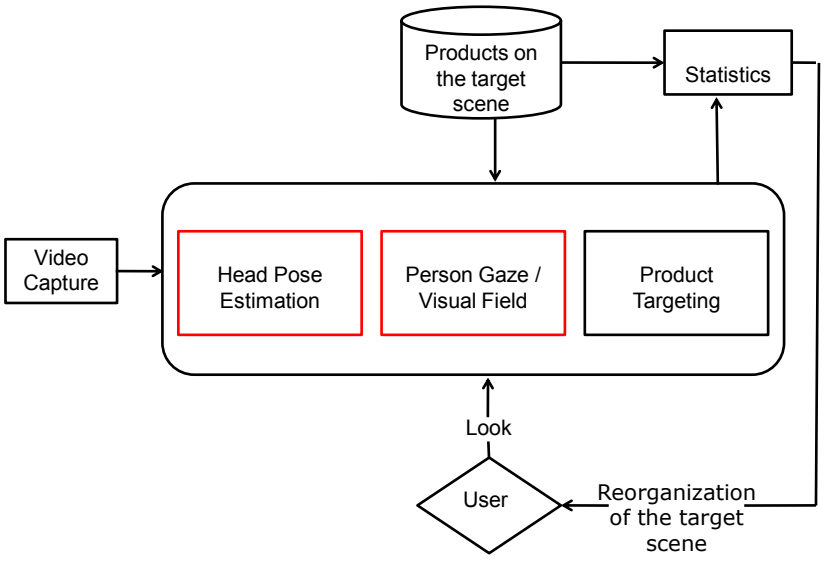


Figure 1.2: General architecture containing our approach.

attract customer interest, civil security for detecting drowsiness or lack of concentration of persons operating machinery such as motor vehicles or air traffic control systems, and in human-computer interaction. In the latter for instance, the user's gaze is used as an additional input device to traditional ones such as a mouse and a keyboard, in particular for disabled users.

In order to detect and track users' gaze, it is necessary to employ a gaze tracking device which is able to determine the fixation point of a user on a screen from the position of her/his eye and her/his head pose. Earlier gaze trackers were very intrusive. In addition to require the user to be totally static, they were in direct contact with him by sticking a reflective white dot directly onto the eye or attaching a number of electrodes around the eye. Nowadays, the most accurate gaze tracking systems consist of head mounted devices which allow detecting the direction of the gaze without having to cope with the pose of the user's head. These trackers are also intrusive ; they consist of devices composed of 3 cameras mounted on a padded headband

(2 eye cameras to allow binocular eye tracking with built-in light sources, and 1 camera to allow accurate tracking of the user's point of gaze).

Non-intrusive gaze tracking systems usually require a static camera recording the face of the user and detecting the direction of their gaze with respect to a known position. A basic gaze tracking system is composed with a static camera, a display device and software to provide an interface between them. The precision of the system can be improved in different ways, such as adding a specific light source (e.g. an infrared beam) in order to create reflections on the eye and produce more accurate tracking information. Figure 1.3 shows a single-camera gaze tracker configuration, based the Pupil-Centre/Corneal-Reflection (PCCR) method to determine the gaze direction. The video camera is located below the computer screen, and monitors the subject's eyes. No attachment to the head is required, but the head still needs to be motionless. A small low power infrared light emitting diode (LED) is embedded in the infrared camera and directed towards the eye. The LED generates the corneal reflection and causes the bright pupil effect, which gives an enhanced image of the pupil. The centers of both the pupil and the corneal reflection are identified and located, and trigonometric calculations allow projecting the gaze point onto the image.

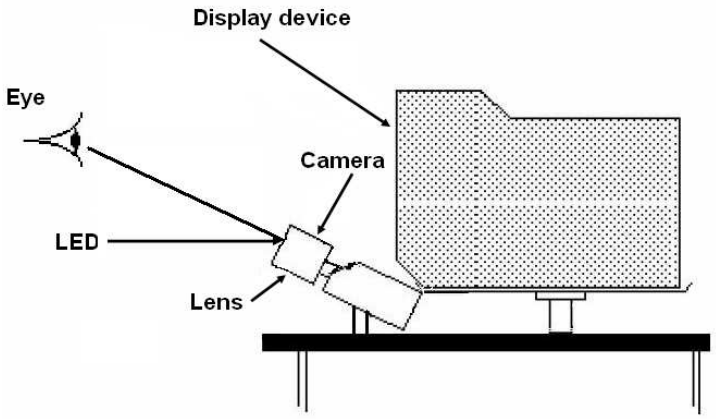


Figure 1.3: Schema of a gaze tracking system.

The following calibration steps are necessary before using the gaze tracker :

– Detection of the eye components (Iris and pupil) and the corneal reflection.

– Calculation of the relative position of the pupil and the corneal reflection.
– Calibration of the tracker.
– Extrapolation of the screen coordinates from the captured position.

Most recent systems still suffer from the following problems : (1) the need for calibration for each session, (2) the large restriction on head motion – especially for single-camera systems, (3) only a single user can be tracked at one time. The main additional difficulty in a single-camera setting is determining the distance of the user from the camera, since a triangulation as in a multi-camera setting cannot be carried out without calibration.

Active research in computer vision is aimed at developing gaze tracking systems which only require a simple webcam. Several approaches attempt to estimate the eye gaze exclusively [ZY02] by determining the position of the eyes for which the iris needs to be located. The methods tend to simplify the problem by assuming that the eye doesn't rotate but just shifts. The eye detection can be based on (among other methods) template matching, appearance classification, or feature detection. In the template matching methods, a generic eye model is first created based on the eye shape, and a template matching process is then used to search eyes in the image. The appearance based methods detect eyes based on their appearance using a classifier trained using a large amount of image patches representing the eyes of several users under different orientations and in different light conditions. The feature detection methods explore the visual characteristics of the eyes (such as edge, intensity of iris, or color distribution) to identify some distinctive features around the eyes. However, head orientation is assumed to be a very reliable indicator of the direction of a person's gaze and the eye gaze tracking system. The basic idea consists to consider the head position and orientation to give a rough initialization of the visual gaze, and then use the information about the eye centers and corners to fine tune the information. The importance of the contribution of the head pose in the gaze direction has been demonstrated by Stiefelhagen [SZ02] and the results of four experiment sessions are summarized in Table 1.1.

The experiments conducted in [SZ02] allow several interesting points to be underlined : (i) Most of the time, the subjects rotate their heads and eyes in the same direction, (ii) head orientation contributes more than half of the overall gaze direction (68.9%), (iii) focus of attention target can be correctly determined with only head orientation data in 88.7% of instances.

Subject	Nb Frames	Eye blinks	Same direction	Head contribution
1	36003	25.4%	83.0%	62.0%
2	35994	22.6%	80.2%	53.0%
3	38071	19.2%	91.9%	63.9%
4	35991	19.5%	92.9%	96.7%
Mean	-	21.7%	87.0%	68.9%

Table 1.1: Contribution of head orientation to the overall gaze.

Our approach is a contribution to the human gaze estimation based on his head pose. This choice has been dictated by the needs of a marketing application (determining where customers of a shop look on shelves) in which the person is not very close to the camera, unlike a personal environment where the person is postitioned in front of the camera.

1.2.2 Head pose estimation

Head pose estimation from a monocular camera or a simple image is a challenging topic. It consists of inferring the orientation of a human head from digital imagery. Several processing steps are performed in order to transform a pixel-based representation of the head into a high-level concept of direction. The head pose is important in a lot of domains like human-computer interfaces, video conferencing or driver monitoring.

Head pose estimation is often linked with visual gaze estimation [LMID09] which is the ability to characterize the direction and focus of attention of a person looking at a poster [SBOGP08] or at another person during meeting scenarios [VS08] for example. The head pose provides a coarse indication of the gaze that can be estimated in situations when the eyes of a person are not visible (like low-resolution imagery, or in the presence of eye-occluding objects like sunglasses). When the eyes are visible, head pose becomes a requirement to accurately predict gaze direction [VYG09].

1.2.2.1 Definition

The head pose estimation consists of locating a person's head and estimating its orientation in a space using the 3 degrees of freedom (see Figure 1.4) which are :

- Tilt (Pitch) : corresponds to a bottom/up head movement, around the x axis.
- Pan (Yaw) : corresponds to a right/left head movement, around the y axis.
- Roll (Slant) : corresponds to a profile head movement, around the z axis.

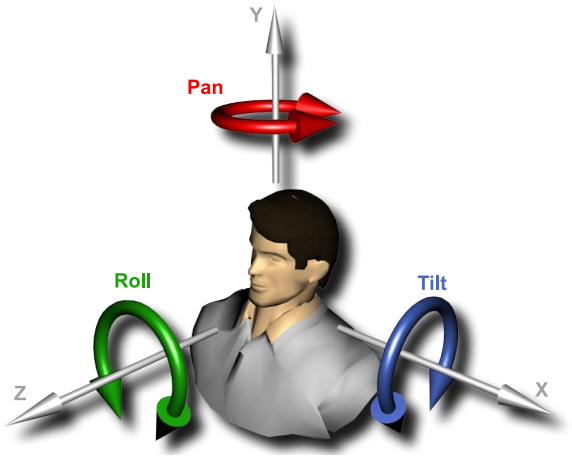


Figure 1.4: Head degrees of freedom model for head pose estimation.

1.2.2.2 Related work

Head pose estimation from a monocular camera or a simple image has received a lot of attention over the years. Various techniques have been proposed, and they can be categorized into two different classes :

1. Feature-based approaches : A set of specific facial features such as the eyes, nose, and mouth are used to estimate the head pose. They can use :
   - a geometric method that determines the head pose from the relative position of the eyes, mouth and nose [PZJ05].

- a flexible model that fits a non-rigid model to the facial structure of each individual in the image plane. The estimation is then performed from feature-level comparisons or from the instantiation of the model parameters. As an example of flexible models, the Active Shape Model (ASM) [CTCG95] which can be augmented with texture information in order to get an Active Appearance Model (AAM) [XBMK04].

2. Appearance based approaches : Instead of concentrating on the specific facial features, the appearance of the entire head image is modeled and learned from the training data. They can use :
   - a template based method which compares a new image of a head to a set of exemplars (each labelled with a discrete pose) in order to find the most similar view such as using multi-dimensional Gaussian distributions [WT00].
   - a detector array method which trains a series of head detectors. Each one is adjusted to a specific pose and assigned to a discrete pose according to the detector that has the greatest support such as using SVM [HSW98].
   - a nonlinear regression method that uses nonlinear regression tools to develop a functional mapping from the image or feature data to a head pose measurement such as using neural networks [RR98].
   - a manifold embedding method which seeks the low-dimensional manifolds that model the continuous variation in head pose. New images can be embedded into these manifolds and then used for embedded template matching or regression such as using Pose-eigenspaces [SB02].

The above two classes may be combined [VBB07] in order to overcome the limitations inherent in any single approach. The temporal information could be also introduced to improve head pose estimation by using the results of head tracking. It is done by recovering the global head pose changes from movement observed between video frames. A reliable and recent survey in head pose estimation can be found in [MCT09].

1.2.2.3 Criteria of the head pose estimation

We have proposed the following criteria as a guide for a head pose estimation module :

- Accuracy : It has to provide a reasonable estimation of the head pose with a small mean absolute error in the presence of good and bad head localization.
- Monocular : It has to estimate the head pose from a single camera.
- Multi-Person : It has to estimate the head pose of several persons in one image.
- Distance independence : It has to work in near-field and far-field images regardless of the resolution.
- Identity independence : It has to work across all identities.
- Computational cost : It has to estimate the head pose of several persons in real-time.

Our approach is composed of two steps : head pose estimation and visual field projection. The first step uses a template method based on SVD and Gabor descriptors. The second step consists of geometric projection of the gaze based on the distance of the person from the target scene and the head pose. We will detail these points in the following sections.

1.2.2.4 Template based approach

In the head pose estimation problem using a template based approach, a training and testing dataset of m subjects with n poses characterized by the tilt and pan angles are pre-processed. The head image pose estimation consists of a discriminating metric learning phase, where the objective is to find a D-dimensional feature vector that allows a learning method to achieve the highest accuracy. The range of a head pose is divided into a limited number of exclusive classes and a classifier is trained. The number of classes defines the accuracy of the final head pose estimation that can be achieved. Using a template based approach our model has the advantage of being suitable in near-field and far-field images, and learned from a training set that can be expandable to a larger size at any time without requiring any negative examples or facial feature points. However, the success of our estimation highly depends upon a correct locating of a person's head, and estimates discrete head poses only.

Head Pose Database : We use the Pointing database [GHC04] to build the head pose model and to test it. It consists of 93 poses for 15 persons with each pose taken twice per person (see Figure 1.5). We divide them into two sets :

– The training dataset : It consists of 20 images for each pose representing 11 persons (9 persons were taken twice and 2 persons were taken once).
– The testing dataset : It consists of 10 images for each pose representing 6 persons (4 persons were taken twice and the second images of the two persons left in the training dataset).

We select five poses : down-left, down-right, front, up-left and up-right which correspond, respectively, to a pair of pan and tilt angles of {(60, -90), (-60, +90), (0, 0), (+60, -90), (+60, +90)}.

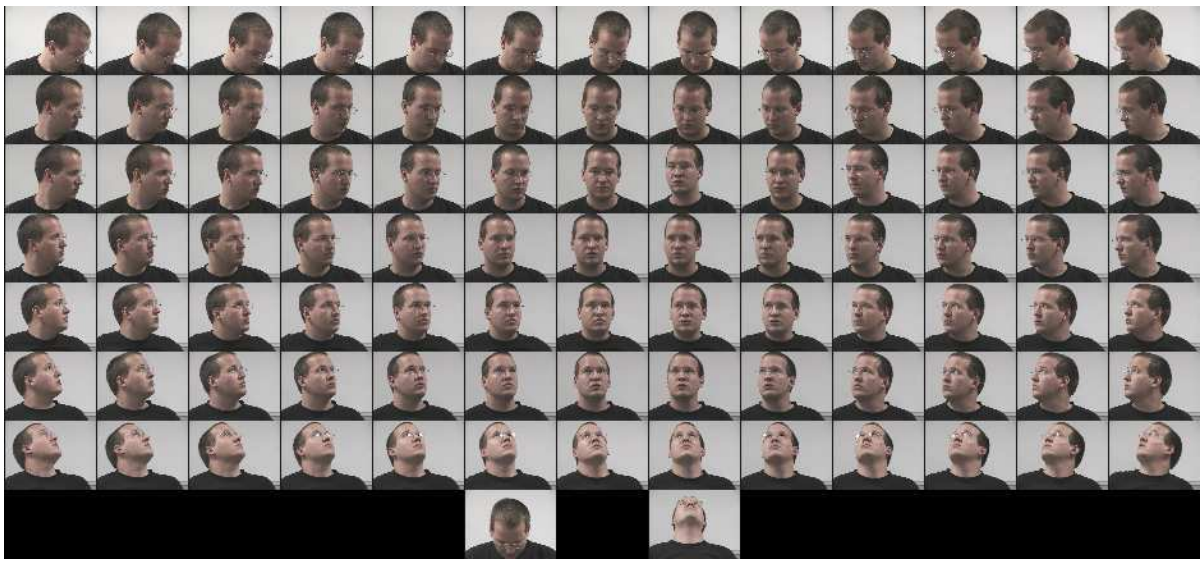


Figure 1.5: The head images of the Person01 in Pointing' 04 dataset.

We make a pre-processing on these images. We start by locating a tight bounding box around the head. Then, we normalize the images in 64x64 size. Finally, we apply a histogram equalization which ensures that two faces taken under different lighting conditions are transformed into two grayscale images with similar brightness levels. We will extract different feature vectors on this transformed database.

We will extract feature vectors on the pre-processed dataset. This is based on the pose similarity assumption that different people at the same pose look more similar than the same

person at different poses. Specifically two methods were chosen :

- Singular Value Decomposition : SVD is applied to the whole pose image to obtain SVD vector ;
- Gabor wavelets : Gabor wavelet coefficients are sampled from the pose image in different scales and orientations ;

The result is the extraction of a feature vector F_i of n elements for each head image i (with n chosen according to the specific technique used for the extraction) :

$$F_i = (F_{i_1}, F_{i_2}, \ldots, F_{i_n})^t \tag{1.1}$$

Feature vector extraction using SVD : The singular value decomposition [Vac91] of an MxN matrix A is its representation of a product of a diagonal matrix and two orthonormal matrices :

$$A = U * W * V^t \tag{1.2}$$

Where W is a diagonal matrix of singular values that can be coded as a 1D vector. All the singular values are non-negative and sorted in descending order. Applying this decomposition to a normalized head image i, it gives us a 1D vector :

$$W_i = (w_{i_1}, w_{i_2}, \ldots, w_{i_{64}})^t \tag{1.3}$$

Every singular value W_{i_j} will be associated with two vectors U_{i_j} and V_{i_j} with $j \in \{1, \ldots, 64\}$:

$$U_{i_j} = (u_{i_{j_1}}, u_{i_{j_2}}, \ldots, u_{i_{j_{64}}})^t \tag{1.4}$$

$$V_{i_j} = (v^t_{i_{j_1}}, v^t_{i_{j_2}}, \ldots, v^t_{i_{j_{64}}})^t \tag{1.5}$$

Then we calculate the norm of W_i :

$$\|W_i\| = \sqrt{w_{i_1}^2 + w_{i_2}^2 + \ldots + w_{i_{64}}^2} \tag{1.6}$$

Finally, we create two kinds of feature vectors of an image i :

– The first one is composed of elements obtained by dividing each element of the vector W by its norm $\|W_i\|$:

$$F_{i_j} = \frac{w_{i_j}}{\|W_i\|}, j \in \{1,\ldots,64\} \tag{1.7}$$

– The second one is composed of the P first singular value W_{i_j} divided by the norm $\|W_i\|$ with their corresponding U_{i_j} and V_{i_j} vectors :

$$F_{i_j} = (\frac{w_{i_j}}{\|W_i\|}, U_{i_j}, V_{i_j}), j \in \{1,\ldots,P\}, P \leq 64 \tag{1.8}$$

In order to select the appropriate value of P, we perform a reconstruction of the input image using the P top components (Figure 1.6).

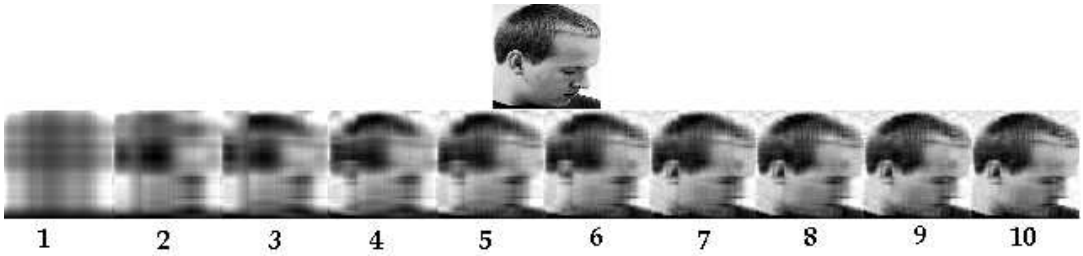


Figure 1.6: Image Reconstruction according to the value of P.

The experiments were done using the two feature vectors according to the value of P and using 3 comparison methods. We have used a support vector machine (SVM) [CST00] with a radial basis function kernel, a K nearest neighbors algorithm (KNN) with K=10 and Frobenius distance. We report in Figures 1.7 and 1.8 the results of the classification rate of the testing dataset using the whole training dataset for learning the classifiers using SVM, KNN and Frobenius distance by varying the value of P on the two feature vectors.

Feature vector extraction using Gabor wavelets : We apply Gabor filters to discriminate different poses due to the evolution of pose estimation in orientation. There is an evaluation

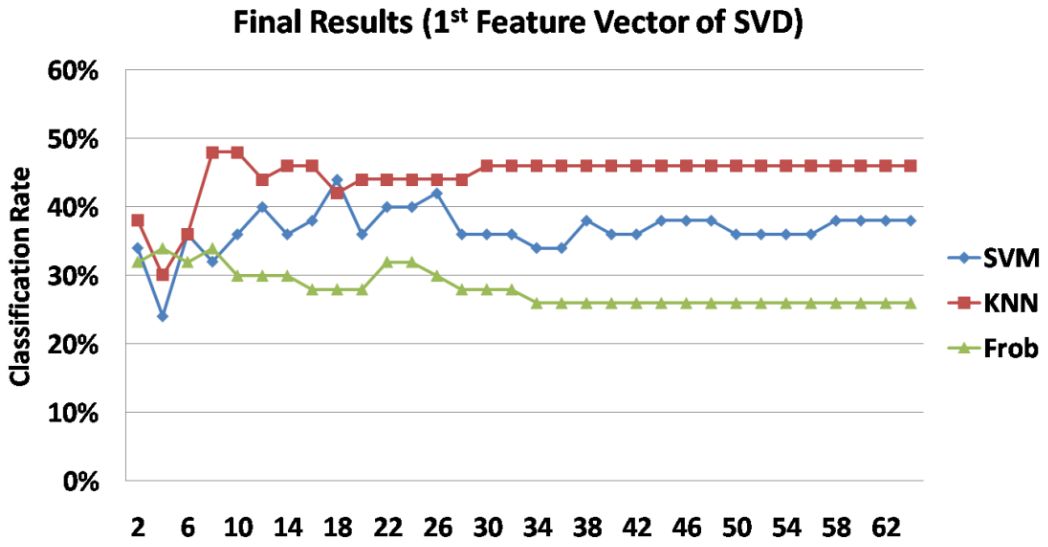


Figure 1.7: Classification rate results using the 1st feature vector of SVD.

of the pose similarity ratio at a fixed pose with varying Gabor filter orientation in [SGO01]. A Gabor wavelet $\psi_{o,s}(z)$ is defined as [ZW06] :

$$\psi_{o,s}(z) = \frac{\|k_{o,s}\|^2}{\sigma^2} e^{-\frac{\|k_{o,s}\|^2 \|z\|^2}{2\sigma^2}} [e^{ik_{o,s}z} - e^{-\frac{\sigma^2}{2}}] \tag{1.9}$$

where $z = (x, y)$ is the point with the horizontal coordinate x and the vertical coordinate y. The parameters o and s define the orientation and scale of the Gabor kernel, $\|\cdot\|$ denotes the norm operator, and σ is related to the standard derivation of the Gaussian window in the kernel and determines the ratio of the Gaussian window width to the wavelength. The wave vector $k_{o,s}$ is defined as follows :

$$k_{o,s} = k_s e^{i\phi_o} \tag{1.10}$$

where $k_s = \frac{k_{max}}{f^s}$ and $\phi_o = \frac{\pi o}{S}$. k_{max} is the maximum frequency, f^s is the spatial frequency between kernels in the frequency domain, and s is the number of the orientations chosen.

For the creation of a feature vector, we use generally eight orientations $\{ 0, \frac{\pi}{8}, \frac{\pi}{4}, \frac{3\pi}{8}, \frac{\pi}{2},$

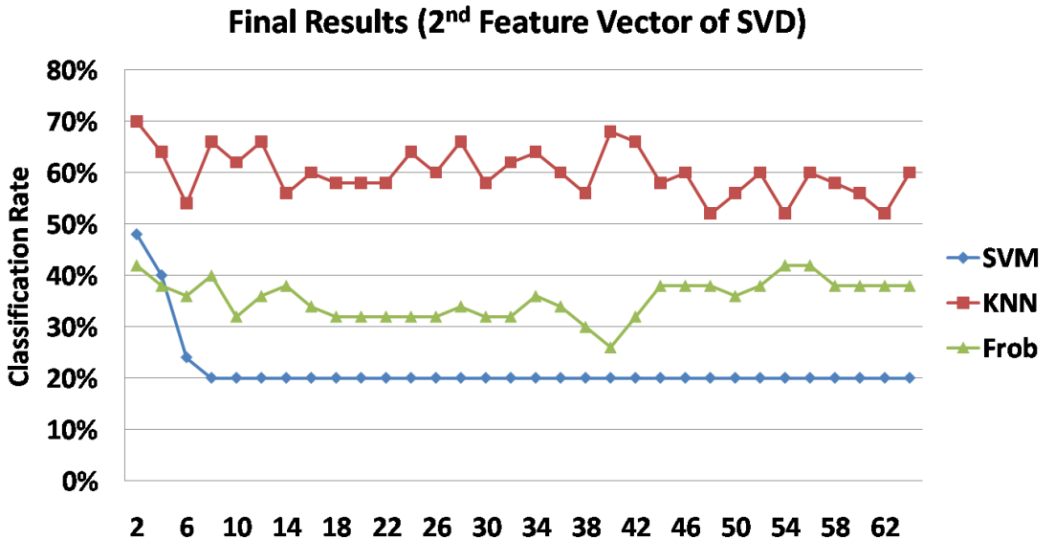


Figure 1.8: Classification rate results using the 2nd feature vector of SVD.

$\frac{5\pi}{8}$, $\frac{3\pi}{4}$, $\frac{7\pi}{8}$ } at five different scales { 0, 1, 2, 3, 4 } of Gabor wavelets with $\sigma = 2\pi$, $k_{max} = \frac{\pi}{2}$, and $f = \sqrt{2}$.

The Gabor wavelet representation of an image is the convolution of the image with a family of Gabor kernels as defined in Equation (1.9) (see Figure 1.9).

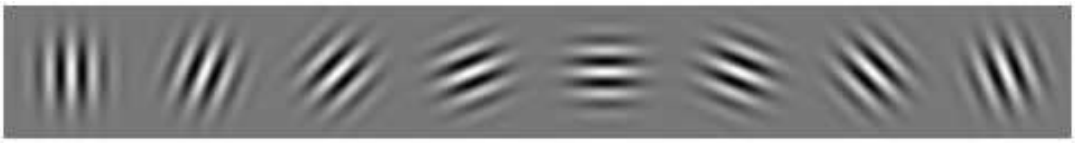


Figure 1.9: A real response of Gabor wavelets using the 8 orientations.

The convolution of an image I and a Gabor kernel $\psi_{o,s}(z)$ is defined as follows :

$$Conv_{o,s}(z) = I(z) * \psi_{o,s}(z) \tag{1.11}$$

The response $Conv_{o,s}(z)$ to each Gabor kernel is a complex function with a real part

$\mathrm{Re}\{Conv_{o,s}(z)\}$ and an imaginary part $\mathrm{Im}\{Conv_{o,s}(z)\}$ defined as :

$$Conv_{o,s}(z) = \mathrm{Re}\{Conv_{o,s}(z)\} + i\mathrm{Im}\{Conv_{o,s}(z)\} \tag{1.12}$$

The magnitude response $\|Conv_{o,s}(z)\|$ is expressed as :

$$\|Conv_{o,s}(z)\| = \sqrt{\mathrm{Re}\{Conv_{o,s}(z)\}^2 + \mathrm{Im}\{Conv_{o,s}(z)\}^2} \tag{1.13}$$

For each image, the outputs are o*s images which record the real, the imaginary or the magnitude of the responses to the Gabor filters. As a feature vector using a specific response, we calculate for each image at a specific scale s and orientation o the mean and the deviation of its pixels intensities.

We finally concatenate the mean and deviation of each image at the o orientations and s scales in a vector. We obtain a feature vector composed of 2*o*s elements for each head image i :

$$F_i = (M_1, D_1, M_2, D_2, ..., M_{O*S}, D_{O*S})^t \tag{1.14}$$

We obtain 3 variations of the feature vectors using Gabor wavelets depending on the responses to the Gabor filters chosen (real, imaginary or magnitude).

In order to test the influence of the scale on the Gabor feature vectors, we conduct the experiments using 3 variations of the feature vectors using Gabor wavelets (real, imaginary and magnitude). We report respectively in Figures 1.10, 1.11 and 1.12 the classification rate of the testing dataset using the whole training dataset for learning the classifiers using KNN, SVM and Frobenius distance by varying the number of the selected scale s for the construction of the feature vector F_i from 1 to 5 and using the 8 following orientations.

Since it appears from the last experiment that it is more suitable to select five scales for the extraction of the feature vectors, we have selected five scales for the construction of the feature vector. We have conducted another experiment by varying the selected number of orientations from 1 to 8. We have reported respectively in Figures 1.13 and 1.14 the classification rate of the testing dataset using the whole training dataset for learning the classifiers using KNN, SVM. We have avoided reporting the results using the Frobenius distance since the classification

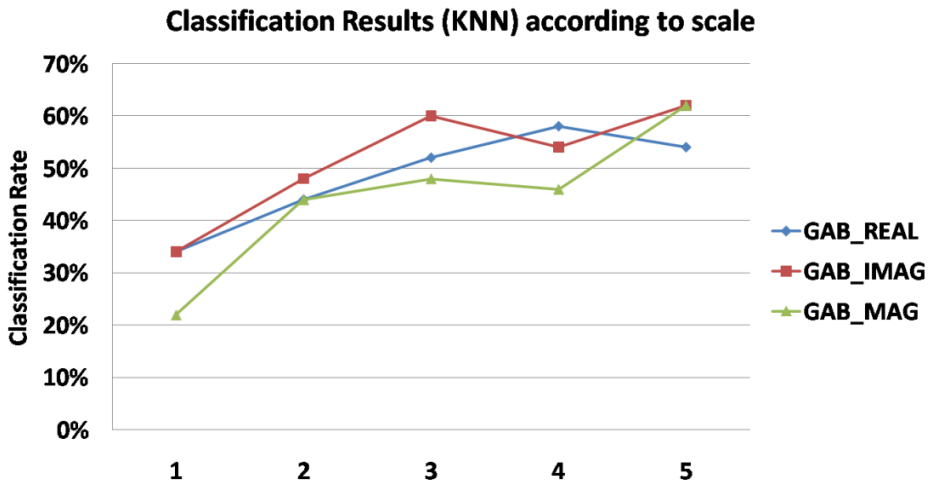


Figure 1.10: Classification rate results according to the number of selected scales using KNN.

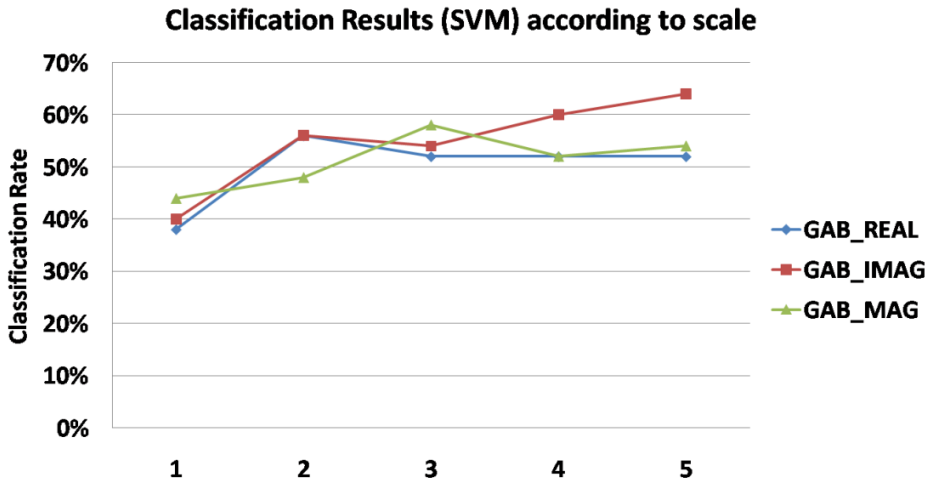


Figure 1.11: Classification rate results according to the number of selected scales using SVM.

results were weak.

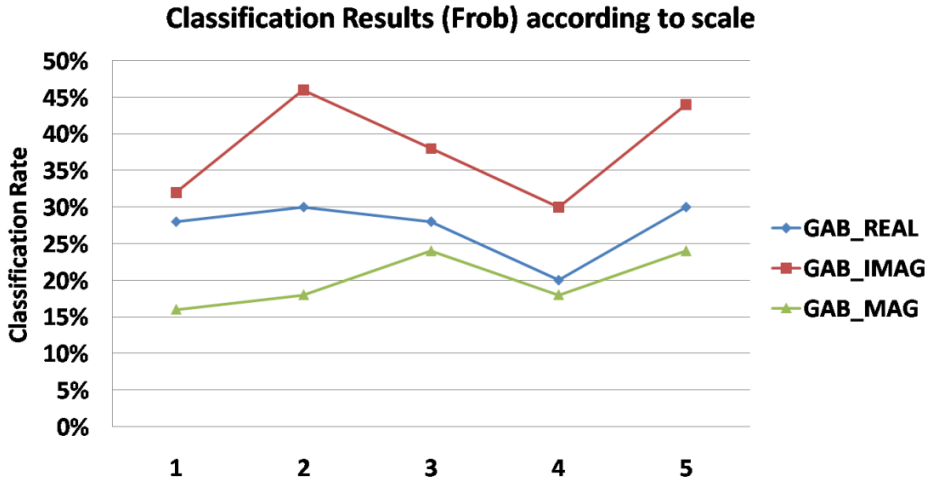


Figure 1.12: Classification rate results according to the number of selected scales using Frobenius distance.

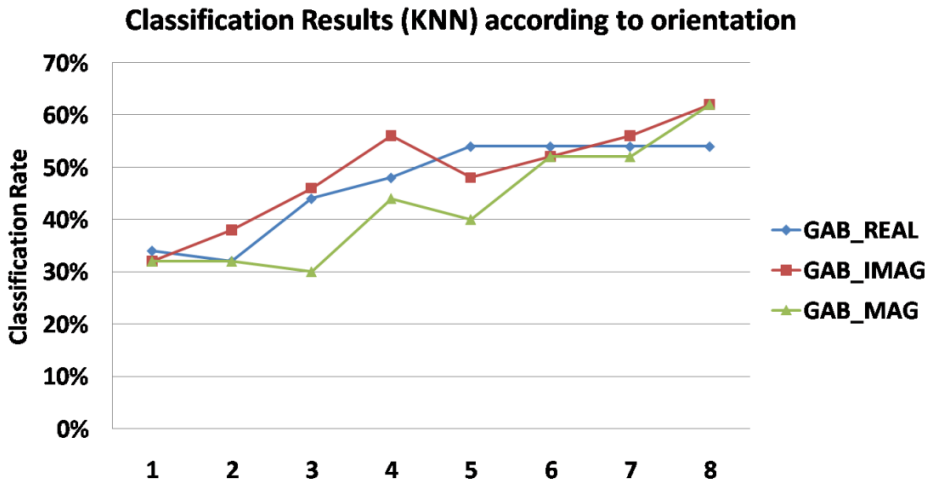


Figure 1.13: Classification rate results according to the 8 selected orientations using KNN.

Discussions : We have used a support vector machine (SVM) with a radial basis function kernel, a K nearest neighbor algorithm (KNN) with $K = 10$ and the Frobenius distance for the

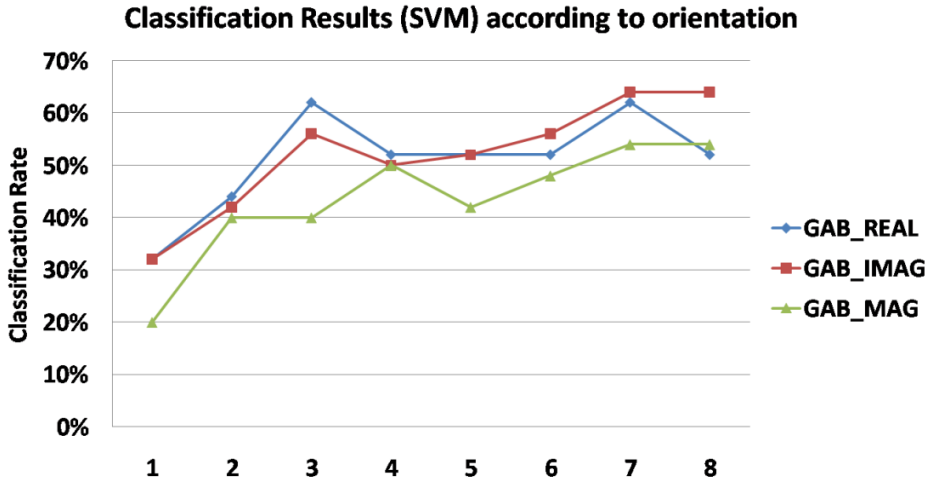


Figure 1.14: Classification rate results according to 8 selected orientations using SVM.

experiments. In [LZD08], we note that the head pose recognition accuracies increase with the number of the training samples which is consistent with the typical supervised learning. Thus, we use the whole learning dataset for learning the classifiers in all experiments.

From the Figures 1.7 and 1.8, it's clear that the information contained in the diagonal matrix of singular values is not sufficient alone. The addition of the information contained on U and V enhances the results. Since the values are ordered, the information contained in the first components is enough to perform the head pose estimation.

In general, we have observed from the three different Gabor wavelet features that the imaginary component features are better than the magnitude and real features. This is probably due to the fact that the majority of the information is typically contained in the phase component.

We have noticed that the Gabor wavelet features perform better than the SVD features. This is partly due to the fact that the Gabor wavelet features are capable of handling different orientations and scales while the SVD features are not. Even if the 2nd feature vector of SVD gets the best result of the experiments using KNN.

1.2.2.5 Head Pose Tracking

In order to correctly estimate the visual gaze of a user in a personal environment, multi-modal information derived from the head pose and the location of the eyes should be taken in consideration. We use the system described in [VLS+10] which is able to accurately determine this information from a video. The Cylindrical Head Model (CHM) pose tracker [XKC02] and the isophote based eye location estimation methods [VG08] have several advantages over the other reported methods.

The 3D cylindrical shaped model is used to perform head tracking. The perspective projection is used to re-construct the focal length, height of the 3D head and to project 2D pixels onto 3D points and vice-versa. The locations of the points on the 3D cylinder are found using the ray-tracing technique. Then, the correct points on the cylinder are projected back to image plane. Through optical flow the new position of the head and its pose are estimated. The Figure 1.15 shows some qualitative examples of head tracking using the CHM.

Figure 1.15: Qualitative examples of the head tracking using CHM.

1.2.3 Visual field projection and extraction of regions of interest

The visual field of view is the total area in which objects can be seen in the peripheral region, while the eyes are focused on a central point. The visual field projection gives an estimation of which regions are of interest to the subject in the target scene. A mapping with the objects that belongs to the target scene can give an explanation of the behavior of the people present in the monitored area. Various interpretations could be derived from these regions. We can construct a heatmap based on the regions of the target scene attracting the most attention, or a trajectory map based on the centroid points of the person's locations of interest for example. We are proposing a methodology that projects the visual field of a person onto a target scene image from a starting point regardless of the method used for the estimation.

1.2.3.1 Visual field estimation

In presence of a target scene, the human visual field is limited where it is freedom in its absence.

Human Vision Features : The visual gaze is the peripheral space seen by the eyes. It extends normally from 60° at the top, 70° at the bottom and 90° approximately laterally, which corresponds to a photographic lens angle of 180°. The brain interprets the images that come from the two eyes enabling objects to be perceived three-dimensionally. Long-distance vision without any obstacle gets two images that are very similar whereas at a shorter distance, the images obtained by the left eye and the right eye are slightly different (the point of observation is not similar). The common field of the two eyes is called the binocular field of view and extends on 120° broad. It is surrounded by two monocular field of view of approximately 30° broad. The Figure 1.16 represents the visual field of a person [PZ79].

The binocular visual gaze of a person is shown on the Figure 1.17 :

Determination of the visual field : We have represented the field of view of a person as a rectangle *ABCD*. Thus the visual gaze is represented by a pyramid *OABCD* with the starting point *O* representing the position of a person's head. The calculation of the width (W) and

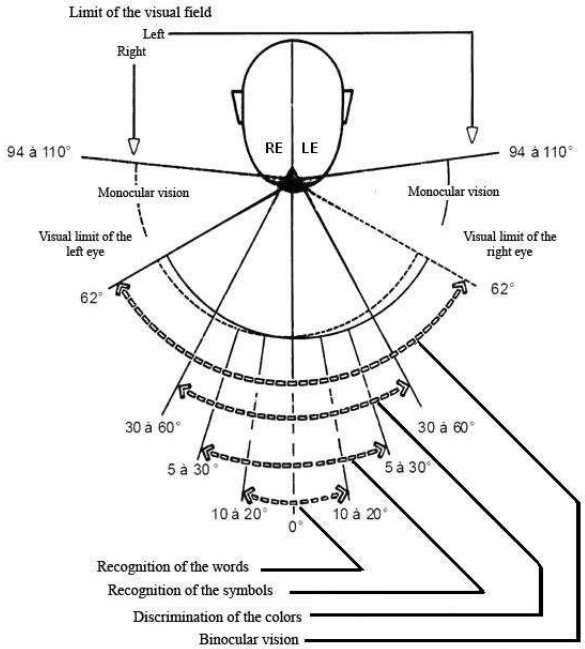


Figure 1.16: Human vision system.

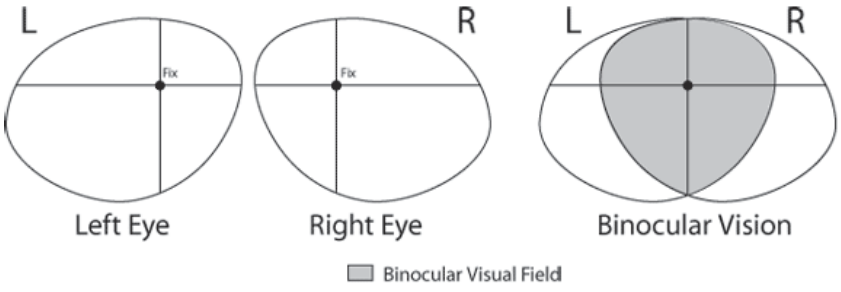


Figure 1.17: Binocular visual gaze.

the height (H) of the visual gaze needs 3 values [LMD08a]. These 3 values are the distance d (distance from the target scene), the angles α and β (which are respectively the horizontal and vertical angles of the visual gaze in binocular human vision). The point M represents the horizontal projection of the point O at a distance d as seen in figure 1.18.

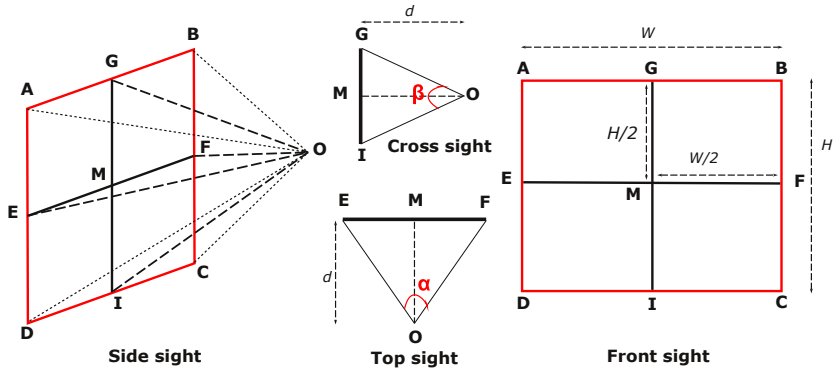


Figure 1.18: Representation of the visual gaze at a distance d.

The width (W) and the height (H) of the visual gaze at a distance d are calculated using

$$W = 2MF = 2d.\tan\frac{\alpha}{2},$$

$$H = 2MG = 2d.\tan\frac{\beta}{2}.$$

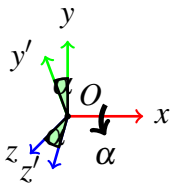
Adaptation of the Visual field to the head pose : Once the calculation of the width and the height of the visual gaze is done, we need the head pose of the person and his/her location in the monitored area in order to adapt the visual to the head pose. We use vectorial rotation around the 3 degrees of freedom of the head.

We choose the x-y-z covention to perform the rotation which corresponds to rotation around the tilt angle following the pan and roll angles as shown in Figure 1.19

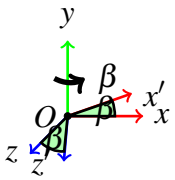
Using this convention, the orientation matix $R_{(\alpha,\beta,\gamma)}$ is obtained by the 3 matrix associated to the 3 degress of freedom of the head as following :

$$R_{(\alpha,\beta,\gamma)} = R_{(z,\gamma)}.R_{(y,\beta)}.R_{(x,\alpha)} \tag{1.15}$$

$$R_{(x,\alpha)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{pmatrix}$$



$$R_{(y,\beta)} = \begin{pmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{pmatrix}$$



$$R_{(z,\gamma)} = \begin{pmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

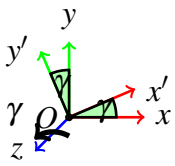


Figure 1.19: The 3 kind of rotations (tilt, pan et roll) and their matrix.

$$R_{(\alpha,\beta,\gamma)} = \begin{pmatrix} \cos\beta\cos\gamma & \cos\gamma\sin\alpha\sin\beta - \cos\alpha\sin\gamma & \cos\alpha\cos\gamma\sin\beta + \sin\alpha\sin\gamma \\ \cos\beta\sin\gamma & \cos\alpha\cos\gamma + \sin\alpha\sin\beta\sin\gamma & \cos\alpha\sin\beta\sin\gamma - \cos\gamma\sin\alpha \\ -\sin\beta & \cos\beta\sin\alpha & \cos\alpha\cos\beta \end{pmatrix} \tag{1.16}$$

1.2.3.2 Visual field projection

Before processing the projection, a calibration step is performed using the method suggested by [BÓ2]. The projection of the visual field is a quadrilateral $A'B'C'D'$ with a central point M' constructed from the intersection between the plane that represents the target scene $P : ax + by + cz + d = 0$ and the lines (OA), (OB), (OC), (OD), and (OM). It represents the region of interest extracted in the target scene. The Figure 1.20 shows the projection of the visual gaze on a shelf. The detailled steps can be found in [LUBD10].

In order to determine what is being viewed in the target scene, a granularity must be chosen

Figure 1.20: Visual gaze projection on a shelf.

(i.e. it is the period of time that several projection areas should share before saying that an area of the target scene is being viewed).

1.2.3.3 Extraction of regions of interest

The Figure 1.21 shows a person from The Boston University dataset [CSA00] represented at an instant t by its visual gaze, its region of interest in the target scene with a central point, a scanpath in the target scene, and a heatmap of the target scene.

Correction of the gaze point The gaze point M' (represented by the blue dot in Figure 1.22) is the projection of the central point M on the target scene. The eye location obtained from the system discussed above should alter this gaze point, according to their distance from the reference eye locations. This modification is only allowed to happen inside the computed visual field. The result is an eye gaze point M_{disp} that represents the displacement of the gaze point of a certain percentage in the visual gaze area. The projection of this point M'_{disp} (represented by the red dot in Figure 1.22) is obtained by the intersection between the plane P and the line (OM_{disp}) by $M_{disp}(M_x + W * disp_x, M_y + H * disp_y)$.

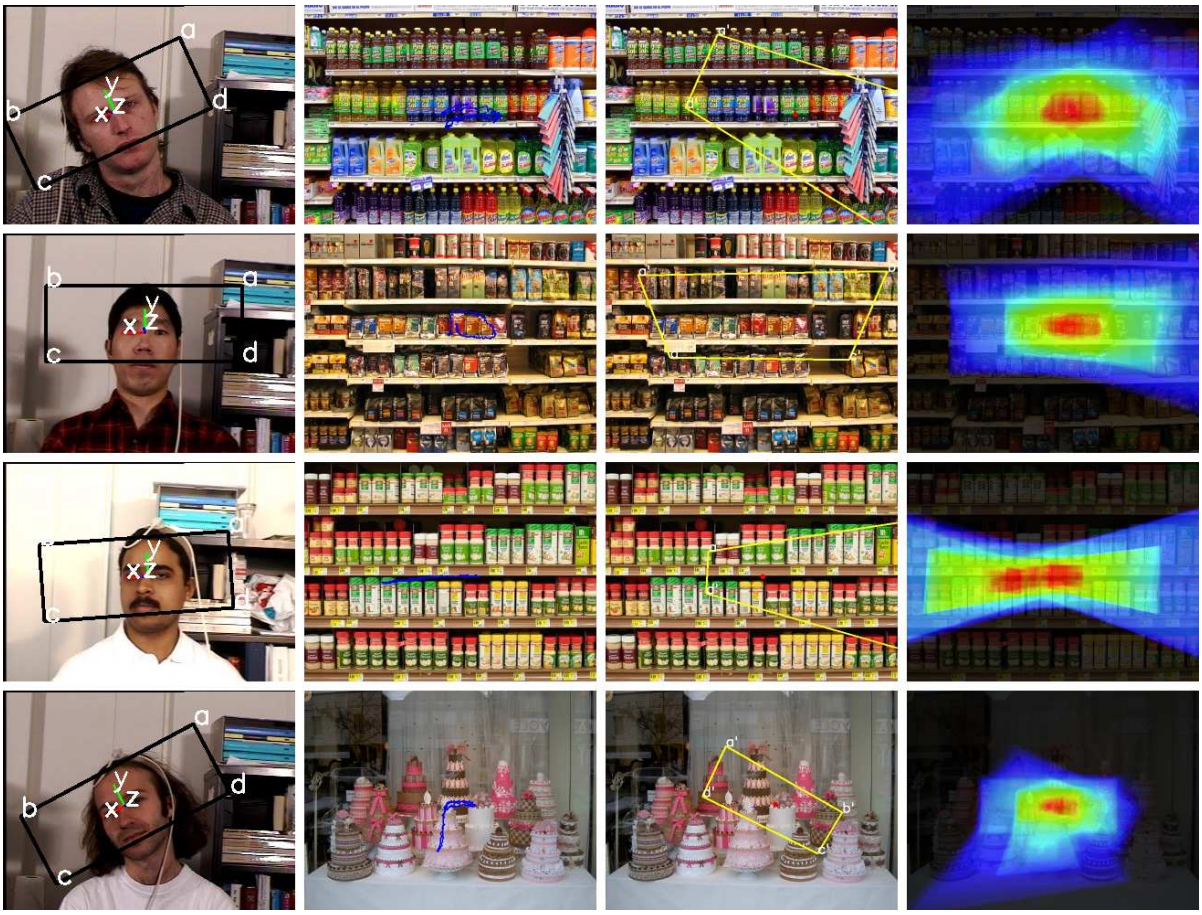


Figure 1.21: Representation of the gaze information.

In order to gain a better insight into the accuracy of the system, we tested it on a subset of the Boston University head pose dataset [CSA00]. The dataset consists of 45 video sequences, where 5 subjects were asked to perform 9 different head motions under uniform illumination in a standard office setting. The head is always visible and there is no occlusion except for some minor self-occlusions. The videos are recorded in low resolution and a Flock of Birds tracker records the pose information coming from the magnetic sensor on the person's head. This system claims a nominal accuracy of 1.8 mm in translation and 0.5 degrees in rotation. Since the subjects in this dataset were not requested to perform any gazing task, the ground

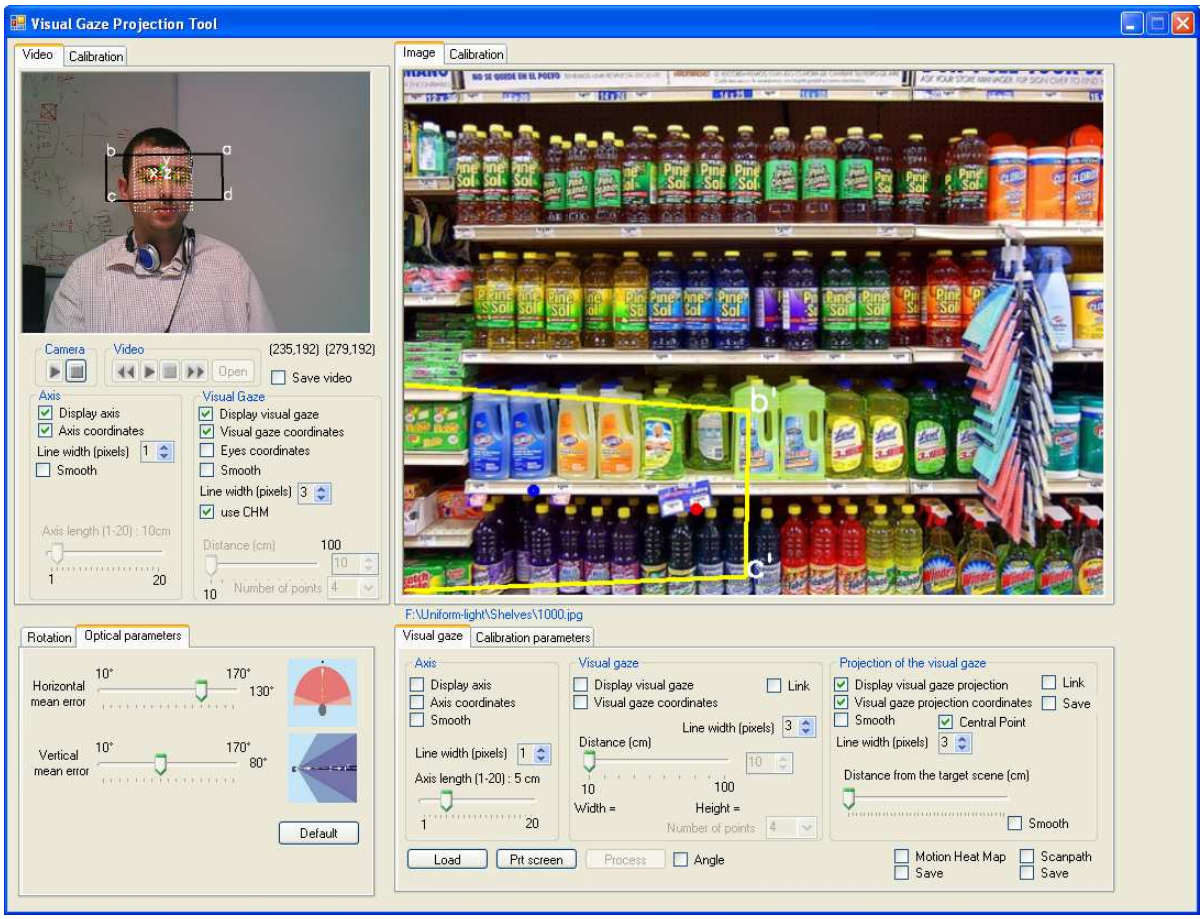


Figure 1.22: A Screenshot of the final system, working in real time using a simple webcam. The yellow rectangle indicates the user's region of interest (defined by head pose only), while the red dot is the combined head and eyes visual gaze projection.

truth of the visual gaze is not available. In most of the videos, however, the subject gazes at the camera as his head pose changes (as displayed in Figure 1.23). As this eye movement is supposed to compensate for the head pose, we expect the coordinates of the projected head and eyes gaze vector to have a minor standard deviation when compared to the one obtained solely by the head pose.

Therefore, we will analyze the deviation of the combined head-and-eye visual gaze with

respect to a the head pose alone, on the subject of the BU dataset that are gazing at the camera.

Figure 1.23: Example of the Boston University head pose dataset : some subjects kept gazing at the camera, even if not instructed to do so.

Regarding the experimentation on the subset of the BU dataset, the standard deviation was reduced by 61.06% on the x dimension and 52.23% in the y dimension, with a respective stds of 20.48% and 19.05%. Figure 1.24 shows one example (for subject jam8) in which the use of eye displacement (in blue dots) significantly helps in reducing the standard deviation of the visual gaze given by only the head pose (in red crosses). Note how the gaze is localized around the same location when using the eye information.

While analyzing our results we noted that, in some of the videos, the head pose estimator would lag behind the eye locator. In this way, a slight displacement between the reference eye points and the current eye location leads the system to deduce that the user is looking in the same direction as the movement. This is an inherent problem of tracking, which is solved in most cases by Kalman filtering. However, making wrong assumptions on the next position of the head might result in wrong head pose estimations, so a conservative tracker is often preferred.

Another point to discuss is how much eye displacements should affect the visual gaze vector generated by the head pose : in our experiments, we empirically found that a weighting factor of 0.2 is the best to achieve better results. We noted that a weighting factor of 0.1 was having little effects on the final estimation, while a weighting factor superior to 0.3 would generate many outliers, leading to a higher standard deviation.

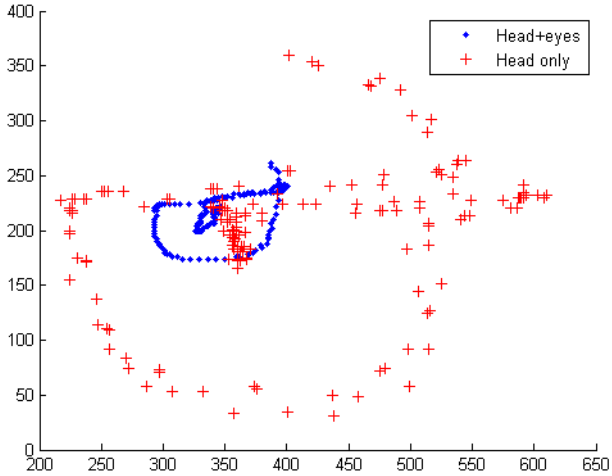


Figure 1.24: Example of gaze correction by eyes, on the movie "jam8.avi".

1.2.3.4 Gaze analysis

Based on the information provided from gaze direction, we studied metrics [MLLD09] to estimate some features about the distribution of the gaze fixation points. We highlighted a validation methodology using a framework for measuring the quality of a visual media, that is to say its ability to transmit the original idea of the creator. This framework is based on the natural human gaze of people watching static images or dynamic scenes. As an example, the Figure 1.25 shows some keyframes taken from broadcasted sporting events (soccer and tennis). It illustrates how our approach can help advertising designers to evaluate the impact of their advertisement during sport events, and how to best distribute it. For this specific application, if we want to focus on the specific advertisement areas in the media, while disregarding other irrelevant parts, the model could benefit from an estimator that counts the hits of the fixations points in advertisement areas.

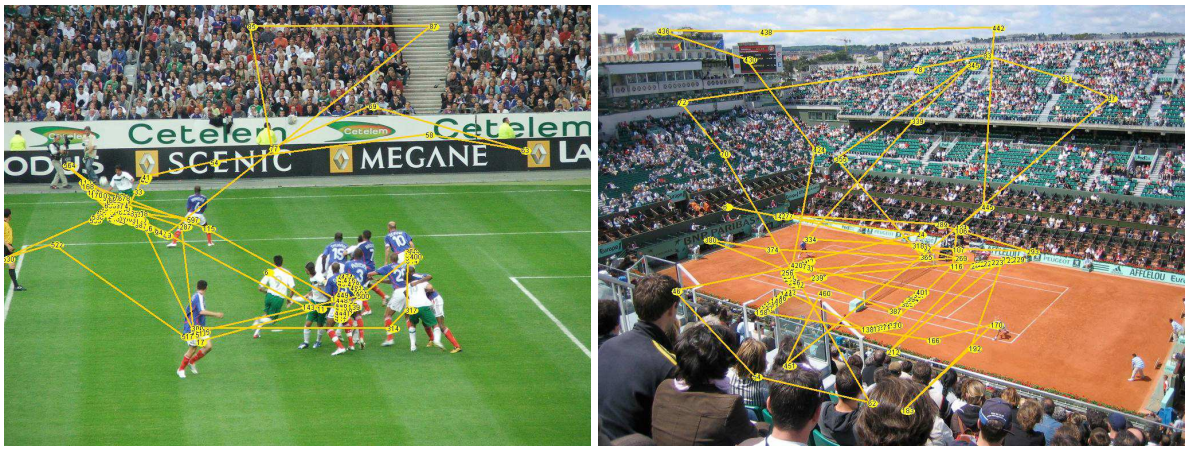


Figure 1.25: Illustration of our approach for evaluating the impact of advertisement campaigns during sporting events.

1.2.4 Multi-person detection and tracking

Based on the environment of the experiments, we have defined 3 modalities which are expressed in the following way :

- Stop/Moving : this modality refers to whether the person is moving or not,
- Looking/Not looking : this modality denotes persons whose visual field is oriented towards the target scene or not,
- Location of interest : determines where the persons are looking at the target scene. This modality provides the exact location of the person's gaze.

In order to extract these modalities we construct a system composed of 3 modules. These 3 modules are Multi-person detection/tracking, Head pose estimation, and Visual field projection (on the target scene) as shown in figure 1.26.

We construct a model which stores the necessary information for representing this architecture. The monitored area may contains several persons, so the model at a time t will vary according to the number of persons and is defined by $I_t = \{X_{i,t} / i \in |I_t|\}$, where $N = |I_t|$ denotes the number of people in the monitored area at an instant t.

The state $X_{i,t}$ of a single person i at an instant t contains 4 components :

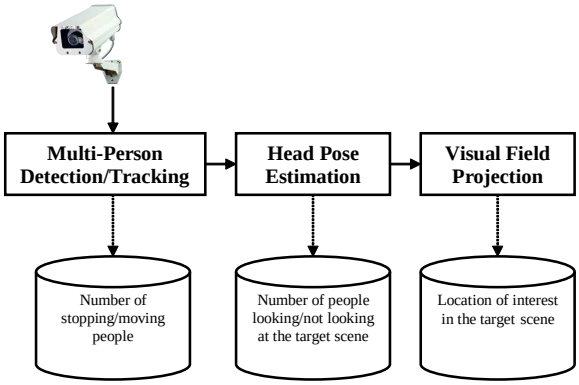


Figure 1.26: Overview of the behaviour detection system.

1. The body component $X_{i,t}^{b}$ contains information about the 2D location of the person and the body color signature.
2. The head pose component $X_{i,t}^{h}$ contains information about 2D location of the head and its pose class.
3. The modality component $X_{i,t}^{m}$ is composed of a set of 2 modalities which are Stop/Moving and Looking/Not looking. They can have one of the three states : Ok, No or Unknown.
4. The location of interest component $X_{i,t}^{l}$ can have one of these two states : 'In' (when the person is looking at the target scene and it corresponds to the coordinates of the rectangle which represents the person's location of interest in the target scene), or 'Out' (when the person is not looking at the target scene or unknown).

The multi-person detection/tracking locates and tracks several persons over time in the monitored area which is useful when counting their numbers. Several approaches were attempted based on serial [ZH05] or parallel [WN07] processing. An overview of the approach used is shown in Figure 1.27.

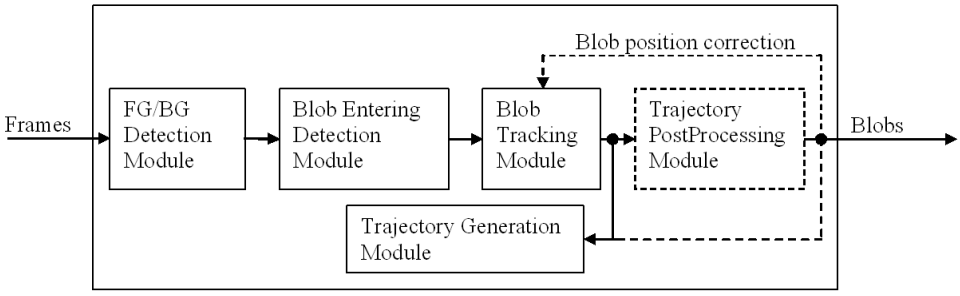


Figure 1.27: Overview of Multi-person detection and tracking.

Person detection is essential in starting the process. It permits us to detect all persons present in the monitored area and to count their number. We also need to track them to analyze their behaviour while in front of the target scene. The approach that has been used is serial. For human detection, we use a background subtraction method based on modeling the background using Mixture of Gaussians [SG99]. It is the most appropriate method for this task due to its robustness when the persons are stopping or moving. The human classification is done on the blobs extracted from the detected foreground using some trivial features related to the environment such as the length (the number of connected foreground pixels), and position (the person is on the ground). Once the persons present in the monitored area has been detected, we use a kernel-based method to match them between the consecutive frames using mean shift algorithm [CRM00]. This method gives us accurate results when there are less than 4 persons present in the monitored area with no overlapping bodies. Finally, we determine the number of persons detected in the monitored area in order to count them and we group them into two classes : stopping or moving. We use temporal and spatial information applied to the body component to determine whether they have stopped or not. A person is considered stopping if there is no movement below the torso.

The experiments were done using videos captured from a camera placed in a shop window and filming outside at face level. We present below some screenshots (see Figure 1.28) which highlights the results of the person detection coupled with the result of the visual field representation taken at the frames 281, 341 and 422.

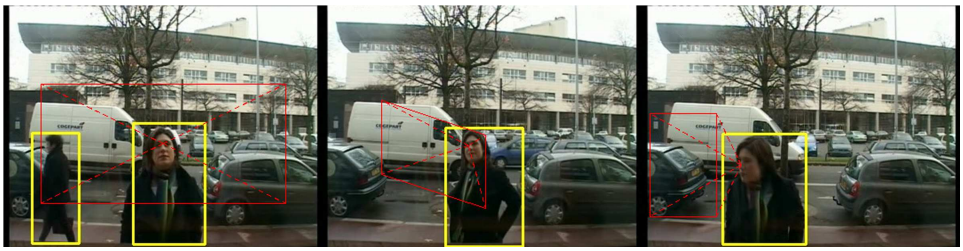


Figure 1.28: Human behaviour detection system in front of a shop window.

The successful portability of the system depicted in Figure 1.29 shows the system tested on standard sequences using a simulated distance from a target scene of 0.5 meter (see http ://groups.inf.ed.ac.uk/vision/CAVIAR).

Figure 1.29: Human behaviour detection system on Caviar sequences.

The results are very sensitive to the results of the head pose estimation.

1.3 Conclusions

We have presented an approach aimed at collecting different types of information about people in the proximity of a target scene. This type of study could be a beneficial application in marketing, graphic design, or Human-Computer Interaction. The analysis of this collected information using data-mining techniques could provide clues about people's interest in items placed in a shop window for example.

We experimented with a support vector machine (SVM), a K nearest neighbor algorithm (KNN) and Frobenius distance in order to create the head pose model applied to several descriptors including SVD and Gabor. In general, we observed from the three different Gabor wavelet features that the imaginary component features are better than the magnitude and real features. We notice that the Gabor wavelet features perform better than the SVD features. However SVD gets the best results of the experiments using KNN. Once the calculation of the width and the height of the visual gaze are completed, we need the subject's head pose and his/her location in the monitored area in order to adapt the visual to the head pose. The extraction of the person's region of interest is done by the projection of the visual field on the target scene.

Our future work will focus on three directions : human detection based on human features rather than trivial features related to environmental information, the detection of the person's gaze at distant range based on their eyes rather than their head pose and using higher resolution cameras, and the exploitation of the collected information as a feedback.

Chapitre 2

Introduction

Sommaire

2.1 Introduction

De nos cinq sens, celui que nous utilisons le plus est la vue. C'est par nos yeux que nous percevons les informations qui nous permettent de comprendre et d'interagir avec notre environnement. La vision par ordinateur (appelée aussi vision artificielle, vision numérique ou vision cognitive) a pour but de reproduire sur un ordinateur la vision humaine. Elle s'inscrit dans le cadre général de la recherche de moyens susceptibles de doter l'ordinateur d'une intelligence (apprentissage, représentation, raisonnement) comparable à celle des êtres humains. C'est ce à quoi s'attelle l'extraction de l'information à partir de la vidéo : un domaine très dynamique et en pleine effervescence, tant du point de vue de la recherche scientifique que des applications de la vie quotidienne.

Ce domaine ne date pas d'aujourd'hui, David Marr [1] a été un des premiers à définir les bases formelles de la vision par ordinateur en intégrant des résultats issus de la psychologie, de l'intelligence artificielle et de la neurophysiologie. Son formalisme est composé de deux visions : la vision brute qui permet à l'homme d'observer ce qu'il détecte, et la vision abstraite qui permet de reconnaître des objets sémantiques (e.g. rivière, montagne, personne, voiture, etc.). Dans ces deux visions, l'homme utilise son intelligence, sa mémoire, et ses connaissances à priori.

Aristote a proposé la conjecture selon laquelle la vision naturelle est un processus actif où les yeux sont reliés à des vrilles sensiblement tactiles qui atteignent et ressentent la scène observée. Au XVII^{e} siècle, René Descartes a démontré que les yeux contenaient une image rétinienne en 2D de la scène. Cependant, à cette époque aucune théorie sur un mécanisme de calcul pour la compréhension de la scène n'existait. Au XVIII^{e} siècle, une approche de calcul a été proposée par Thomas Bayes et Pierre-Simon Laplace pour fournir des explications plausibles sur des données. Ils ont montré l'utilité de la probabilité de modèles de données actualisée pour tenir compte des nouvelles observations (en utilisant le théorème de Bayes).

En 1867, Hermann von Helmholtz a établi une thèse selon laquelle la vision est influencée en plus du système nerveux par des inférences psychologiques basées sur les modèles acquis de

1. David Marr (1945-1980) est un neurophysiologiste et mathématicien anglais qui travaillait au MIT. Il a mis les bases des sciences cognitives.

l'expérience. Il a émis la conjecture selon laquelle le cerveau apprend des modèles sur la façon dont les éléments de la scène sont mis ensemble pour expliquer l'entrée visuelle (c'est défini actuellement dans la littérature comme modèle génératif) et que la vision est l'inférence dans ces modèles. Il est allé jusqu'à conjecturer que l'individu effectue des expériences physiques, telles que le déplacement d'un objet, afin de bâtir un meilleur modèle visuel de l'objet et de ses interactions avec d'autres objets dans la scène. L'apparition des ordinateurs au XX^{e} siècle a permis aux chercheurs d'élaborer des modèles réalistes et d'effectuer des expériences pour évaluer ces modèles donnant ainsi naissance à la thématique de la vision par ordinateur.

Dans le cadre de la thèse, nous nous intéressons à un domaine émergent de la vision par ordinateur, à savoir l'extraction de la direction du regard d'une personne sur une scène cible (e.g. écran, linéaire, affiche publicitaire, etc.).

2.2 Définitions

L'estimation de l'orientation de la tête et la direction du regard sont employés souvent dans le document. Voici une brève description illustrée par la Figure 2.1 :

(a) Les 3 degrés de liberté de la tête. (b) Estimation de la direction du regard.

FIGURE 2.1: Illustration de l'orientation de la tête et de la direction du regard.

L'orientation de la tête détermine la valeur des angles selon les 3 degrés de liberté de la rotation de la tête, à partir d'une image. Les 3 degrés de liberté illustrés par la Figure 2.1(a) sont définis par Tilt, Pan et Roll lors d'un mouvement de la tête autour de l'axe X, Y et Z respectivement.

La direction du regard détermine le sens du regard d'une personne. La Figure 2.1(b) illustre une personne dont la direction du regard couvre une partie du linéaire.

2.3 Objectifs

L'objectif principal de mon travail de thèse est l'analyse du comportement visuel d'une personne en face d'une scène cible, en déterminant l'objet de son attention visuelle dans un environnement contrôlé.

L'environnement est dit "contrôlé" dans le sens où il est composé d'une manière statique d'une zone observée ou scène cible (e.g. un écran plasma, une image projetée sur un mur, une affiche publicitaire, un linéaire, la vitrine d'un magasin) et d'une zone d'observation où se trouve l'observateur (e.g. rue, couloir, allées d'un supermarché). La Figure 2.2 illustre un exemple d'environnement contrôlé à savoir l'entrée d'un magasin. La zone d'observation est le trottoir et la zone observée est la vitrine du magasin. En détectant les zones d'intérêts du client dans la vitrine, il serait possible d'aider le gérant du magasin à organiser la présentation de sa vitrine ou encore d'identifier le profil de la clientèle à laquelle est destiné un produit (e.g. familles, personnes âgées, handicapés, etc.).

2.4 Schéma général de l'approche

Nous avons proposé une approche qui permet d'atteindre les objectifs cités précédemment. Les données vidéo sont collectées dans un environnement contrôlé composé d'une zone d'observation et d'une zone observée. La détection et le suivi des personnes qui passent devant la scène cible (éventuellement ceux qui s'arrêtent) sont effectués afin de déterminer la position de la tête. Ensuite, une estimation de l'orientation de la tête est réalisée. L'orientation de

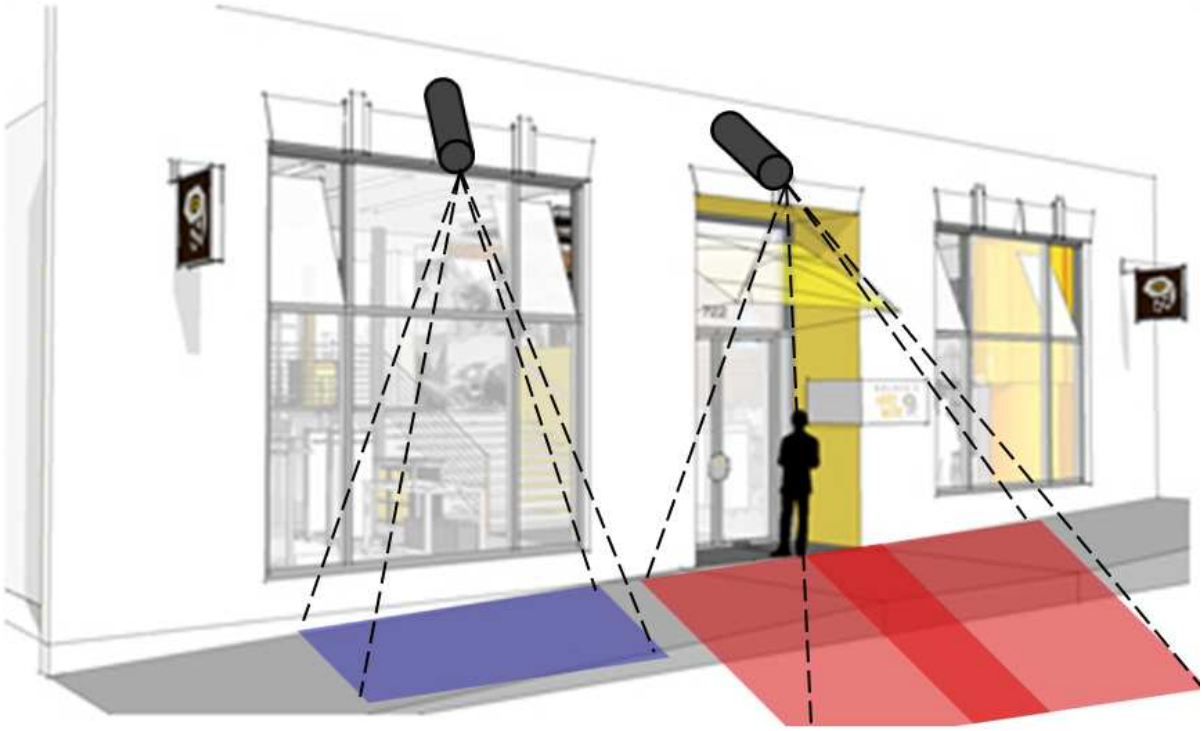


FIGURE 2.2: Exemple d'un environnement contrôlé.

la tête est couplée aux données physiologiques sur la vision humaine permettent d'estimer le champ visuel. La projection de ce champ visuel est utilisée pour extraire la région d'intérêt de la personne dans la scène cible. Lorsque la localisation des yeux peut être effectuée, le point de fixation obtenu à partir de la projection visuelle est corrigé afin d'atteindre une meilleure précision. Enfin, des statistiques calculées sur ces régions d'intérêts sont utilisées pour réorganiser la scène cible. Le schéma général de cette approche est illustré par la Figure 2.3.

2.5 Contribution et originalité

Les principales contributions de la thèse sont les suivantes :

– Orientation de la tête : une méthode d'estimation basée sur l'apparence globale est pro-

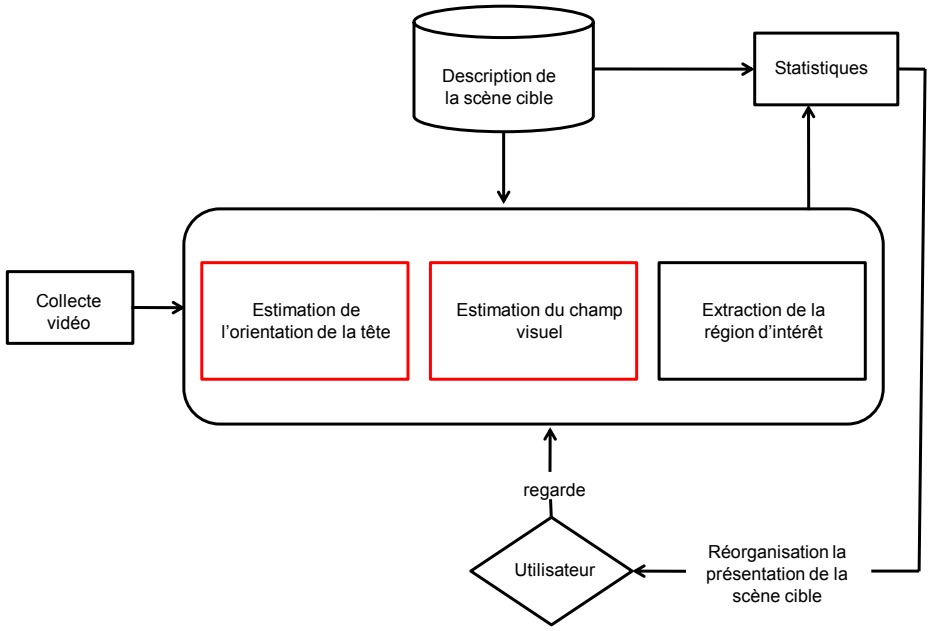


FIGURE 2.3: Schéma général de l'approche.

posée pour estimer l'orientation de la tête lorsque la personne se trouve distant de la caméra. Cette méthode permet aussi la réinitialisation du suivi de la tête (en cas de perte). Enfin, un modèle cylindrique d'estimation de l'orientation de la tête à travers le temps est présenté.

– Champ visuel : des données physiologiques sont utilisées pour déterminer les angles d'ouverture horizontale et verticale du champ visuel. Une méthode de projection géométrique est élaborée pour extraire la région d'intérêt d'une personne dans une scène cible. Cette projection est adaptée à l'estimation du champ visuel basée sur l'orientation de la tête. Elle peut aussi être adaptée à l'estimation du champ visuel basée sur la localisation des yeux voire une combinaison de l'orientation de la tête et de la position des

yeux.

– Analyse des données : un ensemble de métriques est élaboré pour récolter des informations sur la scène cible, sous forme de régions d'intérêts, heatmap, chemins (scanpath), et points de fixation. Les valeurs de ces métriques sont analysables afin de mieux comprendre le comportement visuel des personnes et leurs intérêts dans la scène cible. Cette analyse est effectuée sur des flux vidéo.
– Détection et suivi de personnes : une méthodologie est proposée pour construire et paramétrer le modèle le plus adéquat à la détection et au suivi de personnes. Elle est basée sur l'extraction de l'avant plan à partir d'une modélisation de l'arrière plan. Cette méthode est une contribution du suivi du regard de personnes en mouvement pour pouvoir détecter la tête.

2.6 Plan de la thèse

Après cette introduction de ce mémoire de thèse qui fait ressortir le schéma général de l'approche proposée, le Chapitre 3 présente un état de l'art sur l'estimation de la direction du regard. Nous présentons tout d'abord le système visuel humain et les étapes nécessaires à la formation du regard qui se base sur l'orientation de la tête et la position des yeux. Ensuite, nous détaillons les techniques disponibles pour le suivi du regard en fonction de la catégorie du système utilisé (intrusif ou non intrusif) après un bref historique. Nous déterminons ensuite la contribution de l'orientation de la tête et de la localisation des yeux sur l'estimation de la direction du regard. Enfin, nous présentons différentes applications qui utilisent l'information provenant du regard.

Dans le Chapitre 4, nous présentons tout d'abord l'estimation de l'orientation de la tête. Les capacités humaines à effectuer cette tâche sont ensuite discutées. Ensuite, nous passons en revue les méthodes utilisées pour la collecte de bases d'images nécessaires à la construction des modèles et à leurs validations. Enfin, deux approches pour l'estimation de l'orientation de la tête sont présentées : la première utilise un modèle basé sur l'apparence globale et la seconde basée sur un modèle cylindrique.

Dans le Chapitre 5, nous présentons la méthode utilisée pour estimer le champ visuel d'une

personne basée sur les données physiologiques de la vision humaine. Après adaptation du champ visuel à l'orientation de la tête, la région d'intérêt et le point de regard sur la scène cible sont extraits grâce à la géométrie projective. Enfin, une analyse des régions d'intérêts est effectuée afin de mieux comprendre le comportement visuel des personnes sur la scène cible. Enfin, une comparaison des résultats issus des deux systèmes est effectuée.

Dans le Chapitre 6, nous présentons les caractéristiques de l'environnement qui permettent de calibrer et d'accélérer le traitement, et nous passons en revue brièvement les méthodes de détection et de suivi de personnes. Ensuite, nous présentons la méthode que nous avons utilisée pour détecter les personnes. Celle-ci se base sur une modélisation de l'arrière plan en utilisant un mélange de Gaussienne. Nous présentons ensuite les deux approches utilisées pour le suivi de personnes : une approche basée sur la mise correspondance et une autre approche basée sur le suivi de blobs à l'aide de l'apparence.

Nous concluons ce mémoire de thèse au Chapitre 7 en résumant les contributions scientifiques et les résultats obtenus. Nous présentons aussi des perspectives de recherche.

Chapitre 3

Estimation de la direction du regard

Sommaire

3.1 Introduction

La direction du regard d'une personne apporte une information utile pour l'accomplissement des tâches visuelles. Deux facteurs contribuent à la direction du regard : l'orientation de la tête et la position des yeux. Généralement, l'homme estime la direction du regard à partir de l'orientation de la tête lorsqu'il est à une certaine distance de la cible visuelle [LHT04].

Partant de ce constat général, nous détaillons dans ce chapitre le système visuel humain et les étapes nécessaires à la formation du regard. Ensuite, nous présentons un historique des travaux relatifs au suivi du regard ainsi que les techniques employées selon le type du système utilisé (intrusif ou non intrusif). Nous présentons ensuite quelques applications en illustrant comment les informations provenant du regard sont exploitées. Enfin, la contribution de l'orientation de la tête dans l'estimation de la direction du regard est étudiée, tout comme l'estimation du regard basée uniquement sur la position des yeux.

3.2 Formation du regard

La vision est l'un des cinq sens que possède l'être humain. Elle lui permet de voir et de décrire les objets qui composent le monde qui l'entoure. Cette facilité de perception est obtenue grâce à un processus complexe réalisé par deux organes que sont l'œil et le cerveau. En effet, l'œil reçoit les rayons lumineux qui forment environ un million de points sur la rétine [BGG96]. Ces points contiennent des informations sur la quantité de lumière et de couleur provenant de l'environnement. Ils forment ainsi l'image 2D qui est conduite sous forme de signaux le long des voies neuronales vers un appareil cortical d'analyse. Enfin, le cerveau interprète les différents signaux reçus pour décrire ce qui nous entoure afin de donner une signification à l'image perçue par les yeux. La Figure 3.1 présente schématiquement une coupe horizontale de l'œil humain où le rôle des tissus transparents (la cornée et le cristallin) est de permettre la formation d'une image nette sur la rétine.

Différentes recherches en physiologie ont montré que le regard d'une personne peut être déduit d'une combinaison de l'orientation (ou pose) de la tête et de la position des yeux. En effet, Langton et al. [LHT04] ont établi que l'interprétation du regard effectuée par un

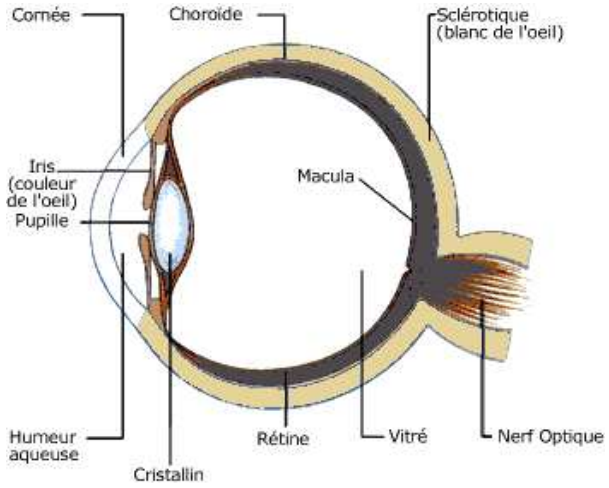


FIGURE 3.1: Coupe schématique horizontale de l'œil humain.

observateur est influencée par la direction de la tête de la personne observée. Pour cela, ils ont utilisé une composition numérique d'images des yeux prises selon différentes directions avec une variété d'orientations de tête. Un exemple frappant de cet effet est illustré par une image datant du XIX^{e} siècle, présentée dans la Figure 3.2 [Wol24].

FIGURE 3.2: Illusion de WOLLASTON : bien que les yeux soient rigoureusement identiques dans les deux vues, la perception de la direction du regard est influencée par l'orientation de la tête.

Dans cette figure, deux vues du visage d'une même personne sont présentées avec une configuration identique de la position des yeux mais sous deux orientations différentes de la tête. L'analyse de ces deux vues illustre l'influence de l'orientation de la tête sur la perception de la direction du regard. En effet, la suppression de la partie inférieure du visage (en gardant uniquement les yeux) fait passer la perception de la direction du regard en une configuration frontale de la tête.

3.3 Historique du suivi du regard

L'étude du regard est une pratique ancienne qui a largement précédé l'apparition des premiers ordinateurs et le développement des technologies de l'information qui s'en est suivi. En effet, les premiers travaux dans le domaine sont recensés vers la fin du XIX^{e} siècle. Javal [Jav78] a utilisé un système de miroirs pour observer le mouvement oculaire de personnes en train de lire. Cependant, ces travaux étaient techniquement limités car les sujets étaient invités à retranscrire eux-mêmes les endroits où ils ont regardés, rendant la validité des résultats ainsi obtenus fortement discutables.

En 1901, Dodge et Cline [DC01] ont conçu un système qui utilise la réflexion sur la cornée d'un faisceau lumineux orienté verticalement pour enregistrer les mouvements horizontaux de l'œil à l'aide d'une plaque photographique placée derrière une fente horizontale. Malgré le fait qu'elle nécessite une immobilité totale du sujet, cette méthode fera date dans l'histoire du suivi des yeux (*eye tracking*) car elle fut utilisée - avec quelques améliorations - jusqu'aux années 1970. En 1905, Judd et al. [JMS05] ont utilisé un système de capture photographique (ancêtre de la camera) pour enregistrer le mouvement des yeux. En 1921, Gilliland [Gil21] a effectué le premier enregistrement du mouvement oculaire en deux dimensions en réussissant à décomposer la réflexion cornéenne d'un faisceau lumineux en ses composantes horizontales et verticales. Cette méthode a permis à Buswelle [Bus35] en 1935 d'obtenir les premières traces oculaires (i.e. la position du regard superposée sur le support visuel qui l'a engendrée à un instant donné). Durant cette période, l'essentiel des travaux se situaient dans le cadre de la physiologie et de la psychologie cognitive, afin d'étudier notamment le mouvement oculaire durant la lecture [Hue00] ou bien le mouvement des yeux des pilotes durant les atterrissages

[FJM50].

Avec l'arrivée de la télévision, le premier système qui permet de superposer une trace oculaire sur une scène visuelle mobile a été fabriqué en 1958 [MM58]. Au début des années 1970, ce domaine a pris un nouvel essor avec l'orientation des recherches vers l'analyse des données collectées au détriment de la conception de nouveaux systèmes d'acquisition. De nouvelles théories reliant le mouvement des yeux à des processus cognitifs ont commencé alors à émerger [MS76]. Le développement de l'informatique vers la fin des années 1970 a enfin permis d'obtenir des capacités de calcul nécessaires au traitement des données en temps réel. Le mouvement des yeux fut alors utilisé comme nouveau moyen d'interaction avec l'ordinateur [Anl76]. Les années 1980 furent le berceau des premières études visant à analyser l'information véhiculée par le regard. En effet, les avancées technologiques ont permis de coupler les systèmes oculométriques à des ordinateurs et ainsi de suivre en temps réel les processus cognitifs et le comportement d'un utilisateur face à une scène visuelle. Les premières applications furent notamment destinées aux personnes handicapées [HJM$^+$89].

Les années 1990 virent émerger de nouveaux domaines d'application orientés vers le Travail Collaboratif Assisté par Ordinateur (TCAO), la réalité virtuelle, et la réalité augmentée [VVS98]. Dans les années 2000, à l'ère du numérique, des dispositifs d'enregistrement des mouvements oculaires précis sont disponibles. Ils sont aussi bien utilisés en ergonomie cognitive pour valider des interfaces informatiques qu'en marketing pour mesurer l'impact de la présentation de produits. Actuellement, la tendance est à l'utilisation d'une simple webcam pour estimer la direction du regard [LMID09].

3.4 Techniques de suivi du regard

Depuis les premières études menées durant les années 1900, les méthodes de mesures des mouvements oculaires ont beaucoup évolué. Il existe plusieurs outils de mesures ayant chacun des caractéristiques adaptées à l'utilisation qui en sera faite. La contrainte la plus discriminante se fait généralement sur l'outil de capture qui ne doit pas gêner l'utilisateur dans ses mouvements. En effet, tout système qui requiert le port d'un équipement par l'utilisateur ou qui génère des perturbations visuelles ou sonores est appelé système intrusif (ou invasif). Les

sections suivantes présentent les techniques d'estimation de la direction du regard en fonction du type de système utilisé : intrusif ou non intrusif.

3.4.1 Systèmes intrusifs

Les systèmes intrusifs permettent généralement d'obtenir des résultats précis. Nous présentons dans ce qui suit certains d'entre eux :

3.4.1.1 Electro-oculographie

Cette technique est basée sur une propriété de l'œil qui est la présence d'un champ électrostatique dont les caractéristiques sont liées à la position de la cornée par rapport à la rétine. Ainsi, en plaçant un certain nombre d'électrodes autour de l'œil (voir Figure 3.3) et en enregistrant les différences de potentiels, il est possible de mesurer l'orientation de l'œil avec une précision de l'ordre de 0.5˚ à 1˚ sur un large domaine d'orientations ($\pm$ 70˚) [Hal86]. Cependant, cette technique ne permet pas d'obtenir la position du point de regard (ou point de fixation) sur un support visuel. En effet, le mouvement est mesuré relativement à la position de la tête et non par rapport à un repère extérieur. Cette méthode est fastidieuse à mettre en place (pose d'électrodes sur le visage) et relativement pénible pour le sujet. Encore très largement répandue dans le domaine médical, elle permet le diagnostic de symptômes ophtalmologiques tels que le strabisme chez l'enfant.

FIGURE 3.3: Électrodes permettant l'électro-oculographie.

3.4.1.2 Lentilles de contacts à bobines magnétiques

Cette technique, inventée par Robinson [Rob63] en 1963, permet de mesurer directement le mouvement des globes oculaires. Elle nécessite le placement de deux lentilles de contacts spécifiques à large diamètre sur les yeux du sujet. Ces lentilles spéciales comportent un petit réceptacle en forme d'anneau dans lequel est placée une fine bobine de fil électrique. Chaque lentille est reliée par un fil électrique très fin à un instrument qui mesure le courant induit dans la bobine, généré par les champs magnétiques émanant des deux grandes bobines électriques placées perpendiculairement l'une par rapport à l'autre autour de la tête du sujet. Il est ainsi possible d'enregistrer la position exacte de la lentille après étalonnage du système (le courant dépend de la position de l'œil par rapport aux deux grandes bobines). Bien que cette mesure soit sans doute la plus précise, elle reste néanmoins réservée à une utilisation clinique. La lentille est extrêmement inconfortable pour le sujet et sa pose reste très délicate (voir Figure 3.4). De plus, elle peut provoquer des réactions allergiques de l'œil.

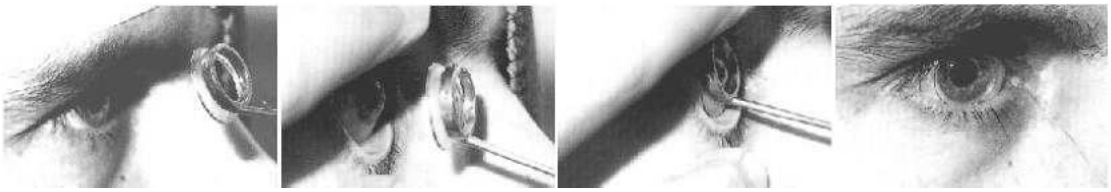


FIGURE 3.4: Pose de lentilles de contacts à bobines magnétiques.

3.4.1.3 Localisation du limbe

Le limbe est la frontière entre le blanc de l'œil (sclère) et l'iris. La différence de luminosité de ces deux parties de l'œil permet de détecter et de suivre facilement ce limbe. A l'aide d'un simple instrument de mesure de luminosité constitué de diodes et de capteurs infrarouges monté sur des lunettes, il est possible de réaliser un système de mesure peu couteux. Des systèmes commercialisés dont la précision est de l'ordre de 0.1° (en mesure horizontale ou verticale seule), ou de 0.5° en horizontal et 1° en vertical (mesurés simultanément) sont disponibles. L'occultation occasionnelle des parties supérieures et inférieures du limbe par les

paupières rend cependant ces mesures instables [GEN95]. De plus, le sujet doit rester immobile car cette technique ne permet pas de mesurer la position de sa tête.

3.4.1.4 Analyse de l'image de l'œil

Cette technique consiste à analyser l'image de l'œil provenant d'une caméra afin de localiser une de ses composantes (e.g. la pupille ou l'iris). Le suivi d'une composante permet de mesurer les mouvements de l'œil réalisés par rapport à la tête du sujet qui doit être fixe par rapport à la caméra. Pour cela, il est possible de fixer la caméra sur la tête du sujet ou d'utiliser un miroir semi-réfléchissant. La technique la plus courante nécessite une lumière infrarouge projetée sur l'œil. Quatre reflets de cette source lumineuse sont engendrés : les deux premiers sont dus à la cornée et les deux suivants au cristallin. Ces réflexions lumineuses portent le nom de celui qui les a découvertes : "images de Purkinje", et sont illustrées par la Figure 3.5.

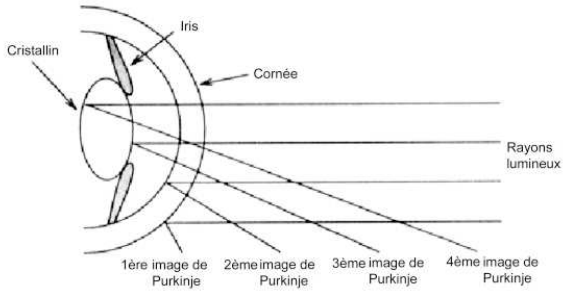


FIGURE 3.5: Un rayon lumineux se reflète de quatre manières différentes sur l'œil. Ces réflexions sont appelées images de Purkinje, du nom du chercheur qui les a mises en évidence en 1832.

Ces reflets sont exploités de deux manières :

1. **Reflet cornéen/pupille** : Cette technique consiste à placer, sous un écran d'ordinateur et en direction des yeux d'un sujet, une caméra équipée d'une diode infrarouge au centre de son objectif. La Figure 3.6 illustre la configuration d'un tel système. A l'aide de cette diode, un rayon lumineux infrarouge (donc invisible) éclaire l'œil de l'utilisateur. Ce

rayon provoque plusieurs phénomènes distincts qui permettent de déterminer la direction du regard.

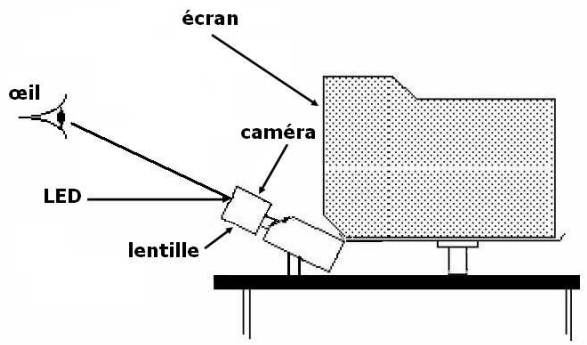


FIGURE 3.6: Schéma d'un système d'*eye tracking*.

Le rayon infrarouge est réfléchi par la cornée, provoquant un reflet relativement intense et de petite taille, facilement détectable, appelé première image de Purkinje ou Glint (voir la Figure 3.7). Ce rayon traverse également les différents constituants du globe oculaire pour être réfléchi par la rétine de l'œil. Il provoque ainsi un rougissement de la pupille que tous les amateurs de photographie connaissent bien. Ce phénomène permet de distinguer la pupille de l'iris qui l'entoure à l'aide de l'augmentation du contraste entre ces deux éléments.

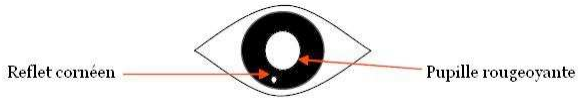


FIGURE 3.7: Première image de Purkinje avec surbrillance de la pupille.

Un programme spécialisé dans le traitement d'images identifie la position de la pupille et de l'iris, et calcule la position relative de l'un par rapport à l'autre (voir Figure 3.8). En effet, un changement dans la direction du regard implique une variation minime de la position du reflet cornéen. Ceci fournit un excellent repère de référence pour le calcul de la position de la pupille. De manière opposée, la direction du regard suit la pupille

et donne un renseignement sur la localisation de l'endroit regardé. Ainsi, le calcul de la position relative entre ces deux éléments à chaque instant permet de déterminer les positions successives de l'œil. Il reste ensuite à associer chaque position (reflet cornéen/pupille) captée à un point de la scène cible ou de l'écran regardé par le sujet. Pour cela, une courte phase de calibration est nécessaire.

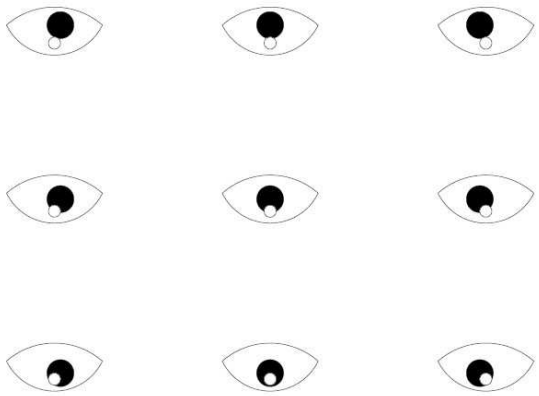


FIGURE 3.8: Calcul de la position du regard par détection de la pupille et de la première image de Purkinje.

La phase de calibration permet d'associer la position relative de la pupille et du reflet cornéen à des coordonnées sur l'écran. Pour cela, l'utilisateur doit fixer un point affiché à différentes positions afin de quadriller l'écran. Une fois la correspondance entre les positions de la pupille et les coordonnées de l'écran qui leurs sont associées connue, il ne reste plus qu'à extrapoler ces coordonnés en fonction des positions intermédiaires de la pupille. La projection de la fovéa (la zone centrale de la macula qui est la zone de la rétine où la vision des détails est la plus précise) sur l'écran couvre environ 1° à 1.5°. La position spatiale de la zone regardée ne peut donc pas être considérée comme un point mais comme une surface, ce qui implique une évaluation limitée par la taille de cette surface. Des travaux récents ont montré la possibilité d'éviter cette phase de calibration par l'utilisation d'une deuxième caméra et de l'image de l'autre œil, à l'aide du calcul du point de convergence des yeux.

Cette technique est certainement la plus largement utilisée dans les systèmes d'*eye tra-*

cking commerciaux appelés également oculométres [MLID08]. Cependant, ces systèmes ne tolèrent que de très légers mouvements de la tête. En effet, le sujet doit rester quasiment immobile. La marge d'erreur est de l'ordre de 2° à 5° selon les modèles, et reste plus grande qu'avec les autres méthodes intrusives. Cette erreur diminue avec l'amélioration des caméras.

2. **Dual Purkinje Image** : Cette technique se base sur le même principe que celui du reflet cornéen sauf que dans cette méthode c'est la position de la première et de la quatrième image de Purkinje qui sont détectées. Cela permet d'évaluer la rotation de l'œil indépendamment des translations horizontales ou verticales. L'utilisation d'un miroir mobile asservi à ces translations permet de garder une image de l'œil centrée par rapport à la caméra. Cependant, la quatrième image de Purkinje est difficile à suivre, notamment quand la pupille est trop petite. Il est donc important de pouvoir contrôler les conditions lumineuses lors des mesures. Les dispositifs basés sur cette technique sont plus précis que ceux basés sur le reflet cornéen/pupille (marge d'erreur de 1°), mais nécessitent un matériel spécifique plus coûteux.

3.4.2 Systèmes non intrusifs

Les systèmes non intrusifs peuvent être classés en fonction du nombre de caméras utilisé. Les systèmes qui utilisent une caméra sont appelés monoculaires alors que ceux qui utilisent deux sont appelés binoculaires. Une étude récente qui présente les travaux dans le domaine est disponible dans [HJ10].

3.5 Applications

L'information provenant du regard d'une personne est importante dans de nombreux domaines. Elle permet d'analyser son comportement et d'avoir une meilleure compréhension de la scène cible. Nous allons présenter dans ce qui suit quelques applications qui utilisent ou analysent l'information provenant du regard d'une personne en mettant l'accent sur celles qui sont utilisées pour valider nos résultats.

Interaction dans une réunion : Le regard joue un rôle important dans l'interaction entre les humains. Ces interactions sont étudiées en psychologie sociale depuis plusieurs décennies [McG84] car le regard accomplit plusieurs tâches (e.g. l'établissement de relations à travers le regard mutuel, l'expression de l'intimité [AD65], et l'exercice d'un contrôle social [LWB00]). Les personnes ont tendance à ne regarder que les objets qui sont d'un intérêt immédiat pour eux. Vertegaal et al. [VSvdVN01] ont montré que dans une conversation regroupant 4 personnes, lorsque le regard de la personne qui s'exprime croise les yeux d'une autre personne, elle est dans 80% des cas la cible de son discours. De plus, quand un auditeur regarde une autre personne dans les yeux, c'est dans 77% des cas la personne qui s'exprime. Ainsi, l'information provenant du regard des autres est utilisée pour déterminer à quel moment ils doivent prendre la parole ou bien si le message prononcé leur est adressé. Par exemple, le regard est un moyen utilisé pour trouver le moment approprié pour demander un tour de parole [NHW96], alors qu'en fin de phrase, il peut être interprété comme une demande de réponse [JodA04]. L'importance de ce thème de recherche est telle que les campagnes d'évaluation 2006 et 2007 de CLEAR [Wora] lui ont été dédiées.

Surveillance des conducteurs : Les systèmes avancés d'assistance à la conduite peuvent sauver de nombreuses vies en aidant les conducteurs à prendre rapidement des décisions sécuritaires lors d'une manœuvre. Chaque année, les accidents de la circulation provoquent environ 1.2 million de morts dans le monde entier, dus pour 80% d'entre eux à l'inattention des conducteurs [AAA+06]. Pour contrer l'effet de l'inattention, des systèmes d'assistance sont conçus pour fournir au conducteur un avertissement lorsqu'une situation dangereuse est imminente, voire pour l'aider à réagir de façon appropriée. Toutefois, le type et l'emplacement des capteurs (voir Figure 3.9) utilisés pour optimiser les performances sont encore à l'étude. Ils sont souvent associés à une application particulière [DT09]. Ces capteurs doivent détecter ce qu'il se passe dans les 3 composantes principales d'un système de conduite : l'environnement, la voiture et le conducteur. Les recherches récentes ont appuyé l'intégration des capteurs à l'intérieur du véhicule dans une approche holistique [TGM07] afin d'observer le comportement du conducteur.

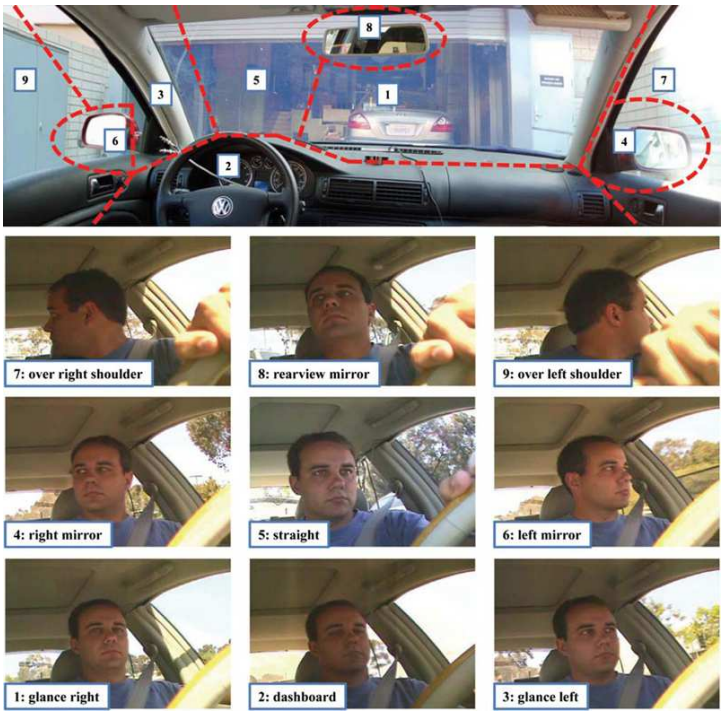


FIGURE 3.9: (Image du haut) Répartition approximative des zones où se pose le regard pour l'étiquetage. (Images du bas) Exemples correspondant au regard se posant sur les 9 zones étiquetées.

Réalité virtuelle : La plupart des approches en réalité virtuelle qui se basent sur le regard comme moyen de navigation et d'interaction utilisent des unités d'affichage montées sur la tête (*Head Mounted Displays*). La connaissance de la direction du regard peut informer en temps réel ou hors ligne les systèmes d'optimisation permettant une amélioration visuelle notable du rendu 3D. Hillaire et al. [HLCC08] ont montré comment le suivi du regard en ligne peut être utilisé pour améliorer les effets visuels (e.g. la profondeur du champ visuel, les mouvements de la caméra, etc.). L'expérience menée par Tanriverdi et Jacob [TJ00] atteste que la sélection d'objets à visualiser sur l'écran s'effectue plus rapidement en utilisant une approche fondée sur le regard par rapport à un geste de pointage avec le doigt. Le regard est très utilisé dans les environnements collaboratifs virtuels afin d'effectuer une tâche précise (e.g. la résolution

d'un puzzle [SOM+09] illustrée par la Figure 3.10). Une évaluation des systèmes binoculaires dédiés au suivi des yeux pour l'interaction avec le regard dans des environnements de réalité virtuelle est disponible dans [PLW08].

FIGURE 3.10: Résolution d'un puzzle en réalité virtuelle.

Interaction Homme-Machine L'utilisation du regard pour interagir avec l'ordinateur est très en vogue. L'estimation du regard est devenue un outil très intéressant pour améliorer notamment les possibilités d'interaction avec l'ordinateur pour les personnes handicapées. Hirotaka et al. [AHI08] ont examiné le processus d'apprentissage de 8 étudiants qui utilisent leur regard pour saisir 110 phrases, alors que Kammerer et al. [KSB08] ont évalué l'efficacité de deux approches utilisées pour effectuer une sélection de menus à l'aide du regard.

3.5.1 Extraction des cartes de saillance (pertinence) dans une image

Dans les applications en relation avec le graphisme, le design, voire même pour l'indexation et la recherche d'information, il est essentiel de pouvoir déterminer où la personne regarde dans une scène, que ce soit pour la réduction de la taille d'une image (*cropping*) [SAD+06], la compression directe d'images et de vidéos en résolutions multiples [WLB03], la détection des détails pour le rendu non réaliste d'images [DS02] (voir Figure 3.11), ou pour le recadrage intelligent de l'image (*seam carving*).

FIGURE 3.11: Rendu non réaliste d'images.

Un système de suivi des yeux est souvent intégré dans le processus pour enregistrer les fixations d'un utilisateur assis devant un ordinateur muni d'un *eye tracker*. Les données récoltées sont ensuite transmise à la méthode choisie qui utilise un modèle d'apprentissage pour créer une carte de pertinence [LXG09, KWSF06]. Cependant, étant donné qu'un système d'*eye tracking* ne peut être toujours disponible, il est nécessaire d'avoir un moyen de prédire où les utilisateurs vont regarder sans utiliser un tel système. Comme alternative, des modèles qui permettent de mesurer la visibilité d'une région ou la probabilité qu'une région puisse attirer l'attention d'un observateur ont été utilisés [AL09, BT09, HZ07, IK00]. Ces modèles s'inspirent de la biologie et d'une méthode de calcul dite *bottom-up*. Pour ce faire, plusieurs caractéristiques de bas niveaux (e.g. l'intensité de couleur, l'orientation, la texture, etc.) sont extraites de l'image à différentes échelles. Une carte de pertinence est construite pour chacune des caractéristiques. Ces cartes sont ensuite normalisées et combinées de manière linéaire ou non linéaire en une seule carte de pertinence générale qui représente l'importance de chaque pixel. Par exemple, Ehinger et al. [EHSTO09] ont proposé un modèle basé sur la combinaison de trois facteurs (carte de pertinence associée à des caractéristiques de bas niveau, les caractéristiques de la cible, et le contexte de la scène) pour prédire les zones où les personnes regardent dans une tâche de recherche de piétons.

3.5.2 Marketing pour les magasins

La concurrence sans cesse croissante incite les magasins détaillants à s'intéresser de plus en plus à la compréhension du comportement et du processus de décision d'achat de leurs clients. Traditionnellement, cette information ne peut être obtenue que par l'observation directe des clients ou indirectement par l'intermédiaire de groupes de discussion ou d'expériences spécialisées dans des environnements contrôlés. En revanche, la vision par ordinateur a le potentiel de répondre à ces questions à partir d'analyses de vidéos qui présentent des clients tests pendant l'exploration d'un magasin. Ce type d'application est plus rentable que celui destiné à la prévention des pertes. L'obtention d'un aperçu de la circulation et du comportement des clients dans un magasin est d'un grand intérêt pour le marketing, les opérations promotionnelles et l'exploitation des données. Il est particulièrement intéressant d'analyser le processus de prise de décision d'un achat et répondre à des questions telles que : Qu'est ce qui a capté l'attention du client ? Quels produits sont passés inaperçus ? Que regarde un client avant de prendre une décision d'achat ? Alors que la plupart des caméras de surveillance présentes dans les magasins sont utilisées pour suivre le parcours des clients, les nouvelles technologies essayent plutôt d'estimer la direction du regard des clients. Cette catégorie d'applications a été utilisée pour valider nos différentes approches.

Plusieurs approches destinées au suivi des clients dans des environnements de vente au détail ont été proposées. Haritaoglu et al. [HBF02] ont utilisé des caméras stéréoscopiques positionnées en haut et pointées vers le bas pour le suivi de la position et des actions des clients. La stéréovision possède l'avantage de pouvoir facilement séparer les clients des caddies parmi les *blobs* en mouvement mais exige la présence de capteurs stéréos dédiés qui sont rarement disponibles dans un magasin. Une autre approche utilisée pour compter les clients utilisant la stéréovision est présentée dans [Bey00]. Krahnstoever [KRT+05] a utilisé un système de stéréovision pour le suivi des interactions entre les clients et les produits à travers le suivi de la position de la tête et de la main. Le système utilise également les informations obtenues de puces RFID placées dans les produits pour les détecter et suivre leurs mouvements. Mustafa et Sethi [MS05] ont utilisé une approche basée sur le déplacement des bords pour le suivi des vendeurs entrant ou sortant de l'arrière boutique d'un magasin. L'attention des clients qui

attendent dans une queue par rapport à des panneaux d'affichage a été étudiée dans [HF02]. Liu et al.[LKYT07] estiment ce que regardent les personnes dans un magasin.

3.6 Contribution de l'orientation de la tête dans la direction du regard

Afin de déterminer le lieu où regarde une personne, il faut estimer la direction de son regard et ensuite le projeter dans cette direction sur la scène cible. Le regard est souvent défini comme étant la direction où les yeux sont dirigés dans l'espace. Il s'agit alors d'une combinaison de l'orientation de la tête et l'orientation des yeux. Cela signifie qu'il faut estimer les deux orientations pour déterminer avec précision la direction du regard.

La détection des yeux n'est pas facilement réalisable lorsqu'une personne est éloignée de la caméra, dépendant de la résolution. Dans ce cas, l'unique information sur le regard provient de l'orientation de la tête. Ceci amène à se poser la question suivante : comment peut-on prédire où une personne regarde en se basant uniquement sur l'orientation de sa tête ?

Pour répondre à cette question, une étude [SZ02] a été menée sur l'analyse du regard de quatre personnes dans une réunion en utilisant un matériel spécial pour mesurer l'orientation de la tête et la direction du regard. Les objectifs de cette étude sont de déterminer la contribution de l'orientation de la tête dans la direction du regard et de calculer la précision avec laquelle il est possible d'affirmer qu'une personne regarde une autre personne en se basant uniquement sur l'orientation de sa tête. Cette étude est détaillée car les résultats obtenus sont utilisés dans le Chapitre 5.

3.6.1 Base de données utilisée

Le scénario proposé pour cette étude réunit quatre personnes autour d'une table ronde. Une session de données d'une dizaine de minutes est recueillie pour chaque participant. Dans chaque session, un participant porte un appareil monté sur sa tête [1]. Le système utilise un

1. fabriqué par ISCAN Inc (http ://www.iscaninc.com)

sous-système de suivi magnétique pour suivre la position et l'orientation de la tête et un sous-système composé d'une caméra montée sur la tête du participant pour enregistrer les images de ses yeux. Le logiciel fourni avec ce système permet d'enregistrer à une fréquence de 60 Hz les données suivantes avec une précision de moins d'un degré : la position de la tête, l'orientation de la tête, l'orientation de l'œil, le clignement des yeux, et la direction globale du regard (ligne de vue). La Figure 3.12 [SZ02] illustre une image prise au cours de l'expérimentation.

FIGURE 3.12: Image de la collecte de données.

3.6.2 Calcul de la contribution de l'orientation de la tête

Le calcul de la contribution de l'orientation de la tête dans la direction du regard consiste à analyser sa contribution par rapport à la contribution des yeux sur l'axe horizontal. Il a été constaté dans 87% des images que la direction de l'orientation de la tête et celle des yeux étaient similaires (droite ou gauche). Pour les images où la direction est similaire, le calcul de la contribution de l'orientation de la tête dans la direction du regard est effectué. La composante horizontale de la ligne de regard los_x est la somme du déplacement horizontal de l'orientation de la tête ho_x et de l'orientation des yeux eo_x. Le pourcentage de la contribution de l'orientation

de la tête *HC* par rapport à la direction globale du regard est calculée comme suit :

$$HC = \frac{ho_x}{los_x}$$

Les résultats obtenus au cours de ces 4 sessions d'expérimentation sont présentés dans le Tableau 3.1 :

Sujet	**Nb Images**	**Clignement des yeux**	**Même sens**	**Contribution de la tête**
1	36003	25.4%	83.0%	62.0%
2	35994	22.6%	80.2%	53.0%
3	38071	19.2%	91.9%	63.9%
4	35991	19.5%	92.9%	96.7%
Moyenne	-	21.7%	87.0%	68.9%

TABLE 3.1: Contribution de l'orientation de la tête dans la direction globale du regard.

A partir des résultats de cette expérimentation, plusieurs remarques intéressantes peuvent être faites :

- La plupart du temps, les participants tournent leurs têtes et leurs yeux dans la même direction pour regarder ce qui les intéresse.
- L'utilisation de l'orientation de la tête pour changer la direction du regard varie beaucoup entre les participants (de 53% pour le participant 2 jusqu'à 96% pour le participant 4 avec une moyenne de 68.9%) mais reste néanmoins assez élevée (plus de la moitié de la direction globale du regard).
- Le clignement des yeux se produit dans environ 20% du nombre total d'images. Cette information est importante car elle montre la limite des équipements commerciaux qui estiment la direction globale du regard à partir des yeux. En effet, ils ne fonctionnent pas pendant environ un cinquième du temps (position des yeux inconnue).

Ces observations permettent de conclure que l'orientation de la tête est un facteur déterminant et parfois même suffisant disponible pour estimer la direction du regard d'une personne.

3.6.3 Prédiction de la cible

La direction du regard est utilisée dans l'expérimentation de Stiefelhagen [Sti02] pour calculer la fréquence à laquelle la personne qui est regardée par le participant est détectée correctement, en se basant uniquement sur l'orientation de la tête. La Figure 3.13 présente les histogrammes de la composante horizontale de la direction du regard de deux participants. Les trois pics présents dans les histogrammes correspondent à l'orientation utilisée par le participant pour regarder l'emplacement où les trois autres participants étaient assis. Ces pics sont déterminés automatiquement en utilisant l'algorithme des K-moyennes (K-means) [Mac67]. Une étiquette est attribuée à chaque image, basée sur la distance la plus courte entre la ligne associée à la direction du regard et celle des directions associées aux 3 cibles (les 3 autres participants).

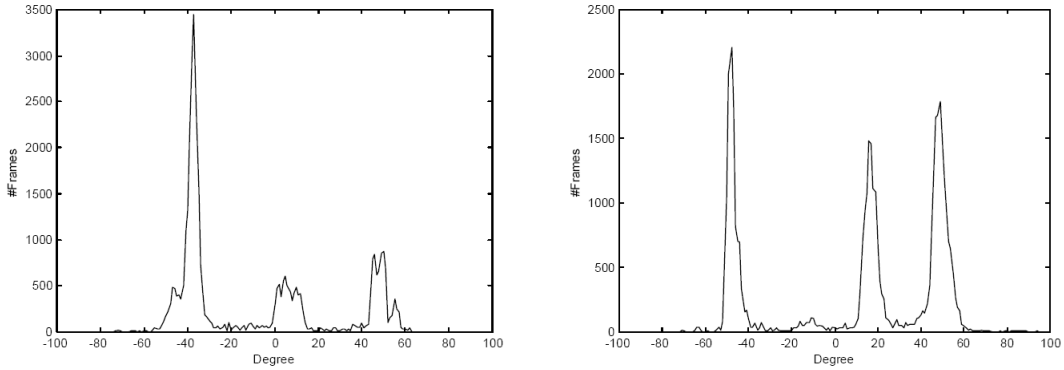


FIGURE 3.13: Histogrammes de la composante horizontale de la direction du regard de deux participants.

Une approche basée sur la modélisation en utilisant une mixture de gaussiennes est utilisée pour estimer la direction du regard, en se basant sur l'orientation de la tête uniquement [SYW99]. Les résultats sont comparés avec les étiquettes obtenues à partir des données de la direction du regard. Le Tableau 3.2 présente la précision obtenue en utilisant uniquement l'orientation de la tête.

Cette expérience montre que l'orientation de la tête est un facteur important pour estimer la direction du regard d'une personne. La précision atteint en moyenne 88.7%, ce qui encourage

Sujet	Précision
1	85.7%
2	82.6%
3	93.2%
4	93.2%
Moyenne	88.7%

TABLE 3.2: Détection de l'attention visuelle basée sur la composante horizontale de l'orientation de la tête.

à utiliser cette information lorsque la position et l'orientation des yeux sont inconnues.

3.7 Estimation de la direction du regard en se basant uniquement sur la localisation des yeux

L'estimation de la direction du regard en se basant uniquement sur les yeux est généralement utilisée lorsque la personne se trouve en face d'un écran d'ordinateur, car les mouvements de sa tête sont limités. Les méthodes proposées dans l'état de l'art tendent à simplifier le problème en supposant que l'œil ne tourne pas mais qu'il se déplace seulement (translation). Cette simplification provient de l'hypothèse que la personne est toujours en pose frontale par rapport à l'écran de sorte à éliminer l'estimation de l'orientation de la tête. L'estimation de l'orientation de la tête est alors ignorée en raison du temps de calcul important nécessaire. L'information concernant la position des yeux et de leurs coins est alors nécessaire [VSG08] et la méthode la plus utilisée est celle suggérée par Zhu et Yang [ZY02]. Ces derniers ont utilisé une méthode basée sur une mise en correspondance linéaire. Une étape de calibration est nécessaire et consiste en l'affichage d'un ensemble donné de points sur l'écran qui doivent être regardés par l'utilisateur. Une mise en correspondance 2D est effectuée à partir du vecteur entre les coins des yeux et le centre de l'iris, et est enregistrée pour des positions connues sur l'écran. Ce vecteur est ensuite utilisé pour effectuer une interpolation sur les points connus de l'écran.

Par exemple, en prenant deux points de calibration $P1$ et $P2$ aux coordonnées sur l'écran α et β, et le vecteur du point central de l'œil x et y, l'interpolation suivante est utilisée pour obtenir les coordonnées sur l'écran d'un nouveau vecteur :

$$\alpha = \alpha_1 + \frac{x - x_1}{x_2 - x_1}(\alpha_2 - \alpha_1)$$

$$\beta = \beta_1 + \frac{y - y_1}{y_2 - y_1}(\beta_2 - \beta_1)$$

L'avantage de cette approche est le coût faible en temps de calcul, ainsi que la précision acceptable par rapport aux systèmes plus complexes. Malheureusement, cette méthode n'autorise pas des mouvements de la tête. Chaque mouvement conséquent horizontal ou vertical nécessite une recalibration du système. Toutefois, si la distance de l'écran et les paramètres intrinsèques de la caméra sont connus, il est possible de compenser ce problème en effectuant une autre mise en correspondance des points utilisés pour le calibrage en fonction du déplacement des yeux. Par conséquent, la précision du système est limitée uniquement par la résolution de la caméra. Cela génère un effet de grille sur données. La précision, quant à elle, est limitée par la qualité de la caméra et sa distance de l'utilisateur. Ce type de système est utilisé pour des applications spécifiques qui ne requièrent pas une grande précision (comme le changement de la fenêtre active ou bien une action particulière quand l'utilisateur regarde en dehors des bords de l'écran).

3.8 Conclusion

Ce chapitre présente un aperçu des différents systèmes utilisés pour l'estimation de la direction du regard. Ces systèmes ont été classés en fonction de la gêne occasionnée chez le sujet par l'outil de capture des données. Les applications qui utilisent l'information provenant du regard selon le domaine ont été présentées en particulier dans le cadre du scénario marketing que nous avons utilisé pour valider nos approches. Enfin, la contribution de la pose de la tête (qui sera détaillée dans le chapitre suivant) et de la position des yeux qui sont les deux facteurs utilisés pour estimer la direction du regard ont été présentés.

Chapitre 4

Estimation de l'orientation de la tête

Sommaire

4.1 Introduction

La capacité d'estimer l'orientation (ou la pose) de la tête d'une personne représente un défi de taille pour les systèmes de vision par ordinateur. A la différence de la détection de visages ou l'identification des personnes à l'aide des caractéristiques faciales, qui ont été les principaux axes de recherche de la communauté "Vision par ordinateur" en relation avec la tête d'une personne, l'estimation de l'orientation de la tête est une thématique récente et très active.

La détection de la *position* de la tête est le procédé permettant de localiser la tête d'une personne dans une image, alors que l'*estimation* de la pose de la tête est le procédé visant à déterminer l'orientation de la tête. Lorsque les yeux d'une personne sont visibles, la détection de la position de la tête et l'estimation de son orientation associées à une analyse de la position des pupilles dans les yeux permettent une estimation plus précise de la direction du regard tel que nous l'avons présenté dans le chapitre précédent.

4.2 État de l'art

Nous allons décrire dans ce qui suit l'estimation de l'orientation de la tête ainsi que les difficultés qui lui sont inhérentes au travers des différentes approches existantes.

4.2.1 Définition

L'estimation de l'orientation de la tête consiste à déterminer la valeur des angles selon les 3 degrés de liberté de la rotation de la tête, à partir d'une image. Les 3 degrés de liberté illustrés par la Figure 4.1 sont définis comme suit :

- **Tilt (Pitch)** : correspond à un mouvement de la tête de haut en bas autour de l'axe X (l'axe des "oui").
- **Pan (Yaw)** : correspond à un mouvement de la tête de gauche à droite autour de l'axe Y (l'axe des "non").

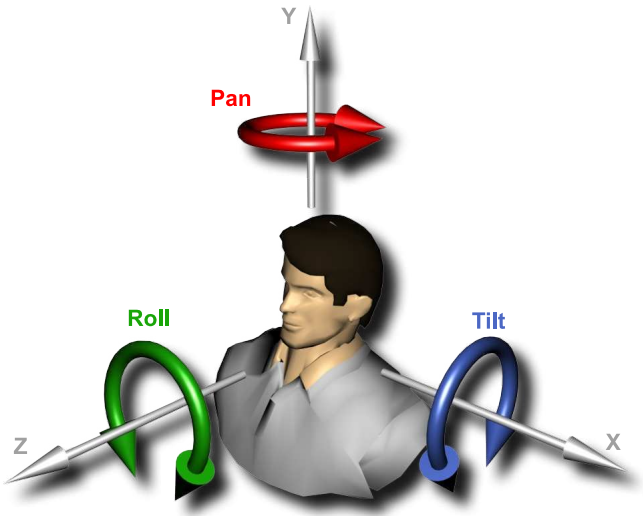


FIGURE 4.1: Les 3 degrés de liberté de la tête.

– **Roll (Slant)** : correspond à un mouvement de la tête de l'épaule gauche vers l'épaule droite autour de l'axe Z (l'axe qui permet de poser une oreille sur l'épaule).

4.2.2 Capacités humaines à estimer l'orientation de la tête

Les fondements psychophysiques des aptitudes humaines à estimer l'orientation de la tête demeure en majeure partie inconnue. On ignore si les humains ont une capacité naturelle à estimer les 3 degrés de liberté de la tête ou s'ils l'acquièrent avec l'expérience. Néanmoins, il existe quelques données qui permettent de mesurer les compétences humaines à accomplir cette tâche. En effet, Kersten et al. [KTB96] utilisent les poses de face et de profil comme poses clés car elles sont activées inconsciemment par le cerveau mais pas les autres. La Figure 4.2 illustre l'image utilisée par les auteurs. Elle représente une projection cylindrique d'un visage utilisée dans les compétitions de poses car toutes les orientations horizontales de la tête y sont présentes.

Gourier [Gou06] a conduit une expérience qui consiste à demander à un groupe de personnes d'estimer l'orientation de la tête sur des images de *Pointing'04 Head Pose Image Data-*

FIGURE 4.2: Projection cylindrique aplatie d'un visage humain.

base [GHC04]. De plus amples détails sur cette base d'images sont disponibles dans la Section B.2. L'expérimentation a été effectuée sur un groupe de 72 personnes composé de 36 hommes et de 36 femmes âgées de 15 à 80 ans. Il a été demandé à chaque personne d'examiner une image contenant un visage et d'entourer l'orientation de la tête qui lui correspond parmi plusieurs suggérées. L'expérience s'est déroulée dans un ordre aléatoire en 2 phases : une pour l'estimation de l'angle *Pan* et l'autre pour l'estimation de l'angle *Tilt*. 65 images pour l'angle Pan et 45 images pour l'angle Tilt (i.e. 5 images par angle) issues de la *Pointing'04 Head Pose Image Database* sont présentées dans un ordre aléatoire à chaque sujet durant 7 secondes par image. La Figure 4.3 montre un exemple d'images présentées durant l'expérimentation où les symboles "+" et "-" indiquent respectivement l'orientation droite et gauche de la tête pour éviter la confusion.

Les sujets étaient divisés aléatoirement en 2 sous-groupes : "Entrainés" et "Non Entrainés" afin d'évaluer l'impact de l'entraînement sur leurs capacité à estimer l'orientation de la tête. Les sujets calibrés ont pu inspecter des images exemples étiquetées en orientation aussi longtemps qu'ils le souhaitaient avant de commencer l'expérience. Par contre, les sujets non calibrés n'ont vu aucune image d'entraînement.

La fin de l'expérimentation se déroule par la présentation de l'image issue des travaux de Kersten illustrée précédemment dans la Figure 4.2. Il a été demandé aux sujets d'entourer les angles qu'ils voient sur l'image. Le but de cette question est de confirmer l'utilisation des

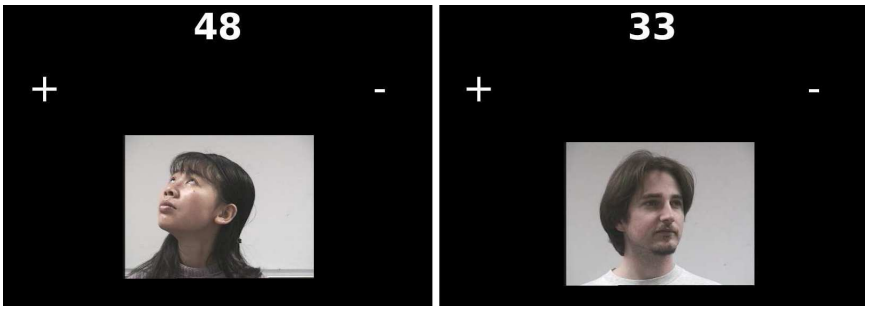


FIGURE 4.3: Exemples d'images de test présentées pendant l'expérimentation.

poses de face et de profil comme poses clés par le cerveau humain car tous les angles Pan sont visibles sur cette image.

Trois métriques ont été choisies pour mesurer la capacité humaine à estimer l'orientation de la tête. La pose théorique de l'image k est notée par $p(k)$ alors que la pose estimée par le sujet est notée par $p^*(k)$. Elles sont exprimées en degré. Le nombre total d'images sur chaque axe est noté par N.

$$\text{Erreur Moyenne} = \frac{1}{N}\sum_{k=1}^{N} \|p(k) - p^*(k)\|$$

$$\text{Erreur Maximale} = Max_k \|p(k) - p^*(k)\|$$

$$\text{Classification Correcte} = \frac{Card\{\text{Images classifiées correctement}\}}{Card\{\text{Images}\}}$$

Les résultats de ces métriques appliquées à l'estimation des angles Pan et Tilt sont reportés respectivement dans les Tableaux 4.1 et 4.2.

Mesures	**Erreur Moyenne**	**Erreur Maximale**	**Classification Correcte**
Tous les sujets	11.85°	44.79°	41.58%
Sujets Entrainés	11.79°	42.50°	40.73%
Sujets Non Entrainés	11.91°	47.08°	42.44%

TABLE 4.1: Résultats de l'estimation de l'angle Pan.

Mesures	Erreur Moyenne	Erreur Maximale	Classification Correcte
Tous les sujets	11.04°	45.10°	53.55%
Sujets Entrainés	9.45°	39.58°	59.14%
Sujets Non Entrainés	12.63°	50.63°	47.96%

TABLE 4.2: Résultats de l'estimation de l'angle Tilt.

Les sujets entrainés sont significativement meilleurs que les sujets non entrainés pour l'estimation de l'angle Tilt alors qu'ils sont comparables pour l'angle Pan. L'estimation de l'angle Pan semble être plus naturelle que celle de l'angle Tilt. Ceci est dû au fait que les gens tournent plus souvent la tête à gauche et à droite qu'en haut et en bas durant les interactions sociales. La pose la mieux reconnue est la pose frontale. C'est confirmé par la présentation de l'image cylindrique de visage de Kersten à la fin de l'expérience. En effet, 81% des sujets n'ont pas vu de pose autre que face et profil sur cette image.

Ces résultats montrent que les poses de face et de profil sont utilisées par le système visuel humain comme des poses clés. En outre, les humains sont plus aptes à estimer l'orientation horizontale.

4.2.3 Problématique de l'estimation de l'orientation de la tête

L'estimation de l'orientation de la tête d'une personne à partir d'une image est une tâche complexe étant donnée l'immense variabilité des paramètres physionomiques, de posture, d'angle de vue, de conditions d'éclairage, etc. En effet, la conception d'un système robuste capable d'accomplir cette tâche doit faire face à la variation de ces paramètres liés aux personnes, à l'environnement et aux images.

- **La variation d'échelle** : c'est le changement de la taille du visage dans l'image lorsque la personne s'approche ou s'éloigne de la caméra.
- **Les conditions d'éclairage** : il s'agit de l'influence des sources lumineuses sur l'estimation de l'orientation de la tête d'une personne se trouvant à un endroit donné. L'analyse d'une texture de peau par exemple est différente sous deux lumières distinctes alors qu'un éclairage droit ou gauche peut influencer de façon notoire sur l'estimation à cause

des problèmes d'asymétries de l'image du visage.

- **Les variations intra-personnes** : pour une même personne, la tête et surtout le visage (mais aussi l'agencement des cheveux, la présence de la barbe ou d'une moustache, etc.) sont sujets à des variations. La présence d'expressions ou de maquillage influence la capacité d'un système à déterminer l'orientation de la tête voire même sa position (surtout pour les approches basées sur l'apparence globale).
- **Les variations inter-personnes** : les caractéristiques anthropométriques du visage sont propres à chaque personne. L'estimation de l'orientation de la tête se base sur l'hypothèse suivante : plusieurs personnes ayant la même pose sont plus similaires que les modélisations faites pour une personne selon des poses différentes (hypothèse de similarité inter-personnes). Il faut alors décorréler dans la modélisation les caractéristiques individuelles de celles qui sont communes à toutes les personnes par rapport à une pose donnée.
- **Les (auto)occlusions** : des gestes ordinaires tels qu'une personne se retournant (i.e. présentant le dos de la tête), passant sa main sur son visage, ou réajustant ses lunettes suffisent à produire des erreurs d'estimation. De même, la présence d'un bonnet, d'une écharpe, ou d'un pansement peut empêcher la détection du visage dans l'image.
- **La surdimensionnalité de l'image** : il s'agit de la difficulté à représenter le nombre important de données disponibles pour une image. Par exemple, une image de résolution 300×200 possède 60 000 pixels auxquels sont associés pour la caractéristique couleur 180 000 valeurs dans l'espace de couleur RGB.

Nous proposons de prendre en considération les pré-requis suivants [Lab08] lors de l'analyse d'une méthode d'estimation de l'orientation de la tête :

- **Précision** : l'erreur moyenne absolue doit être la plus petite possible en présence d'une bonne ou d'une mauvaise localisation de la tête.
- **Type de caméra** : l'estimation de l'orientation de la tête doit être effectuée à partir d'images fournies par une seule caméra.
- **Nombre de personnes** : l'estimation de l'orientation de la tête de plusieurs personnes doit être effectuée à partir d'une seule image.
- **Distance** : l'estimation de l'orientation de la tête doit être effectuée sur des images prises

de prés ou de loin indépendamment de leurs résolution.
- **Identité de la personne** : l'estimation de l'orientation doit être effectuée indépendamment de l'identité des personnes.
- **Temps de calcul** : l'estimation de l'orientation de la tête de plusieurs personnes doit être effectuée en temps réel.

4.2.4 Taxonomie des méthodes

Il n'existe à notre connaissance aucune taxonomie largement acceptée des différentes approches dédiées à l'estimation de l'orientation de la tête. Certains systèmes estiment l'orientation de la tête automatiquement à partir de l'image alors que d'autres ont besoin de certaines caractéristiques telles que la position du visage dans l'image. Certains algorithmes et techniques de classification sont utilisés par plusieurs méthodes pour l'estimation de l'orientation de la tête. On se rend ainsi compte de la difficulté à établir une taxonomie universelle. Selon l'étude bibliographique de Murphy-Chutorian et Trivedi [MCT09], il est possible de distinguer huit groupes de méthodes prépondérantes permettant d'estimer l'orientation de la tête. Ba [Ba07], de son coté, a indiqué dans sa thèse l'existence de deux catégories distinctes de méthodes : discriminatives et génératives. Une autre taxonomie basée sur 5 catégories a également été proposée par Balasubramanian et al. [BKP08]. Toutefois, il nous semble toutefois que bien que toutes ces classifications sont justifiables, aucune n'est véritablement bien équilibrée. Nous avons proposé dans [LMD09] une nouvelle classification de méthodes en trois catégories :

- Approches basées sur la forme : construisent un modèle à partir de caractéristiques spécifiques du visage.
- Approches basées sur l'apparence globale : modélisent l'apparence globale de la tête à partir d'une base d'apprentissage.
- Approches hybrides : combinent deux ou plusieurs des méthodes issues des deux approches précédentes afin de palier aux limites inhérentes à l'utilisation d'une seule méthode.

Il faut noter que certains systèmes nécessitent une détection préalable de la tête. Ces méthodes nécessitent un pré-traitement et peuvent donc être considérées comme hybrides, ce qui rend complexe la classification des méthodes. Les méthodes étudiées sont aussi caractérisées par les concepts de discriminativité et de générativité. Les modèles génératifs sont efficaces pour estimer l'orientation de la tête, mais ne sont pas recommandés pour estimer la position de celle-ci dans une image. Ceci a amené à l'apparition de méthodes issues du chaînage d'une méthode discriminative et d'une ou plusieurs méthodes génératives.

4.2.4.1 Approches basées sur la forme

Un ensemble de caractéristiques spécifiques au visage telles que les yeux, le nez et bouche sont utilisées pour estimer l'orientation de la tête. Le modèle construit peut être soit flexible, soit géométrique.

Les modèles flexibles utilisent une méthode qui adapte un modèle standard non rigide (un masque) à l'image analysée. Le modèle utilise comme nœuds des points caractéristiques tels que les coins de la bouche, la position des yeux ainsi que certains angles du visage. En plus d'une annotation des différentes poses, une base d'apprentissage contenant les caractéristiques locales du visage doit être créée. Trois sous groupes de modèles peuvent être distingués : Elastic Graph Matching (EGM), Active Shape Model (ASM) et Active Appearance Model (AAM). Ces modèles sont génératifs de par leur nature car ils supposent l'application d'un masque virtuel sur l'image pour s'adapter aux caractéristiques physiologiques et de l'orientation de la tête.

Le modèle EGM [KPvdM97] possède la capacité à représenter des objets déformés ou non rigides. L'estimation de l'orientation de la tête est effectuée par la création d'un graphe pour chaque pose. Chacun d'eux est comparé avec l'image contenant un visage à l'aide d'une déformation itérative qui permet de trouver les distances minimum entre chaque nœud et chaque point caractéristique détecté. La pose associée au graphe qui maximise la similarité est choisie [WT08]. L'EGM est utilisé étant donné que les variations anthropométriques inter-personnes sont négligeables devant les variations des points caractéristiques induites par la pose de la tête. En effet tous les individus ont à quelques millimètres (voire centimètres) près le même

écart entre les yeux, la même hauteur de visage (environ 23 cm), etc. [GBC+88]

Le modèle ASM [CTCG95] est le plus largement utilisé. Il exploite un modèle à distribution de points qui permet d'analyser et de représenter une forme. Il consiste en un prototype d'une forme moyenne (voir Figure 4.4) doté de modes de variations combinables appris à partir d'un ensemble de données d'apprentissage constitué d'instances de la forme étudiée. La construction d'un ASM se décompose en deux parties. Lors de la première partie, qui correspond à l'initialisation, un ensemble d'images est sélectionné en entrée du système. Ce sont des images normalisées où les positions moyennes des points pertinents se trouvent dans des zones prédites. L'ASM se fonde sur un algorithme de détection des points pertinents, et garde en mémoire leurs positions. Ainsi, grâce à l'analyse d'un nombre d'images relativement important, il est possible d'estimer des corrélations entre les points, et de calculer les valeurs moyennes de leurs positions afin d'établir une position moyenne du modèle par rapport à l'image. La seconde partie, qui est la partie active, prend une suite d'images en entrée (une séquence vidéo par exemple). Les positions des points pertinents sont ensuite détectées. L'estimation de l'orientation de la tête est finalement effectuée à l'aide d'un algorithme de calcul géométrique qui utilise ces données pour déduire l'orientation de la tête.

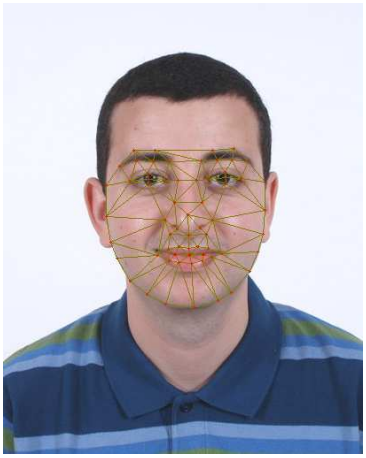


FIGURE 4.4: Exemple de l'application d'un ASM sur un visage.

Le modèle AAM [CET01] est le modèle flexible le plus évolué. L'ASM apprend unique-

ment les premiers modes de variation de la forme du visage. L'AAM apprend en plus ceux de la texture. Ceci permet de palier au problème de perspectives induites par la localisation des points pertinents sur le visage. La construction d'un AAM commence par la génération d'un ASM à partir d'un ensemble de données d'apprentissage. Ensuite, les images des visages sont déformées de telle sorte que les points caractéristiques correspondent à ceux de la forme moyenne. Les images sont ensuite normalisées pour construire le modèle. Enfin, les corrélations entre la forme et la texture sont apprises pour générer un modèle d'apparence qui combine les deux. Une fois que le modèle converge vers l'emplacement des points caractéristiques, une estimation de l'orientation de la tête est obtenue par la mise en correspondance des paramètres de l'apparence avec l'orientation de la tête.

Avantages
- Invariance aux erreurs de localisation de la tête.
- Estimation précise de l'orientation de la tête.
- Moins de variations inter-personnes.

Inconvénients
- Phase d'initialisation très longue.
- Peu de robustesse aux occlusions (l'estimation des orientations nécessite généralement la présence sur l'image des coins des deux yeux).
- Difficulté d'utilisation lorsque le sujet se trouve distant de la caméra ou pour des images en basse résolution.

Les modèles géométriques s'appuient sur la détection de points caractéristiques du visage, tout comme les modèles flexibles. La différence se situe dans l'analyse de la position de ces points. En effet, un modèle géométrique est créé en ayant pour axe symétrique un segment tracé depuis le milieu des yeux jusqu'au milieu de la bouche. A cet axe sont ajoutés les positions des coins des yeux et de la bouche, ainsi que la position du nez. Il est ainsi possible d'estimer les 3 degrés de liberté à l'aide de calculs géométriques. La difficulté notable de ce modèle est de localiser les points caractéristiques du visage avec précision. D'autres problèmes peuvent

aussi survenir en présence d'occlusions (e.g. port de lunettes, présence d'une écharpe, etc.). Une étape de détection des points caractéristiques est d'abord réalisée. Ensuite, la différence entre le visage et le modèle est calculée, basée sur des informations statistiques. L'exemple suivant [PZJ05] illustre l'estimation de l'angle Roll γ. Il est calculé à l'aide de la formule suivante :

$$\gamma = \frac{\gamma_1 + \gamma_2 + \gamma_3}{3}$$

Avec :

$$\gamma_1 = \arctan(\frac{y_{ell} - y_{elr}}{x_{ell} - x_{elr}})$$

$$\gamma_2 = \arctan(\frac{y_{erl} - y_{err}}{x_{erl} - x_{err}})$$

$$\gamma_3 = \arctan(\frac{(y_{ell} + y_{elr})/2 + (y_{erl} - y_{err})/2}{(x_{ell} + x_{elr})/2 - (x_{erl} + x_{err})/2})$$

Les points *ell*, *elr*, *erl*, *err*, *em*, *nm* et *mm* utilisés dans les calculs sont identifiés sur la Figure 4.5.

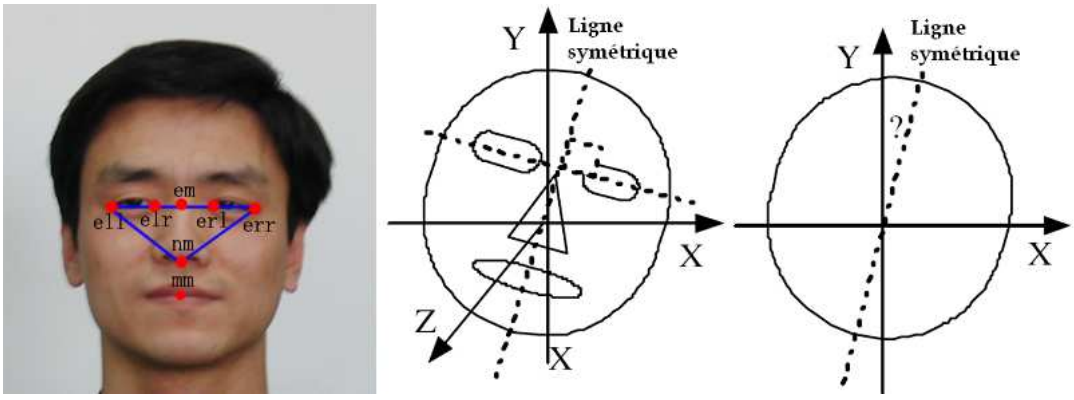


FIGURE 4.5: Estimation de l'angle Roll.

Avantages

– Rapidité de calcul.
– Simplicité.
– Estimation raisonnable avec peu d'informations.

Inconvénients
– La localisation des points caractéristiques est une source d'erreurs.
– L'orientation de la tête ne peut être estimée que si elle est proche d'une vue frontale.
– Difficulté d'identifier les points caractéristiques en cas d'occlusions.

4.2.4.2 Approches basées sur l'apparence globale

Au lieu de se concentrer sur les caractéristiques faciales, l'apparence globale de la tête est modélisée à partir de la base d'images d'apprentissage. Pour cela plusieurs méthodes peuvent être utilisées :

Les méthodes basées sur un patron (Template based methods) comparent l'image reçue en entrée contenant un visage à une collection d'images étiquetées dans le but de trouver la pose la plus similaire. La base d'images utilisée est un ensemble d'images correspondant à plusieurs poses prises par une même personne. Cette base est élargie par la même séquence de poses à un nombre important de personnes en leur attribuant une même étiquette pour une pose identique. Des métriques de comparaison d'images telles que la Minimisation de l'Erreur Quadratique (MEQ) [NF96] ou les machines à vecteurs de support (SVM) peuvent être utilisées pour mettre en correspondance l'image analysée avec une image présente dans la base d'images. Et donc, en déduire ainsi l'orientation probable de la tête.

Avantages
– Simplicité.
– Facilité d'extension de la base d'images.
– Pas d'apprentissage négatif requis.
– Indépendance de la résolution.

Inconvénients

- La région de l'image où se situe la tête doit être détectée au préalable.
- Une erreur sur la localisation de la tête implique de fortes dégradations sur la précision de l'estimation de l'orientation de la tête.
- Une grande base d'images entraine un nombre de calculs important.
- Le plus gros inconvénient est que deux personnes différentes analysées avec la même position auront une probabilité moindre d'être mises en correspondance qu'une même personne avec deux poses différentes (non-respect de l'hypothèse de similarité). Cependant, pour remédier à ce problème, une convolution des images avec un filtre Laplacien peut mettre en évidence les contours du visage les plus similaires, tout en retirant les textures spécifiques à l'identité.

Les méthodes basées sur un ensemble de détecteurs (Detector Arrays) utilisent un ensemble de détecteurs associés à une orientation spécifique de la tête pour affiner la qualité de classification (un détecteur = une orientation de la tête). Cette méthode est un prolongement naturel des nombreux modèles utilisés avec succès pour la détection frontale du visage [VJ01]. Il s'agit alors de spécialiser plusieurs détecteurs de visage correspondant à des orientations discrètes de la tête. La pose est choisie selon un seul degré de liberté. Les premiers prototypes [HSW98] pour ce type de méthodes se contentaient de reconnaître 3 orientations différentes de la tête (3 détecteurs) selon l'axe vertical. Actuellement une douzaine de détecteurs (au maximum) parviennent à décomposer les différentes orientations de la tête. Le procédé est identique pour les trois degrés de liberté (il est aisé d'imaginer la conséquence dans le volume des données générées). Ces détecteurs sont entrainés en utilisant un algorithme d'apprentissage (SVM ou Adaboost) ou des réseaux de neurones. La pose associée au détecteur ayant le plus grand support (la plus grande pertinence) est assignée à l'image. Afin de réduire le temps de calcul, un routeur de classification est souvent utilisé.

Avantages

- La détection de la tête et l'estimation de son orientation peuvent se faire en parallèle par un même classificateur.

- Très bons résultats sur un seul degré de liberté.
- Immobilité aux changements d'apparence ne correspondant pas à un changement de pose.
- Indépendance de la résolution.

Inconvénients

- Nécessité d'apprentissage sur une base négative (exemple d'images ne contenant pas de têtes).
- Taille importante de la base d'images avec lourdeur de création (même pose pour des personnes différentes).
- Augmentation des problèmes de classification de manière proportionnelle au nombre de détecteurs.
- Temps de calcul conséquent.
- Difficulté d'estimer les trois degrés de liberté simultanément (estimations généralement limitées à un degré de liberté avec moins de 12 détecteurs).
- Possibilité d'ambiguïté lorsque plusieurs détecteurs classifient une image comme positive.

Les méthodes basées sur l'intégration de variété (Manifold Embedding) cherchent un ensemble réduit de dimensions qui modélise la variation continue de l'orientation de la tête. Une variété est un espace topologique abstrait construit par recollement d'autres espaces simples. Cela permet de changer la représentation de l'image dans d'autres dimensions. Les poses ont une valeur bien définie dans un intervalle de cette nouvelle dimension. Un sous dimensionnement est alors nécessaire pour estimer l'orientation de la tête tout en ignorant les autres variations dans l'image. La variété peut être linéaire, et dans ce cas l'Analyse en Composantes Principales (ACP) [SGO01] est la méthode la plus utilisée. Par contre, elle ne tient pas compte de l'étiquette disponible, car la méthode utilise un apprentissage non supervisé.

Une autre approche consiste à normaliser l'image et à la projeter sur chacun des espaces propre de pose (*pose-eigenspaces*) et de trouver la pose induisant la plus grande énergie de projection [SB02]. Une approche basée sur une approximation linéaire telle que l'analyse lo-

cale intégrée (Locally Embedded Analysis _ LEA) [FH06]. Il existe une variante des méthodes linéaires qui se base sur le noyau tel que l'Analyse en Composantes Principales (en KPCA) ou bien l'Analyse Discriminante Linéaire (en KLDA) [WT08]. Un autre sous dimensionnement, non linéaire, parait être plus adéquat étant données les variations dans l'image causées par le changement de pose. Il existe plusieurs techniques illustrées par la Figure 4.6 telles que : Isometric feature mapping (Isomap) [HHR05, RYS04], Locally Linear Embedding (LLE) [RS00], et Laplacian Eigenmaps (LE) [BN03]

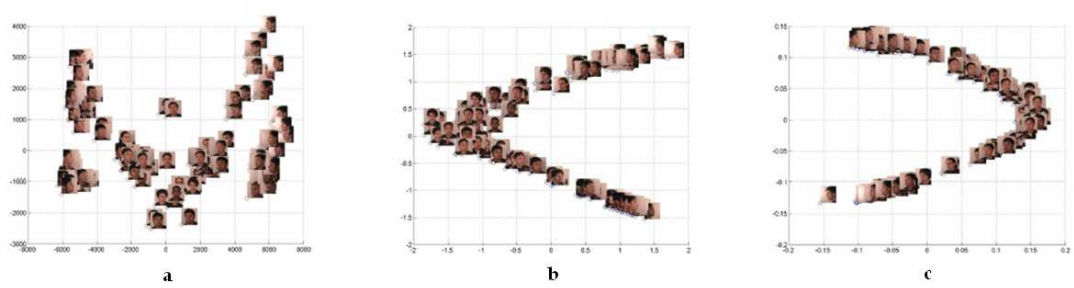


FIGURE 4.6: Intégration d'images de visages avec des orientations différentes de la tête sur 2 dimensions. (a) Isomap, (b) LLE, (c) LE.

Avantages

– Très bons résultats lors de l'application d'un filtre de Gabor à l'image.

Inconvénients

– Tendance à créer des applications injectives aussi bien pour l'identité que pour l'orientation de la tête.
– Hétérogénéité de la base de données.
– Capacité d'adaptation limitée pour les techniques non linéaires.

Les méthodes basées sur la régression non-linéaire utilisent des outils de régression non-linéaire pour effectuer une mise en correspondance fonctionnelle entre l'image (ou plutôt ses

caractéristiques) et une mesure de l'orientation de la tête. En effet, le nombre élevé de dimensions d'une image présente un défi pour les outils de régression. Le succès de ces méthodes a été démontré en utilisant les machines à vecteurs de support régressif (Support Vector Regressors _ SVR) après une réduction de dimension effectuée par l'ACP [LGSL04], ou sur les caractéristiques faciales du visage [MKK+06].

Les outils de régression non-linéaire les plus utilisés sont les réseaux de neurones, et notamment le MLP (Multi-Layer Perceptron) [SYW02]. Ces outils peuvent être entrainé sur un intervalle continu de poses avec une sortie pour chaque degré de liberté (l'activation d'une sortie est proportionnelle à l'orientation qui lui correspond) ou bien avec un ensemble de MLP avec une sortie pour chaque nœud appris pour chaque degré de liberté. Une combinaison séquentielle d'un filtre qui sélectionne des caractéristiques et un réseau de neurones généralisé peut aussi être utilisée par le biais d'un processus de boosting [BM09].

L'approche nommée Locally-Linear Map (LLM) [RR98] est un autre réseau de neurones composé de plusieurs cartes linéaires. La construction du réseau nécessite une comparaison des données en entrée avec un échantillon pris du barycentre de chaque carte. Celui-ci est utilisé pour apprendre une matrice de poids. Un réseau convolutif peut être aussi utilisé [OCM07].

Avantages

- Rapidité.
- Efficacité.
- Insensibilité au zoom.

Inconvénients

- Erreurs d'estimation importantes en cas de mauvaise localisation de la tête.
- Difficulté d'apprendre une mise en correspondance en se basant sur un outil de régression spécifique.

4.2.4.3 Approches hybrides

Ce type d'approches combine deux ou plusieurs des méthodes citées précédemment afin de palier aux limites inhérentes à l'utilisation d'une seule méthode. La démarche qui est la

plus couramment utilisée est de compléter une estimation statique de l'orientation de la tête par un système de suivi. Le système statique est chargé de l'initialisation et le système de suivi est chargé de maintenir l'estimation de l'orientation de la tête à travers le temps. Si le système de suivi commence à dériver, le système statique peut réinitialiser le suivi. Ces méthodes permettent d'obtenir une plus grande précision que les approches utilisant le suivi uniquement. De nombreuses combinaisons ont été utilisées pour l'estimation de l'orientation de la tête en utilisant : une méthode géométrique avec une méthode de suivi de points [HCZZ04], une mise en correspondance basée sur un modèle intégré qui utilise l'ACP avec le flux optique [ZF04] ou avec un modèle de Markov caché basé sur une densité continue [HT04], un modèle d'apparence basé sur la couleur et la texture avec un filtre à particules [BO04], ou une mise en correspondance des images clés basée sur un modèle intégré qui utilise l'ACP avec un suivi en stéréovision basé sur le niveau de gris et la constance de la profondeur [MRD03].

Les méthodes de suivi (tracking) opèrent en suivant le mouvement de la tête entre des images consécutives d'un flux vidéo. Ces modèles peuvent utiliser une approche basée sur le modèle pour se centrer sur une position connue. Les méthodes de suivi sont dans la pratique toujours secondées d'une étape de réinitialisation pour palier aux effets d'occlusion ou la perte de l'objet suivi (e.g. une main qui vient réajuster des lunettes).

Avantages

- Haut niveau de précision.
- Détecte les moindres variations dans les images analysées.
- Moins d'erreurs dues aux variations d'apparences.

Inconvénients

- Initialisation nécessaire : le sujet doit maintenir une position frontale avant que le système ne se mette en marche et il doit se recentrer si cette position est perdue (l'utilisation d'un détecteur de visage est nécessaire).
- Manque de robustesse

4.2.4.4 Synthèse

Nous avons regroupé les avantages et les inconvénients des approches basées sur la forme et celles basées sur l'apparence globale dans le Tableau 4.3. Ceci nous a permis de choisir une méthodologie pour effectuer l'estimation de l'orientation de la tête.

	Forme	**Apparence globale**
Basse résolution	-	+
Luminosité	+	-
Occlusion partielle	-	+
Bonne localisation du visage	-	+
Localisation des points du visage	+	-
Orientations extrêmes de la tête	-	+
Distant de la caméra	-	+

TABLE 4.3: Comparaison entre les approches basées sur la forme et celles basées sur l'apparence globale.

Après analyse des différentes méthodes utilisées dans l'état de l'art en termes d'avantages et d'inconvénients, nous avons décidé d'utiliser une approche hybride. Cette approche hybride sera composée : (i) d'une méthode d'estimation de l'orientation de la tête basée sur l'apparence globale car ce type de méthodes est approprié lorsque l'utilisateur est distant de la caméra, et (ii) d'une méthode de suivi de la tête basée sur un modèle cylindrique. La première méthode servira à la fois à l'initialisation du suivi de la tête et à sa ré-initialisation en cas de dérive.

4.3 Bases d'images

Pour construire et évaluer les systèmes d'estimation de l'orientation de la tête, une vérité-terrain (*ground truth*) est nécessaire. Nous présentons d'abord les méthodes utilisées pour acquérir les bases d'images. Nous présenterons ensuite les bases d'images qui sont le plus utilisées dans l'état de l'art.

4.3.1 Construction des bases d'images

Les méthodologies utilisées pour construire une vérité-terrain sont présentées ici selon un ordre de précision croissant (i.e. des moins précises aux plus précises).

4.3.1.1 Suggestion directionnelle

Une série de marqueurs est placée dans des endroits précis de la scène cible par rapport à un repère où se situe le sujet. Chaque sujet est invité à regarder ces marqueurs successivement en utilisant le mouvement de sa tête plutôt qu'en bougeant les yeux. Une caméra placée en face du sujet enregistre des images distinctes de sa tête pour chaque marqueur qu'il regarde. Cette méthode génère une vérité-terrain peu fiable. En effet, cette méthode suppose que chaque tête se situe dans le même emplacement physique dans l'espace 3D pour que toutes les directions de la tête correspondent à la même pose. De plus, elle suppose que chaque sujet possède la capacité de diriger avec précision sa tête vers un objet. Cependant, il s'agit d'une tâche imprécise que les sujets effectuent avec difficulté, notamment en déplaçant uniquement les yeux.

4.3.1.2 Suggestion directionnelle avec un pointeur laser

Cette méthode utilise le même procédé que la suggestion directionnelle en ajoutant l'utilisation d'un pointeur laser qui est fixé sur la tête du sujet. Cela permet au sujet de repérer les emplacements qu'il doit regarder avec une précision beaucoup plus élevée grâce au retour visuel. Cette méthode suppose également que chaque tête se situe dans le même emplacement physique de l'espace 3D, ce qui est difficile à assurer car les sujets ont tendance à bouger naturellement durant l'enregistrement des images.

4.3.1.3 Annotation manuelle

Des images contenant des visages sont regardées par une personne. L'annotateur assigne une pose au visage présent dans l'image en fonction de sa propre perception de l'orientation de la tête. Ce procédé est souvent utilisé pour l'estimation de l'orientation de la tête sur un

seul degré de liberté avec un ensemble approximatif de poses. Cependant, elle est inappropriée pour une estimation précise de l'orientation de la tête.

4.3.1.4 Série de caméras

Plusieurs caméras placées à des positions données enregistrent des images du visage d'une personne simultanément à partir d'angles différents. Cette méthode offre une vérité-terrain très précise si la tête de chaque sujet se trouve au même endroit. Par contre, elle est adaptée uniquement aux images prises à courte distance. Elle est difficilement applicable aux vidéos provenant de cas réels.

4.3.1.5 Capteurs magnétiques

Ces capteurs mesurent le champ magnétique qu'ils émettent. Le capteur peut être fixé sur la tête du sujet et détermine la position et l'orientation de la tête. La précision théorique de ces capteurs est élevée (moins de 1° d'erreur). Néanmoins, ils sont très sensibles au bruit environnant et à la présence de métaux (même en quantité infimes) durant la collecte de données. C'est une méthode très largement utilisée. Les capteurs les plus connus sont le *Polhemus FastTrak* et le *Ascension Flock of Birds*.

4.3.1.6 Systèmes de capture de mouvement optique

Ces systèmes sont robustes mais leurs déploiement est coûteux. Ils sont plus souvent utilisés pour la capture cinématographique professionnelle des mouvements d'un corps articulé. Une série de caméras infrarouges se trouvant à courte distance utilisent plusieurs vues stéréo pour suivre les marqueurs attachés à une personne. Ces marqueurs peuvent être fixés à l'arrière de la tête d'un sujet. Le système le plus connu est le *Vicon MX*.

4.3.2 Bases d'images utilisées

Au cours de ces dernières années plusieurs bases d'images dédiées à l'estimation de l'orientation de la tête ont été créées. Dans ce qui suit, nous présentons quelques bases d'images

disponibles tout en détaillant les bases utilisées dans cette thèse.

4.3.2.1 Pointing'04 Head Pose Image Database

La construction de la base d'images *Pointing'04 Head Pose Image Database* [GHC04] a nécessitée 15 personnes. Chacun a été pris deux fois sous 93 poses différentes selon deux degrés de liberté. Elle est échantillonnée tous les 15 degrés en pan, tous les 15/30 degrés en tilt et couvre une demi-sphère de poses allant de -90° à +90° sur les 2 axes. L'angle pan peut donc prendre les valeurs (0, $\pm$15, $\pm$30, $\pm$45, $\pm$60, $\pm$75, $\pm$90), où les valeurs négatives correspondent aux poses droites et les valeurs positives correspondent aux poses gauches. L'angle tilt peut prendre les valeurs (0, $\pm$15, $\pm$30, $\pm$60, $\pm$90), où les valeurs négatives correspondent aux poses basses et les valeurs positives correspondent aux poses hautes. La Figure 4.7 présente les 93 poses que peut prendre une personne dans cette base d'images.

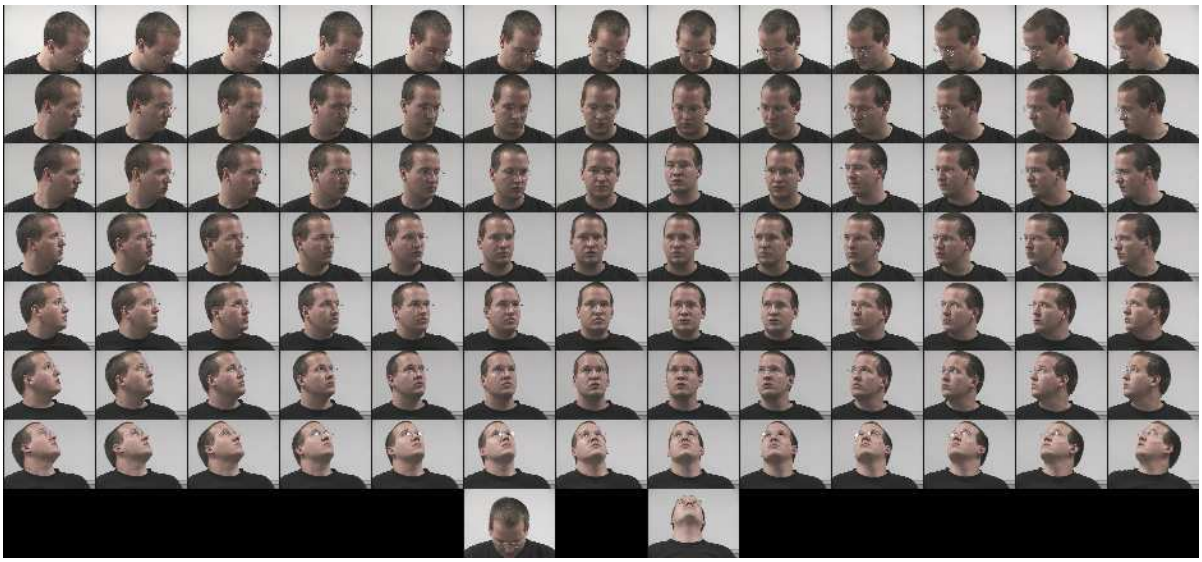


FIGURE 4.7: Exemple de toutes les orientations de la tête associées à une personne.

4.3.2.2 Boston University Head Pose Dataset

La base d'images *Boston University Head Pose Dataset* [CSA00] se compose de 45 séquences vidéos. Il a été demandé à 5 sujets de procéder à 9 mouvements différents de leurs têtes dans un environnement de bureau standard sous un éclairage uniforme. La tête est toujours visible avec l'existence de certaines auto-occlusions insignifiantes. Les vidéos sont en basse résolution (320x240 pixels). Un capteur magnétique de type *Flock of Birds* est utilisé pour enregistrer la position et l'orientation de la tête. Ce système possède une précision de 1.8 mm dans la translation et de 0.5° en rotation. Durant la collecte de données, du bruit électromagnétique à interférer sur la précision du système de capture. Néanmoins, les mesures enregistrées sont assez fiables pour être utilisées comme vérité-terrain. La Figure 4.8 présente les 5 personnes qui ont effectué l'expérience sous des orientations différentes.

FIGURE 4.8: Exemples des personnes filmées lors de la collecte de données.

4.3.2.3 Autres bases d'images

- La base *FacePix* [LKBP05] se compose de 181 images pour chacun des 30 sujets sur l'axe Pan dans l'intervalle [-90°, 90°]. Les images capturées à l'aide d'une caméra placée sur une plate-forme tournante ont été recadrées manuellement pour s'assurer que les yeux et le visage apparaissent à la même position dans chaque vue.
- La base *CMU PIE* [SBB03] se compose de 68 images faciales de personnes en utilisant 13 poses sous 43 conditions d'éclairage avec 4 expressions différentes. Une série de 13 caméras est utilisée : 9 caméras à 22.5° d'intervalle sur l'axe Pan, 1 au dessus du centre, 1 au dessous du centre, et 2 caméras chaque coin de la pièce.

– La base *Hermes Head direction dataset* [Pro] contient 4 vidéos (2 filmées en intérieur et 2 en extérieur). Deux personnes différentes apparaissent dans les vidéos (une seule personne est visible à la fois). En outre, 25 images d'apprentissage pour chaque personne est disponible à raison de 5 images par pose. Les poses sont à 45° d'intervalle dans l'axe Pan entre -90° et 90°.
– Les bases *CHIL-CLEAR 06 et 07* créées pour la campagne d'évaluation CLEAR [Wora]. Les séquences vidéo proviennent de 4 caméras synchronisées placées dans chacun des coins d'une salle de séminaire. L'orientation de la tête est fournie manuellement dans la base de 2006 alors qu'un capteur magnétique est utilisé pour la base de 2007.
– La base *Idiap Head Pose database* [BO04] se compose de 8 séquences d'une minute enregistrées durant une réunion à partir d'une seule caméra où deux sujets sont visibles. L'annotation de la position et de l'orientation de la tête s'est faite en utilisant un capteur magnétique.

4.4 Estimation de l'orientation de la tête basée sur l'apparence globale

Nous présentons dans ce qui suit la méthode utilisée pour l'estimation de l'orientation de la tête basée sur l'apparence globale [LZD08]. Cette méthode considère cette tâche comme un problème de classification d'images qui consiste à convertir l'image de la tête reçue en entrée en un vecteur de caractéristiques. Ces vecteurs issus d'images de plusieurs personnes prises sous la même pose sert à apprendre un classificateur qui permet d'estimer l'orientation de la tête. Ces images proviennent d'une base composée de N poses associées à des valeurs discrètes des angles Pan et Tilt pour M personnes. Cette base est ensuite divisée après un prétraitement en deux ensembles : apprentissage et test. L'objectif est de déterminer une mesure discriminative à appliquer sur un vecteur de caractéristiques de dimension n. Pendant la phase d'apprentissage un classificateur est construit sur un nombre limité de poses exclusives qui définissent la précision de l'estimation qui doit être atteinte.

4.4.1 Base d'images utilisée

La base d'images Pointing [GHC04] décrite dans la Section B.2 est utilisée pour la construction du modèle associé à l'orientation d'une tête ainsi que pour le tester. La base est découpée en deux ensembles :

- L'ensemble d'apprentissage : il se compose de 20 images par pose associées à 11 personnes (9 personnes ont été prises deux fois et 2 personnes ont été prises une seule fois).
- L'ensemble de test : il se compose de 10 images par pose associées à 6 personnes (4 personnes ont été prises deux fois et les 2 personnes qui ont utilisées qu'une seule fois pour l'apprentissage ont été prises une fois).

Cinq poses ont été sélectionnées : Bas-gauche, Bas-droit, En face, Haut-gauche et Haut-droit qui correspondent respectivement aux paires d'angles Pan et Tilt suivants {(-60, -90), (-60, +90), (0, 0), (+60, -90), (+60, +90)} illustrées par la Figure 4.9.

FIGURE 4.9: 5 poses sélectionnées de la base d'images Pointing'04.

Un pré-traitement sur ces images est nécessaire afin d'en extraire certaines caractéristiques. Un rectangle serré autour de la tête est d'abord localisé. Ensuite, les images sont normalisées à la même taille 64x64. Enfin, une égalisation d'histogrammes (histogram equalization) est appliquée afin d'assurer que deux images de visages pris sous des conditions d'éclairage différentes soient transformées en deux images en niveaux de gris de luminosité similaire. Plusieurs vecteurs de caractéristiques sont sélectionnés sur cette base.

4.4.2 Sélection des caractéristiques

Dans ce qui suit, nous allons présenter l'extraction des vecteurs de caractéristiques sur la base d'images pré-traitées. Cette extraction est basée sur l'hypothèse suivante : Plusieurs

personnes sous la même pose sont plus similaires qu'une personne sous des poses différentes. En particulier deux méthodes sont retenues :

- Décomposition en valeurs singulières (SVD) : elle est appliquée sur la totalité de l'image pour en extraire un vecteur ;
- Utilisation des ondelettes de Gabor : permet d'extraire un vecteur composé de coefficients qui sont échantillonnés en utilisant des échelles et des orientations différentes sur une image qui contient un visage sous une pose particulière.

Le résultat pour chaque image I est un vecteur de caractéristiques F_I de taille n (n est choisi en fonction de la technique utilisée pour la sélection des caractéristiques) :

$$F_I = (F_{I_1}, F_{I_2}, \ldots, F_{I_n})^t$$

4.4.2.1 Décomposition en valeurs singulières (SVD)

La décomposition en valeurs singulières [Vac91] d'une matrice A est sa représentation en un produit d'une matrice diagonale et de deux matrices de base orthonormés :

$$A = U * W * V^t$$

W est une matrice diagonale composée d'éléments qui sont représentés par un vecteur à n dimensions. Toutes les valeurs singulières sont positives et triées dans un ordre décroissant. L'application de cette décomposition sur les intensités des pixels d'une image normalisée I (n=64) d'un visage permet d'obtenir le vecteur suivant :

$$W_I = (w_{I_1}, w_{I_2}, \ldots, w_{I_{64}})^t$$

Deux vecteurs U_{I_j} et V_{I_j} sont associés à chaque valeur singulière w_{I_j} avec $j \in \{1, \ldots, 64\}$:

$$U_{I_j} = (u_{I_{j_1}}, u_{I_{j_2}}, \ldots, u_{I_{j_{64}}})^t$$

$$V_{I_j} = (v^t_{I_{j_1}}, v^t_{I_{j_2}}, \ldots, v^t_{I_{j_{64}}})^t$$

Ensuite la norme $\|W_I\|$ est calculée :

$$\|W_I\| = \sqrt{w_{I_1}^2 + w_{I_2}^2 + \ldots + w_{I_{64}}^2}$$

Enfin, deux type de vecteurs de caractéristiques sont créés pour une image i :

– Le premier se compose d'éléments obtenus de la division des P premiers éléments du vecteur W par sa norme $\|W_i\|$:

$$F_{I_j} = \frac{w_{I_j}}{\|W_I\|}, j \in \{1, \ldots, 64\}, P \leq 64$$

– Le second se compose des P premières valeurs singulières w_{I_j} divisées par la norme $\|W_I\|$ ainsi que les vecteurs U_{I_j} et V_{I_j} qui leurs sont associés :

$$F_{I_j} = (\frac{w_{I_j}}{\|W_I\|}, U_{I_j}, V_{I_j}), j \in \{1, \ldots, P\}, P \leq 64$$

Afin de choisir une valeur appropriée pour P, l'image reçue en entrée est reconstruite en utilisant les P premières valeurs singulières (voir Figure 4.10).

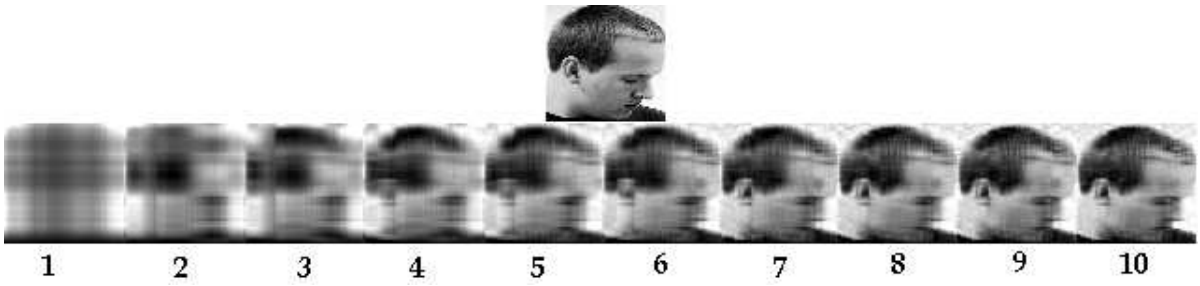


FIGURE 4.10: Reconstruction d'une image en fonction de la valeur de P.

Les deux vecteurs de caractéristiques sont utilisés pour effectuer l'estimation de l'orientation de la tête. Trois méthodes sont utilisées pour construire le classificateur : SVM avec un noyau RBF (Fonction de Base Radiale), K plus proches voisins (KNN) avec $K = 10$, et la distance de Frobenius. Les Figures 4.11 et 4.12 rapportent les résultats obtenus sur l'ensemble

de test en variant la valeur de P lors de la construction des deux vecteurs de caractéristiques.

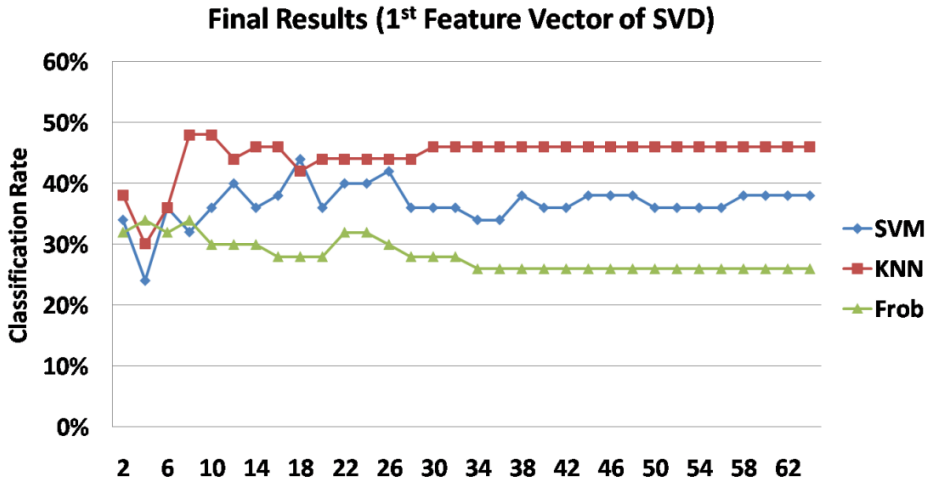


FIGURE 4.11: Taux de classification en utilisant le premier vecteur de caractéristiques de SVD.

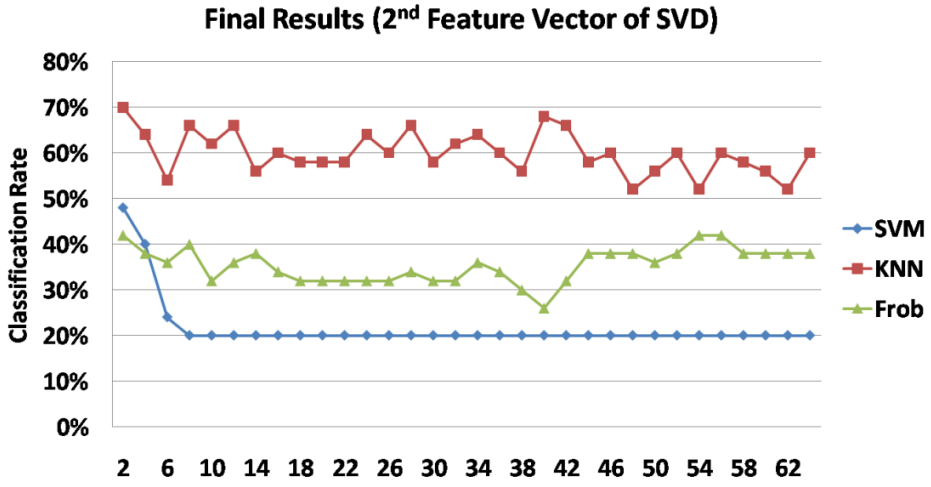


FIGURE 4.12: Taux de classification en utilisant le second vecteur de caractéristiques de SVD.

Ces résultats permettent de déduire que l'utilisation des valeurs singulières contenues dans

la matrice diagonale ne suffit pas. En effet, l'ajout des informations contenues dans les matrices U et V améliorent la qualité des résultats. Étant donné que les valeurs singulières sont triées, l'information contenue dans les premières valeurs est suffisante pour effectuer une bonne estimation de l'orientation de la tête.

4.4.2.2 Utilisation des ondelettes de Gabor

Les ondelettes de Gabor (*Gabor Wavelets*) sont utilisées pour différentier les orientations de la tête entre elles. Une évaluation du ratio de similarité entre les poses est disponible dans [SGO01]. Elle a été réalisée en faisant varier l'orientation des filtres de Gabor à une pose donnée. Une ondelette de Gabor $\psi_{o,s}(z)$ est définie comme suit [ZW06] :

$$\psi_{o,s}(z) = \frac{\|k_{o,s}\|^2}{\sigma^2} e^{-\frac{\|k_{o,s}\|^2 \|z\|^2}{2\sigma^2}} [e^{ik_{o,s}z} - e^{-\frac{\sigma^2}{2}}] \tag{4.1}$$

où le point $z = (x,y)$ possède la coordonnée horizontale x et la coordonnée verticale y. Les paramètres o et s définissent l'orientation et l'échelle du noyau de Gabor alors que $\|\cdot\|$ désigne l'opérateur de norme. σ est lié à l'écart type de la fenêtre Gaussienne dans le noyau et détermine le ratio de la largeur de la fenêtre Gaussienne sur la longueur d'ondelette. Le vecteur d'onde $k_{o,s}$ est définit comme suit :

$$k_{o,s} = k_s e^{i\phi_o}$$

avec $k_s = \frac{k_{max}}{f^s}$ et $\phi_o = \frac{\pi o}{O}$. k_{max} est la fréquence maximale, f^s est la fréquence spatiale entre les noyaux dans le domaine fréquentiel, et O est le nombre d'orientations choisi.

Pour la création du vecteur de caractéristiques, huit orientations $\{$ 0, $\frac{\pi}{8}$, $\frac{\pi}{4}$, $\frac{3\pi}{8}$, $\frac{\pi}{2}$, $\frac{5\pi}{8}$, $\frac{3\pi}{4}$, $\frac{7\pi}{8}$ $\}$ ont été choisies à cinq échelles différentes $\{$ 0, 1, 2, 3, 4 $\}$ avec $\sigma = 2\pi$, $k_{max} = \frac{\pi}{2}$, et $f = \sqrt{2}$.

La partie réelle d'une ondelette de Gabor en utilisant les 8 orientations est illustrée par la Figure 4.13 :

En utilisant les ondelettes de Gabor, une image est représentée par la convolution de celle-ci avec une série de noyaux de Gabor. La Figure 4.14 représente la réponse réelle d'une image d'une personne sous une pose en utilisant 8 orientations et 5 échelles.

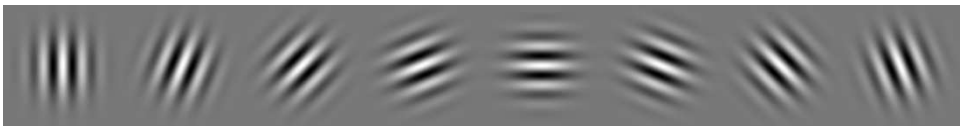


FIGURE 4.13: Réponse réelle des ondelettes de Gabor sur les 8 orientations choisies.

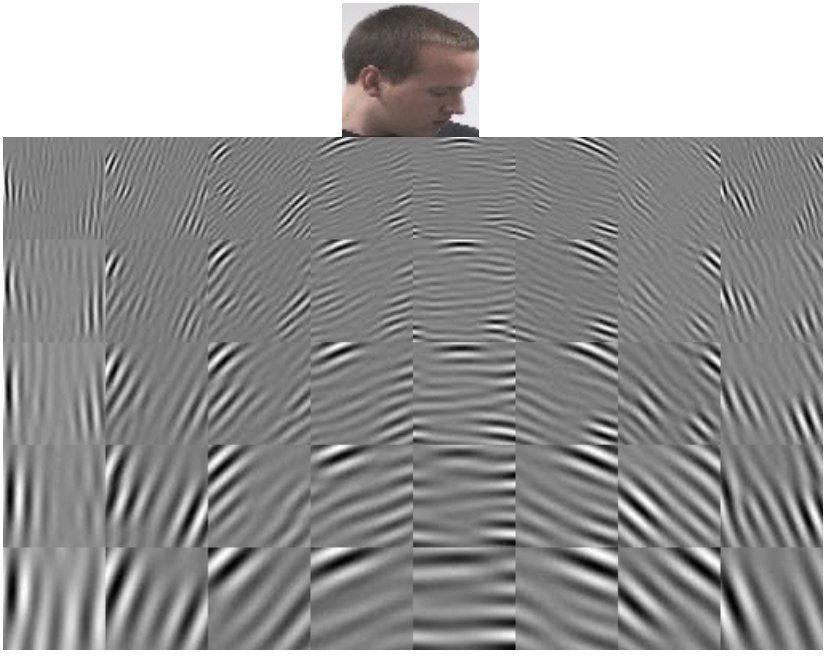


FIGURE 4.14: Réponse réelle d'une image d'une tête en utilisant 8 orientations et 5 échelles.

La convolution d'une image I et d'un noyau de Gabor $\psi_{o,s}(z)$ est définie comme suit :

$$Conv_{o,s}(z) = I(z) * \psi_{o,s}(z)$$

La réponse $Conv_{o,s}(z)$ de chaque noyau de Gabor est une fonction complexe munie d'une partie réelle $\mathrm{Re}\{Conv_{o,s}(z)\}$ et d'une partie imaginaire $\mathrm{Im}\{Conv_{o,s}(z)\}$ définie comme suit :

$$Conv_{o,s}(z) = \mathrm{Re}\{Conv_{o,s}(z)\} + i.\mathrm{Im}\{Conv_{o,s}(z)\}$$

La réponse de magnitude $\|Conv_{o,s}(z)\|$ s'exprime comme suit :

$$\|Conv_{o,s}(z)\| = \sqrt{\mathrm{Re}\{Conv_{o,s}(z)\}^2 + \mathrm{Im}\{Conv_{o,s}(z)\}^2}$$

Chaque image reçue en entrée engendre $O * S$ images qui enregistrent les réponses réelles, imaginaires, ou de magnitude au filtre de Gabor. La moyenne et la déviation par rapport à l'intensité des pixels de l'image sont ensuite calculées pour chaque réponse au filtre de Gabor. Le vecteur de caractéristiques F_I associé à une réponse spécifique d'une image I est alors composé de $2 * O * S$ éléments :

$$F_I = (M_1, D_1, M_2, D_2, ..., M_{O*S}, D_{O*S})^t$$

Ainsi 3 variations du vecteur de caractéristiques sont obtenues en utilisant les ondelettes de Gabor en fonction de la réponse choisie (réelle, imaginaire ou de magnitude).

Une évaluation a été effectuée afin de déterminer l'influence de l'échelle sur les 3 variations du vecteur de caractéristiques. Trois méthodes sont utilisées pour construire le classificateur : SVM avec un noyau RBF, KNN avec $K = 10$, et la distance de Frobenius. Les taux de classification obtenus par ces 3 méthodes sur l'ensemble de test sont illustrés respectivement par les Figures 4.15, 4.16 et 4.17. Les 8 orientations ont été utilisées alors que l'échelle a variée de 0 à 4.

L'évaluation effectuée sur l'échelle permet de déduire qu'il faut utiliser 5 échelles pour l'extraction du vecteur de caractéristiques. La seconde évaluation consiste alors à étudier l'influence de l'orientation sur l'estimation de l'orientation de la tête. Les Figures 4.18 et 4.19 présentent respectivement le taux de classification en utilisant KNN et SVM. Les résultats obtenus en utilisant la distance de Frobenius n'ont pas été reportés (taux de classification très faible).

4.4.3 Résultats expérimentaux

Afin de démontrer l'influence du nombre d'images utilisées durant l'apprentissage, { 5, 10, 15 et 20 } images ont été sélectionnées aléatoirement de l'ensemble d'apprentissage. Les

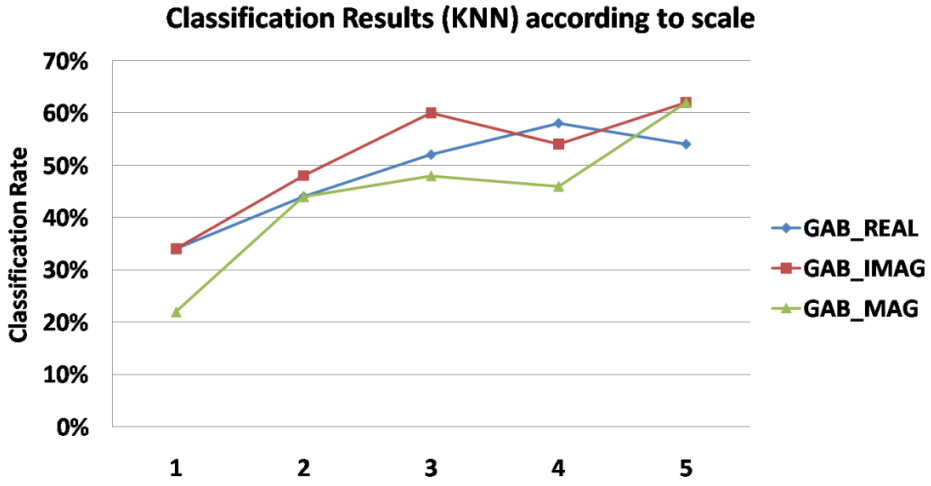


FIGURE 4.15: Taux de classification en fonction de l'échelle en utilisant KNN.

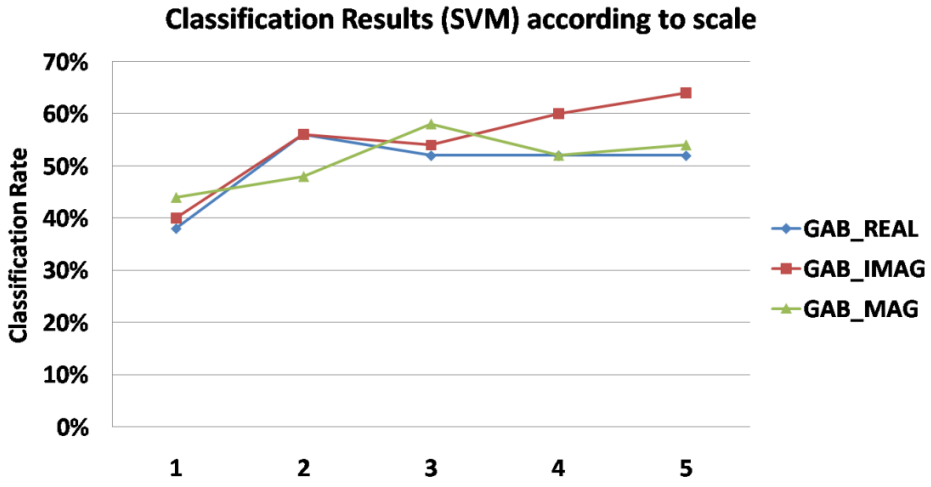


FIGURE 4.16: Taux de classification en fonction de l'échelle en utilisant SVM.

taux de classification obtenus sur l'ensemble de test en utilisant KNN et SVM sont illustrés respectivement par les Figures 4.20 et 4.21.

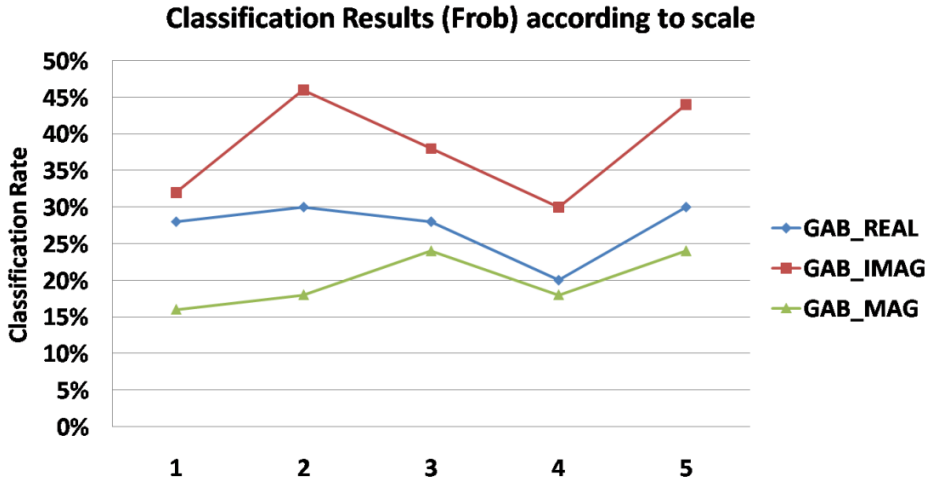


FIGURE 4.17: Taux de classification en fonction de l'échelle en utilisant la distance de Frobenius.

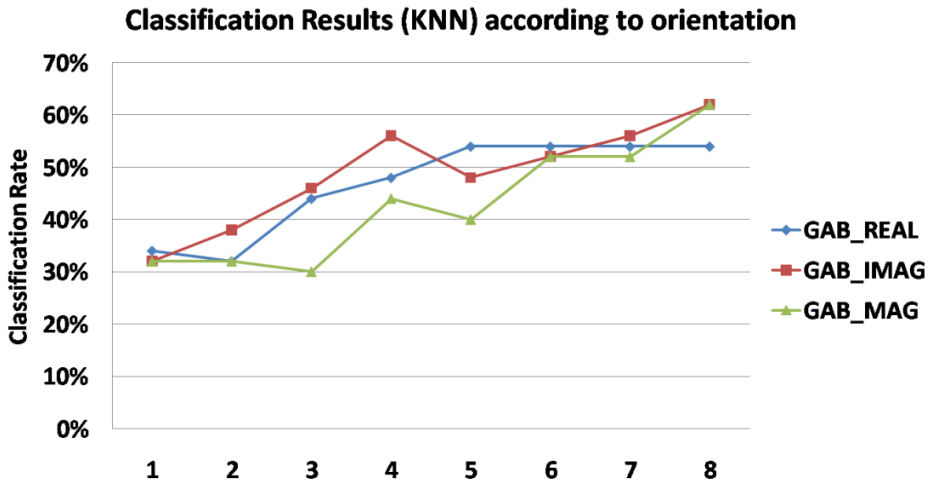


FIGURE 4.18: Taux de classification en fonction de l'orientation en utilisant KNN.

Les Figures 4.22 et 4.23 présentent respectivement les taux de classification en utilisant KNN et SVM de 5 poses. Les poses étiquetées de 1 à 5 correspondent aux paires d'angles Pan

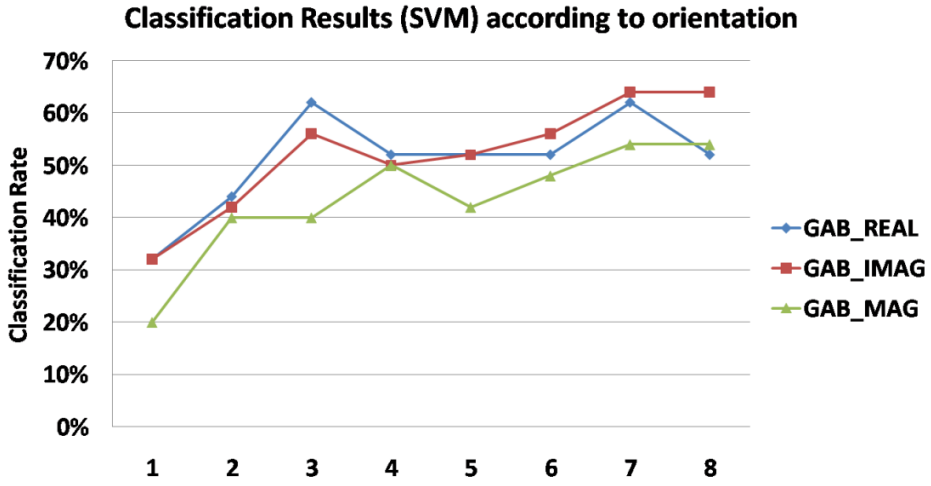


FIGURE 4.19: Taux de classification en fonction de l'orientation en utilisant SVM.

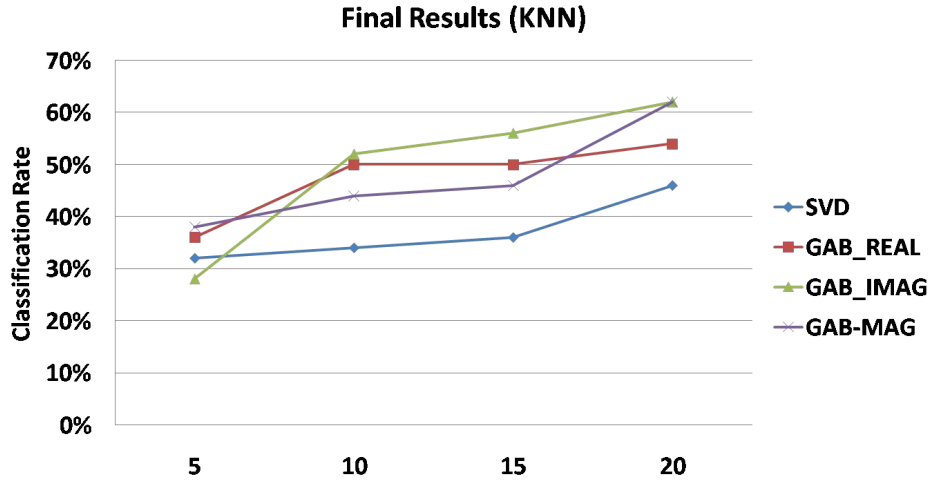


FIGURE 4.20: Taux de classification en fonction du nombre d'images utilisées durant l'apprentissage en utilisant KNN.

et Tilt suivantes :{(-60, -90), (-60, +90), (0, 0), (+60, -90), (+60, +90)}.

À partir de ces figures, les observations suivantes peuvent être émises :

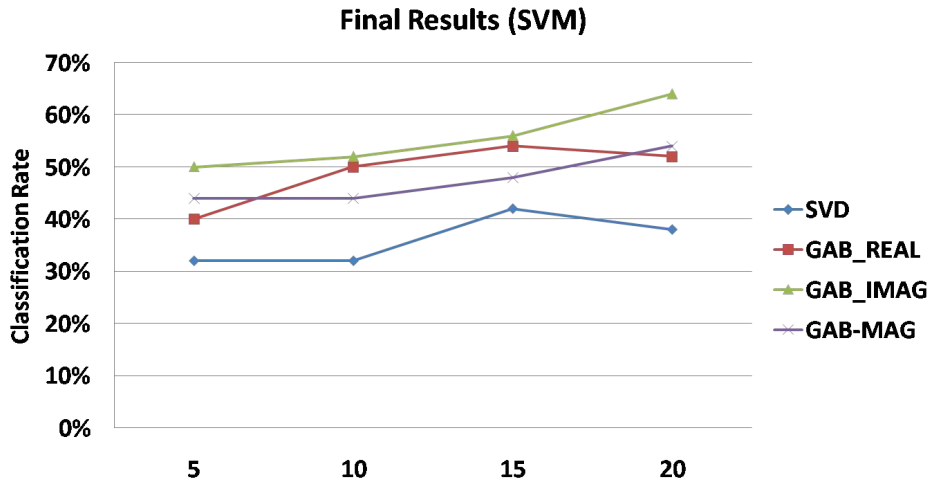


FIGURE 4.21: Taux de classification en fonction du nombre d'images utilisées durant l'apprentissage en utilisant SVM.

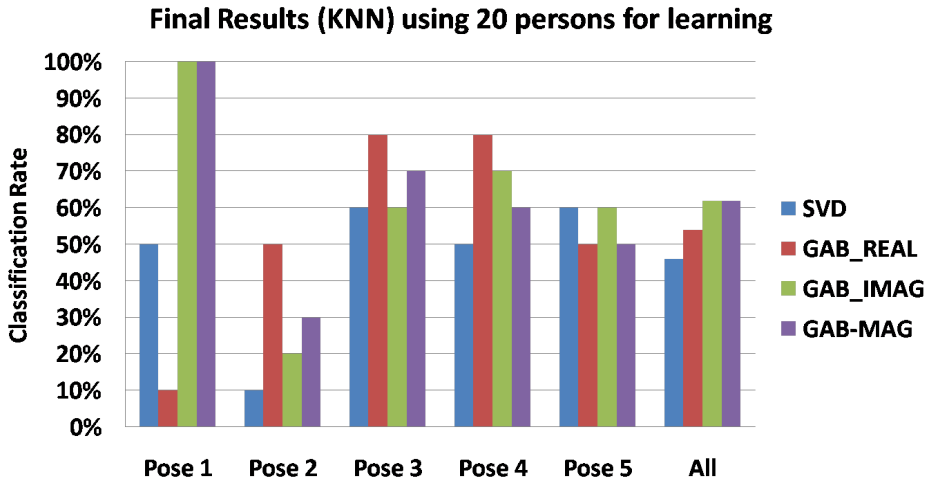


FIGURE 4.22: Taux de classification de 5 poses en utilisant KNN.

– La précision de la classification augmente en fonction du nombre d'images utilisées pour l'apprentissage. Ceci est une propriété typique de l'apprentissage supervisé.

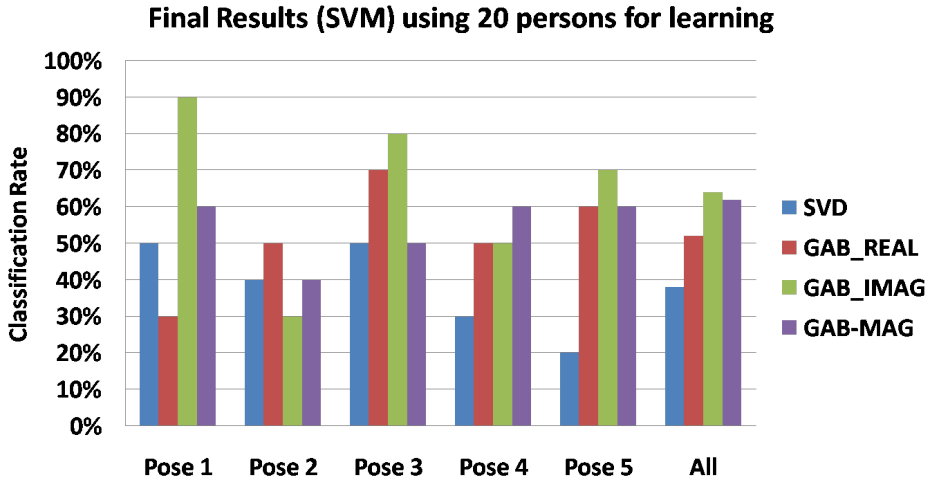


FIGURE 4.23: Taux de classification de 5 poses en utilisant SVM.

– En général, le vecteur de caractéristiques obtenu en utilisant les ondelettes de Gabor donne de meilleurs résultats que celui obtenu en utilisant SVD. Ceci est dû à la capacité des ondelettes de Gabor à manipuler différentes orientations et échelles par rapport à SVD qui ne le fait pas.
– Des 3 réponses de Gabor, il apparait qu'en général les performances de la composante réelle et imaginaire sont meilleures que celles de la réponse de magnitude. C'est dû au fait que la majorité de l'information est contenue dans la phase.

4.5 Modèle cylindrique pour le suivi de la tête

Pour être en mesure d'estimer correctement le direction du regard d'un utilisateur dans un environnement personnel, les informations provenant de l'orientation de la tête et de la position des yeux doivent être prises en considération. Nous utilisons le système décrit dans [VLS+10] qui permet d'extraire ces informations avec une grande précision à partir d'un flux vidéo. La modèle cylindrique (Cylindrical Head Model _ CHM) pour le suivi de la tête est utilisé [XKC02] alors que la position des yeux est déterminée à l'aide d'une technique basée

sur les propriétés des isophotes [VG08].

Le modèle cylindrique est utilisé pour effectuer le suivi de la tête. La projection perspective est utilisée pour reconstruire la distance focale, la hauteur de la tête et projette les pixels de l'image en points 3D et vice-versa. Les emplacements des points sur le cylindre sont trouvé avec la technique du *ray-tracing*. Ensuite, les points corrects du cylindre sont projetés en arrière sur le plan de l'image. Par le biais du flux optique, la nouvelle position de la tête et son orientation sont estimées. La Figure 4.24 illustre quelques exemples qualitatives pour le suivi de la tête en utilisant le CHM.

FIGURE 4.24: Exemples qualitatifs du suivi de la tête en utilisant le CHM.

L'inconvénient du premier sous-système (suivi de la tête) est qu'il peut converger à tort vers des minima locaux avec l'incapacité de récupérer le suivi correctement alors que le second (localisation des yeux) suppose une pose semi-frontale de la tête pour la détection des tendances circulaire des isophotes et échoue dans la localisation des yeux en présence des poses extrêmes. La combinaison de ces deux sous-systèmes permet d'obtenir de meilleurs résultats par rapport à une utilisation séquentielle. En effet, l'intégration leurs permet d'utiliser les matrices de transformation obtenues par les deux systèmes d'une manière entrelacés. Ça

permet de détecter l'emplacement des yeux étant donné la pose de la tête alors que la pose de la tête est ajustée compte tenu de la position des yeux. Les coordonnées 2D des yeux détectées dans la première image sont utilisés comme points de références (points jaunes de la Figure 4.24). Ces points de référence sont projetés sur le modèle cylindrique afin d'être utilisés pour l'estimation des emplacements successifs des yeux. La position 3D des yeux est ainsi utilisée pour mettre à jour le modèle cylindrique lorsqu'il devient instable.

4.6 Conclusion

Dans ce chapitre, nous avons présenté une taxonomie complète des différentes approches utilisées pour l'estimation de l'orientation ainsi que les difficultés qui leurs sont inhérentes. Nous avons aussi présenté la capacité humaine à effectuer cette tâche. Enfin, nous avons présenté deux approches pour l'estimation de l'orientation de la tête (un modèle basée sur l'apparence globale et un modèle cylindrique). Pour obtenir une solution robuste, il faut se placer dans un contexte bien précis en utilisant le schéma général suivant : Localisation + Suivi + Estimation. Cependant, il n'existe aucune solution universelle mais cela risque fort d'arriver d'ici peu.

Chapitre 5

Projection du champ visuel et analyse des régions d'intérêts

Sommaire

5.1 Introduction

Le champ visuel détermine ce qu'une personne est en mesure de voir. En présence d'une scène cible devant la personne, la projection du champ visuel permet de caractériser le foyer de son attention par la localisation d'une région d'intérêt. L'analyse des régions d'intérêts obtenues de plusieurs personnes à travers le temps permet une meilleure compréhension du comportement visuel des personnes observées. En plus, ça permet de mieux comprendre la scène cible et de la réorganiser si ça s'impose.

Dans ce chapitre, nous allons déterminer les caractéristiques qui permettent d'estimer le champ visuel d'une personne à partir d'un flux vidéo. Cette estimation est d'abord effectuée lorsque le sujet est en pose frontale. Quand les yeux sont visibles dans le flux vidéo à un instant donné, le point de fixation est corrigé en calculant le déplacement des yeux par rapport au point de référence. Ensuite, le champ visuel est adapté à l'orientation de la tête grâce aux valeurs des 3 degrés de liberté de la tête calculées précédemment. En présence d'une scène cible sur laquelle le sujet porte son attention, le champ visuel (ou le volume de perception) associé au sujet est projeté sur celle-ci. Cette projection permet d'extraire un point de fixation et une région d'intérêt sur la scène cible. Différentes représentations sont utilisées pour illustrer ces informations (e.g. le point de fixation, le chemin suivi par le point de fixation, la forme associée à la région d'intérêt ou une *heatmap*). Les flux vidéo sont collectés en utilisant une webcam dirigée vers le sujet. Une expérimentation est ensuite conduite impliquant des personnes qui regardent des images et des vidéos afin d'analyser la distribution des points de fixation. Cette analyse est conduite en utilisant un ensemble de métriques.

5.2 Estimation du champ visuel

Le champ visuel est l'ensemble de l'espace vu par les yeux. En présence d'une scène cible placée devant une personne, son champ visuel englobe la scène qui est explorée par une succession de points de fixation. Par contre, le champ visuel est libre en absence d'une scène cible sur laquelle le sujet pose son regard. Nous allons présenter dans ce qui suit la méthode utilisée pour l'estimer [LMD08a].

5.2.1 Données physiologiques

Le champ visuel s'étend normalement par rapport à l'axe oculaire à 60˚ en haut, 70˚ en bas, et à 90˚ environ latéralement, correspondant à un objectif photographique grand angle de 180˚. Le champ visuel d'un œil, appelé champ monoculaire, est l'ensemble de tous les points (objets, surfaces) de l'espace qui sont vus simultanément par cet œil en fixant un point (le point de fixation). La partie centrale de la rétine, appelée fovéa, permet d'obtenir une vision nette (champ visuel central) alors que sur le reste de la rétine, la vision est floue (champ visuel périphérique). Les images provenant des 2 yeux sont très semblables lors d'une vision à l'infini alors qu'elles sont légèrement différentes lors de la vision d'un objet plus rapproché car le point d'observation pour chaque œil est différent. Le cerveau interprète ces deux images et leurs différences, permettant une perception tridimensionnelle de l'objet. Le champ commun aux deux yeux, appelé *vision binoculaire*, s'étend sur environ 120˚. La Figure 5.1 illustre la division du champ visuel d'un être humain dans un plan horizontal, d'après Panero et Zelnik [PZ79].

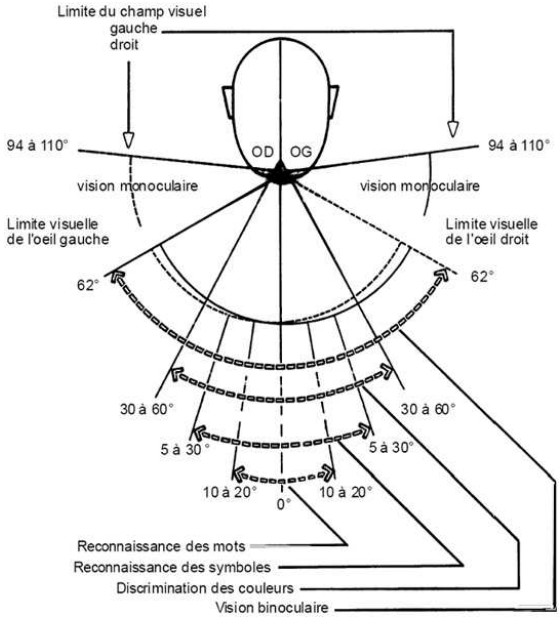


FIGURE 5.1: Division du champ visuel dans un plan horizontal [PZ79].

Les limites du champ visuel ne sont pas strictement circulaires. Il présente un aplatissement dans le secteur supérieur correspondant au relief de l'arcade sourcilière et une encoche nasale inférieure correspondant au relief du nez. Le champ visuel périphérique est mesuré en utilisant un stimulus tel que le mouvement des doigts de l'examinateur ou bien une boule placée au bout d'une tige. L'examinateur qui est placé derrière le patient (qui garde un point de fixation statique) amène ses doigts ou la boule d'avant en arrière jusqu'à ce que le patient lui annonce qu'il commence à percevoir le stimulus. La Figure 5.2 illustre la vision binoculaire humaine qui représente la partie commune aux deux champs monoculaires (droit et gauche) en fixant un point.

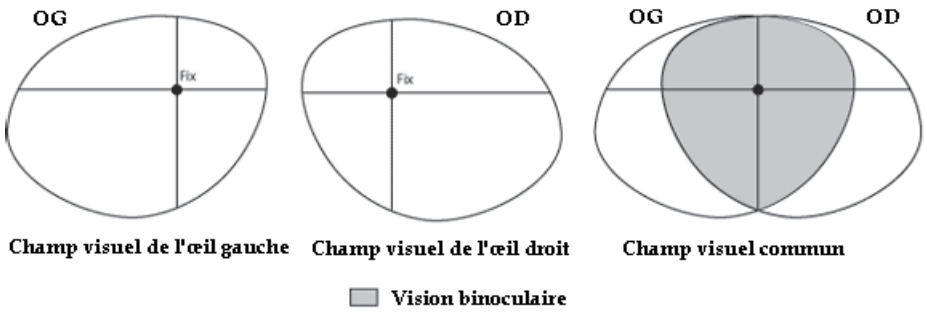


FIGURE 5.2: Représentation de la vision binoculaire.

Nous allons utiliser la vision binoculaire pour représenter le champ visuel d'une personne. En effet, cette partie centrale du champ visuel contient le point fixé par les deux yeux et permet une vision nette. Ainsi, les valeurs associées aux angles d'ouvertures horizontale et verticale seront respectivement égales à 120° et 60°.

5.2.2 Estimation du champ visuel et du point de fixation en pose frontale

L'objectif est de calculer la longueur et la hauteur du champ visuel à une certaine distance afin de déterminer les coordonnées des limites du champ visuel. Ces calculs sont effectués dans un premier temps en pose frontale (i.e. tous les degrés de liberté de la tête sont égaux à 0). Pour les besoins de la modélisation, le champ visuel qui est associé à la vision binoculaire

d'une personne est représenté par un rectangle défini par quatre points A, B, C et D. Le volume de perception du sujet est alors représenté par la pyramide $OABCD$ dont le point de départ (point d'observation) est O (souvent associé aux yeux). La Figure 5.3 illustre le champ visuel selon plusieurs vues.

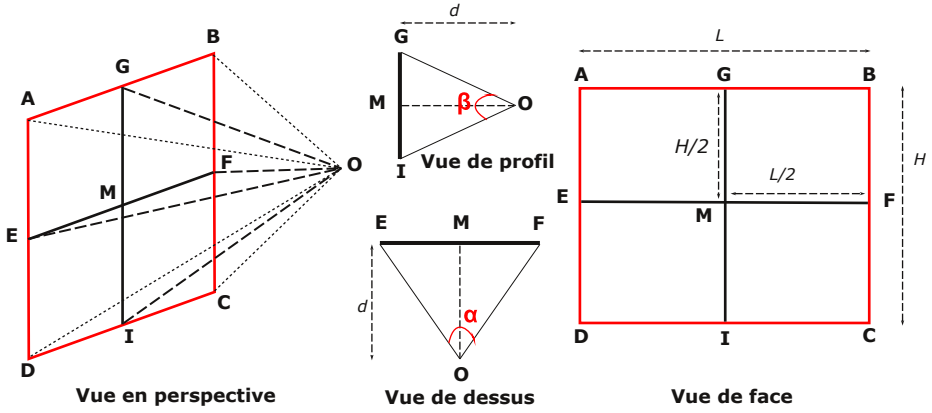


FIGURE 5.3: Différentes vues d'un même volume de perception d'une personne.

A une distance d en face du point d'observation O se situe le point de fixation M. En supposant que la personne voit autant de l'œil gauche que de l'œil droit, le point M est le centre du rectangle $ABCD$. Les milieux respectifs des segments $[AD]$ et $[BC]$ sont nommés E et F. De même, les milieux respectifs des segments $[AB]$ et $[CD]$ sont nommés G et I. Selon la vue de face du champ visuel, le point M est aussi le milieu de $[GI]$ et $[EF]$. Les angles α et β correspondent aux angles d'ouvertures horizontale et verticale respectivement. Il est possible de déduire à partir des vues de profil et de dessus les relations suivantes :

$$\widehat{MOF} = \widehat{MOE} = \frac{\alpha}{2} \quad \text{et} \quad \widehat{MOG} = \widehat{MOI} = \frac{\beta}{2}$$

Étant donné que $MO = d$, les relations trigonométriques suivantes sont déduites :

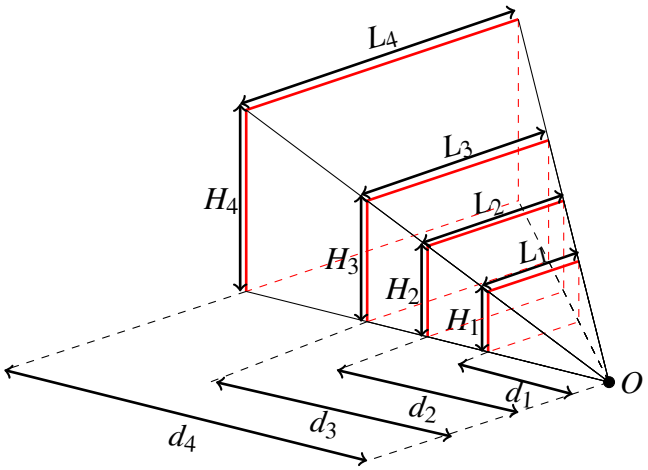
$$\begin{cases} \tan(\frac{\alpha}{2}) = \frac{MF}{MO} = \frac{MF}{d} & \Longleftrightarrow MF = d.\tan(\frac{\alpha}{2}) \\ \tan(\frac{\beta}{2}) = \frac{MG}{MO} = \frac{MG}{d} & \Longleftrightarrow MG = d.\tan(\frac{\beta}{2}) \end{cases} \tag{5.1}$$

La longueur du champ visuel est notée par L et sa hauteur par H. Les longueurs MF et MG sont connues en fonction de d, α et β. L'équation 5.1 permet donc de déduire que :

$$L = 2MF = 2.d.\tan(\frac{\alpha}{2}) \tag{5.2}$$

$$H = 2MG = 2.d.\tan(\frac{\beta}{2}) \tag{5.3}$$

Les valeurs des angles d'ouvertures horizontale et verticale peuvent être choisies en fonction de la configuration voulue. La Figure 5.4 présente les valeurs de longueurs et de hauteurs du champ visuel à différentes distances par rapport au point d'observation. La valeur de α est égale à 120° et celle de β est égale à 60° (vision binoculaire humaine).



d_i	L_i	H_i
1	3.46	1.15
1.5	5.19	1.73
2	6.92	2.30
2.5	10.39	3.46

FIGURE 5.4: Valeurs des longueurs L_i et des hauteurs H_i du champ visuel à différentes distances exprimées en mètres.

Dans le repère dont l'origine est le point O, les points A, B, C et D ont respectivement ces coordonnées à une distance d_i :

$$\begin{cases} A(-\frac{L_i}{2}, \frac{H_i}{2}, d_i) \\ B(\frac{L_i}{2}, \frac{H_i}{2}, d_i) \\ C(\frac{L_i}{2}, -\frac{H_i}{2}, d_i) \\ D(-\frac{L_i}{2}, -\frac{H_i}{2}, d_i) \end{cases} \tag{5.4}$$

Les coordonnées de ces quatre points lorsque la tête est en pose frontale sont donc déterminées uniquement à partir de la distance d, et des angles d'ouvertures horizontale et verticale α et β.

5.2.3 Adaptation du champ visuel à l'orientation de la tête

Les coordonnées des points A, B, C et D, qui forment avec le point O le volume de perception d'un sujet à une certaine distance d, définies précédemment sont correctes lorsque la personne est en pose frontale. Cependant ces coordonnées changent car la personne se déplace (translation) et effectue divers mouvements de la tête (rotations selon les 3 degrés de liberté). La Figure 5.5 présente une personne sous différentes poses devant la caméra.

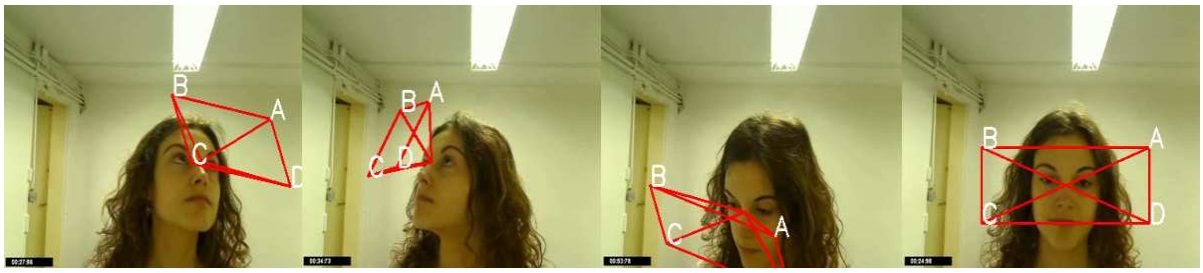


FIGURE 5.5: Représentations du champ visuel d'une personne sous différentes poses devant la caméra.

Cette adaptation du champ visuel à l'orientation de la tête peut être effectuée en utilisant une approche matricielle ou quaternionique. Chacune de ces méthodes est appropriée pour une application précise.

5.2.3.1 Approche matricielle

Les coordonnées d'un point P dans l'espace peuvent être représentées sous la forme d'un vecteur V_P de taille 3. Le point issu de la rotation de P d'un angle φ selon un des 3 axes du repère est noté P'. Les coordonnées de ce point P' sont calculées comme suit :

$$V_{P'} = R.V_P \tag{5.5}$$

Où V_P et $V_{P'}$ sont les vecteurs de taille 3 contenant les coordonnées respectives de P et P'. R est une matrice de rotation de taille 3×3 (correspondant aux 3 types de rotations possibles : tilt, pan et roll). R peut prendre la valeur d'une des 3 matrices suivantes (voir Figure 5.6) :

$$R_{(x,\alpha)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{pmatrix}$$

$$R_{(y,\beta)} = \begin{pmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{pmatrix}$$

$$R_{(z,\gamma)} = \begin{pmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

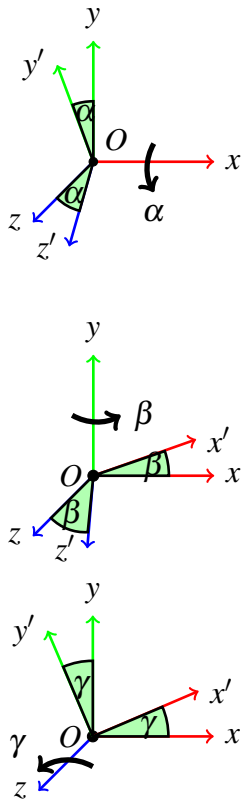


FIGURE 5.6: Trois types de rotations (tilt, pan et roll) et les matrices qui leurs sont associées.

– $R_{(x,\alpha)}$: la matrice de rotation d'angle α autour de l'axe x.
– $R_{(y,\beta)}$: la matrice de rotation d'angle β autour de l'axe y.
– $R_{(z,\gamma)}$: la matrice de rotation d'angle γ autour de l'axe z.

Cependant, les rotations dans l'espace à trois dimensions ne sont pas commutatives. En effet, une rotation autour de l'axe x suivie d'une autre autour de l'axe z donne un résultat différent de celui obtenu en effectuant d'abord la rotation autour de l'axe z, puis une autour de l'axe x. Cet exemple est illustré par la Figure 5.7.

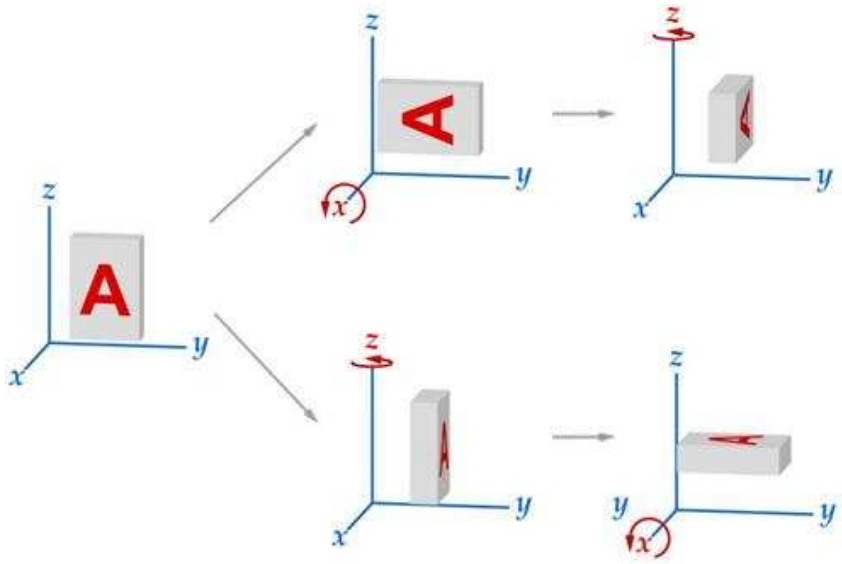


FIGURE 5.7: Non commutativité de la composition des rotations dans l'espace tridimensionnel.

Afin de palier à ce problème de commutativité, il existe 12 conventions possibles classées en deux catégories :

– **Angles de Cardan** : 6 conventions pour les rotations selon les trois axes : $x-z-y$, $x-y-z$, $y-x-z$, $y-z-x$, $z-y-x$ et $z-x-y$.
– **Angles d'Euler** : 6 autres conventions pour les rotations suivantes : $x-z-x$, $x-y-x$, $y-x-y$, $y-z-y$, $z-y-z$ et $z-x-z$.

Une convention doit être choisie pour adapter le champ visuel à l'orientation de la tête. Une matrice d'orientation est associée à chaque convention. Nous avons utilisé la convention

"$x-y-z$" (rotation autour de l'axe x puis rotation autour de l'axe y puis rotation autour de l'axe z). La matrice de rotation associée à cette convention est obtenue comme suit :

$$R_{(\alpha,\beta,\gamma)} = R_{(z,\gamma)}.R_{(y,\beta)}.R_{(x,\alpha)} \tag{5.6}$$

$$R_{(\alpha,\beta,\gamma)} = \begin{pmatrix} \cos\beta\cos\gamma & \cos\gamma\sin\alpha\sin\beta - \cos\alpha\sin\gamma & \cos\alpha\cos\gamma\sin\beta + \sin\alpha\sin\gamma \\ \cos\beta\sin\gamma & \cos\alpha\cos\gamma + \sin\alpha\sin\beta\sin\gamma & \cos\alpha\sin\beta\sin\gamma - \cos\gamma\sin\alpha \\ -\sin\beta & \cos\beta\sin\alpha & \cos\alpha\cos\beta \end{pmatrix} \tag{5.7}$$

Lorsque le sujet se déplace, il effectue des mouvements de translation par rapport au repère de la caméra décrits par le vecteur $T = (t_x, t_y, t_z)^t$ avec : t_x, t_y et t_z les translations selon les axes x, y et z respectivement.

Ainsi, pour déterminer les coordonnées d'un point P'' issu de la rotation du point P autour d'un ou de plusieurs axes suivie de sa translation (voir la Figure 5.8), l'opération suivante est effectuée :

$$V_{P''} = R.V_P + T \tag{5.8}$$

Avec V_P et $V_{P''}$ deux vecteurs de taille 3 contenant respectivement les coordonnées des points P et P''. R est la matrice de rotation et T est le vecteur de translation.

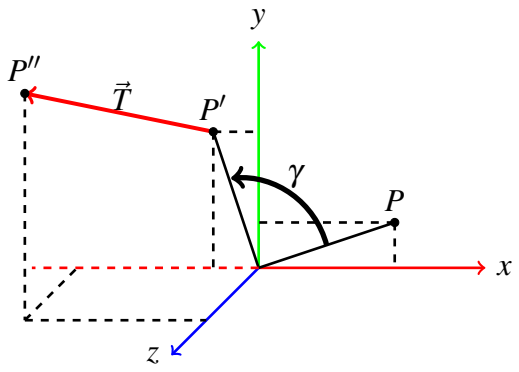


FIGURE 5.8: Rotation puis translation d'un point P.

Une transformation linéaire (multiplication par une matrice de taille 3×3) suivie d'une translation (addition d'une matrice de taille 1×3) est une application affine. Au lieu d'utiliser deux matrices (une pour la rotation et une autre pour la translation), ces deux informations peuvent être stockées dans une matrice de taille 4×4. Pour ce faire, une coordonnée factice $w \neq 0$ (généralement fixée à 1) est ajoutée. Le vecteur de taille 4 ainsi obtenu représente les coordonnées homogènes du point. Le changement réciproque de coordonnées 3D vers des coordonnées homogènes s'effectue de la manière suivante :

$$\begin{array}{rcl} \text{coordonnées 3D} & & \text{coordonnées homogènes} \\ (x,y,z) & \rightarrow & (x,y,z,w) \\ \text{coordonnées homogènes} & & \text{coordonnées 3D} \\ (x,y,z,w) & \rightarrow & (x/w, y/w, z/w) \end{array}$$

Les mouvements de rotation et de translation sont alors exprimés dans une seule matrice R de taille 4×4 :

$$R = \begin{pmatrix} a_{11} & a_{12} & a_{13} & t_x \\ a_{21} & a_{22} & a_{23} & t_y \\ a_{31} & a_{32} & a_{33} & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{5.9}$$

Où les coefficients a_{ij} sont calculés en fonction des diverses rotations effectuées (voir équation 5.7). t_x, t_y et t_z sont les valeurs du vecteur de translation T.

5.2.3.2 Approche quaternionique

Les quaternions [Gir04] notés $\mathbb{H}$ appartiennent aux nombres hypercomplexes. Ils constituent une extension des nombres complexes. Un quaternion Q peut s'écrire sous la forme suivante :

$$Q = a.1 + b.i + c.j + d.k \quad \text{avec} \quad (a,b,c,d) \in \mathbb{R}^4 \quad \text{et} \quad i^2 = j^2 = k^2 = i.j.k = -1$$

Les nombres a, b, c et d sont les caractéristiques de Q alors que i, j et k sont des coefficients

imaginaires. Il n'existe qu'une seule façon d'écrire Q sous cette forme et tout quaternion comportant les mêmes caractéristiques est nécessairement égal à Q (la réciproque est vraie). Le quaternion Q peut être décomposé de manière unique en un couple formé d'un réel a et d'un vecteur $\vec{V} \in \mathbb{R}^3$ dont les coordonnées sont (b,c,d). La notation Q = (a,$\vec{V}$) permet de définir :

- le conjugué de Q par : Q^*=(a,-$\vec{V}$)
- la norme de Q par : $\|Q\|$=$\sqrt{a^2 + \|\vec{V}\|}$

Le produit de deux quaternions Q_1=(a_1,$\vec{V_1}$) et Q_2=(a_2,$\vec{V_2}$) est calculé comme suit :

$$Q_1.Q_2 = (a_1.a_2 - \vec{V_1} \bullet \vec{V_2}, a_1.\vec{V_2} + a_2.\vec{V_1} + \vec{V}1 \wedge \vec{V_2}) \tag{5.10}$$

Où $\vec{V_1} \bullet \vec{V_2}$ et $\vec{V}1 \wedge \vec{V_2}$ désignent respectivement le produit scalaire et le produit vectoriel de $\vec{V_1}$ et $\vec{V_2}$.

L'utilisation des quaternions peut être mise en correspondance avec la composition de rotations vectorielles. En effet, une rotation vectorielle s'effectue autour d'un vecteur $\vec{N}$ dont l'origine est celle du repère (0,x,y,z). La rotation d'angle α effectuée autour d'un axe orienté selon le vecteur $\vec{N}$ de coordonnées $(N_x, N_y, N_z)^t$ (illustrée par la Figure 5.9) est associée au quaternion normalisé Q représenté par :

$$Q = \cos\frac{\alpha}{2} + \sin\frac{\alpha}{2}[i.N_x + j.N_y + k.N_z] \tag{5.11}$$

L'image du vecteur $\vec{V}$ obtenue par la rotation d'angle α autour du vecteur $\vec{N}$ est notée $\vec{V'}$. Ainsi l'expression de $\vec{V}$ est obtenue par l'égalité quaternionique suivante :

$$(0, Rot_{[\alpha,\vec{N}]}(\vec{V})) = Q.V.Q^* = (\cos\frac{\alpha}{2}, \sin\frac{\alpha}{2}.\vec{N}).(0,\vec{V}).(\cos\frac{\alpha}{2}, -\sin\frac{\alpha}{2}.\vec{N}) \tag{5.12}$$

avec Q le quaternion associé à une rotation d'angle α autour du vecteur $\vec{N}$.

L'image d'un vecteur issu de la composition de rotations vectorielles dirigées respectivement par les vecteurs $\vec{N_1}, \vec{N_2} \ldots \vec{N_n}$ d'angles respectifs $\alpha_1, \alpha_2, \ldots \alpha_n$ est obtenue comme suit :

$$(0, Rot_{[\alpha_n,\vec{N_n}]}(Rot_{[\alpha_{n-1},\vec{N}_{n-1}]}(\ldots(Rot_{[\alpha_1,\vec{N_1}]}(\vec{V}))))) = (\prod_{i=1}^{n} Q_{n-i+1}).(0,\vec{V}).(\prod_{i=1}^{n} Q_i^*) \tag{5.13}$$

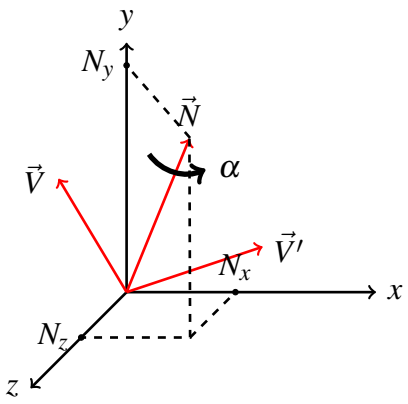


FIGURE 5.9: Rotation d'angle α d'un vecteur $\vec{V}$ autour du vecteur $\vec{N}$

avec $Q_i = (\cos\frac{\alpha_i}{2}, \sin\frac{\alpha_i}{2}.\vec{N_i})$ et $Q_i^* = (\cos\frac{\alpha_i}{2}, -\sin\frac{\alpha_i}{2}.\vec{N_i})$.

En effectuant des rotations autour des 3 degrés de liberté de la tête, l'image du vecteur $\vec{V}$ qui est issu d'une rotation d'angle α (Tilt) dirigée par un vecteur $\vec{X}$ puis d'une rotation d'angle β (Pan) dirigée par un vecteur $\vec{Y}$ puis d'une rotation d'angle γ (Roll) dirigée par un vecteur $\vec{Z}$ est déterminée comme suit :

$$(0, Rot_{[\gamma,\vec{Z}]}(Rot_{[\beta,\vec{Y}]}(Rot_{[\alpha,\vec{X}]}(\vec{V})))) = Q_3.Q_2.Q_1.(0,\vec{V}).Q_1^*.Q_2^*.Q_3^*$$

$$\begin{array}{lll} \text{avec :} & Q_1 = (\cos\frac{\alpha}{2}, \sin\frac{\alpha}{2}.\vec{X}) & , \quad Q_1^* = (\cos\frac{\alpha}{2}, -\sin\frac{\alpha}{2}.\vec{X}), \\ & Q_2 = (\cos\frac{\beta}{2}, \sin\frac{\beta}{2}.\vec{Y}) & , \quad Q_2^* = (\cos\frac{\beta}{2}, -\sin\frac{\beta}{2}.\vec{Y}), \\ & Q_3 = (\cos\frac{\gamma}{2}, \sin\frac{\gamma}{2}.\vec{Z}) & \text{et} \quad Q_3^* = (\cos\frac{\gamma}{2}, -\sin\frac{\gamma}{2}.\vec{Z}) \end{array}$$

Un exemple de la rotation du champ visuel avec un angle de π radian autour de l'axe (Oy) est illustrée par la Figure 5.10. Les calculs effectués pour obtenir les coordonnées du point A après cette rotation sont présentés en Annexe C.

Nous avons présenté dans cette section deux approches qui permettent de déterminer les coordonnées des points A, B, C et D du champ visuel après déplacement du sujet et en fonction de l'orientation de sa tête. L'approche choisie est celle des quaternions car ils sont plus simple à composer et évitent le problème du blocage de cardan. Ils sont aussi plus stables numériquement et peuvent se révéler plus efficaces étant donné qu'ils sont adoptés dans des applications

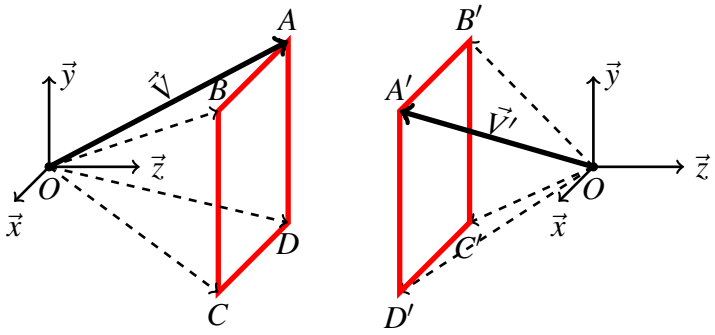


FIGURE 5.10: Rotation de π radian du vecteur $\vec{V}$ autour de l'axe Y.

en infographie, robotique et la mécanique spatiale des satellites.

5.3 Projection du champ visuel

Le but de la projection du champ visuel [LMID09] est de déterminer ce qu'une personne est en mesure de regarder sur une scène cible (e.g. un linéaire ou une vitrine d'un magasin). Cette scène cible est définie par un plan $\mathscr{P}$. La disposition du plan par rapport à la caméra déterminera les coordonnées des points A', B', C' et D' qui représentent respectivement la projection des points A, B, C et D sur la scène cible. La Figure 5.11 illustre la projection des points A, B, C et D sur deux plans différents.

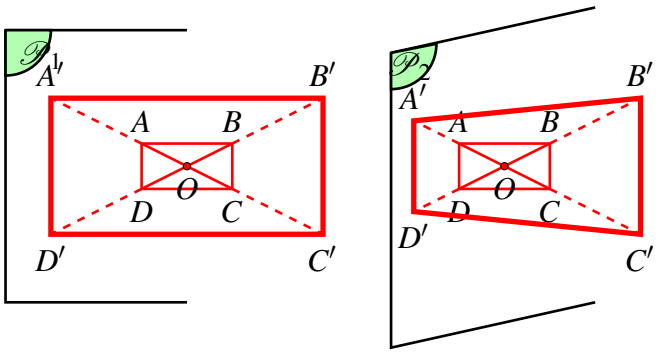


FIGURE 5.11: Projection des points A, B, C et D sur deux plans différents $\mathscr{P}_1$ et $\mathscr{P}_2$.

5.3.1 Projection d'un point

Les matrices de projections sont utilisées afin de calculer le projeté P' d'un point P sur un plan $\mathscr{P}$ d'équation $ax+by+cz+d=0$ à partir du point d'observation L. Pour réaliser la projection, le point L est considéré comme une source lumineuse ponctuelle (voir Figure 5.12). La relation suivante est alors déduite :

$$P \in [LP'] \Rightarrow \overrightarrow{PP'} = \lambda \overrightarrow{LP} \qquad (\lambda \in \mathbb{R}) \tag{5.14}$$

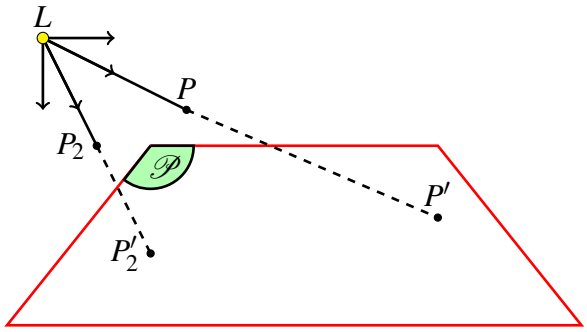


FIGURE 5.12: Projection de points sur un plan $\mathscr{P}$ via une source lumineuse ponctuelle L.

Les points P, P' et L possèdent des coordonnées en x, y et z. Elles sont notées P_x, P_y et P_z pour le point P (de même pour les coordonnées des points P' et L). D'après l'équation 5.14 :

$$\begin{cases} P'_x - P_x = \lambda (P_x - L_x) \\ P'_y - P_y = \lambda (P_y - L_y) \\ P'_z - P_z = \lambda (P_z - L_z) \end{cases} \iff \begin{cases} P'_x = \lambda (P_x - L_x) + P_x \\ P'_y = \lambda (P_y - L_y) + P_y \\ P'_z = \lambda (P_z - L_z) + P_z \end{cases} \tag{5.15}$$

Étant donné que le point P' appartient au plan $\mathscr{P}$ d'équation $ax+by+cz+d=0$, il est possible d'obtenir à partir de l'équation 5.15 l'expression suivante pour λ :

$$
\begin{aligned}
(5.15) \quad &\Rightarrow \quad a(\lambda(P_x - L_x) + P_x) + b(\lambda(P_y - L_y) + P_y) + c(\lambda(P_z - L_z) + P_z) + d = 0 \\
&\Rightarrow \quad a\lambda(P_x - L_x) + aP_x + b\lambda(P_y - L_y) + bP_y + c\lambda(P_z - L_z) + cP_z + d = 0 \\
&\Rightarrow \quad \lambda(a(P_x - L_x) + b(P_y - L_y) + c(P_z - L_z)) + aP_x + bP_y + cP_z + d = 0 \\
&\Rightarrow \quad \lambda(a(P_x - L_x) + b(P_y - L_y) + c(P_z - L_z)) = -(aP_x + bP_y + cP_z + d) \qquad (5.16) \\
&\Rightarrow \quad \lambda = \frac{-(aP_x + bP_y + cP_z + d)}{a(P_x - L_x) + b(P_y - L_y) + c(P_z - L_z)} \\
&\Rightarrow \quad \lambda = \frac{aP_x + bP_y + cP_z + d}{a(P_x - L_x) + b(P_y - L_y) + c(P_z - L_z)}
\end{aligned}
$$

La valeur de λ est réinjectée dans l'équation 5.15 afin d'obtenir la matrice de projection M_{Proj} :

$$
M_{Proj} = \begin{pmatrix}
b.L_y + c.L_z + d & -b.L_x & -c.L_x & -d.Lx \\
-a.L_y & a.L_x + c.L_z + d & -c.L_y & -d.L_y \\
-a.L_z & -b.L_z & a.L_x + b.L_y + d & -d.L_z \\
-a & -b & -c & a.Lx + b.L_y + c.L_z
\end{pmatrix} \qquad (5.17)
$$

La multiplication de cette matrice de projection avec un vecteur qui contient les coordonnées homogènes d'un point P permet d'obtenir un vecteur regroupant les coordonnées du point P' comme suit :

$$
V_{P'} = M_{Proj}.V_P \qquad (5.18)
$$

Ainsi, les coordonnées des points A', B', C' et D' qui représentent la projection des points A, B, C et D sur la scène cible à partir du point d'observation O sont calculées.

5.3.2 Projection du volume de perception

En prenant le point d'observation O comme une source lumineuse ponctuelle, le quadrilatère formé des points A', B', C' et D' ne représente pas forcément la région d'intérêt de

l'utilisateur sur la scène cible. En effet, une incohérence de la région d'intérêt extraite de la projection des points A, B, C et D sur la scène cible se produit si au moins le projeté de l'un de ces 4 points est *incorrect*. Le projeté P' d'un point P sur le plan $\mathscr{P}$ est *incorrect* si $P \notin [OP']$. Pour mieux comprendre cela, la Figure 5.13 présente une vue de derrière et une vue de profil de la projection des points A, B, C et D du champ visuel sur le plan $\mathscr{P}$ en fonction de différentes valeurs de l'angle de rotation tilt. La scène cible est délimitée en rouge sur le plan $\mathscr{P}$ dont la surface est de couleur noire.

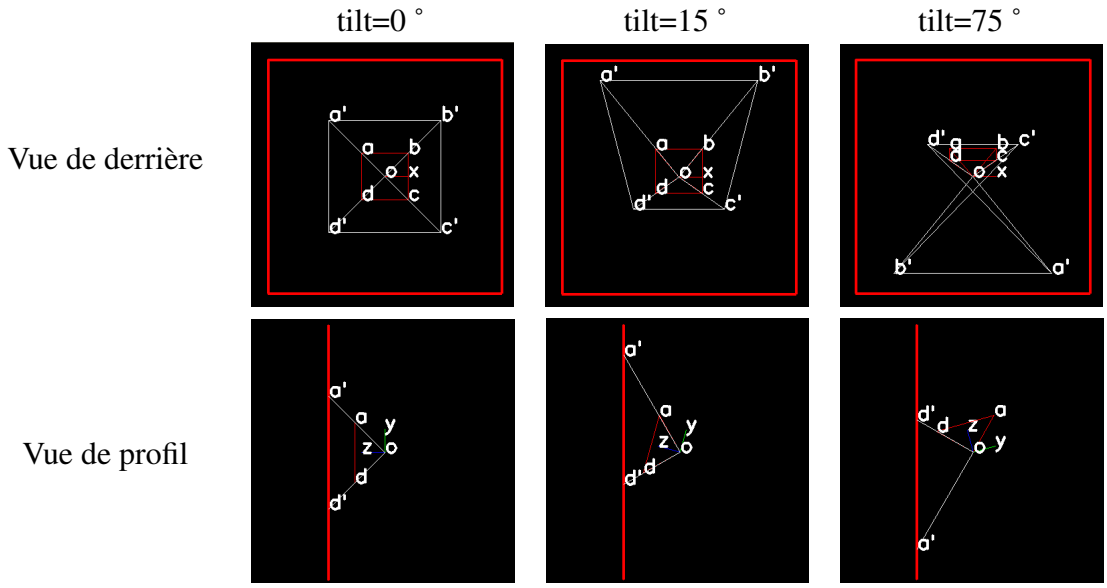


FIGURE 5.13: Projection des points A, B, C et D sur le plan $\mathscr{P}$ selon différents valeurs de l'angle tilt.

- Tilt=0° : la région d'intérêt extraite est correcte car $A' \in \mathscr{P}$ et $A \in [OA']$.
- Tilt=15° : la région d'intérêt extraite est correcte car $A' \in \mathscr{P}$ et $A \in [OA']$.
- Tilt=75° : la région d'intérêt extraite est incorrecte car la condition $A \in [OA']$ n'est pas respectée. Il résulte du fait que le point A se situe derrière la source lumineuse.

Les conditions nécessaires pour que P' le projeté du point P sur le plan $\mathscr{P}$ ne soit pas *incorrect* sont : $P' \in \mathscr{P}$, $P' \in (OP)$ et $P \in [OP']$. La Figure 5.14 illustre la projection du segment $[AD]$ sous deux orientations différentes. Le segment rouge représente la projection théorique

de $[AD]$ alors que le segment vert représente le segment liant les deux points A' et D'. Dans la partie gauche de la figure, les points A' et D' représentent les limites de l'espace de projection théorique correspondant à une projection correcte. Par contre, dans la partie droite le point A' n'est pas projeté correctement. Une correction du point A' s'impose pour obtenir un affichage correct de la projection du champ visuel sur le plan $\mathscr{P}$.

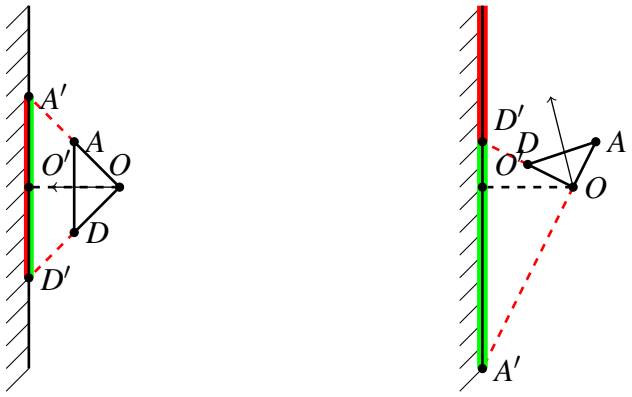


FIGURE 5.14: Projection sur un plan du segment $[AD]$ du champ visuel selon des angles de rotations différents.

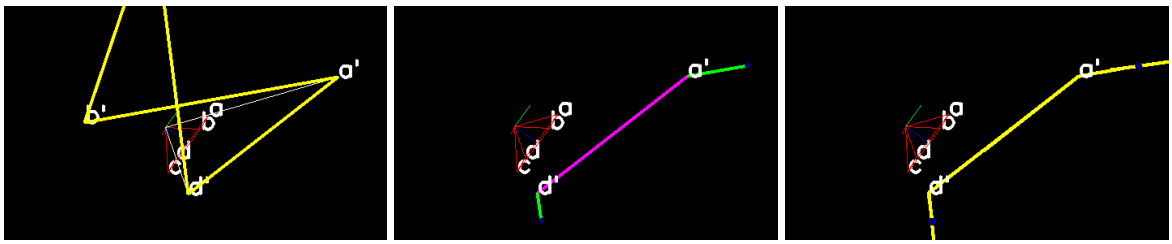
La projection du champ visuel n'est plus fermée et s'étend à l'infini lorsque le projeté de l'un des 4 points (A, B, C et D) est *incorrect* (car l'utilisateur voit une partie de la scène cible s'étalant à l'infini). Cette ouverture de la projection nécessite l'introduction de points intermédiaires situés entre les segments reliant les limites du champ visuel. Ces points permettent de corriger les points de type *incorrect* car l'alignement des points est conservé lors de la projection.

La correction d'un point de type *incorrect* P' s'effectue de la manière suivante :

1. Vérifier si ces voisins sont de type *incorrect*.
2. Pour chaque voisin P_v qui ne soit pas de type *incorrect* effectuer les étapes suivantes :
   - Chercher un point intermédiaire P_t appartenant au segment qui relie le point à corriger P' et son voisin P_v qui ne soit pas de type *incorrect*.
   - Tracer la demi-droite $[P_vP_t)$.

Cet algorithme est utilisé pour chaque point (A', B', C' et D') de type *incorrect*. Le nombre maximum de points *incorrect* est 3 (car 4 signifie que le sujet ne fixe plus la scène cible).

La Figure 5.15 montre un cas où la correction de la projection du champ visuel est nécessaire. La Figure 5.15(a) illustre une projection incohérente car les points B' et C' sont *incorrect*. Dans la Figure 5.15(b), des points intermédiaires sont utilisés pour corriger la projection (les nouveaux points sont affichés en bleu). Dans la Figure 5.15(c), la région d'intérêt obtenue est correcte (issue du prolongeant des points intermédiaires).



(a) Projection originale (b) Calcul des nouveaux points (c) Correction de la projection

FIGURE 5.15: Correction de la projection du champ visuel lorsque les points B et C sont *incorrect*.

5.4 Affichage du champ visuel et de sa projection sur une image

Une fois le champ visuel d'une personne déterminé et sa projection effectuée, il faut l'afficher sur une image afin de visualiser le champ visuel du sujet ainsi que sa projection sur une image de la scène cible.

L'un des modèles décrivant le processus de formation d'image le plus simple est le modèle du sténopé (*pin-hole*) [BÓ2]. Ce modèle est une représentation linéaire de la projection perspective qui permet de simplifier considérablement le calcul. Cependant, il n'est qu'une approximation. Ce modèle va être adapté pour l'affichage du champ visuel. Il utilise 3 repères (voir Figure 5.16) :

- Le repère univers $(R_u, \overrightarrow{x_u}, \overrightarrow{y_u}, \overrightarrow{z_u})$ dans lequel un point a pour coordonnées (x,y,z).
- Le repère de la caméra $(R_c, \overrightarrow{x_c}, \overrightarrow{y_c}, \overrightarrow{z_c})$ avec R_c le centre optique de la caméra.
- Le repère associé à l'image visualisée $(R, \overrightarrow{u}, \overrightarrow{v})$ dans lequel un point a pour coordonnées (u,v).

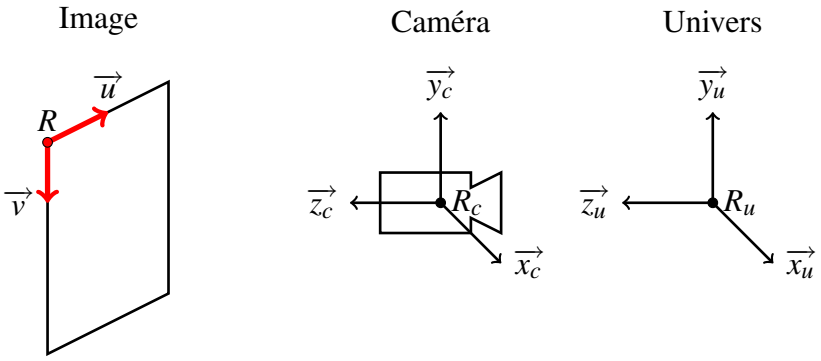


FIGURE 5.16: Repères du modèle sténopé.

À des fins de simplification, l'axe optique de la caméra (axe $\overrightarrow{z_c}$ du repère R_c) est considéré orthogonal au plan de projection. Le but est de calculer les coordonnées u et v (en pixels) résultant de la projection d'un point de l'espace sur l'image (dans le repère R). Pour ce faire, il faut définir les deux types de paramètres du modèle : les paramètres extrinsèques et les paramètres intrinsèques.

Paramètres extrinsèques : sont relatifs à la position d'un point de l'espace dans le repère de la caméra. Ces paramètres ont été déterminés dans les parties relatives aux rotations et translations de la Section 5.2.3.

Paramètres intrinsèques : les paramètres intrinsèques sont :

- La distance focale f : la distance (en millimètres) entre l'image et la caméra.
- Le rapport d'échelle (exprimé en pixels/mm) : représente l'équivalence entre la distance entre deux points de l'image (en pixels) et la distance entre ces deux points dans le monde réel (en millimètres). Cette distance est mesurée en longueur (p_x) et en hauteur (p_y).

– Les coordonnées (u_0,v_0) en pixels : représentent la projection du centre optique de la caméra sur le plan image.

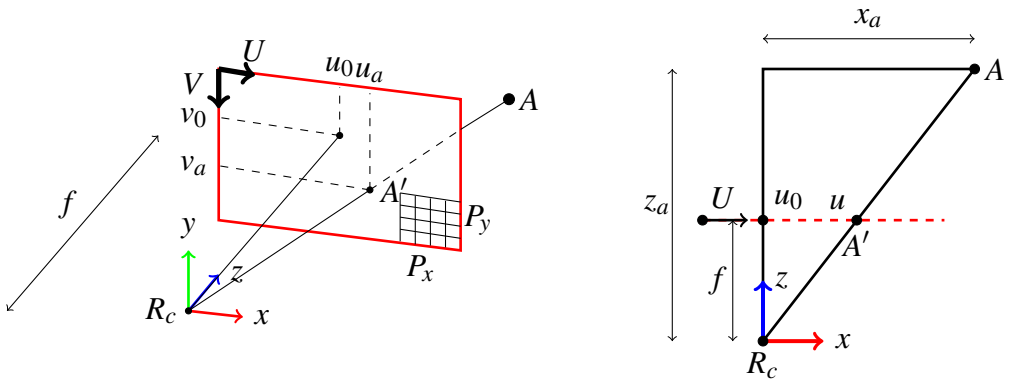


FIGURE 5.17: Vue latérale (à gauche) et vue de dessus (à droite) d'une projection perspective avec le modèle du sténopé.

La projection perspective d'un point sur le plan image (voir Figure 5.17) est donnée par l'expression canonique suivante :

$$u = u_0 + f.\frac{y_a}{p_x.z_a} \tag{5.19}$$

$$v = v_0 + f.\frac{y_a}{p_y.z_a} \tag{5.20}$$

Ces deux expressions ont été obtenues en utilisant la propriété des triangles semblables comme suit :

$$\frac{z_a}{f} = \frac{x_a}{p_x.(u-u_0)} \Rightarrow p_x.z_a.(u-u_0) = f.x_a \Rightarrow u-u_0 = \frac{f.x_a}{p_x.z_a} \Rightarrow u = u_0 + \frac{f.x_a}{p_x.z_a}. \tag{5.21}$$

$$\frac{z_a}{f} = \frac{y_a}{p_y.(v-v_0)} \Rightarrow p_y.z_a.(v-v_0) = f.y_a \Rightarrow v-v_0 = \frac{f.y_a}{p_y.z_a} \Rightarrow v = v_0 + \frac{f.y_a}{p_y.z_a}. \tag{5.22}$$

5.5 Extraction des régions d'intérêts

L'objectif est d'extraire les régions d'intérêts d'un ou de plusieurs utilisateurs dans la scène cible [LMID09].

5.5.1 Représentation des informations sur le regard

Les informations issues du regard peuvent être représentées de différentes manières (voir Figure 5.18).

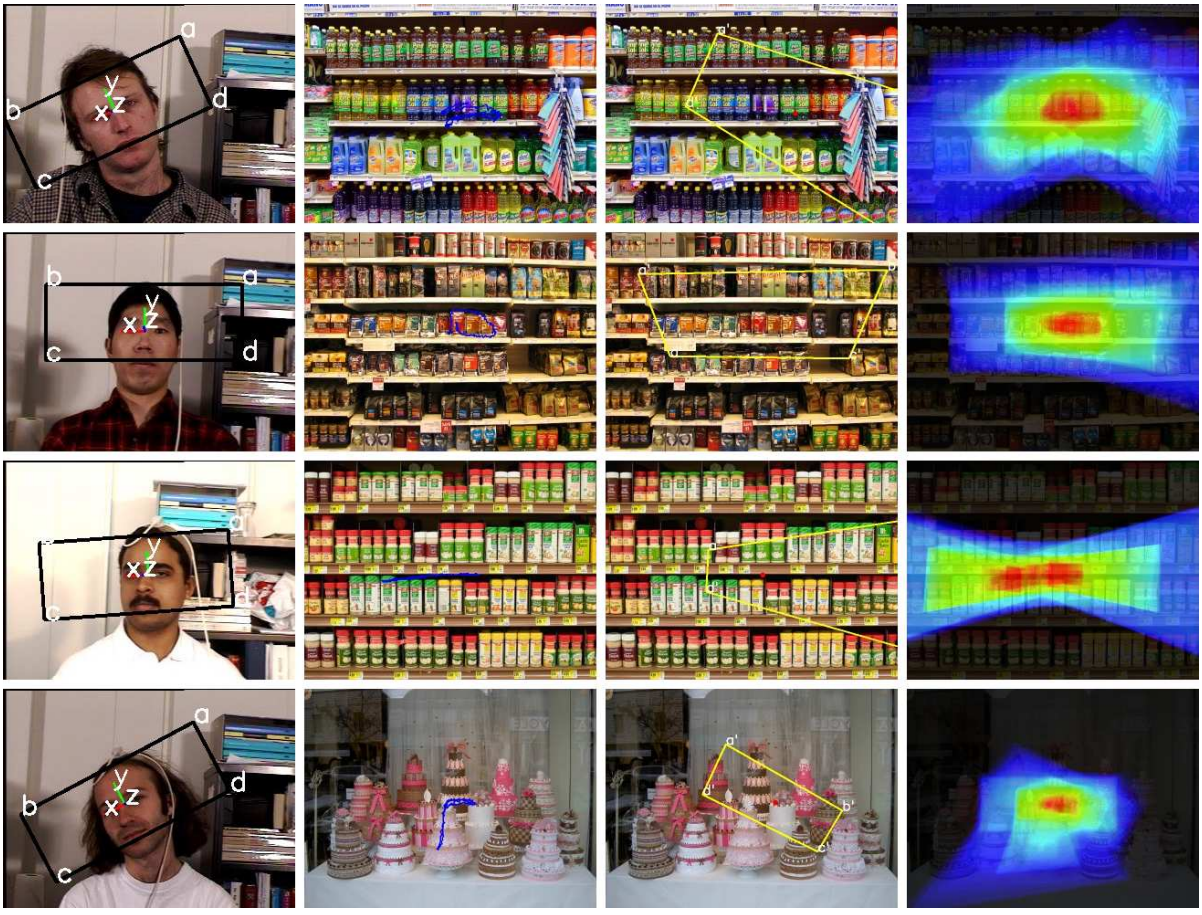


FIGURE 5.18: Représentation des informations sur le regard.

La base d'images de l'Université de Boston [CSA00] est utilisée pour valider l'approche proposée. Nous avons utilisé les représentations suivantes qui ont été illustrées par la Figure 5.18.

- Le champ visuel d'une personne qui regarde un linéaire.
- Le chemin suivi par le centre de la région d'intérêt (i.e. le point de fixation) à travers le temps.
- La région d'intérêt extraite lors de la projection du champ visuel.
- Les régions d'intérêts cumulées à travers le temps pour une même personne avec une plus grande importance donnée à la partie centrale de la région.

5.5.2 Correction du point de regard

Le point de regard M' (représenté par le point bleu dans la Figure 5.19) est la projection du point central M sur la scène cible. Lorsque la localisation des yeux est possible en utilisant la méthode proposée dans [VG08], la position du point de regard est altérée (en fonction de la distance par rapport au repère associé à la position des yeux). Le résultat de cette modification est un nouveau point de regard M_{disp} qui représente le déplacement du point M d'un certain pourcentage à l'intérieur du champ visuel. La projection du point M_{disp} notée M'_{disp} (représenté par le point rouge dans la Figure 5.19) est obtenue par l'intersection entre le plan $\mathscr{P}$ et la droite (OM_{disp}) avec $M_{disp}(M_x + L * disp_x, M_y + H * disp_y)$.

Étant donné que le mouvement des yeux est censé compenser celui de la tête, l'écart type entre les points de regard obtenus avec la combinaison de la pose de la tête et de la position des yeux devrait être plus petit par rapport à celui obtenu de la pose de la tête uniquement. Afin de le démontrer, nous avons analysé l'écart type [VLS+10] sur un sous ensemble de la base de l'Université de Boston. L'expérimentation a permis de réduire l'écart-type de 61.06% sur X et de 52.23% sur Y avec des écart-type respectifs de 20.48% et 19.05%. La Figure 5.20 illustre un exemple (du sujet "jam8") dans lequel l'utilisation du déplacement des yeux (en points bleus) aide significativement à réduire l'écart-type des points de regard donnés par la pose de la tête uniquement (en croix rouges). Cependant, le point de regard est localisé autour de la même position en utilisant l'information provenant des yeux.

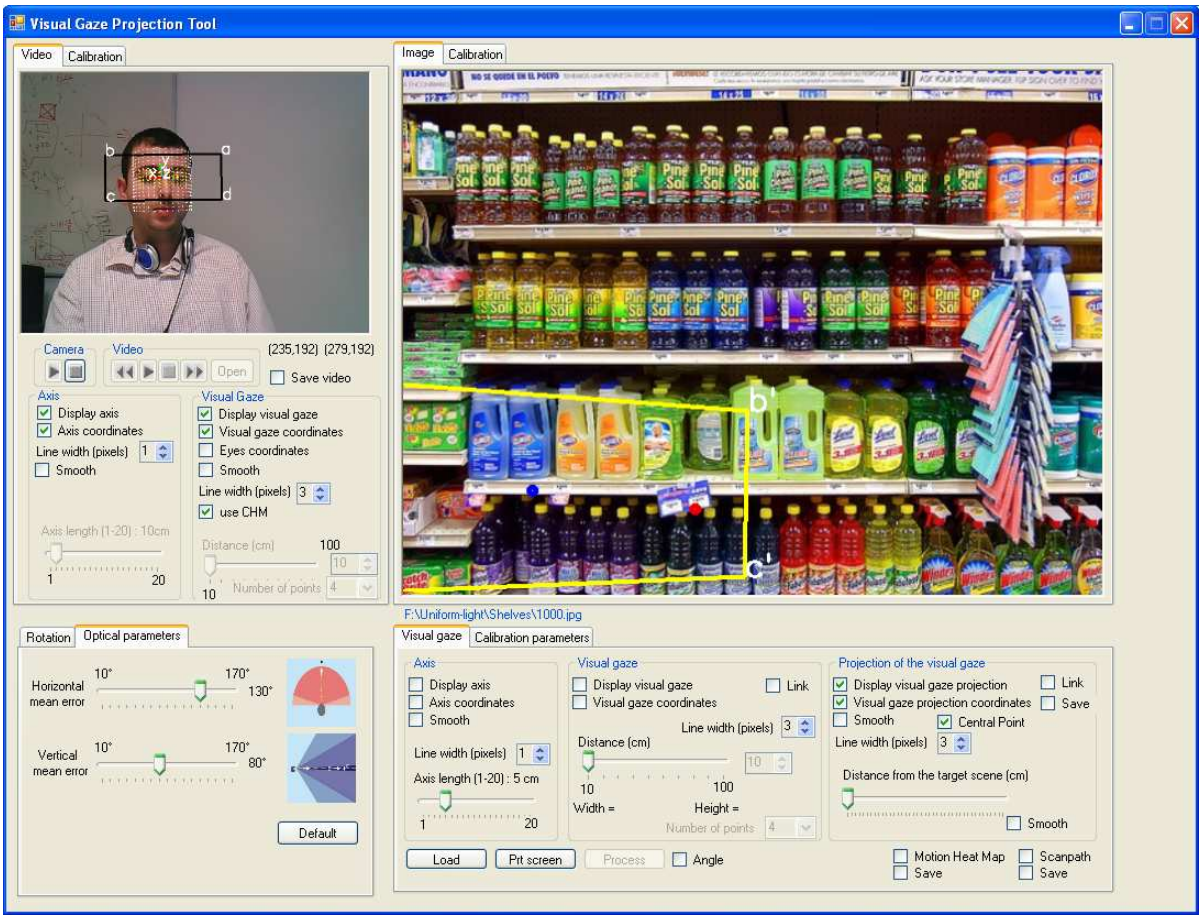


FIGURE 5.19: Capture d'écran du système exécuté en temps réel en utilisant une simple webcam. Le rectangle jaune indique la région d'intérêt de l'utilisateur (définie par l'orientation de la tête uniquement) alors que le point rouge représente la projection du point de regard (définie par l'orientation de la tête et la position des yeux).

L'analyse des résultats permet de noter que dans certaines vidéos l'estimation de l'orientation de la tête est décalée par rapport à la localisation des yeux. En effet, un léger déplacement entre la position actuelle des yeux et leurs références fait croire au système que l'utilisateur regarde dans la même direction que le mouvement. Il s'agit d'un problème inhérent du suivi qui peut être résolu par un filtre de Kalman. Cependant, une supposition erronée sur la prochaine

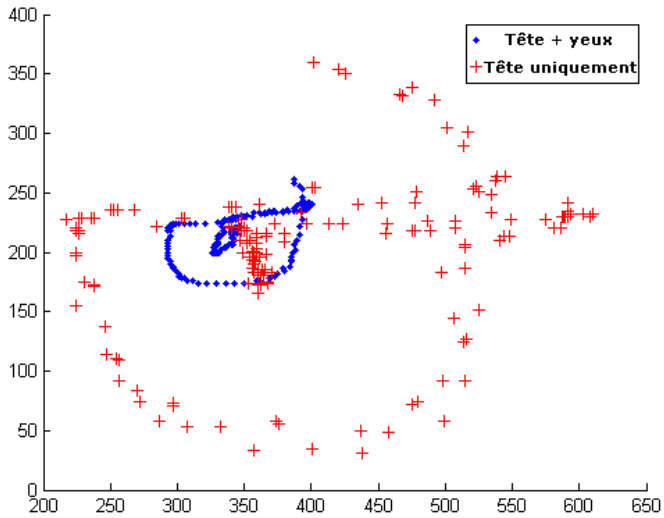


FIGURE 5.20: Exemple de la correction du point de regard par la position des yeux sur la vidéo "jam8.avi".

position de la tête implique une mauvaise estimation de l'orientation de la tête.

Afin d'évaluer l'influence du déplacement des yeux sur le point de regard, un facteur de pondération de 0.2 a été trouvé de manière empirique. Il est à noté qu'un facteur de pondération de 0.1 avait peu d'effets sur l'estimation finale alors qu'un facteur supérieur à 0.3 génère de nombreuses valeurs aberrantes conduisant à un écart-type large.

5.5.3 Calcul des angles *tilt* et *pan* correspondant à un point de regard

Dans certaines applications, il est nécessaire d'obtenir les angles de rotations utilisés par une personne se trouvant à une distance d de la scène cible pour regarder un point de celle-ci. L'angle roll ne peut être estimé en utilisant cette méthode. L'objectif est de déterminer les valeurs des angles de rotations tilt (α) et pan (β) pour aller du point O_p (projection orthogonale du point O appartenant au volume de perception du sujet sur l'image) vers un point P quelconque (voir Figure 5.21).

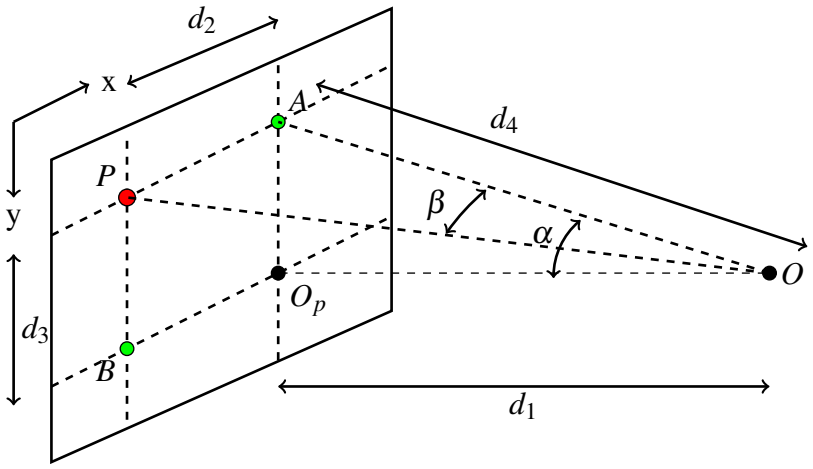


FIGURE 5.21: Calcul du tilt et du pan pour passer du point O_p vers le point P. Ici le tilt vaut α et le pan vaut β

Soit A un point dans l'image ayant pour coordonnées $(O_p.x, P.y)$ et d_2=$|A.x - O_p.x|$ et d_3= $|A.y - O_p.y|$ des longueurs exprimées en pixels. La longueur d_1 qui exprime la distance entre le sujet et la cible est exprimée en mètres. Afin d'utiliser les mêmes unités de mesure, il faut convertir les valeurs de d_2 et d_3 en mètres en les multipliant respectivement par les valeurs p_x et p_y (les paramètres intrinsèques du modèle du sténopé). L'angle tilt (α) est alors calculé comme suit :

$$tan(\alpha) = \frac{d_3 * p_y}{d_1} \Longleftrightarrow \alpha = atan\left(\frac{d_3}{d_1}\right) \Longleftrightarrow \alpha = atan\left(\frac{|A.y - O_p.y| * p_y}{d_1}\right) \tag{5.23}$$

L'angle α permet de passer du point O_p vers le point A. Il faut ensuite déterminer l'angle β afin de passer de A vers P. Pour cela la longueur $d4$=$\sqrt{{d_1}^2 + {d_3}^2}$ est utilisée et l'angle pan (β) est calculé comme suit :

$$tan(\beta) = \frac{d_4}{d_2 * p_x} \Longleftrightarrow \beta = atan\left(\frac{d_4}{d_2 * p_x}\right) \Longleftrightarrow \beta = atan\left(\frac{\sqrt{{d_1}^2 + (|A.y - O_p.y| * p_y)^2}}{|A.x - O_p.x| * px}\right) \tag{5.24}$$

5.6 Métriques pour l'analyse du regard

Nous proposons d'analyser les points de fixation obtenus pour mieux comprendre le comportement visuel des personnes. Afin de valider notre approche, nous avons choisi de construire un système qui mesure la pertinence (ou la qualité) d'un média visuel [MLLD09]. Cette analyse se base sur plusieurs métriques [Lew06].

5.6.1 Construction d'un système de mesure de pertinence d'un média

A titre d'exemple, nous proposons dans les sections suivantes un système mesurant la qualité d'un média visuel. Ce système permet de déterminer l'aptitude à transmettre l'idée originale du créateur (e.g. un message publicitaire) afin de préconiser des recommandations sur la création du média. Les métriques qualitatives utiles à l'évaluation de la perception du média sont obtenues par l'analyse des données issues de la direction du regard des utilisateurs.

5.6.1.1 Collecte des données brutes

A partir de l'estimation du champ visuel, les points de regard sont déterminés. Cette opération permet de récupérer en sortie les coordonnées horizontales et verticales des points de regard par rapport à la scène cible (e.g. écran, linéaire, etc.). Chaque point P_i est représenté par le triplet (x_i, y_i, t_i). La séquence de points (voir l'exemple Figure 5.22) associée aux points de regard d'un utilisateur forme le *chemin* ou *scanpath* du regard. Le chemin oculaire se compose d'une séquence de points de fixation séparés par des saccades oculaires. Les saccades sont des mouvements des yeux rapides permettant au regard de se déplacer d'une zone d'intérêt à une autre. Les points de fixation sur ces zones d'intérêt, périodes où les mouvements des yeux sont stationnaires, permettent au cerveau d'analyser l'information perçue. La durée de ces points de fixation varie selon les auteurs autour de 70ms à 100ms [Ray98, PBP04]. L'ensemble des points de regard P_u d'un utilisateur u qui participe à l'expérience sont classés en deux catégories : points de fixation ou saccades.

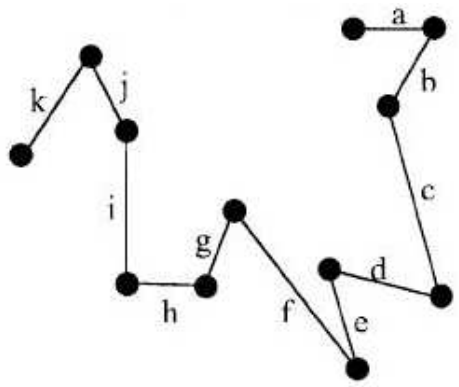


FIGURE 5.22: Scanpath associé à un utilisateur.

5.6.1.2 Identification des fixations

Afin d'analyser et d'effectuer des traitements sur les *scanpaths*, l'identification des fixations et des saccades est nécessaire. En effet, les fixations et les saccades oculaires sont souvent utilisées comme des informations principales pour les différentes métriques qui sont utilisées pour interpréter les mouvements des yeux (e.g. le nombre de fixations, de saccades, la durée de la première fixation, l'amplitude moyenne des saccades, etc.). La technique d'identification la plus répandue se base sur le calcul de la vélocité de chaque point. La vélocité d'un point est la vitesse angulaire de l'œil (calculée en degrés par seconde) et correspond à la distance qui le sépare de son prédécesseur ou son suivant. L'identification de la catégorie d'un point se fait en fonction d'un seuil. Deux points successifs séparés par une distance inférieure à ce seuil sont considérés comme une fixation. Un autre seuil correspondant à la durée minimale d'une fixation permet d'éliminer les groupes insignifiants.

5.6.2 Métriques relatives à la distribution des fixations

Les fixations de tous les utilisateurs sont regroupées dans l'ensemble F. Le processus de regroupement permet de réduire les caractéristiques spatiales des fixations en un ensemble limité de classes K_i. Le regroupement est réalisé par une technique de classification non supervisée

telle que *K-means*. Ces classes définissent les régions spatiales où la plupart des utilisateurs ont porté leur attention. Sur la base de l'ensemble des classes K_i, les métriques suivantes sont définies :

- **Le nombre moyen de fixations** $\bar{n}$ pour tous les participants est défini par :

$$\bar{n} = \frac{1}{\|U\|} \sum_{u} \|F_u\|$$

  U désigne l'ensemble de tous les participants et son cardinal $\|U\|$ représente leurs nombre. $\|F_u\|$ est le nombre de points de fixation d'un participant u.
- **La durée moyenne des fixations** $\bar{\Delta}$ est obtenue en utilisant la formule suivante :

$$\bar{\Delta} = \frac{1}{\|F\|} \sum_{f \in F} \Delta(f)$$

  $\Delta(f)$ est la durée de la fixation f obtenue en soustrayant l'instant associé au premier point de fixation de celui associé au dernier point de la fixation. $\|F\|$ est le nombre total des points de fixation.
- **La durée maximale d'une fixation** Δ_{max} est définie par :

$$\Delta_{max} = Max_{f \in F}(\Delta(f))$$

- **La durée moyenne de la première fixation** $\bar{\Delta}_1$ est définie par :

$$\bar{\Delta}_1 = \frac{1}{\|U\|} \sum_{u} \Delta(f_u^1)$$

  f_u^1 correspond à la première fixation du participant u.
- **La durée moyenne d'un chemin (sur des images)** $\bar{D}$ représente le temps moyen passé par un participant à explorer la scène cible. Elle est définie comme suit :

$$\bar{D} = \frac{1}{\|U\|} \sum_{u} D_u$$

– **La longueur moyenne d'un chemin** $\bar{L}$ est définie comme suit :

$$\bar{L} = \frac{1}{\|U\|} \sum_{u} L_u$$

L_u correspond à la longueur d'un chemin d'un participant u calculée grâce à la somme des longueurs des segments qui relient ses points de fixation :

$$L_u = \sum_{i=1}^{\|F_u\|-1} Dist(f_i, f_{i+1})$$

– **Le nombre moyen de visites** $\bar{H}^r$ dans une région spécifique de l'image ou de la vidéo r est défini comme suit :

$$\bar{H}^r = \frac{1}{\|U\|} \sum_{u} H_u^r$$

H_u^r est le nombre de visites dans une région spécifique r pour un participant u. Il est défini comme suit :

$$H_u^r = \|\{f_i | f_i \in r\}\|$$

– **La zone convexe moyenne correspondant à un chemin** $\bar{S}$ est définie comme suit :

$$\bar{S} = \frac{1}{\|U\|} \sum_{u} S_u$$

S_u correspond à la surface de l'enveloppe convexe qui englobe les fixations d'un participant u. Elle est définie comme suit :

$$S_u = convexHullSurface(F_u)$$

– **Le nombre moyen de régressions** $\bar{R}$ est défini comme suit :

$$\bar{R} = \frac{1}{\|U\|} \sum_{u} R_u$$

R_u correspond au nombre de régressions qui se sont produites pour un participant u. Il

est défini comme suit :

$$R_u = \|\{f_i | f_{i-1}\widehat{f_i f_{i+1}} < 90^\circ\}\|$$

– **Le coefficient de Gini** G est une mesure statistique de dispersion largement utilisée en économie pour estimer l'inégalité des richesses ou la distribution des revenus. La définition suivante du coefficient de Gini est utilisée sur la scène cible partitionnée en rectangles :

$$G = \sum_{i,j} (\frac{\phi_{i,j}}{\|F\|})^2$$

$\phi_{i,j}$ est le nombre de points de fixations se trouvant dans le bloc (i, j). La valeur du coefficient de Gini appartient à l'intervalle [0, 1]. Il quantifie le degré de concentration (ou de dispersion) de l'ensemble des points de fixation d'une image. Une valeur proche de 1 indique une forte dispersion des points alors qu'une valeur proche de 0 signifie que les points sont fortement concentrés dans quelques parties de l'image.

5.6.3 Expérimentation

L'expérimentation a permis d'enregistrer l'information issue du regard de 10 personnes. Les participants ont été invités à regarder de manière attentive plusieurs images successives et des séquences vidéo. Toutes les informations sur le regard ont été enregistrées et traitées. Voici les résultats obtenus sur des images et sur des vidéos.

5.6.3.1 Sur des images

La Figure 5.23 représente deux affiches (l'une pour un événement social et l'autre pour un événement scientifique). La représentation de la première affiche s'effectue par une superposition du chemin sur l'image. Il est composé de plusieurs points (chacun étant lié au précédent pour représenter le chemin d'accès). La représentation de la deuxième affiche s'effectue par la superposition sur l'image d'une heatmap (carte de chaleur). Le résultat du regroupement des points de fixation défini des régions chaudes et des régions froides en fonction du nombre de fois que la zone a été regardée. Ceci donne une indication pour savoir si les informations principales sont vues ou non. Ces informations sont mises en correspondance avec les exigences

du producteur du média (e.g. vérifier si la liste des sponsors est bien mise en évidence).

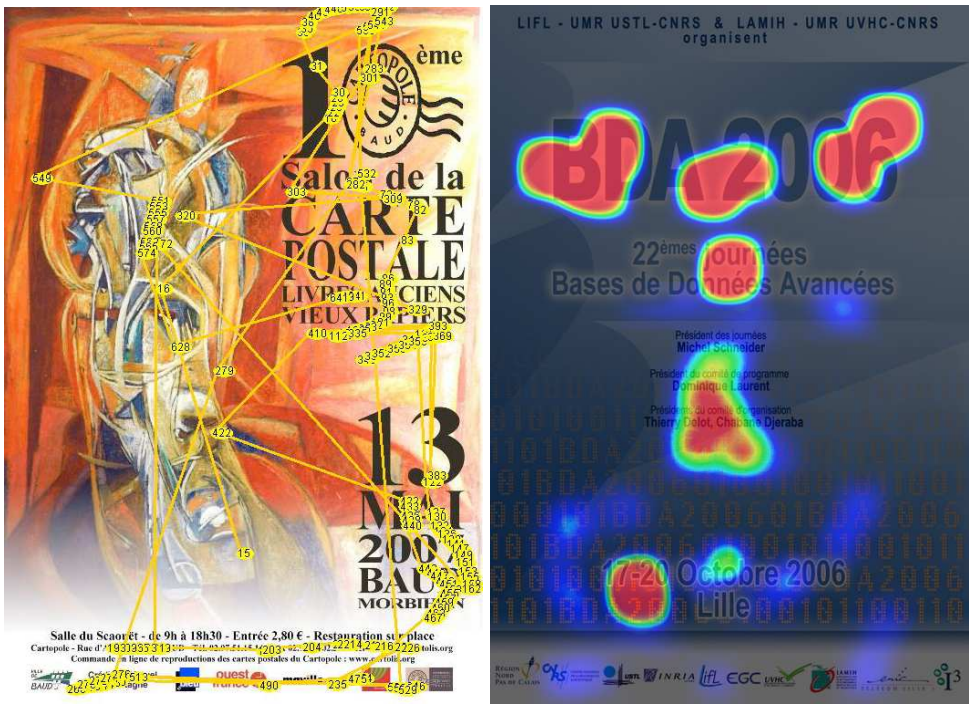


FIGURE 5.23: Exemple d'images représentant des affiches publicitaire pour des événements sociaux ou scientifiques.

La Figure 5.24 quant à elle montre comment cette approche permet aux concepteurs de publicités d'évaluer l'impact de leurs publicités durant un événement sportif pour lui trouver l'emplacement le plus adéquat. Pour cette application particulière, il est possible de se focaliser uniquement sur les zones de l'image qui contiennent de la publicité. Le modèle peut alors bénéficier d'un compteur de fixations dans les zones qui contiennent de la publicité.

5.6.3.2 Sur des vidéos

L'approche est validée en évaluant une vidéo d'une campagne publicitaire de la boisson Coca Cola. Elle est affichée sur un écran de 25 pouces. La Figure 5.25 montre un échantillon

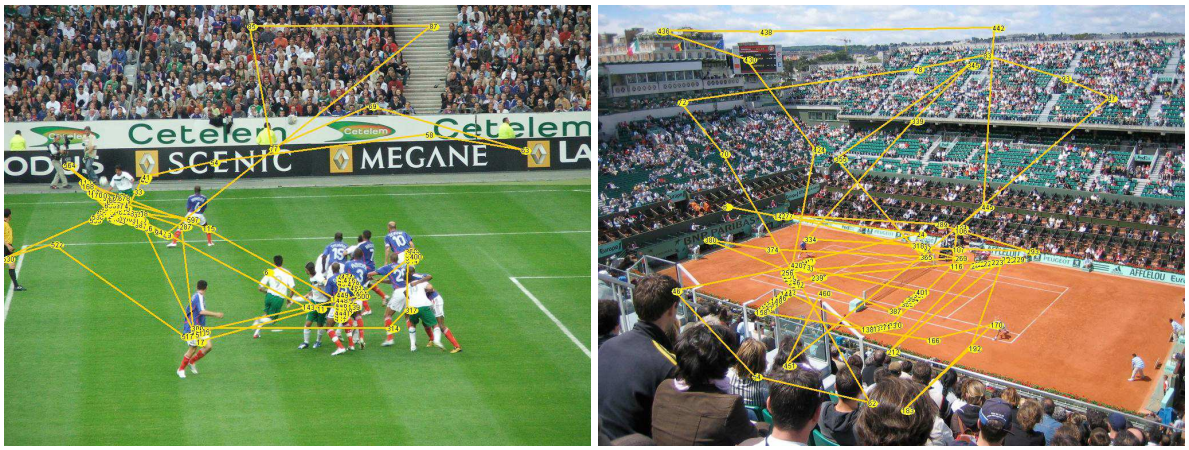


FIGURE 5.24: Illustration de l'approche pour évaluer une campagne publicitaire lors d'un évènement sportif.

d'images clés de la séquence vidéo qui contient les points de fixation de 10 participants superposés sur la vidéo.

FIGURE 5.25: Illustration de l'approche pour évaluer une vidéo publicitaire d'une boisson.

Sur la base des métriques définies précédemment, une estimation de la valeur de la dispersion pour chaque image de la vidéo est effectuée indiquant la répartition des points de fixations de l'ensemble des participants. Une très grande valeur de dispersion indique un manque d'information sur le foyer d'attention alors qu'une petite valeur indique une estimation du foyer d'attention d'une manière précise. La Figure 5.26 présente l'évolution de la valeur de dispersion à travers le temps appliquée à la vidéo. Les intervalles de temps sont mis en évidence sous

le graphe. Certaines images clés qui correspondent à ces moments extraits de la vidéo sont également illustrées.

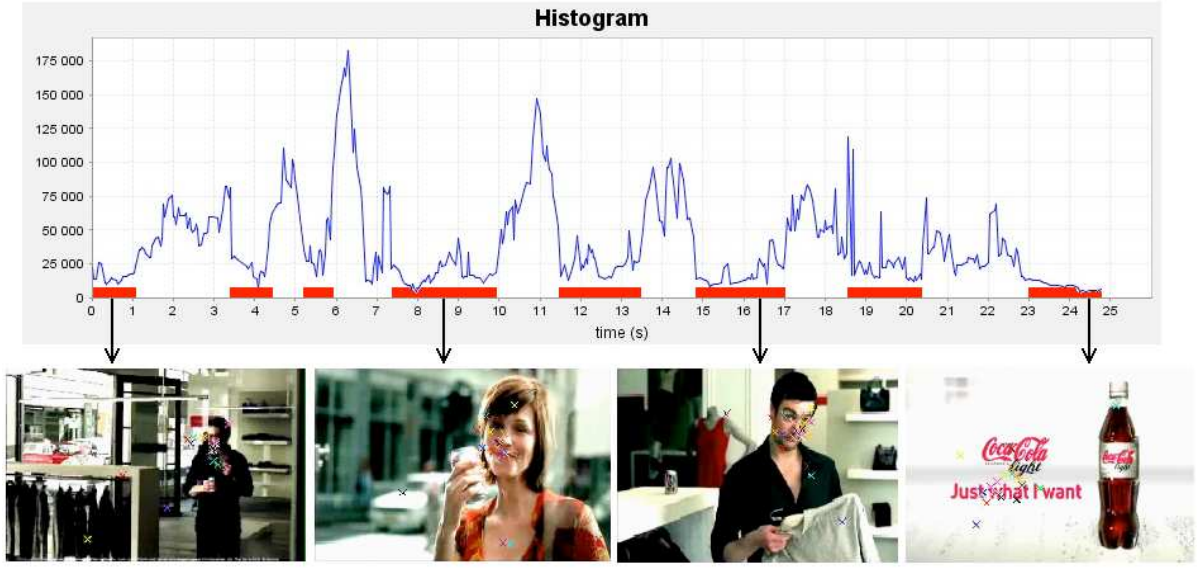


FIGURE 5.26: Évolution de la valeur de la dispersion à travers le temps appliquée à une vidéo publicitaire d'une boisson.

La valeur de la dispersion révèle la structure de la vidéo. Elle est composée d'une séquence rapide contenant des plans de courte durée avec quelques plans de plus longue durée entre les deux où les visages des acteurs peuvent clairement être vus. Les intervalles à faible dispersion (i.e. les moments de concentration) correspondent à des plans longs, et les intervalles de haute dispersion (i.e. concentration incertaine) correspondent à plusieurs plans de courte durée avec les coupures fortes lors de la transition d'un plan vers un autre. Dans de telles situations, le foyer d'attention se déplace de la dernière position pertinente dans un plan vers la première du plan suivant. En raison des variations physiologique individuelle, la durée utilisée pour changer la direction du regard n'est pas la même pour tous les participants ce qui explique de grandes valeurs de dispersions. Nous constatons sur le dernier plan qui correspond à un affichage statique d'un produit de couleur rouge sur fond blanc que la valeur de la dispersion est très faible.

5.6.4 Discussions

Une analyse des résultats de la séquence vidéo permet de révéler sa structure et de formuler la recommandation suivante à un concepteur de vidéos publicitaires : le foyer d'attention est mieux préservé quand les emplacements saillant coïncident entre deux plans consécutifs. En effet, lorsque les emplacements sont différents, le regard des personnes est naturellement déplacé avec des durées différentes.

L'évaluation de la qualité d'un support visuel à partir de points de fixation de regard demeure difficile. Cependant, la satisfaction des producteurs de médias qui ont utilisée cette démarche constitue une première validation qualitative [ANA]. La comparaison d'une carte de pertinence avec la fusion de tous les points de fixation de tous les participants peut être une solution.

5.7 Conclusion

Dans ce chapitre, nous avons présenté l'approche utilisée pour extraire la direction du regard d'une personne sur une scène cible. L'approche est basée essentiellement sur les données physiologiques de la vision humaine et la géométrie projective. L'extraction de la région d'intérêt de l'utilisateur dans la scène cible utilise les données issues de l'estimation de la pose de la tête. En présence d'une position précise des yeux dans l'image, le point de fixation est mieux déterminer et l'écart type entre les points de fixations est réduit. Une méthode d'analyse des points de fixation basée sur des métriques a été proposée et validée avec succès sur des données collectées sur le regard.

Chapitre 6

Détection et suivi de personnes

Sommaire

6.1 Introduction

L'objectif optimal de notre travail est de détecter et de suivre la direction du regard d'une personne en situation dynamique. A savoir, la direction du regard est estimée pendant le déplacement de la personne devant la scène cible. Tout au long des chapitres précédents, nous avons supposé que la personne est en situation fixe devant la scène cible. Dans ce chapitre, nous présentons un système qui étudie le comportement visuel d'une personne dans un environnement précis. Pour cela un scénario est construit autour de l'analyse du regard d'une personne qui se déplace dans la scène. Premièrement, l'environnement est défini à travers la procédure suivie pour l'acquisition des données et la description des éléments qui composent la scène en termes de décor et de zones. Ensuite, la détection et le suivi des personnes présentes dans la scène permettent d'extraire les informations nécessaires à la compréhension de leurs comportements dans un contexte particulier. Enfin, les informations récoltées lors de l'analyse du comportement des personnes sont exploitées par le système à travers un retour direct/indirect de l'information.

6.2 Analyse du comportement des personnes

L'approche la plus courante pour modéliser le comportement d'une personne est de décrire chacun de ses mouvements. Un modèle basé sur les états (e.g. un modèle de Markov caché) [DB03] est utilisé pour convertir une série de mouvements en une description d'activité. Oliver et al. [ORP99] ont proposé un système qui permet de déterminer les interactions entre deux personnes (e.g. discuter ensemble, approcher de l'autre personne) alors que Siskind et Morris [SM96] ont proposé un système qui reconnaît des actions simples (e.g. ramasser, poser, pousser, tirer, etc.). La détection de comportement anormal est abordée par Nair et Clark [NC02] dans le cadre de la surveillance d'un couloir qui mène à des bureaux (e.g. détecter des personnes qui flânent ou qui forcent l'entrée d'un bureau). Nous allons nous intéresser à l'analyse du comportement d'une personne qui regarde une scène cible.

En 2002, la campagne d'évaluation PETS [Worb] a définie un certain nombre de tâches en relation avec le passage de personnes devant une vitrine (e.g. déterminer le nombre de

personnes dans la scène, ceux qui sont en face de la vitrine, et ceux qui regardant en direction de la vitrine). Plusieurs méthodes ont été proposées pour réaliser ces tâches [MMR02, Pec02, Sen02]. Ces travaux se sont basés sur l'hypothèse suivante : une personne qui se trouve en face d'une vitrine regarde forcément dans sa direction. Toutefois, ils n'ont pas abordé le problème de l'estimation de l'orientation de la tête.

Dans le cadre de l'analyse de comportement visuel d'une personne qui regarde une scène cible, nous proposons d'extraire les modalités suivantes :

- A l'arrêt / En mouvement : permet de savoir si la personne est en mouvement,
- Regarde / Ne regarde pas : indique si la personne regarde en direction de la scène cible,
- Région d'intérêt : détermine l'emplacement exact de la région d'intérêt dans la scène cible.

Ces modalités sont extraites à l'aide d'un système composé de 3 modules : détection et suivi de personnes, estimation de l'orientation de la tête, et projection du champ visuel (sur la scène cible). L'architecture de ce système est illustrée par la Figure 6.1.

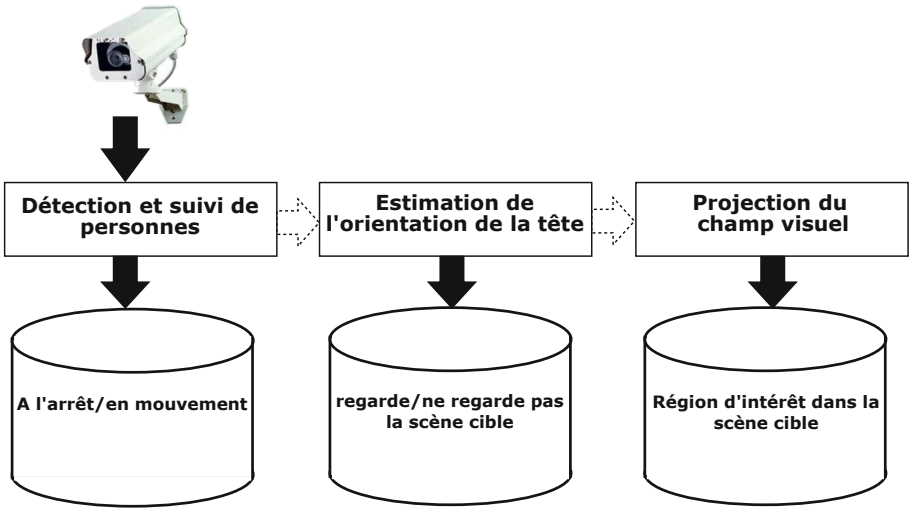


FIGURE 6.1: Architecture du système d'analyse de comportement visuel d'une personne qui regarde une scène cible.

Nous avons construit un modèle qui permet d'enregistrer les informations nécessaires à ce

système [LD08]. La zone surveillée peut contenir plusieurs personnes à un instant t. Le modèle est défini par $I_t = \{X_{i,t}/i \in |I_t|\}$ où $N = |I_t|$ indique le nombre de personnes présentes dans la zone surveillée à un instant t.

L'état $X_{i,t}$ d'une personne i à un instant donné t est défini par les 4 composants suivants :

1. Le composant corps $X^b_{i,t}$: contient les informations sur la localisation 2D de la personne et la signature couleur du corps.
2. Le composant tête $X^h_{i,t}$: contient les informations sur la localisation 2D de la tête et de son orientation selon les 3 degrés de liberté.
3. Le composant modalité $X^m_{i,t}$: est un ensemble de deux modalités (à l'arrêt / en mouvement et regarde / ne regarde pas). Elles peuvent avoir l'une des trois valeurs suivantes : Oui, Non ou Inconnu.
4. Le composant région d'intérêt $X^l_{i,t}$: possède deux états qui sont 'Intérieur' (lorsque la personne regarde en direction de la scène cible et correspond aux coordonnées de la région d'intérêt) ou 'Extérieur' (lorsque la personne ne regarde pas en direction de la scène cible ou que la réponse à cette question est inconnue).

6.3 Description de l'environnement

La notion d'environnement contrôlé (ou contexte) est très importante en vidéosurveillance. Une description précise des caractéristiques de l'environnement est critique [WB06] car elle permet d'améliorer les résultats de la détection, du suivi et de l'analyse du comportement des personnes. En effet, l'apparence et l'environnement sont les sources d'informations utiles pour effectuer la détection d'un objet. L'analyse de l'apparence utilise les informations qui proviennent de l'histogramme de couleur, des contours, de la texture, etc. L'environnement quant à lui peut utiliser la nature des objets proches [SF91, HR78], leurs position relative [SLZ03], ou bien les statistiques sur les caractéristiques visuelles de bas niveau de la scène [TS01].

L'environnement est défini comme l'ensemble des informations pertinentes pour détecter l'instance d'une classe particulière et qui ne soit pas liées son apparence physique. En général,

les objets sont fortement liés à la scène à l'aide d'une certaine relation. La connaissance de l'environnement permet d'accéder à cette relation [BAB+05]. Dans la détection de personnes par exemple, il permet de réduire le temps de traitement et de distinguer entre les ambigüités issues d'une mauvaise qualité d'image.

Plusieurs systèmes utilisent l'information issue de l'environnement. Torralba et al. [TMF04] ont utilisé des Boosted Random Fields pour apprendre les associations entre les objets et leurs positions relatives. Fink et Perona [FP03] ont décrit un système similaire qui détecte les objets présents dans la scène en utilisant une architecture modifiée de l'algorithme de cascades proposé par Viola et Jones [VJ01]. Ce type de composition dense permet d'exploiter toute l'information mais nécessite néanmoins le calcul des dépendances en mode itératif. D'autres systèmes segmentent l'image avant de modéliser les relations entre les segments voisins [CdFB04, KH03]. Cependant, ils souffrent de l'absence de sources d'informations contextuelles lorsque l'objet marqué est seul dans l'image.

L'environnement est défini à travers la procédure suivie pour l'acquisition des données et par la description des éléments qui composent la scène en termes de décor et de zones. Nous présentons dans ce qui suit ces deux notions :

6.3.1 Procédure d'acquisition

La phase d'acquisition d'images est très importante. Le signal vidéo produit par la caméra est converti en images. Les propriétés du périphérique de capture qui doivent être connues sont les suivantes :

1. **Type** : définit la façon avec laquelle la caméra est placée. Elle peut être soit fixée soit en mouvement :
   - Fixée : la caméra est placée dans un endroit précis qui restera inchangé durant toute la phase d'acquisition. C'est le type le plus souvent utilisé lorsque l'information sur l'environnement est utilisée car il permet de délimiter les zones visibles de la scène et les objets qui la compose.
   - En mouvement : la caméra peut pivoter librement sur un ou plusieurs axes. C'est utilisé dans le suivi d'objets en temps réel : manuellement à l'aide d'un agent qui

commande la caméra à distance ou automatiquement grâce à un détecteur de mouvement.

2. **La position** : définit l'emplacement de la caméra dans la scène par rapport à un repère. Cela revient à déterminer la matrice qui permet de passer du repère de la scène vers celui de la caméra.
3. **Le nombre d'images par seconde** : représente le nombre d'images capturées durant une seconde. Ce nombre dépend de l'intervalle de temps qui sépare l'acquisition de deux images successives (Δt). La fréquence est souvent choisie de manière à effectuer les traitements en temps réel.
4. **Propriétés des images** : représentent les informations sur les images (e.g. le format, la résolution, etc.).

6.3.2 Description de la scène

La description des caractéristiques de la scène qui est sous surveillance consiste à déterminer les propriétés du décor et des zones. Certaines propriétés sont communes à ces deux catégories et sont :

- **La forme** : représente la forme géométrique des zones ou des objets (e.g. rectangle, cercle, polygone, triangle, droite, etc.).
- **La taille** : représente la taille de la zone ou de l'objet mais dépend toutefois de la forme (e.g. taille d'un rectangle est donnée à l'aide de la longueur et la largeur).
- **La position** : définit l'emplacement spatiale de la zone ou de l'objet par rapport au repère de la scène.
- **Le nombre de caméras qui couvrent la scène** : définit le nombre de caméras utilisées pour les traitements. En présence d'une caméra (vision monoculaire) la notion de profondeur est perdue alors qu'en présence de deux caméras (vision binoculaire) la scène 3D est reconstruite à partir de deux images 2D après une phase de calibration.

Le décor est constitué des éléments qui composent la scène tels que : les murs, les portes, les escaliers, les piliers, les miroirs, les ascenseurs, etc. Les propriétés cités précédemment sont

appliquées à ces éléments en fonction de leurs présence/absence dans le champ de la caméra.

Les zones représente une partie délimitée de la scène. En plus des propriétés citées précédemment, il faut s'intéresser au type d'accès (libre ou restreint) et l'intérêt de la zone (aucun, élevé). Le type d'accès possède les propriétés suivantes :

- Libre : toutes les personnes peuvent accéder à la zone.
- Restreint : certaines personnes peuvent accéder à la zone sous certaines conditions. Deux types de restrictions sont possibles :
  - Totale : aucune personne ne peut accéder à cette zone.
  - Limitée : des personnes sont autorisées à y accéder sous certaines conditions telles que le durée de présence, la fonction de la personne, ou le nombre de personnes présentes dans la zone, etc.

6.4 Détection de personnes

La détection consiste à localiser et identifier toutes les instances d'une classe particulière. De façon générale, la détection d'objets est utilisée dans plusieurs domaines tels que le diagnostic médical [BNPS03], la télédétection [LDZ00], la reconnaissance de caractères [TSW90], la réalité virtuelle augmentée [ZCHS03], l'indexation automatique d'images et de vidéos [MLID08], les bases de données visuelles [FPZ04], les systèmes de sécurité et de surveillance [CLK$^+$00], etc. La détection d'objets est un processus complexe à cause des facteurs suivants :

- Perte d'information causée par la projection du monde 3D en une image 2D.
- Le bruit dans les images.
- Mouvement complexe des objets.
- La non-rigidité des objets.
- La forme complexe des objets.
- Les changements d'éclairage de la scène.
- La contrainte de réaliser les traitements en temps réel.

Il existe plusieurs méthodes dans la littérature pour la détection des personnes dans la vidéo. Nous allons détailler dans ce qui suit les méthodes proposées en fonction l'utilisation de l'arrière plan ou non.

6.4.1 Méthodes pour la détection de personnes

Les méthodes peuvent être organisées en celles qui utilisent l'extraction de l'arrière plan et celles qui opèrent directement sur l'image. Dans la première catégorie, les techniques sont classées en fonction des méthodes employées pour l'extraction de l'arrière plan et celles utilisées pour la classification de l'avant-plan. Dans la deuxième catégorie, les techniques sont classées en fonction du modèle utilisé pour représenter la personne ainsi que la méthode de classification utilisée.

Le Tableau 6.1 présente les techniques qui effectuent la détection des personnes après l'extraction de l'arrière plan.

Article	**Extraction de l'arrière plan**	**Classification de l'avant plan**
[WADP97]	Couleur et Image de référence	Couleur et Contour
[BFB04]	Couleur et Image de référence	Modèle de la région
[HSY04]	Couleur et Image de référence	F1, F2 et F3
[EWKY04]	Couleur et Image de référence	Couleur
[ELW03]	Mouvement et Différence entre les images	Ondelettes
[TA03]	Mouvement et Différence entre les images	Forme de Fourrier
[ZH05]	Mouvement et Différence entre les images	Forme
[YK04]	Mouvement et Couleur	Valeur géométrique du pixel
[XF03]	Profondeur	Mouvement
[LGS$^+$04]	Profondeur	Forme
[HB03]	Infrarouge	IR et Couleur
[JTS$^+$04]	Infrarouge	IR et Couleur

TABLE 6.1: Méthodes basées sur l'extraction de l'arrière plan.

Le Tableau 6.2 présente certains travaux basées sur la classification directe des éléments d'une image en Personne ou autre.

Article	Modèle de la personne	Méthode de classification
[CD00]	Mouvement périodique	Similarité du mouvement
[UT02]	Valeur géométrique du pixel	Distance
[GG02]	Modèle de la forme	Distance de Chamfer
[VJS03]	Forme et Mouvement	Adaboost cascade
[Sid04]	Flux optique	SVM (RBF)
[DT05]	Histogramme de gradients	SVM (Linéaire)

TABLE 6.2: Méthodes basées sur l'extraction automatique des personnes.

En fonction du type d'application choisis, une méthode est choisie. Dans ce qui suit nous présentons la méthode utilisée :

6.4.2 Détection de personnes basée sur l'extraction de l'arrière plan

Cette méthode utilise une mixture de gaussiennes pour extraire l'arrière plan. Cette technique permet de détecter les blobs dans la scène indifféremment de l'activité de la personne (mouvement ou à l'arrêt).

6.4.2.1 Extraction de l'arrière plan

La modélisation de l'arrière plan est utilisée dans différentes applications telles que la vidéosurveillance [CK05, TKBM99] ou le multimédia [EBBV07, PVM07]. La manière la plus simple consiste à acquérir une image de l'arrière plan qui ne contient aucun objet en mouvement. Une soustraction d'images par rapport à un seuil et ensuite effectuée entre chaque nouvelle image acquise et cette image de l'arrière plan. Cependant une image unique de l'arrière plan est souvent indisponible car il est modifié en continu par divers événements (e.g. changements d'éclairage, rajout d'objets, etc.). Pour palier aux problèmes de robustesse et d'adaptation, de nombreux travaux relatifs à la modélisation de l'arrière plan ont été proposés et peuvent être trouvés dans les études récentes [BEBV08, Pic04, EESA08].

Les méthodes peuvent être classées dans les catégories suivantes : Modélisation basique de l'arrière plan [LH02], Modélisation statistique de l'arrière plan [WADP97], Modélisation floue

de l'arrière plan [SMP08] et estimation de l'arrière plan [MMSZ05]. Une autre classification est disponible en termes de prédiction [WS06], de récursivité [CK05], d'adaptation [Por03], ou de modalité [PT05]. Toutes ces méthodes basées sur la modélisation de l'arrière plan suivent les étapes suivantes : modélisation de l'arrière plan, initialisation de l'arrière plan, maintenance de l'arrière plan, détection de l'avant plan, choix de la taille de la caractéristique (pixel, bloc ou groupe), et choix du type de la caractéristique (e.g. couleur, contour, texture).

Dans le contexte d'une application de vidéo surveillance du trafic routier, Friedman et Russell [FR97] ont proposé de modéliser chaque pixel à l'aide d'un mélange (mixture) de trois Gaussiennes qui correspondent à la route, les véhicules et les ombres. Ce modèle est initialisé au moyen d'un algorithme d'Espérance-Maximisation (EM) [DLR77]. Les gaussiennes sont ensuite marquées de manière heuristique comme suit : le composant le plus sombre est marqué comme ombre alors que les deux autres composants sont différencier grâce à la variance (la variance la plus large est étiqueté en tant que véhicule et l'autre en tant que route). La classification de l'avant plan se fait alors en associant chaque pixel à la gaussienne qui lui correspond. Un algorithme d'EM incrémental est utilisé pour la maintenance de l'arrière plan afin d'effectuer le traitement en temps réel. Cependant, ce processus souffre d'un manque d'adaptation aux changements qui apparaissent dans la scène à travers temps. Stauffer et Grimson [SG99] ont généralisé cette idée par la modélisation de l'intensité de couleur de chaque pixel sur l'intervalle de temps précédent $\{X_1, ..., X_t\}$ par un mélange de K Gaussiennes.

Dans ce qui suit, nous allons détailler les différentes étapes utilisées pour modéliser l'arrière plan. L'intensité de couleur dans l'espace RGB (Rouge, Vert et Bleu) de chaque pixel est prise en considération. Une probabilité d'observation est associée au pixel courant selon la formule suivante :

$$P(X_t) = \sum_{i=1}^{K} \omega_{i,t} \eta(X_t, \mu_{i,t}, \Sigma_{i,t})$$

avec K le nombre de distributions, $\omega_{i,t}$ le poids associé à la i$^{\text{ème}}$ gaussienne à un instant t avec une moyenne $\mu_{i,t}$ et un écart type $\Sigma_{i,t}$. η est une fonction gaussienne définie comme suit :

$$\eta(X_t, \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(X_t-\mu)^T \Sigma^{-1}(X_t-\mu)}$$

Pour des raisons de calcul, les composantes RGB de l'intensité de couleur sont considérées être indépendantes et possèdent la même variance. La matrice de covariance est sous la forme :

$$\Sigma_{i,t} = \sigma_{i,t}^2 I$$

Le nombre de gaussiennes K détermine la multi-modalité de l'arrière plan. Stauffer et Grimson [SG99] ont proposé de choisir une valeur entre 3 et 5 alors que l'initialisation du poids, de la moyenne et de la matrice de covariance est effectuée à l'aide de l'algorithme EM (l'algorithme des k moyennes peut être utilisé pour effectuer les traitements en temps réel).

Une fois les paramètres d'initialisation sélectionnés, une première détection de l'arrière plan peut être effectuée en mettant à jour les paramètres. Tout d'abord, le ratio $r_j = \omega_j/\sigma_j$ est utilisé pour ordonner les K gaussiennes. Cet ordre induit qu'un pixel appartient à l'arrière plan s'il possède un poids élevé avec une variance faible (l'arrière plan est plus présent que les objets en mouvement et sa valeur est presque constante). Les B premières gaussiennes qui dépassent un certain seuil T sont conservées par la distribution associée à l'arrière plan :

$$B = Argmin_b(\sum_{i=1}^{b} \omega_{i,t} > T)$$

Les autres distributions sont considérées comme avant plan. Dans l'image suivante (i.e. prise à l'instant $t+1$), un test de mise en correspondance est effectué pour chaque pixel. Un pixel est alors associé à une distribution de Gaussienne si la distance de Mahalanobis [Mah36] est inférieure à $k\sigma_{i,t}$ avec k un seuil constant égal à 2.5 :

$$\sqrt{(X_{t+1} - \mu_{i,t})^T \Sigma_{i,t}^{-1} (X_{t+1} - \mu_{i,t})} < k\sigma_{i,t}$$

Deux cas peuvent se produire :

- Une correspondance avec une des K gaussiennes est trouvée. Si la distribution gaussienne est identifiée comme arrière plan alors le pixel est classé en tant que tel. Dans le cas contraire, il fait partie de l'avant plan.
- Aucune correspondance avec une des K gaussiennes n'est trouvée. Le pixel fait partie alors de l'avant plan.

Ainsi, un masque binaire est obtenu correspondant à arrière plan / avant plan. Cependant une mise à jour des paramètres doit être effectuée en fonction de :

– Une correspondance avec une des K gaussiennes est trouvée. La mise à jour de la composante concernée est effectuée de la façon suivante :

$$\omega_{i,t+1} = (1-\alpha)\omega_{i,t} + \alpha$$

$$\mu_{i,t+1} = (1-\rho)\mu_{i,t} + \rho X_{t+1}$$

$$\sigma^2_{i,t+1} = (1-\rho)\sigma^2_{i,t} + \rho(X_{t+1} - \mu_{i,t+1})(X_{t+1} - \mu_{i,t+1})^T$$

avec α une constante issue de l'apprentissage et $\rho = \alpha\eta(X_{t+1}, \mu_i, \Sigma_i)$. μ et Σ des autres composantes restent inchangés. Par contre, le poids est mis à jour comme suit :

$$\omega_{j,t+1} = (1-\alpha)\omega_{j,t}$$

– Aucune correspondance avec une des K gaussiennes n'est trouvée. Dans ce cas, la distribution est remplacée par :

$$\omega_{k,t+1} = \text{Le poids précédent le plus petit}$$

$$\mu_{k,t+1} = X_{t+1}$$

$$\sigma^2_{k,t+1} = \text{La variance initiale la plus large}$$

6.4.2.2 Classification de l'avant plan

Une fois l'extraction de l'arrière plan effectuée, l'avant plan doit être classé dans l'une des deux catégories suivantes :

– Personne : inclut une personne qui se trouve toute seule ou bien un groupe de personnes.

– Objet : correspond aux différents objets qui peuvent apparaitre dans la scène.

Deux méthodes distinctes pour la classification de l'avant plan sont généralement utilisées : triviale et apprentissage.

Méthode Triviale utilise des règles simples pour détecter une personne. Ces règles sont basées sur certains attributs tels que :

- **La taille** : représente le nombre de pixels associés à une personne. En effet, une personne occupe souvent une région assez conséquente dans l'image avec une taille plus importante que celle des autres objets présents dans la vidéo.
- **La position** : indique les positions qui peuvent être occupées par une personne dans l'image. En supposant que la partie inférieure de la personne est visible, elle sera certainement sur le sol.
- **La direction** : indique la direction suivie par la personne pour se déplacer (e.g. dans un escalier mécanique la direction suivie est unique : celle du sens de la marche).
- **La forme** : est associée au rapport entre la longueur et la largeur. En effet, une personne est plus longue que large. Toutefois, la difficulté survient dans certains cas tels que : une personne accroupie, allongée sur le sol ou qui apparaît partiellement [SRM05].
- **La couleur** : est une propriété très intéressante lorsque la couleur dominante qui caractérise la personne est uniforme dans la séquence vidéo [SMBP05].

Ces règles peuvent s'appliquer à certaines vidéos mais pas à d'autres. C'est pour cela que la connaissance de l'environnement permet de choisir les facteurs pertinents auxquels un poids est associé.

Méthode basée sur l'apprentissage utilise une classification des blobs détectés. Un mécanisme d'apprentissage est utilisé sur des images qui représentent des personnes et des objets. La méthode d'apprentissage génère une fonction qui fait correspondre à une image reçue en entrée une étiquette. Les caractéristiques jouent un rôle important dans la classification. C'est pour cela qu'il faut choisir celles qui peuvent discriminer au mieux une classe d'une autre. Différents type de caractéristiques peuvent être utilisées telles que : la couleur, la bordure, la texture, le flux optique, etc. Différentes méthodes d'apprentissage sont disponibles telles que les arbres de décision [GK95], les réseaux de neurones [RBK96], adabosst (Adaptive Boosting) [VJS03], ou les machines à vecteurs de support (SVM) [POP98]. Certaines méthodes utilisent de l'apprentissage semi-supervisé pour réduire la quantité de données qu'il faut annoter manuellement lors de la collecte des données. L'idée est de former deux classificateurs à l'aide

d'un petit ensemble de données étiquetées. A la fin de la phase d'apprentissage, chaque classificateur est utilisé pour assigner des données non étiqueté pour l'échantillon d'entrainement de l'autre. Cette méthode a été utilisée avec succès pour réduire la quantité de l'interaction manuelle nécessaire pour l'entrainement en utilisant *AdaBoost* [LVF03] et SVM [KLS03].

AdaBoost est un méta-algorithme qui peut être utilisé conjointement avec de nombreux autres algorithmes d'apprentissage afin d'améliorer leurs performances. Le boosting est une méthode itérative qui trouve une classification très précise, en combinant un grand nombre de classificateurs de base [FS95]. Dans la phase d'entrainement de l'algorithme Adaboost, une première distribution de poids est construite sur l'ensemble d'entrainement. Le mécanisme de boosting sélectionne ensuite le classificateur de base qui donne le minimum d'erreurs. Ainsi, l'algorithme favorise le choix d'un autre classificateur qui est plus performant sur les données restantes au cours de la prochaine itération. Dans le domaine de la détection de personnes, AdaBoost a été utilisé par Viola et al.[VJS03] pour détecter les piétons.

Les Machine à Vecteurs de Support (SVM) sont utilisées pour regrouper des données en plusieurs classes à l'aide des marges maximales de l'hyperplan qui séparent une classe de l'autre [BGV92]. La marge de l'hyperplan, qui est maximisé, est définie par la distance entre l'hyperplan et les points de données les plus proches. Les points de données qui se trouvent à la limite de la marge de l'hyperplan sont appelés vecteurs. Pour la détections de personnes, deux classes sont utilisées : la classe *personne* (les échantillons positifs) et classe *pas une personne* (échantillons négatifs). A partir d'exemples d'entrainement annotés manuellement suivant ces deux classes, le calcul de l'hyperplan parmi une infinité d'hyperplans possibles est effectué. SVM a été utilisé par Papageorgiou et al. [POP98] pour la détection de piétons et de visages dans les images.

6.5 Suivi des personnes

Le suivi d'une personne permet de générer la trajectoire qu'elle a suivie à travers le temps. Ceci nécessite la localisation de sa position dans chaque image de la vidéo. Les tâches de

détection et de suivi de personnes peuvent être effectuées séparément ou conjointement. Dans le premier cas, un algorithme de détection de personne est lancé. Il est suivi par une étape de mise en correspondance entre les images successives. Dans le deuxième cas, la détection et la mise en correspondance sont effectuées conjointement de manière itérative sur l'emplacement de la personne dans les images précédentes. Dans chacune de ces approches, le modèle qui représente la personne utilise la forme et/ou l'apparence.

6.5.1 Méthodes de suivi génériques

Les méthodes de suivi d'objets ou de personnes peuvent être regroupées en 3 catégories [YJS06] :

- **Suivi de points (Point Tracking)** : effectue une association entre les points qui représentent l'objet détectés dans des images successives. La position de l'objet en mouvement peut aussi être utilisée.
- **Suivi de noyau (Kernel Tracking)** : calcule le mouvement du noyau dans des images successives. Le mot noyau fait allusion à la forme ou l'apparence de l'objet suivi (e.g. un modèle de forme rectangulaire ou d'une forme elliptique associé à un histogramme)
- **Suivi de silhouette (Silhouette Tracking)** : effectue une estimation de la région occupée par l'objet dans chaque image en utilisant l'information contenue dans cette région (e.g. densité d'apparition, forme du modèle, etc.).

Le Tableau 6.3 présente certains travaux de l'état de l'art sur le suivi d'objets.

6.5.2 Localisation d'une personne

Une localisation précise des personnes suivies est nécessaire dans les systèmes de vidéosurveillance [FK05] car elle permet de construire leurs trajectoires. Cette tâche devient difficile dans un environnement encombré où les personnes se chevauchent entre elles dans le plan image et sont partiellement cachées. Contrairement à de nombreux systèmes qui utilisent plusieurs caméras [Bat04] ou des approches stéréo [MD03] pour surmonter ces difficultés, une vue monoculaire de la scène est utilisée.

Catégorie	Méthodes	Références
Suivi de points	Méthodes déterministes	Suivi avec MGE [SS90] Suivi avec GOA [VRB01]
	Méthodes probabilistes	JPDAF [BSF87] Filtre de Kalman [BC86]
Suivi de noyau	Modèles d'apparence basés sur un patron	Mean-shift [CRM03] KLT [ST94] Layering [TSK02]
	Modèles d'apparence à vues multiples	Eigentracking [BJ98] Suivi avec SVM [Avi01]
Suivi de silhouette	Évolution du contour	Représentation d'état [IB98] Méthodes variationnelles [BSR00] Méthodes heuristiques [Ron94]
	Mise en correspondance de la forme	Transformée de Hough [SA04] Histogrammes [KCM04]

TABLE 6.3: Méthodes de suivi d'objets.

Le problème de chevauchement entre les personnes est abordé de manière diverses mais ne permet pas de traiter l'occlusion partielle ou totale par les objets présents dans la scène qui est important dans des environnements complexes (e.g. aéroports, supermarchés, etc.). Pour cela, une compréhension de la scène dans son intégralité est nécessaire à laquelle sont ajoutées les connaissances sur le comportement humain, la structure du corps et les lois de la physique [BR05]. La Figure 6.2 illustre une image prise par une caméra dans placée dans un magasin. La position des pieds sur le sol est l'information la plus fiable (représentée par un rectangle rouge dans la figure). Lorsque cette information est indisponible, la position de la tête est combinée avec une estimation de la hauteur du corps d'une personne (représentée par un rectangle bleu dans la figure). Dans le cas d'occlusions supérieure et inférieure de la silhouette d'une personne (représentée par un rectangle vert dans la figure), l'observateur reconstruit la silhouette de la personne dans son esprit et s'intéresse à l'extrémité inférieure de la partie visible du corps pour déduire la position.

L'estimation de la position d'une personne est donc basée sur la prédiction, la position

FIGURE 6.2: Détermination de la position d'une personne.

calculée à partir des coordonnées des pieds ou de la tête dans l'image. En plus, le modèle associé à la caméra doit être connu afin d'effectuer la transformation des coordonnées entre l'image et le monde réel [ZN04]. Il faut donc prendre en considération l'environnement et le modèle associé à la forme humaine :

- **Le modèle associé à la caméra** : transforme les coordonnées 3D (x, y, z) de la scène en coordonnées 2D (x_i, y_i) dans l'image et vice-versa. La transformation 2D en 3D nécessite des informations supplémentaires pour palier à l'indisponibilité d'une dimension. La hauteur au-dessus du sol est souvent utilisée et sera alors égal à 0 si les coordonnées des pieds dans l'image sont détectées, ou sera égale à la hauteur de la personne si les coordonnées de la tête sont détectées. La position de la personne peut alors être déduite des coordonnées des pieds ou de la tête dans l'image. La hauteur moyenne d'une personne est calculée à travers le temps quand la tête et les pieds sont détectés simultanément. L'estimation de la position d'une personne devient moins précise lorsque sa distance de la caméra est grande.
- **Les connaissances sur la scène** : prennent la forme de la structure de la scène et de sa profondeur. La structure de l'espace est représentée par une carte composée de positions

calculées lorsque la personne se déplace dans la scène. Les points d'entrée/sortie, où une personne peut entrer dans le champ de vision de la caméra ou le quitter sont aussi indiqués. La profondeur est quant à elle utiliser pour la reconstruction de la silhouette. La silhouette est également utile pour calculer la profondeur de la personne dans la scène. La profondeur de la scène est calculée à l'aide d'une représentation 3D ou 2D $\frac{1}{2}$. La représentation 2D $\frac{1}{2}$ est une carte composée de points d'intérêts auxquels est associée une hauteur.

– **Le modèle associé à la forme humaine** : est utilisé pour détecter les coordonnées de la tête et des pieds des personnes d'une manière robuste. Il permet en outre de reconstruire la silhouette d'une personne même en présence d'occlusions partielles. La silhouette moyenne est utile pour détecter la forme humaine. La qualité de la forme détectée dépend de la pose de la personne (de face ou de profil). La description basique de la forme humaine possède cinq paramètres (les coordonnées x et y, la hauteur, l'échelle et la largeur relative).

6.5.3 Mise en correspondance

C'est une étape très importante dans le processus de suivi et dépend fortement des caractéristiques sélectionnées. La propriété recherchée d'une caractéristique est son unicité afin de pouvoir aisément distinguer l'objet suivi. La sélection des caractéristiques est étroitement liée à la représentation de l'objet (e.g. la couleur avec une apparence basée sur l'histogramme, ou les bords avec une représentation basée sur les contours). En général, une combinaison de caractéristiques choisies en fonction du domaine d'application est utilisée. Les caractéristiques visuelles les plus utilisées sont les suivantes :

– **La couleur** : est influencée principalement par deux facteurs physiques : la distribution spectrale de la luminosité, et les propriétés de réflexion de la surface de l'objet. En traitement d'images, l'espace de couleur RVB (Rouge, Vert, Bleu) est généralement utilisé pour représenter les couleurs. Cependant, il n'est pas un espace de couleurs uniforme où les dimensions sont fortement corrélées [Pas01]. Par contre, l'espace de couleur TSV (Teinte, Saturation, Valeur) est assez uniforme même s'il est sensible au bruit [SKP96].

- **Les bords (Edges)** : provoquent généralement un changement d'intensité assez conséquent dans l'image. Ils sont moins sensibles aux changements de lumières par rapport à la couleur. En raison de sa simplicité et de sa précision, l'approche la plus populaire est le détecteur de bords proposé par Canny [Can86]. Une évaluation des algorithmes de détection de bords est disponible [BKD01].
- **Le flux optique (Optical flow)** : est un champ dense de vecteurs qui définit le déplacement de chaque pixel dans une région. Il est calculé en utilisant la contrainte de luminosité, ce qui suppose que la luminosité des pixels reste constante dans des images consécutives [HS81]. Le flux optique est couramment utilisé comme caractéristique dans les applications qui se basent sur le mouvement [LK81, BA96, ILD09]. Une évaluation des performances des méthodes qui utilisent le flux optique est disponible [BFBB94].
- **La texture** : est la mesure de la variation d'intensité d'une surface qui quantifie des propriétés telles que la fluidité et la régularité. Une étape de traitement est nécessaire pour générer des descripteurs. Il existe plusieurs descripteurs de texture : Matrices de co-occurrence en niveau de gris (Gray-Level Cooccurrence Matrices (GLCM's)) [HSD73] qui consiste en un histogramme 2D qui indique les co-occurrences des intensités dans une direction et une distance spécifiée, la loi des mesures de texture [Law80] qui consiste en 25 filtres 2D générés à partir de cinq filtres 1D, et pyramides orientables [GBG+94]. La texture est une caractéristique qui est moins sensibles aux changements de luminosité mais qui n'est pas adaptée aux processus en temps réel.

6.5.4 Représentation de la personne

Les personnes peuvent être représentées par leurs formes et leurs apparences. En général, il y a une forte relation entre la représentation de l'objet et l'algorithme de suivi utilisé. La représentation d'objets est choisie en fonction du domaine d'application. D'abord, nous allons présenter les représentations basées sur la forme couramment utilisées pour le suivi. Ensuite, nous présenterons celles qui utilisent une représentation commune de la forme et de l'apparence. Les représentations basées sur la forme illustrées par la Figure 6.3 sont définis comme suit :

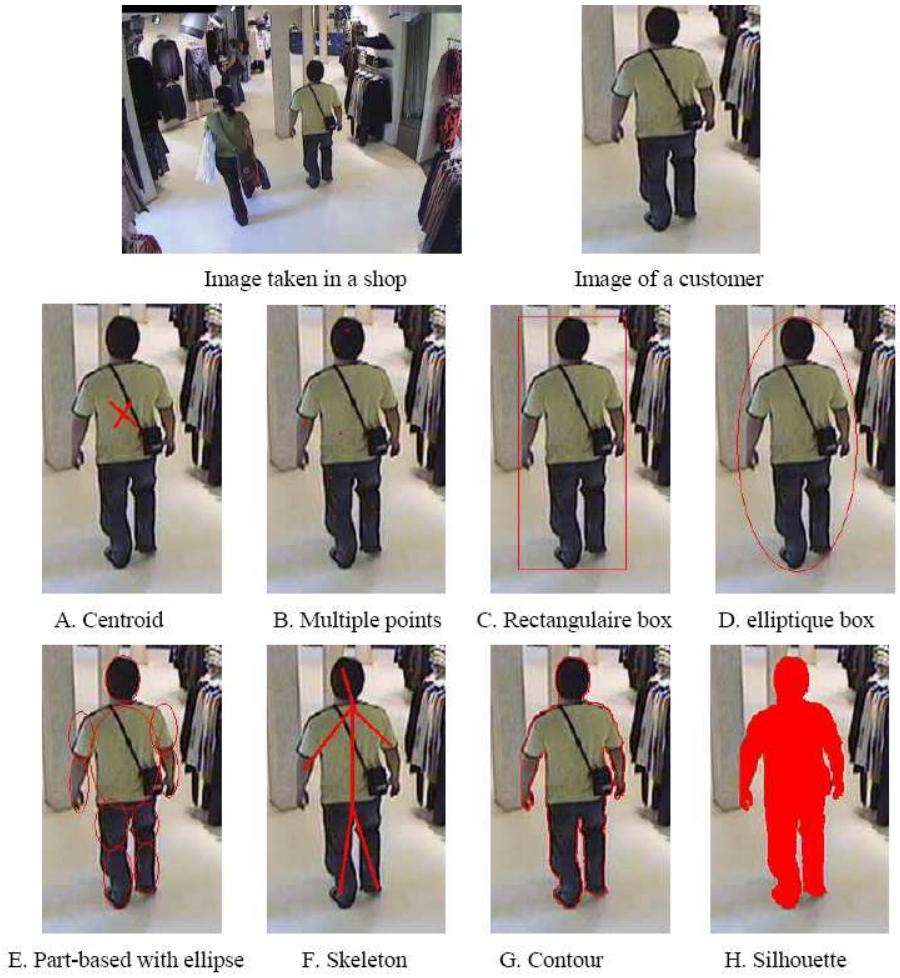


FIGURE 6.3: Différents types de représentations d'une personne.

– Points : L'objet est représenté par un point qui est souvent le centre (A) [VRB01] ou par un ensemble de points (B) [SKMG04]. Cette représentation est adaptée pour le suivi de petits objets.
– Formes géométriques primitives : La forme est représentée par un rectangle (C), une ellipse (D) [CRM03], etc. Les formes géométriques primitives permettent de représenter à la fois des mouvements rigides et non-rigides.

– Modèle en forme articulée : Les objets articulés sont composés de parties reliées entre elles [HHD98, Sen03]. Par exemple, le corps humain est un objet articulé avec le torse, les jambes, les mains, la tête et les pieds reliés entre eux. Les relations entre les parties sont influencées par la cinématique de mouvement. Afin de représenter un objet articulé, ses parties sont modélisées à l'aide de cylindres ou d'ellipses (E).
– Modèle en forme de squelette : Ce modèle peut être extrait par l'application d'une transformation de l'axe médian de la silhouette de l'objet [BB82]. Il est couramment utilisé pour la reconnaissance d'objets [AA01]. Cette représentation est utilisée pour modéliser les objets rigides et articulés (F).
– Modèle basé sur la silhouette de l'objet et de son contour : Le contour définit la frontière d'un objet (G) alors que la région localisé à l'intérieur du contour est la silhouette (H). Ces modèles de représentation sont adaptés pour le suivi de formes complexes non rigides [YLS04].

De nombreuses représentations des caractéristiques de l'apparence d'une personne sont disponibles. Toutefois, ils peuvent également être combinés avec une représentation de l'apparence [CET01] pour le suivi de personnes. Les modèles de représentation basés sur l'apparence sont :

– Basés sur un modèle (Template models) : formés en utilisant des formes géométriques ou des silhouettes [FT97]. Son avantage est qu'il traite à la fois les informations spatiales et celles liées à l'apparence. Cependant, l'apparence de l'objet est générée à partir d'une seule vue. Ces modèles sont adaptés au suivi d'objets dont la pose ne varie pas considérablement à travers le temps.
– Basés sur une apparence multiple (Multi-view appearance models) : permettent d'encoder différentes vues d'un objet. L'approche utilisée pour représenter les différentes vues d'un objet est de générer un sous-espace de points sous une vue donnée. Ces modèles sont souvent utiliser pour représenter la main et le visage [MP97].
– Basés sur une apparence active (Active appearance models) : sont générés par la modélisation simultané de la forme et de l'apparence de l'objet [ETC98]. La forme d'un objet est généralement définie par un ensemble de points de repère. Ces points sont souvent localisés sur les frontières de l'objet suivi. L'apparence est représentée par un vecteur

composé de la couleur, de la texture ou de l'amplitude du gradient. Une phase d'apprentissage sur un ensemble de données associées à la forme et à l'apparence est nécessaire.
– basés sur les densités de probabilité (Probability densities) : évaluent l'aspect de l'objet en utilisant par exemple une gaussienne [ZY96], une mixture de gaussiennes [PD02], ou un histogramme de couleurs [CRM03].

6.6 Expérimentation

Deux approches ont été utilisées pour effectuer la détection et le suivi des personnes :
– Approche 1 : est basée sur détection des blobs dans chaque image grâce à la modélisation de l'arrière plan. Les blobs obtenus à instant t sont mis en correspondance avec ceux détectés à l'instant $t-1$. L'architecture globale de cette approche est illustrée dans la Figure 6.4.

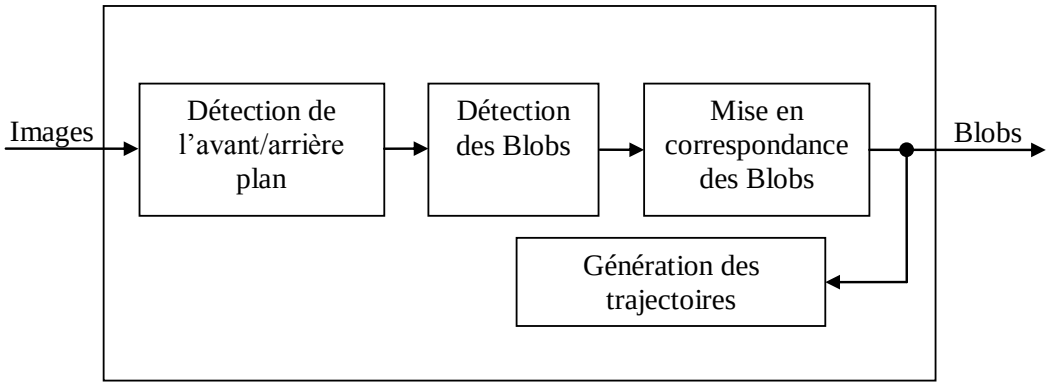


FIGURE 6.4: Architecture globale de l'approche 1 pour la détection et le suivi de personnes.

– Approche 2 : est basée sur détection des blobs qui entrent dans la scène dans l'image courante grâce à la modélisation de l'arrière plan. Chacun de ces blobs est ensuite suivi à travers le temps. L'architecture globale de cette approche est illustrée dans la Figure 6.5.

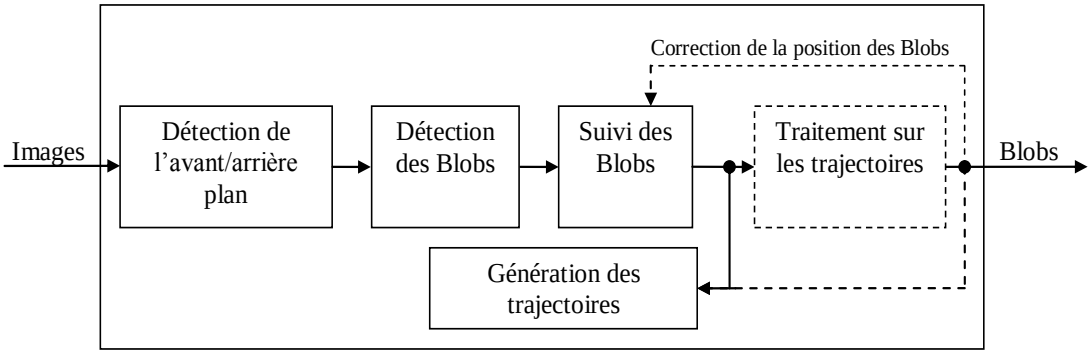


FIGURE 6.5: Architecture globale de l'approche 2 pour la détection et le suivi de personnes.

Les Figures 6.6 et 6.7 illustrent quelques capture d'écrans de la séquence "cam3-Seq1" présentée en Annexe B.1 en utilisant les approche 1 et 2 respectivement :

FIGURE 6.6: Capture d'écran de la séquence "cam3-Seq1" en utilisant l'approche 1.

FIGURE 6.7: Capture d'écran de la séquence "cam3-Seq1" en utilisant l'approche 2.

Enfin, une expérimentation du système complet a été effectuée en utilisant les vidéos présentées en Annexe B.1. Nous présentons ci-dessous quelques captures d'écran (voir Figure 6.8) qui souligne les résultats obtenues pour les images 281, 341 et 422 de la séquence "cam3-Seq1" provenant d'une caméra placée dans une vitrine en pose frontale.

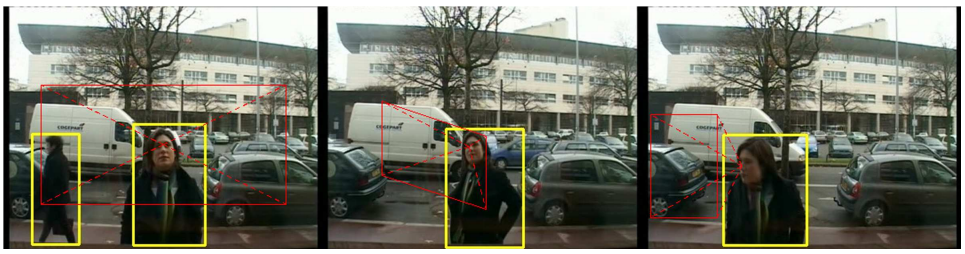


FIGURE 6.8: Extraction du comportement visuel d'une personne qui regarde la vitrine d'un magasin.

Le système a été testé sur une séquence de la base vidéo CAVIAR [Dat] et la Figure 6.9 présente les résultats obtenus.

FIGURE 6.9: Extraction du comportement visuel d'une personne sur les séquences de la base Caviar.

6.7 Exploitation et retour de l'information

Le retour de l'information, exploite les informations collectées sur les événements qui se produisent. Il lance les actions appropriées en fonction du scénario établi. Certaines actions sont lancées en temps réel alors que d'autres le sont en hors ligne. Il est souvent nécessaire d'exploiter les informations collectées en temps réel dans des scénarios en relation avec la sécurité pour accomplir l'action appropriée (e.g. détecter un objet abandonné) alors que dans les scénarios liées au Marketing l'exploitation se fait en hors ligne (e.g. compter le nombre de personnes dans une file d'attente).

6.7.1 Exploitation de l'information

Selon le scénario étudié, deux types d'exploitation peuvent être effectués en fonction du temps et du matériel utilisé pour détecter un événement.

En temps réel effectuée généralement sur les événements qui ont besoin d'une intervention rapide et qui devraient fournir une alerte dans les plus brefs délais. En fonction du scénario défini, une action est déclenchée lors de la détection d'un événement particulier. C'est un défi pour les systèmes automatiques de surveillance qui doivent accomplir toutes les tâches (de l'acquisition jusqu'au déclenchement de l'action) dans le délai le plus court possible.

En hors ligne se produit lorsque le temps de calcul est important, ou la priorité de l'information est faible. C'est surtout utiliser afin d'effectuer une analyse sur les vidéos stockées au vue de la grande quantité d'information disponible et qui n'est pas exploitée. Lors d'un vol par exemple, il est fréquent que les policiers regardent de nombreuses séquences vidéo de caméras placées dans le périmètre du lieu où s'est déroulé l'événement à la recherche d'images du délinquant. Le changement de stockage des données vidéos de cassettes à enregistrements numériques présente un défi à relever pour automatiser la surveillance et tout particulièrement dans le domaine du vidéomining [RDD03].

6.7.2 Retour de l'information

Selon le scénario étudié, deux types de retour d'information peuvent être effectués en fonction de l'instant auquel le résultat est retourné.

Retour direct de l'information s'effectue de manière automatique lorsqu'un événement particulier se produit. Ceci se traduit par exemple par l'arrêt automatique d'un escalier mécanique avec comme événement la détection d'une personne qui tombe, et comme action l'arrêt de l'escalier mécanique.

Retour indirect de l'information se produit lorsque la décision du contenu à retourner se fait de manière immédiate à l'aide du lancement d'une tache par la personne qui observe les caméras et les résultats obtenus, ou bien se fait de manière différée. Le lancement de l'action appropriée peut se faire manuellement ou automatiquement. Dans le scénario d'analyse du comportement des clients d'un magasin, ça peu se traduire par une réorganisation des étalages par exemple. Le retour indirect de l'information permet d'exploiter principalement les données obtenues en hors ligne.

6.8 Conclusion

Dans ce chapitre, nous avons présenté un système qui permet d'analyser le comportement visuel d'une personne dans un environnement précis. La connaissance des caractéristiques de l'environnement aident à accélérer le traitement et à augmenter la précision du système. La méthode utilisée pour la détection des personnes présentes dans la scène se base sur une modélisation de l'arrière plan en utilisant un mélange de Gaussienne. La sélection des paramètres nécessaire à la modélisation est effectuée de sorte à améliorer les résultats de la détection des blobs. Deux approches pour le suivi de personnes sont présentées (une basée sur la mise en correspondance des blobs entre 2 images successives et la deuxième basée sur le suivi d'une personne lors de son apparition dans la scène grâce à son apparence). L'exploitation des différentes modalités extraites permettent d'obtenir un retour direct/indirect de l'information pertinent.

Chapitre 7

Conclusions et perspectives

Sommaire

7.1 Résultats principaux

Ce mémoire de thèse présente le travail réalisé basé sur différentes approches de vision par ordinateur pour construire un système qui détecte plusieurs modalités relatives à l'extraction de la direction du regard d'une personne à partir d'un flux vidéo dans un environnement contrôlé. L'environnement utilisé qui peut être en intérieur, en extérieur ou personnel se compose d'une zone surveillée et d'une scène cible. La scène cible est une région d'intérêt définie pour être analysée alors que la zone surveillée est l'emplacement d'où les personnes regardent cette scène cible.

Nous avons proposé une approche qui permet d'extraire l'attention visuelle d'une personne qui regarde une scène cible. Pour cela, nous nous sommes intéressés à l'estimation de l'orientation de la tête qui constitue le facteur le plus important pour l'estimation du regard. Pour cela une méthode hybride a été proposée et permet de palier aux inconvénients liés à l'utilisation d'une seule méthode. La première méthode basée sur l'apparence globale est utilisée afin d'obtenir une précision élevée de l'estimation de la pose sur des images prises à courte ou lointaine distance. La deuxième méthode est basée sur un modèle cylindrique pour le suivi de la tête à travers le temps afin d'exploiter les informations temporelles. Ce modèle est adapté lorsque le visage est proche de la caméra et en absence de poses extrêmes de la tête. La méthode basée sur l'apparence globale est aussi utilisée pour la réinitialisation du suivi de visages. Ensuite, le champ visuel d'une personne déterminé en se basant sur les données physiologiques de la vision humaine est adapté à la pose de la tête. En présence d'une scène cible en face de la personne suivie, une méthode de projection géométrique a été utilisée pour extraire les points de fixation et les régions d'intérêts. La localisation des yeux dans l'image nous a permis de détecter plus précisément les points de fixation et ainsi réduire sensiblement leurs écart-type.

Plusieurs représentations de l'information issues du regard ont été utilisées (e.g. régions d'intérêts, points de fixation, heatmap, chemins) et ont servies pour catégoriser les observateurs et pour réorganiser la scène cible. La validation de notre approche a été effectuée dans plusieurs environnements. Une méthode basée sur un certain nombre de métriques a ainsi été proposée pour évaluer la qualité d'une scène cible.

7.2 Perspectives

Bien que nous ayons effectué des recherches approfondies pour analyser le comportement visuel d'une personne dans un environnement contrôlé dans cette thèse, d'autres investigations peuvent être menées dans chacun des 3 composants de notre approche. Nous avons analysé certaines des limites de notre travail nous proposons dans ce qui suit les orientations de recherche futures pour chaque composant : détection et suivi de personnes, estimation de la pose de la tête, projection du champ visuel et extraction des régions d'intérêts.

L'extension majeure se situe dans le composant de détection et de suivi de personnes. La méthode proposée a été appliquée dans des environnements intérieurs ou extérieurs en présence de 4 personnes au maximum. Il serait intéressant de pouvoir suivre plusieurs personne qui regardent la scène cible dans un environnement complexe (e.g. en présence de foule, arrière plan en changement continue, etc.). L'influence des conditions d'éclairage n'a pas été abordée dans notre thèse car la plupart des bases vidéo utilisées pour valider notre approche sont filmées dans des conditions d'éclairage stables. Il faudrait aussi mieux utiliser l'information provenant de l'environnement de la scène afin d'améliorer les résultats du suivi des personnes. En effet, une localisation dans la scène 3D de la personne permettrait de coupler les données provenant des zones observées et des zones observables afin de mieux comprendre les personnes présentes dans la scène et pouvoir les suivre en cas d'occlusions.

En ce qui concerne le composant d'estimation de la pose de la tête, le modèle d'apparence qui a été appris de la base d'images dépend de la vue caméra utilisée lors de l'enregistrement des données d'apprentissage. Il faudrait effectuer une normalisation du modèle pour pouvoir l'utiliser indépendamment de la position de la caméra. Une autre piste qui mérite d'être exploitée consiste en l'utilisation de l'apprentissage semi-supervisé. En effet, beaucoup de bases d'images utilisées pour la reconnaissance ou la détection de visages sont disponibles. Par contre, la pose associée à ces visages est inconnue et une annotation manuelle devient alors nécessaire pour obtenir cette information. L'exploitation du parallélisme des traitements pour effectuer une estimation de l'orientation de la tête par apparence globale et par apparence locale permettrait d'améliorer l'estimation en donnant plus de poids à l'apparence locale à courte distance.

Dans la partie projection du champ visuel, le modèle associé à la scène cible est perpendiculaire au sol. L'utilisation de la géométrie projective permettrait d'extraire les régions d'intérêts sur une scène cible inclinée par exemple. L'apprentissage d'un modèle pour détecter les points saillant d'une scène cible à partir des points de fixations obtenus pourrait s'avérer utile pour analyser une scène cible lorsque l'estimation du regard devient impossible. Une méthode d'évaluation de ces points de fixation est en cours de réalisation afin de permettre de déterminer la précision d'un système de suivi du regard à partir de l'information issue du regard et non pas grâce à l'orientation de la tête et la localisation des yeux.

L'approche que nous avons proposée a été essentiellement validée dans le domaine commercial. D'autres aspects applicatifs tels que dans le domaine de la surveillance (domestique ou publique) et dans le domaine des transports (applications de vigilance pour les conducteurs de machines ou dans le domaine aérien) peuvent être utilisés afin de mettre encore plus en valeur notre approche. Un vaste champ d'applications s'ouvre donc et pourrait tirer rapidement profit des concepts définis et mis en œuvre dans cette thèse.

Enfin, nous avons étudié le comportement visuel d'une personne qui regarde une scène cible à partir d'un flux vidéo issu d'une seule caméra. En fonction de l'environnement utilisé, nous pouvons exploiter des vues de plusieurs caméras pour améliorer la détection et le suivi de personnes. Nous pouvons aussi utiliser une caméra placée de manière orthogonale afin d'obtenir des informations complémentaires qui serait fusionnées pour estimer au mieux la distance qui sépare l'utilisateur à la scène cible ainsi que l'orientation de sa tête. L'utilisation de caméras haute résolution au lieu de webcams permettrait de mieux détecter les yeux des personnes à une grande distance. L'extraction des caractéristiques serait aussi améliorée en utilisant de telles caméras.

Publications

Chapitre de livre avec comité de programme et actes

[LMD09] Adel Lablack, Jean Martinet et Chabane Djeraba. Head pose estimation using a texture model based on Gabor Wavelets. *Chapter of the book "Pattern Recognition"*, I-Tech Education and Publishing, 2009.

[LMD08b] Adel Lablack, Jean Martinet et Chabane Djeraba. Multimodal analysis of human behavior. *Chapter of the book "Encyclopedia of Multimedia"*, Springer, 2008.

[MLID08] Jean Martinet, Adel Lablack, Nacim Ihaddadene et Chabane Djeraba. Gaze tracking applied to image indexing. *Chapter of the book "Encyclopedia of Multimedia"*, Springer, 2008.

Conférences Internationales avec comité de programme et actes

[VLSDG09] Roberto Valenti, Adel Lablack, Nicu Sebe, Chabane Djeraba et Theo Gevers. Visual Gaze Estimation by Joint Head and Eye Information. *20th International Conference on Pattern Recognition (ICPR 2010)*, 23 - 26 August, 2010. Istanbul - Turkey. **(submitted)**

[LMID09] Adel Lablack, Frédéric Maquet, Nacim Ihaddadene et Chabane Djeraba Visual gaze projection in front of a target scene. *2009 IEEE International Conference on Multimedia and Expo (ICME 2009)*, June 28 - July 3, 2009. New York City, NY - USA.

[ILD09] Nacim Ihaddadene, Adel Lablack et Chabane Djeraba. Analysing complex videos for public safety and monitoring. *9th International Symposium on Programming and Systems (ISPS 2009)*, 25 - 27 May, 2009. Algiers - Algeria.

[LD08] Adel Lablack et Chabane Djeraba. Analysis of human behaviour in front of a target scene. *19th International Conference on Pattern Recognition (ICPR 2008)*, 08 - 11 December, 2008. Tampa, FL - USA.

[Lab08] Adel Lablack. Head pose estimation for visual field projection. *16th ACM Conference on Multimedia (ACM MM 2008)*, 26 - 31 October, 2008. Vancouver, BC - Canada.

[LMD08a] Adel Lablack, Frédéric Maquet et Chabane Djeraba. Determination of the visual field of persons in a scene. *3rd International Conference on Computer Vision Theory and Applications (VISAPP 2008)*, 22 - 25 January, 2008. Funchal, Madeira - Portugal.

Workshops Internationaux avec comité de programme et actes

[MLLD09] Jean Martinet, Adel Lablack, Stanislas Lew et Chabane Djeraba. Gaze based quality assessment of visual media understanding. *1st International Workshop on Computer Vision and Its Application to Image Media Processing (WCVIM) in conjunction with the 3rd Pacific-Rim Symposium on Image and Video Technology (PSIVT 2009)*, 13 - 16 January, 2009. Tokyo - Japan.

[LZD08] Adel Lablack, Zhongfei (Mark) Zhang et Chabane Djeraba. Supervised learning for head pose estimation using SVD and Gabor Wavelets. *1st International Workshop on Multimedia Analysis of User Behaviour and Interactions (MAUBI) in conjunction with the 10th IEEE International Symposium on Multimedia (ISM 2008)*, 15 - 17 December, 2008. Berkeley, CA - USA.

Conférences Nationales avec comité de programme et actes

[LUBD10] Adel Lablack, Thierry Urruty, Yassine Benabbas et Chabane Djeraba. Extraction de la région d'intérêt d'une personne sur un obstacle. *10ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC 2010)*, 26 - 29 January, 2010. Hammamet - Tunisia.

[UELFJ10] Thierry Urruty, Yue Feng, Adel Lablack, Joemon Jose et Ismail Elsayad. Classification et sélection de caractéristique basées sur les concepts sémantiques pour la recherche d'information multimédia. *10ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC 2010)*, 26 - 29 January, 2010. Hammamet - Tunisia.

Bibliographie

[AA01] A. Ali and J.K. Aggarwal. Segmentation and recognition of continuous human activity. In *IEEE Workshop on Detection and Recognition of Events in Video*, pages 28–35, 2001.

[AAA+06] L. Angell, J. Auflick, P.A. Austria, D. Kochhar, L. Tijerina, W. Biever, T. Diptiman, J. Hogsett, and S. Kiger. Driver workload metrics task 2 final report. *Nat. Highway Traffic Safety Admin., U.S. Dept. Transp.*, November 2006.

[AD65] Michael Argyle and Janet Dean. Eye-contact, distance and affiliation. *Sociometry*, 28(3) :289–304, 1965.

[AHI08] Hirotaka Aoki, John Paulin Hansen, and Kenji Itoh. Learning to interact with a computer by gaze. *Behaviour & Information Technology*, 27(4) :339–344, 2008.

[AL09] Tamar Avraham and Michael Lindenbaum. Esaliency (extended saliency) : Meaningful attention using stochastic image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 99(1), 2009.

[ANA] Projet ANAFIX. http ://www.ouestaudio.com/anafix.

[Anl76] J. Anliker. *Eye movement : On-line measurement, analysis and control*. Eye mouvement and Psychological Processes, 1976.

[AS01] Douglas Ayers and Mubarak Shah. Monitoring human behavior from video taken in an office environment. *Image and Vision Computing*, 19 :833–846, 2001.

[Avi01] Shai Avidan. Support vector tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 184–191, 2001.

[BÓ2] Mohamed Bénallal. *Système de calibration de Camera : Localisation de forme polyedrique par vision monoculaire*. PhD thesis, Ecole des Mines de Paris, 2002.

[BA96] Michael J. Black and P. Anandan. The robust estimation of multiple motions : parametric and piecewise-smooth flow fields. *Computer Vision Image Understanding (CVIU)*, 63(1) :75–104, 1996.

[Ba07] Sileye Oumar Ba. *Joint Head Tracking and Pose Estimation for Visual Focus of Attention Recognition*. PhD thesis, Ecole Polytechnique Federale de Lausanne, February 2007.

[BAB+05] Moshe Bar, Elissa Aminoff, Jasmine Boshyan, Mark Fenske, Nurit Gronauo, and Karim Kassam. The contribution of context to visual object recognition. *Journal of Vision (JOV)*, 5(8), September 2005.

[Bat04] Jorge P. Batista. Tracking pedestrians under occlusion using multiple cameras. In *International Conference on Image Analysis and Recognition*, pages 552–562, Porto - Portugal, 2004.

[BB82] Dana H. Ballard and Christopher M. Brown. *Computer Vision*. Prentice-Hall, 1982.

[BC86] T.J Broida and R. Chellappa. Estimation of object motion parameters from noisy images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 8(1) :90–99, 1986.

[BEBV08] Thierry Bouwmans, Fida El Baf, and Bertrand Vachon. Background Modeling using Mixture of Gaussians for Foreground Detection - A Survey. *Recent Patents on Computer Science*, 1(3) :219–237, 11 2008.

[Bey00] David Beymer. Person counting using stereo. In *Proceedings of the Workshop on Human Motion (HUMO)*, Austin, TX - USA, December 2000.

[BFB04] Csaba Beleznai, Bernhard Frühstück, and Horst Bischof. Human detection in groups using a fast mean shift procedure. In *International Conference on Image Processing (ICIP)*, pages 349–352, Singapore, 2004.

[BFBB94] J. L. Barron, D. J. Fleet, S. S. Beauchemin, and T. A. Burkitt. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1) :43–77, 1994.

[BG95] Hilary Buxton and Shaogang Gong. Visual surveillance in a dynamic and uncertain world. *Artificial Intelligence Journal*, 78(1-2) :431–459, 1995.

[BGG96] Vicki Bruce, Patrick R. Green, and Mark A. Georgeson. *Visual Perception : Physiology, psychology and ecology*. Psychology Press, 1996.

[BGV92] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, 1992.

[BJ98] Michael J. Black and Allan D. Jepson. Eigentracking : Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision (IJCV)*, pages 329–342, 1998.

[BKD01] Kevin Bowyer, Christine Kranenburg, and Sean Dougherty. Edge detector evaluation using empirical roc curves. *Computer Vision Image Understanding (CVIU)*, 84(1) :77–103, 2001.

[BKP08] Vineeth Nallure Balasubramanian, Sreekar Krishna, and Sethuraman Panchanathan. Person-independent head pose estimation using biased manifold embedding. *EURASIP J. Adv. Signal Process*, 2008 :1–15, 2008.

[BM98] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *11th annual conference on COmputational Learning Theory (COLT)*, pages 92–100, 1998.

[BM09] Kevin Bailly and Maurice Milgram. Head pan angle estimation by a nonlinear regression on selected features. In *International Conference on Image Processing (ICIP)*, pages 3589–3592, Cairo - Egypt, 2009.

[BMR82] Irving Biederman, Robert J. Mezzanotte, and Jan C. Rabinowitz. Scene perception : Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2) :143–177, April 1982.

[BN03] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6) :1373–1396, June 2003.

[BNPS03] S. D. Bona, H. Niemann, G. Pieri, and O. Salvetti. Brain volumes characterisation using hierarchical neural networks. *Artificial Intelligence in Medicine*, 28(3) :307–322, 2003.

[BO04] Sileye O. Ba and Jean-Marc Odobez. A probabilistic framework for joint head tracking and pose estimation. In *17th International Conference on Pattern Recognition (ICPR)*, volume 4, pages 264–267, 2004.

[BR05] A. Bovyrin and K. Rodyushkin. Human height prediction and roads estimation for advanced video surveillance systems. In *Advanced Video and Signal Based Surveillance (AVSS)*, pages 219–223, 2005.

[BSF87] Yaakov Bar-Shalom and Thomas E. Fortmann. *Tracking and data association*, volume 179 of *Mathematics in Science and Engineering*. Academic Press Professional, Inc., 1987.

[BSR00] Marcelo Bertalmío, Guillermo Sapiro, and Gregory Randall. Morphing active contours. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22(7) :733–737, 2000.

[BSW06] Li Bai, Linlin Shen, and Yan Wang. A novel eye location algorithm based on radial symmetry transform. In *18th International Conference on Pattern Recognition (ICPR)*, volume 3, pages 511–514, Hong Kong, 2006.

[BT09] Neil D. B. Bruce and John K. Tsotsos. Saliency, attention, and visual search : An information theoretic approach. *Journal of Vision*, 9(3) :1–24, March 2009.

[Bus35] G. T. Buswell. *How People Look at Pictures : A Study of The Psychology of Perception in Art*. University of Chicago Press, 1935.

[Can86] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 8(6) :679–698, November 1986.

[CD00] R. Cutler and L. S. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22(8) :781–796, 2000.

[CdFB04] Peter Carbonetto, Nando de Freitas, and Kobus Barnard. A statistical model for general contextual object recognition. In *European Conference on Computer Vision (ECCV)*, pages 350–362, Prague, Czech Republic, May 2004.

[CET01] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 23(6) :681–685, 2001.

[CK05] S. Cheung and C. Kamath. Robust background subtraction with foreground validation for urban traffic video. *EURASIP Journal on Applied Signal Processing, Special Issue on Advances in Intelligent Vision Systems : Methods and Applications*, 14 :2330–2340, 2005.

[CLK+00] R.T. Collins, A.J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, and O. Hasegawa. A system for video surveillance and monitoring. Technical report, CMU, 2000.

[COG06] *Cogain : communication by gaze interaction, gazing into the future*, http ://www.cogain.org, September 2006.

[CRM00] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of nonrigid objects using mean shift. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 142–149, June 2000.

[CRM03] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 25(5) :564–577, 2003.

[CSA00] M. La Cascia, S. Sclaroff, and V. Athitsos. 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 (TPAMI)*, 22(4) :322–336, 2000.

[CST00] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[CTCG95] Timothy F. Cootes, Christopher J. Taylor, David H. Cooper, and Jim Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1) :38–59, 1995.

[Dat] CAVIAR Dataset. http ://groups.inf.ed.ac.uk/vision/CAVIAR.

[DB03] Anthony Robert Dick and Michael John Brooks. Issues in automated visual surveillance. In *Digital Image Computing : Techniques and Application (DICTA)*, pages 195–204, Sydney, Australia, 2003.

[DC01] R. Dodge and T.S. Cline. *The angle velocity of eye movements*. Psychological Review, 1901.

[DLR77] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1) :1–38, 1977.

[Dog04] Neil A. Dogson. Variation and extrema of human interpupillary distance. In *SPIE*, pages 36–46, San Jose, USA, 2004.

[Dou94] C. Dousson. *Suivi d'Évolutions et Reconnaissance de Chroniques*. PhD thesis, LAAS/CNRS - Université Paul Sabatier, Toulouse, 1994.

[DS02] Doug DeCarlo and Anthony Santella. Stylization and abstraction of photographs. In *29th annual conference on Computer graphics and interactive techniques (SIGGRAPH)*, pages 769–776, San Antonio, Texas, 2002.

[DT05] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893, June 2005.

[DT09] Anup Doshi and Mohan M. Trivedi. On the roles of eye gaze and head pose in predicting driver's intent to change lanes. *IEEE Transactions on Intelligent Transportation Systems*, 10(3) :453–462, September 2009.

[Dun72] Starkey Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23 :283–292, 1972.

[EBBV07] Fida El Baf, Thierry Bouwmans, and Bertrand Vachon. Comparison of background subtraction methods for a multimedia learning space. In *International Conference on Signal Processing and Multimedia (SIGMAP)*, Barcelona - Spain, July 2007.

[EESA08] S. Elhabian, K. El-Sayed, and S. Ahmed. Moving object detection in spatial domain using background removal techniques - state-of-art. *Recent Patents on Computer Science*, 1(1) :32–54, 2008.

[EHSTO09] Krista Ehinger, Barbara Hidalgo-Sotelo, Antonio Torralba, and Aude Oliva. Modeling search for people in 900 scenes : A combined source model of eye guidance. *Visual Cognition*, 17(6-7) :945–978, August 2009.

[ELW03] H. Elzein, S. Lakshmanan, and P. Watta. A motion and shape based pedestrian detection algorithm. In *IEEE Intelligent Vehicles Symposium*, page 500–504, Columbus, Ohio - USa, 2003.

[ETC98] G.J. Edwards, C.J. Taylor, and T.F. Cootes. Interpreting face images using active appearance models. In *3rd IEEE International Conference on Automatic Face and Gesture Recognition*, pages 300–305, 1998.

[EWKY04] How-Lung Eng, Junxian Wang, Alvin Harvey Kam, and Wei-Yun Yau. A bayesian framework for robust human detection and occlusion handling using human shape model. In *17th International Conference on Pattern Recognition (ICPR)*, pages 257–260, 2004.

[FH06] Y. Fu and T.S. Huang. Graph embedded analysis for head pose estimation. In *International Conference Automatic Face and Gesture Recognition (AFGR)*, pages 3–8, 2006.

[FJM50] P.M. Fitts, R.E. Jones, and J.L. Milton. *Eye movements of aircraft pilots during instrument-landing approaches*. Aeronautical Engineering Review, 1950.

[FK05] H. Fillbrandt and K.F. Kraiss. Simultaneous localization and tracking of persons in a cluttered scene with a single camera. In *5th WSEAS International Conference on Signal, Speech and Image Processing (SSIP)*, pages 236–241, Corfu - Greece, August 2005.

[FP03] M. Fink and P. Perona. Mutual boosting for contextual inference. In *Neural Information Processing Systems (NIPS)*, 2003.

[FPZ04] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. In *8th IEEE European Conference on Computer Vision (ECCV)*, Prague, Czech Republic, 2004.

[FR97] N. Friedman and S. Russell. Image segmentation in video sequences : A probabilistic approach. In *13th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 175–181, 1997.

[FS95] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory*, pages 23–37, 1995.

[FT97] Paul Fieguth and Demetri Terzopoulos. Color-based tracking of heads and other mobile objects at video frame rates. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21–27, 1997.

[GBC+88] C. C. Gordon, B. Bradtmiller, T. Churchill, C. E. Clauser, J. T. McConville, I. O. Tebbetts, and R. A. Walker. Anthropometric survey of us army personnel : Methods and summary statistics. Technical report, United States Army Natick Research, 1988.

[GBG+94] H. Greenspan, S. Belongie, R. Goodman, P. Perona, S. Rakshit, and C. H. Anderson. Overcomplete steerable pyramid filters and rotation invariance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 222–228, 1994.

[GEN95] Arne John Glenstrup and Theo Engell-Nielsen. Eye controlled media : Present and future state. Technical report, Laboratory of Psychology, 1995.

[GG02] D. M. Gavrila and J. Giebel. Shape-based pedestrian detection and tracking. In *IEEE Intelligent Vehicle Symposium*, volume 1, pages 8–14, 2002.

[GHC04] N. Gourier, D. Hall, and J. L. Crowley. Estimating face orientation from robust detection of salient facial features. In *Pointing 2004, ICPR Workshop on Visual Observation of Deictic Gestures*, 2004.

[Gil21] A.R. Gilliland. *Photographic methods for studying reading*. Visual Education, 1921.

[Gir04] P.K. Girard. *Quaternions, algèbre de Clifford et physique relativiste*. PPUR, 2004.

[GK95] Lynne Grewe and Avinash C. Kak. Interactive learning of a multiple-attribute hash table classifier for fast object recognition. *Computer Vision and Image Understanding (CVIU)*, 61(3) :387–416, 1995.

[Gou06] Nicolas Gourier. *Machine Observation of the Direction of Human Visual Focus of Attention*. PhD thesis, Iinstitut National Polytechnique de Grenoble, 2006.

[Gri96] J. Gribbin. *Schrodinger's Kittens and the Search for Reality : Solving the Quantum Mysteries*. Little Brown et Co., 1996.

[Hal86] P. Hallett. *Eye movements*. Handbook of perception and human performance, 1986.

[HB03] J. Han and B. Bhanu. Detecting moving humans using color and infrared video. In *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, page 228–233, 2003.

[HBF02] Ismail Haritaoglu, David Beymer, and Myron Flickner. Ghost3d : Detecting body posture and parts using stereo. In *Proceedings of the Workshop on Motion and Video Computing (MOTION)*, 2002.

[HCZZ04] Y. Hu, L. Chen, Y. Zhou, and H. Zhang. Estimating face pose by facial asymmetry and geometry. In *6th IEEE International Conference on Automatic Face and Gesture Recognition (AFGR)*, Seoul - Korea, May 2004.

[HF02] Ismail Haritaoglu and Myron Flickner. Attentive billboards : Towards to video based customer behavior. In *Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision (WACV)*, 2002.

[HHD98] Ismail Haritaoglu, David Harwood, and Larry S. Davis. W4 : A real time system for detecting and tracking people. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Santa Barbara, CA - USA, 1998.

[HHR05] Nan Hu, Weimin Huang, and Surendra Ranganath. Head pose estimation by non-linear embedding and mapping. In *International Conference on Image Processing (ICIP)*, volume 2, pages 342–345, Genoa - Italy, 2005.

[HJ10] Dan Witzner Hansen and Qiang Ji. In the eye of the beholder : A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[HJM+89] T.E. Hutchinson, K.P. White Jr, W.N. Martin, K.C. Reichert, and L.A. Frey. Human-computer interaction using eye-gaze input. *IEEE Transactions on Systems, Man and Cybernetics (SMC)*, 19(6) :1527–1534, 1989.

[HLCC08] Sébastien Hillaire, Anatole Lécuyer, Rémi Cozot, and Géry Casiez. Using an eye-tracking system to improve camera motions and depth-of-field blur effects in virtual environments. *IEEE Virtual Reality Conference*, pages 47–55, 2008.

[HR78] A.R. Hanson and E.M. Riseman. Visions : A computer system for interpreting scenes. In *Computer Vision Systems (CVS)*, pages 303–333. Academic Press, 1978.

[HS81] B.K.P. Horn and B.G. Schunk. Determining optical flow. *Artificial Intelligence*, 17 :185–203, 1981.

[HSD73] R.M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6) :610–621, 1973.

[HSW98] J. Huang, X. Shao, and H. Wechsler. Face pose discrimination using support vector machines (svm). In *International Conference on Pattern Recognition (ICPR)*, volume 1, pages 154–156, Brisbane - Australia, 1998.

[HSY04] T. Haga, K. Sumi, and Y. Yagi. Human detection in outdoor scene using spatio-temporal motion analysis. In *International Conference on Pattern Recognition (ICPR)*, volume 4, page 331–334, 2004.

[HT04] K.S. Huang and M.M. Trivedi. Robust real-time detection, tracking, and pose estimation of faces in video streams. In *17th International Conference on Pattern Recognition (ICPR)*, volume 3, pages 965–968, 2004.

[Hue00] E.B. Huey. On the psychology and physiology of reading. *The American Journal of Psychology*, 11(3), 1900.

[HZ07] Xiaodi Hou and Liqing Zhang. Saliency detection : A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007.

[IB98] Michael Isard and Andrew Blake. Condensation - conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29 :5–28, 1998.

[IB00] Yuri A. Ivanov and Aaron F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22(8) :852–872, 2000.

[IK00] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40 :1489–1506, 2000.

[ILD09] Nacim Ihaddadene, Adel Lablack, and Chabane Djeraba. Analysing complex videos for public safety and monitoring. In *9th International Symposium on Programming and Systems (ISPS)*, Algiers - Algeria, May 2009.

[Int99] S.S. Intille. *Visual Recognition of Multi-Agent Action*. PhD thesis, Massachusetts Institute of Technology, 1999.

[Jav78] Emile Javal. *Essai sur la physiologie de la lecture*. Annales d'Oculistique, 1878.

[JEDT09] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Modelling activity global temporal dependencies using time delayed probabilistic graphical model. In *International Conference on Computer Vision (ICCV)*, Kyoto - Japan, 2009.

[JMS05] C.H. Judd, C.N. McAllister, and W.M. Steel. *General introduction to a series of studies of eye movements by means of kinetoscopic photographss*. Psychological Review, 1905.

[JodA04] Natasa Jovanovic and Rieks op den Akker. Towards automatic addressee identification in multi-party dialogues. In *5th SIGdial Workshop on Discourse and Dialogue*, page 89–92, Cambridge, MA -USA, 2004.

[JTS+04] Lijun Jiang, Feng Tian, Lim Ee Shen, Shiqian Wu, Susu Yao, Zhongkang Lu, and Lijun Xu. Perceptual-based fusion of ir and visual images for human detection. In *International Symposium on Intelligent Multimedia, Video and Speech Processing*, page 514–517, 2004.

[KCM04] Jinman Kang, Isaac Cohen, and Gérard G. Medioni. Object reacquisition using invariant appearance model. In *International Conference on Pattern Recognition (ICPR)*, volume 4, pages 759–762, 2004.

[KH03] S. Kumar and M. Hebert. Discriminative random fields : a discriminative framework for contextual interaction in classification. In *International Conference on Computer Vision (ICCV)*, volume 2, pages 1150–1157, Nice, France, 2003.

[KHT02] Charles Kervrann, Mark Hoebeke, and Alain Trubuil. Isophotes selection and reaction-diffusion model for object boundaries estimation. *International Journal of Computer Vision*, 50(1) :63–94, 2002.

[KK01] Hiromi Kobayashi and Shiro Kohshima. Unique morphology of the human eye and its adaptive meaning : comparative studies on external morphology of the primate eye. *Journal of Human Evolution*, 40(5) :419–435, May 2001.

[KLS03] Michael Kockelkorn, Andreas Lüneburg, and Tobias Scheffer. Using transduction and multi-view learning to answer emails. In *7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 266–277, Cavtat-Dubrovnik, Croatia, September 2003.

[KPvdM97] Norbert Krüger, Michael Pötzsch, and Christoph von der Malsburg. Determination of face position and pose with a learned representation based on labelled graphs. *Image Vision Computing*, 15(8) :665–673, 1997.

[KRT+05] N. Krahnstoever, J. Rittscher, P. Tu, K. Chean, and T. Tomlinson. Activity recognition using visual tracking and rfid. In *Proceedings of the Seventh IEEE Workshops on Application of Computer Vision (WACV/MOTION)*, volume 1, pages 494–500, 2005.

[KS03] Hannes Kruppa and Bernt Schiele. Using local context to improve face detection. In *British Machine Vision Conference (BMVC)*, Norwich, UK, 2003.

[KSB08] Yvonne Kammerer, Katharina Scheiter, and Wolfgang Beinhauer. Looking my way through the menu : the impact of menu design and multimodal input on gaze-based menu selection. In *Eye Tracking Research & Application Symposium (ETRA)*, pages 213–220, Savannah, Georgia - USA, 2008.

[KTB96] D. Kersten, N.F. Troje, and H.H. Bülthoff. *Phenomenal competition for poses of the human head*, volume 25. Perception, 1996.

[KTF02] Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal on Computer Vision (IJCV)*, 50(2) :171–184, 2002.

[KWSF06] Wolf Kienzle, Felix Wichmann, Bernhard Schölkopf, and Matthias Franz. A nonparametric approach to bottom-up visual saliency. In *Conference on Neural Information Processing Systems (NIPS)*, page 689–696, 2006.

[Lab08] Adel Lablack. Head pose estimation for visual field projection. In *16th ACM International Conference on Multimedia (MM '08)*, Vancouver, BC - Canada, October 2008.

[Law80] Kenneth Ivan Laws. *Textured Image Segmentation*. PhD thesis, University of Southern California, 1980.

[LD08] Adel Lablack and Chabane Djeraba. Analysis of human behaviour in front of a target scene. In *19th International Conference on Pattern Recognition (ICPR)*, Tampa, Florida - USA, December 2008.

[LDZ00] Anne Lorette, Xavier Descombes, and Josiane Zerubia. Texture analysis through a markovian modelling and fuzzy classification : Application to urban area extraction from satellite images. *International Journal of Computer Vision*, 36(3) :221–236, 2000.

[Lew06] Stanislas Lew. Extraction de connaissances à partir du suivi des positions du regard. Master's thesis, Laboratoire d'Informatique Fondamentale de Lille, 2006.

[LGS+04] Liyuan Li, Shuzhi Sam Ge, T. Sim, Ying Ting Koh, and Xiaoyu Hunag. Object-oriented scale-adaptive filtering for human detection from stereo images. In *IEEE Conference on Cybernetics and Intelligent Systems*, volume 1, pages 135–140, 2004.

[LGSL04] Y.M. Li, S.G. Gong, J. Sherrah, and H. Liddell. Support vector machine based multi-view face detection and recognition. *Image and Vision Computing (IVC)*, 22(5) :413–427, May 2004.

[LH02] B. Lee and M. Hedley. Background estimation for video surveillance. In *Image and Vision Computing New Zealand (IVCNZ)*, pages 315–320, 2002.

[LHR05] Jeroen Lichtenauer, Emile Hendriks, and Marcel Reinders. Isophote properties as features for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 649–654, 2005.

[LHT04] Stephen R.H. Langton, Helen Honeyman, and Emma Tessler. The influence of head contour and nose angle on the perception of eye-gaze direction. *Perception & psychophysics*, 66(5) :752–771, 2004.

[LK81] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 674–679, 1981.

[LKBP05] D. Little, S. Krishna, J. Black, and S. Panchanathan. A methodology for evaluating robustness of face recognition algorithms with respect to variations in pose angle and illumination angle. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 89–92, 2005.

[LKYT07] X. Liu, N. Krahnstoever, T. Yu, and P. Tu. What are customers looking at ? In *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS)*, London, 2007.

[LLTH02] Jiangung Lou, Qifeng Liu, Tieniu Tan, and Weiming Hu. Semantic interpretation of object activities in a surveillance system. In *16th International Conference on Pattern Recognition (ICPR)*, volume 3, pages 777–780, 2002.

[LMD08a] Adel Lablack, Frédéric Maquet, and Chabane Djeraba. Determination of the visual field of persons in a scene. In *3rd International Conference on Computer Vision Theory and Applications*, Funchal, Portugal, January 2008.

[LMD08b] Adel Lablack, Jean Martinet, and Chabane Djeraba. *Multimodal analysis of human behavior*. Springer, 2008.

[LMD09] Adel Lablack, Jean Martinet, and Chabane Djeraba. *Head pose estimation using a texture model based on Gabor Wavelets*. I-Tech Education and Publishing, 2009.

[LMID09] Adel Lablack, Frédéric Maquet, Nacim Ihaddadene, and Chabane Djeraba. Visual gaze projection in front of a target scene. In *2009 IEEE International Conference on Multimedia and Expo (ICME)*, New York City, NY - USA, 2009.

[LT07] A. Leykin and M. Tuceryan. Detecting shopper groups in video sequences. In *Advanced Video and Signal Based Surveillance (AVSS)*, pages 417–422, September 2007.

[LUBD10] Adel Lablack, Thierry Urruty, Yassine Benabbas, and Chabane Djeraba. Extraction de la région d'intérêt d'une personne sur un obstacle. In *10 ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC)*, Hammamet - Tunisia, January 2010.

[LVF03] A. Levin, P. Viola, and Y. Freund. Unsupervised improvement of visual detectors using co-training. In *International Conference on Computer Vision (ICCV)*, volume I, pages 626–633, 2003.

[LWB00] S.R.H. Langton, R.J. Watt, and V. Bruce. Do the eyes have it ? cues to the direction of social attention. *Trends in Cognitive Sciences*, 4(2) :50–59, February 2000.

[LXG09] Chen Change Loy, Tao Xiang, and Shaogang Gong. Learning to predict where humans look. In *International Conference on Computer Vision (ICCV)*, Kyoto - Japan, 2009.

[LZD08] Adel Lablack, Zhongfei (Mark) Zhang, and Chabane Djeraba. Supervised learning for head pose estimation using svd and gabor wavelets. In *1st International Workshop on Multimedia Analysis of User Behaviour and Interactions (MAUBI) in conjunction with the 10th IEEE International Symposium on Multimedia (ISM)*, pages 592–596, Berkeley, California - USA, December 2008.

[LZTS04] D.J. Lee, P. Zhan, A. Thomas, and R. Schoenberger. Shape-based human intrusion detection. In *SPIE International Symposium on Defense and Security, Visual Information Processing XIII*, page 81–91, 2004.

[Mac67] J.B. Macqueen. Some methods for classification and analysis of multivariate observations. In *Procedings of the Fifth Berkeley Symposium on Math, Statistics, and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[Mah36] P.C. Mahalanobis. On the generalised distance in statistics. *National Institute of Sciences of India*, 2(1) :49–55, 1936.

[McG84] Joseph E. McGrath. *Groups : Interaction and Performance*. Prentice Hall College Div, 1984.

[MCT09] Erik Murphy-Chutorian and Mohan Trivedi. Head pose estimation in computer vision : A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 31(4) :607–626, 2009.

[MD03] Anurag Mittal and Larry S. Davis. M2tracker : A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo. *International Journal of Computer Vision*, 51(3) :189–203, 2003.

[MKK+06] Yong Ma, Yoshinori Konishi, Koichi Kinoshita, Shihong Lao, and Masato Kawade. Sparse bayesian regression for head pose estimation. In *18th International Conference on Pattern Recognition (ICPR)*, pages 507–510, 2006.

[MLID08] Jean Martinet, Adel Lablack, Nacim Ihaddadene, and Chabane Djeraba. *Gaze tracking applied to image indexing*. Springer, 2008.

[MLLD09] Jean Martinet, Adel Lablack, Stanislas Lew, and Chabane Djeraba. Gaze based quality assessment of visual media understanding. In *1st International Workshop on Computer*

Vision and Its Application to Image Media Processing (WCVIM) in conjunction with the 3rd Pacific-Rim Symposium on Image and Video Technology (PSIVT), Tokyo - Japan, January 2009.

[MM58] J.F. Mackworth and N.H. Mackworth. Eye fixations recorded on changing visual scenes by the television eye-marker. *Journal of the Optical Society of America*, page 439–445, 1958.

[MMR02] L. Marcenaro, L. Marchesotti, and C. Regazzoni. Tracking and counting multiple interacting people in indoor scenes. In *Performance Evaluation of Tracking and Surveillance (PETS) Workshop*, Copenhagen, Denmark, 2002.

[MMSZ05] S. Messelodi, C. Modena, N. Segata, and M. Zanin. A kalman filter based background updating algorithm robust to sharp illumination changes. In *13th International Conference on Image Analysis and Processing (ICIAP)*, pages 163–170, Cagliari - Italy, 2005.

[MP97] Baback Moghaddam and Alex Pentland. Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 19(7) :696–710, 1997.

[MRD03] Louis-Philippe Morency, Ali Rahimi, and Trevor Darrell. Adaptive view-based appearance models. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1 :803–810, 2003.

[MS76] R.A. Monty and J.W. Senders. Eye movements and psychological processes. *Journal of Experimental Psychology : Human Perception and Performance*, 1976.

[MS05] A. Mustafa and I. Sethi. Detecting retail events using moving edges. In *Advanced Video and Signal Based Surveillance (AVSS)*, pages 626–631, 2005.

[MTF03] Kevin Murphy, Antonio Torralba, and William T. Freeman. Using the forest to see the trees : a graphical model relating features, objects and scenes. In *NIPS*. MIT Press, 2003.

[NC02] V. Nair and J.J. Clark. Automated visual surveillance using hidden markov models. In *International Conference on Vision Interface*, pages 88–93, 2002.

[NF96] S. Niyogi and W.T. Freeman. Example-based head tracking. In *2nd International Conference on Automatic Face and Gesture Recognition (AFGR)*, pages 374–378, 1996.

[NFK+02] Y. Nakanishi, T. Fujii, K. Kitajima, Y. Sato, and H. Koike. Vision-based face tracking system for large displays. In *4th international conference on ubiquitous Computing*, 2002.

[NHW96] David G. Novick, Brian Hansen, and Karen Ward. Coordinating turn-taking with gaze. In *International Conference on Spoken Language Processing (ICSLP)*, pages 1888–1891, 1996.

[OCM07] Margarita Osadchy, Yann Le Cun, and Matthew L. Miller. Synergistic face detection and pose estimation with energy-based models. *Journal Machine Learning Research*, 8 :1197–1215, 2007.

[ORP99] Nuria Oliver, Barbara Rosario, and Alex Pentland. A bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 :831–843, 1999.

[Pas01] George Paschos. Perceptually uniform color spaces for color texture analysis : an empirical evaluation. *IEEE Transactions on Image Processing*, 10(6) :932–937, June 2001.

[PBP04] Alex Poole, Linden J. Ball, and Peter Phillips. In search of salience : A response-time and eye-movement analysis of bookmark recognition. In *BCS HCI*, pages 363–378, Leeds – UK, 2004.

[PD02] Nikos Paragios and Rachid Deriche. Geodesic active regions and level set methods for supervised texture segmentation. *International Journal on Computer Vision*, 46(3) :223–247, 2002.

[Pec02] A.E.C. Pece. From cluster tracking to people counting. In *Performance Evaluation of Tracking and Surveillance (PETS) Workshop*, Copenhagen, Denmark, 2002.

[Pic04] M. Piccardi. Background subtraction techniques : A review. In *International Conference on Systems, Man and Cybernetics (SMC)*, The Hague, The Netherlands, 2004.

[PLW08] Thies Pfeiffer, Marc E. Latoschik, and Ipke Wachsmuth. Evaluation of binocular eye trackers and algorithms for 3d gaze interaction in virtual reality environments. *Journal of Virtual Reality and Broadcasting*, 5(16), December 2008.

[POP98] C.P. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *International Conference on Computer Vision (ICCV)*, pages 555–562, 1998.

[Por03] F. Porikli. Human body tracking by adaptive background models and mean-shift analysis. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS 2003)*, 2003.

[Pro] Hermes Project. http ://www.cvmt.dk/projects/Hermes/head-data.html.

[PSS+04] M. Peden, R. Scurfield, D. Sleet, D. Mohan, A.A. Hyder, E. Jarawan, and C. Mathers, editors. *World Report on Road Traffic Injury Prevention*, Geneva, Switzerland, 2004. World Health Org.

[PT05] F. Porikli and O. Tuzel. Bayesian background modeling for foreground detection. In *ACM Int Workshop on Video Surveillance and Sensor Networks (VSSN)*, pages 55–58, November 2005.

[PVM07] A. Pande, A. Verma, and A. Mittal. Network aware optimal resource allocation for e-learning videos. In *6th International Conference on mobile Learning*, Melbourne - Australia, 2007.

[PZ79] Julius Panero and Martin Zelnik. *Human Dimension and Interior Space : A Source Book of Design Reference Standards*. Watson-Guptill, 1979.

[PZJ05] Y. Pan, H. Zhu, and R. Ji. *3-D Head Pose Estimation for Monocular Image*. Fuzzy Systems and Knowledge Discovery. Springer, 2005.

[Ray98] K. Rayner. Eye movements in reading and information processing : 20 years of research. *Psychol Bull*, 124(3) :372–422, November 1998.

[RBK96] Henry A. Rowley, Shumeet Baluja, and Takeo Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 20 :23–38, 1996.

[RDD03] Azriel Rosenfeld, David Doermann, and Daniel DeMenthon. *Video Mining*. Kluwer Academic Publisher, 2003.

[Rob63] David A. Robinson. A method of measuring eye movement using a scleral search coil in a magnetic field. *IEEE Transactions on Bio-Medical Electronics*, page 137–145, 1963.

[Ron94] Remi Ronfard. Region based strategies for active contour models. *International Journal of Computer Vision*, 13(2) :229–251, October 1994.

[RR98] R. Rae and H. Ritter. Recognition of human head orientation based on artificial neural networks. *IEEE Trans. on Neural Networks*, 9(2) :257–265, 1998.

[RR06] N.M. Robertson and I.D. Reid. Estimating gaze direction from low-resolution faces in video. In *9th IEEE European Conference on Computer Vision (ECCV)*, Graz, Austria, 2006.

[RS00] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500) :2323–2326, December 2000.

[RSA08] Michael Rubinstein, Ariel Shamir, and Shai Avidan. Improved seam carving for video retargeting. *ACM Transactions on Graphics*, 27(3) :1–9, 2008.

[RYS04] Bisser Raytchev, Ikushi Yoda, and Katsuhiko Sakaue. Head pose estimation by nonlinear manifold learning. In *17th International Conference on Pattern Recognition (ICPR)*, volume 4, pages 462–466, 2004.

[SA04] Koichi Sato and Jake K. Aggarwal. Temporal spatio-velocity transform and its application to tracking and interaction. *Computer Vision Image Understanding*, 96(2) :100–128, 2004.

[SAD+06] Anthony Santella, Maneesh Agrawala, Doug DeCarlo, David Salesin, and Michael Cohen. Gaze-based interaction for semi-automatic photo cropping. In *CHI '06 : Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 771–780. ACM, 2006.

[SB02] S. Srinivasan and K. Boyer. Head-pose estimation using view based eigenspaces. In *16th International Conference on Pattern Recognition (ICPR)*, Quebec City - Canada, 2002.

[SBB03] Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 25 :1615–1618, 2003.

[SBGPO06] K. Smith, S.O. Ba, D. Gatica-Perez, and J-M. Odobez. Tracking the multi-person wandering visual focus of attention. In *International Conference on Multimodal Interfaces (ICMI)*, Banff, Canada, November 2006.

[SBH+07] A.W. Senior, L. Brown, A. Hampapur, C.F. Shu, Y. Zhai, R.S. Feris, Y.L. Tian, S. Borger, and C. Carlson. Video analytics for retail. In *Advanced Video and Signal Based Surveillance (AVSS)*, pages 423–428, September 2007.

[SBOGP08] Kevin Smith, Sileye O. Ba, Jean-Marc Odobez, and Daniel Gatica-Perez. Tracking the visual focus of attention for a varying number of wandering people. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(7) :1212–1229, July 2008.

[Sen02] A.W. Senior. Tracking with probabilistic appearance model. In *Performance Evaluation of Tracking and Surveillance (PETS) Workshop*, Copenhagen, Denmark, 2002.

[Sen03] A.W. Senior. Real-time articulated human body tracking using silhouette information. In *Performance Evaluation of Tracking and Surveillance (PETS) Workshop*, October 2003.

[SF91] Thomas M. Strat and Martin A. Fischler. Context-based vision : recognizing objects using information from both 2-d and 3-d imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 13 :1050–1065, 1991.

[SG99] C. Stauffer and W.E.L. Grimson. Adaptative background mixture models for real-time tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 246–252, 1999.

[SGO01] J. Sherrah, S. Gong, and E. J. Ong. Face distributions in similarity space under varying head pose. *Image and Vision Computing*, 19(12) :807–819, October 2001.

[Sid04] H. Sidenbladh. Detecting human motion with support vector machines. In *International Conference on Pattern Recognition (ICPR)*, volume 2, page 188–191, 2004.

[SKMG04] David Serby, Esther Koller-Meier, and Luc Van Gool. Probabilistic object tracking using multiple features. In *17th International Conference on Pattern Recognition (ICPR)*, volume 2, pages 184–187, 2004.

[SKP96] K.Y. Song, J.V. Kittler, and M. Petrou. Defect detection in random color textures. *Image and Vision Computing (IVC)*, 14(9) :667–683, October 1996.

[SLZ03] A. Singhal, J.B. Luo, and W.Y. Zhu. Probabilistic spatial context models for scene content understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages I : 235–241, Madison - Wisconsin, USA, June 2003.

[SM96] J.M. Siskind and Q. Morris. A maximum-likelihood approach to visual event classification. In *4th European Conference on Computer Vision (ECCV)*, volume II, page 347–360, 1996.

[SMBP05] K. Schwerdt, D. Maman, P. Bernas, and E. Paul. Target segmentation and event detection at video-rate : the eagle project. In *IEEE Conference on Advanced Video and Signal based Surveillance (AVSS)*, pages 183–188, Côme, Italie, September 2005.

[SMP08] M. Sigari, N. Mozayani, and H. Pourreza. Fuzzy running average and fuzzy background subtraction : concepts and application. *Int J Comput Sci Network Security*, 8(2) :138–143, 2008.

[SOM+09] William Steptoe, Oyewole Oyekoya, Alessio Murgia, Robin Wolff, John Rae, Estefania Guimaraes, David Roberts, and Anthony Steed. Eye tracking for avatar eye gaze control during object-focused multiparty interaction in immersive collaborative virtual environments. In *2009 IEEE Virtual Reality Conference (VR)*, pages 83–90, 2009.

[SRM05] M. Spirito, C.S. Regazzoni, and L. Marcenaro. Automatic detection of dangerous events for underground surveillance. In *IEEE Conference on Advanced Video and Signal based Surveillance (AVSS)*, pages 195–200, Côme, Italie, September 2005.

[SS90] V. Salari and I.K. Sethi. Feature point correspondence in the presence of occlusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 12(1) :87–91, 1990.

[ST94] Jianbo Shi and Carlo Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 593–600, 1994.

[Sti02] Rainer Stiefelhagen. *Tracking and Modeling Focus of Attention in Meetings*. PhD thesis, Université de Karlsruhe, 2002.

[SYW99] Rainer Stiefelhagen, Jie Yang, and Alex Waibel. Modeling focus of attention for meeting indexing. In *7th ACM International Conference on Multimedia*, volume 1, pages 3–10, 1999.

[SYW02] Rainer Stiefelhagen, Jie Yang, and Alex Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, 13 :928–938, 2002.

[SZ02] R. Stiefelhagen and J. Zhu. Head orientation and gaze direction in meetings. In *Conference on Human Factors in Computing Systems*, Minneapolis, Minnesota, 2002.

[TA03] D. Toth and T. Aach. Detection and recognition of moving objects using statistical motion detection and fourier descriptors. In *International Conference on Image Analysis and Processing (ICIAP)*, pages 430–435, 2003.

[TGM07] M.M. Trivedi, T. Gandhi, and J.C. McCall. Looking-in and looking-out of a vehicle : Computer-vision-based enhanced vehicle safety. *IEEE Transactions on Intelligent Transportation Systems*, 8(1) :108–120, March 2007.

[TJ00] Vildan Tanriverdi and Robert J.K. Jacob. Interacting with eye movements in virtual environments. In *SIGCHI conference on Human factors in computing systems*, pages 265–272, 2000.

[TKBM99] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower : Principles and practice of background maintenance. In *International Conference on Computer Vision (ICCV)*, pages 255–261, Corfu - Greece, 1999.

[TMF04] A. Torralba, K. P. Murphy, and W. T. Freeman. Contextual models for object detection using boosted random fields. In *Neural Information Processing Systems (NIPS)*, pages 1401–1408, 2004.

[TS01] A. Torralba and P. Sinha. Statistical context priming for object detection. In *International Conference on Computer Vision (ICCV)*, pages 763–770, Vancouver - BC, Canada, 2001.

[TSK02] Hai Tao, Harpreet S. Sawhney, and Rakesh Kumar. Object tracking with bayesian estimation of dynamic layer representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 24(1) :75–89, 2002.

[TSW90] C.C. Tappert, C.Y. Suen, and T. Wakahara. The state of the art in online handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 12(8) :787–808, 1990.

[UT02] Akira Utsumi and Nobuji Tetsutani. Human detection using geometrical pixel value structures. In *5th IEEE International Conference on Automatic Face and Gesture Recognition (AFGR)*, pages 39–44, 2002.

[Vac91] Richard J. Vaccaro. *SVD and Signal Processing II : Algorithms, Analysis and Applications*. Elsevier Science Inc., New York, NY, USA, 1991.

[Vap99] V.N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1999.

[VBB07] Teodora Vatahska, Maren Bennewitz, and Sven Behnke. Feature-based head pose estimation from images. In *IEEE-RAS 7th International Conference on Humanoid Robots (Humanoids)*, Pittsburgh - USA, December 2007.

[VG08] Roberto Valenti and Theo Gevers. Accurate eye center location and tracking using isophote curvature. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[VHAC06] Arantxa Villanueva, Dan Witzner Hansen, Javier San Agustin, and Rafael Cabeza. Basics of gaze estimation. In *2nd Conference on Communication by Gaze Interaction (COGAIN)*, 2006.

[VJ01] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages I : 511–518, Lihue, Hawaii, USA, 2001.

[VJS03] Paul Viola, Michael Jones, and Daniel Snow. Detecting pedestrians using patterns of motion and appearance. In *International Conference on Computer Vision (ICCV)*, pages 734–741, 2003.

[VLS+10] Roberto Valenti, Adel Lablack, Nicu Sebe, Chabane Djeraba, and Theo Gevers. Visual gaze estimation by joint head and eye information. In *20th International Conference on Pattern Recognition (ICPR)*, Istanbul - Turkey (submitted), August 2010.

[VRB01] C.J. Veenman, M.J.T. Reinders, and E. Backer. Resolving motion correspondence for densely moving points. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 23(1) :54–72, 2001.

[VS08] M. Voit and R. Stiefelhagen. Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In *International Conference on Multimodal interfaces (ICMI)*, pages 173–180, Chania – Grece, 2008.

[VSG08] R. Valenti, N. Sebe, and T. Gevers. Simple and efficient visual gaze estimation. In *Workshop on Multimodal Interactions Analysis of Users in a Controlled Environment*, 2008.

[VSvdVN01] Roel Vertegaal, Robert Slagter, Gerrit van der Veer, and Anton Nijholt. Eye gaze patterns in conversations : there is more to conversational agents than meets the eyes. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 301–308, 2001.

[VvdVV00] Roel Vertegaal, Gerrit van der Veer, and Harro Vons. Effects of gaze on multiparty mediated communication. In *Graphics Interface*, pages 95–102. Morgan Kaufmann, 2000.

[VVS98] Roel Vertegaal, Harro Vons, and Robert Slagter. Look who's talking : the gaze groupware system. In *SIGCHI conference summary on Human factors in computing systems*, pages 293–294, 1998.

[VYG09] R. Valenti, Z. Yucel, and T. Gevers. Robustifying eye center localization by head pose cues. In *21st International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 612–618, Miami, FL - USA, 2009.

[WADP97] Christopher Richard Wren, Ali Azarbayejani, Trevor Darrell, and Alex Pentland. Pfinder : Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 19(7) :780–785, 1997.

[WB06] Lior Wolf and Stanley Bileschi. A critical view of context. *International Journal of Computer Vision (IJCV)*, 69(2) :251–261, 2006.

[WLB03] Zhou Wang, Ligang Lu, and Alan C. Bovik. Foveation scalable video coding with automatic fixation selection. *IEEE Trans. Image Processing*, 12 :243–254, 2003.

[WM00] Toshikazu Wada and Takashi Matsuyama. Multiobject behavior recognition by event driven selective attention method. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22(8) :873–887, 2000.

[WN07] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75(2) :247–266, 2007.

[Wol24] W.H. Wollaston. *On the apparent direction of eyes in a portrait*, volume 114. Philosophical Transactions of the Royal Society of London, 1824.

[Wora] CLEAR Workshop. http ://www.clear-evaluation.org.

[Worb] PETS Workshop. http ://www.cvg.rdg.ac.uk/slides/pets.html.

[WS06] H. Wang and D. Suter. A novel robust statistical method for background initialization and visual surveillance. In *Asian Conference on Computer Vision (ACCV)*, pages 328–337, Hyderabad - India, January 2006.

[WT00] Y. Wu and K. Toyama. Wide range illumination insensitive head orientation estimation. In *Automatic Face and Gesture Recognition (AFGR)*, pages 183–188, Grenoble - France, 2000.

[WT08] Junwen Wu and Mohan M. Trivedi. A two-stage head pose estimation framework and evaluation. *Pattern Recognition*, 41(3) :1138–1158, 2008.

[XBMK04] J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-time combined 2d+3d active appearance models. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 535–542, 2004.

[XF03] Fengliang Xu and Kikuo Fujimura. Human detection using depth and gray images. In *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 115–121, 2003.

[XKC02] J. Xiao, T. Kanade, and J. Cohn. Robust full motion recovery of head by dynamic templates and re-registration techniques. In *Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 156–162, Washington, DC - USA, May 2002.

[YB99] Yaser Yacoob and Michael J. Black. Parameterized modeling and recognition of activities. *Computer Vision Image Understanding (CVIU)*, 73(2) :232–247, 1999.

[YJS06] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking : A survey. *ACM Computing Surveys*, 38(4), 2006.

[YK04] Sang Min Yoon and Hyunwoo Kim. Real-time multiple people detection using skin color, motion and appearance information. In *International Workshop on Robot and Human Interactive Communication*, pages 331–334, 2004.

[YLS04] Alper Yilmaz, Xin Li, and Mubarak Shah. Contour based object tracking with occlusion handling in video acquired using mobile cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26 :1531–1536, 2004.

[ZCHS03] Li Zhang, Brian Curless, Aaron Hertzmann, and Steven M. Seitz. Shape and motion under varying illumination : Unifying structure from motion, photometric stereo, and multi-view stereo. In *9th IEEE International Conference on Computer Vision (ICCV)*, pages 618–625, 2003.

[ZF04] Y.D. Zhu and K. Fujimura. Head pose estimation for driver monitoring. In *IEEE Intelligent Vehicles Symposium (IVS)*, pages 501–506, 2004.

[ZH05] Jianpeng Zhou and Jack Hoang. Real time robust human detection and tracking system. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, pages 149–156, 2005.

[ZN04] Tao Zhao and Ram Nevatia. Tracking multiple humans in complex situations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9) :1208–1221, 2004.

[ZW06] Mian Zhou and Hong Wei. Face verification using gaborwavelets and adaboost. In *18th International Conference on Pattern Recognition (ICPR)*, volume 1, pages 404–407, Hong Kong, August 2006.

[ZY96] Song Chun Zhu and Alan Yuille. Region competition : Unifying snakes, region growing, and bayes/mdl for multi-band image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18 :884–900, 1996.

[ZY02] Jie Zhu and Jie Yang. Subpixel eye gaze tracking. In *5th IEEE International Conference on Automatic Face and Gesture Recognition (AFGR)*, pages 124–129, 2002.

Annexe A

Description du projet MIAUCE

Le projet européen MIAUCE (Multi-modal Interactions Analysis and exploration of Users within a Controlled Environment) vise à étudier et développer des techniques pour analyser le comportement multimodal des utilisateurs dans un contexte applicatif. Le comportement multimodal prend la forme de regard fixe, de clignotement de l'oeil ou de mouvements de corps.

Le projet MIAUCE est un projet européen qui regroupe 8 partenaires : CNRS (Lille, France), Université d'Amsterdam (Pays-Bas), Université de Glasgow (Ecosse), Université de Trento (Italie), Université de Namur (Belgique), Sylis (Nantes, France), Visual Tools (Madrid, Espagne) et Tilde (Riga, Lettonie). Les différents partenaires du projet MIAUCE A.1 étudient et développent des techniques qui capturent et analysent le comportement multimodale dans les environnements contrôlés. En raison d'une telle analyse, l'information est adaptée aux besoins d'utilisateur et à la situation donnée. Ils étudient l'utilisation et l'efficacité de leur technique dans trois applications différentes telles que la sécurité, la vente adaptée aux besoins du client, et la TV interactive sur le Web. L'objectif est donc de développer des techniques d'analyse d'interactions de l'homme dans un environnement contrôlé (supermarché, etc.), plutôt que des interactions de l'homme avec un ordinateur.

Les techniques sont alors développées et validées dans trois différents domaines d'applications. Ceci permettra de développer des techniques généralisables et d'ouvrir des voies pour l'exploitation industrielle. Les applications développées seront faites par le biais d'une aide industrielle étant donné que ce projet est réalisé pour des raisons marketing, mais aussi de sécurité. Dans ce contexte, le CRID se concentrera sur les problèmes légaux que de telles technologies impliquent, avec pour objectif d'adresser des recommandations sur le design du système d'information et sa gestion. L'aspect éthique et

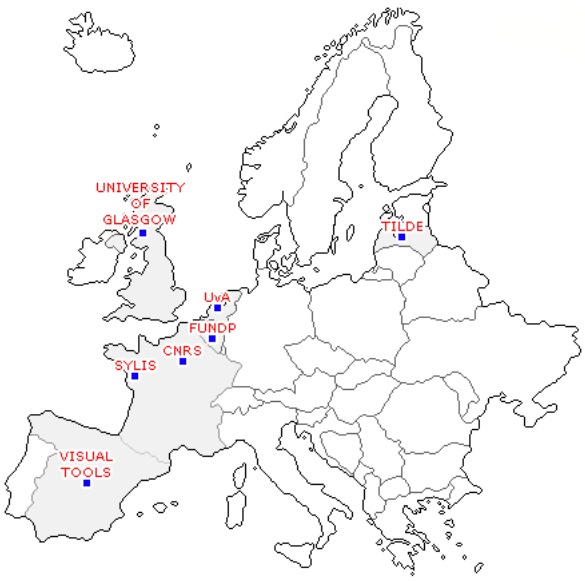


FIGURE A.1: Partenaires du projet MIAUCE.

sociologique sera étudié au sein de la CITA. En conclusion, le projet étudie les conditions légales et morales nécessaires pour concevoir ces nouvelles applications. En outre, l'acceptabilité éthique de ces nouvelles technologies d'intelligences que sont la capture multimodale de comportement, sera étudiée afin de garantir le succès de leur exécution.

L'objectif Principale des 3 partenaires industriels (Sylis, Visual Tools, Tilde) est de valider et expérimenter les technologies développées par les partenaires universitaires dans des applications déployées dans un environnement grandeur nature. Trois tâches qui correspondent à 3 différentes applications sont supportées par des industriels afin d'étudier la généralité de la technologie développée. Les applications ciblées par ce projet sont :

- Sécurité : La détection de personnes qui tombent à la sortie d'un escalier mécanique dans un aéroport.
- Marketing : L'estimation des produits regardé par les personnes qui passent devant une vitrine.
- Interactive WebTV : Proposer du contenu vidéo à une personne qui regarde une WebTV en se basant sur l'émotion.

Annexe B

Protocole d'acquisition des données

B.1 La base SYLIS

Des vidéos ont été capturées de personnes qui passent en face d'une vitrine. Pour cela une caméra a été placée à trois niveaux différents illustrée par la figure B.1

FIGURE B.1: 3 différentes vues prises par les caméras placées dans la vitrine.

Les informations relatives au contenu de ces vidéos est disponible dans le Tableau B.1 :

Caméra ID	Séquence ID	Durée	Nombre d'images
Vue de dessous	cam1-Seq1	4m13	6 337
Vue de dessous	cam1-Seq2	14m07	21 186
Vue de dessus	cam2-Seq1	4m06	6 154
Vue de dessus	cam2-Seq2	10m00	15 000
Vue de dessus	cam2-Seq3	3m46	5 667
Vue de face	cam3-Seq1	4m14	6 373
Vue de face	cam3-Seq2	4m20	6 507
Vue de face	cam3-Seq3	6m56	10 403
Vue de face	cam3-Seq4	2m31	3 794

TABLE B.1: Description des vidéos.

B.2 La base MIAUCE

La base MIAUCE a été créée par SYLIS sous les recommandations du CNRS et de l'université d'Amsterdam. L'acquisition de l'information concernant l'orientation de la tête a été obtenue en effectuant une combinaison d'une méthode de suggestion directionnelle et d'annotation automatique comme présentée dans la Section du Chapitre 4. La description de l'installation du linéaire et du processus d'acquisition des vidéos est la suivante :

- 4 linéaires possédant chacun 5 points.
- Boites de céréales utilisées comme produit du linéaire.
- Papier en couleur fixé sur les boites de céréales pour visualiser les numéros associés aux boites.
- Un point noir sur chaque étiquette qui sert comme point de fixation à être regardé par la personne.
- La longueur du linéaire et de 1m20.
- Position de la caméra1 : entre le second et le quatrième étage.
- Position de la caméra2 : en face du linéaire.
- Arrière plan : Un mur de couleur blanche.
- Participants : 4 personnes (2 hommes et 2 femmes)
- Des points repères sur le sol à chaque position que doit utilisé une personne avec 12 positions prises en compte en fonction de la présence / absence de la personne du champ de la caméra :
  - 4 distances disponibles en face du linéaire (0m50, 1m, 1m50 et 2m).
  - 3 positions (centrale et deux positions orientées vers la droite : à une distance d'environ 50cm)

La synchronisation des données est effectuée à l'aide d'un minuteur utilisé sous PowerPoint pour identifier l'instant de fixation. A chaque trame, on enregistre les informations suivantes :

– La position de la personne entre les positions A et H.

– Le point de fixation qui est compris entre 1 et 20.

– L'identificateur de la personne qui regarde le linéaire ainsi que les informations qui le caractérise.

La Figure B.2 représente l'image entière des 4 personnes qui ont effectuées l'expérimentation se trouvant à la même position et regardant le même point.

FIGURE B.2: 4 personnes ayant participé à l'expérimentation.

La Figure B.3 illustre une personne qui regarde tous les points à une position donnée.

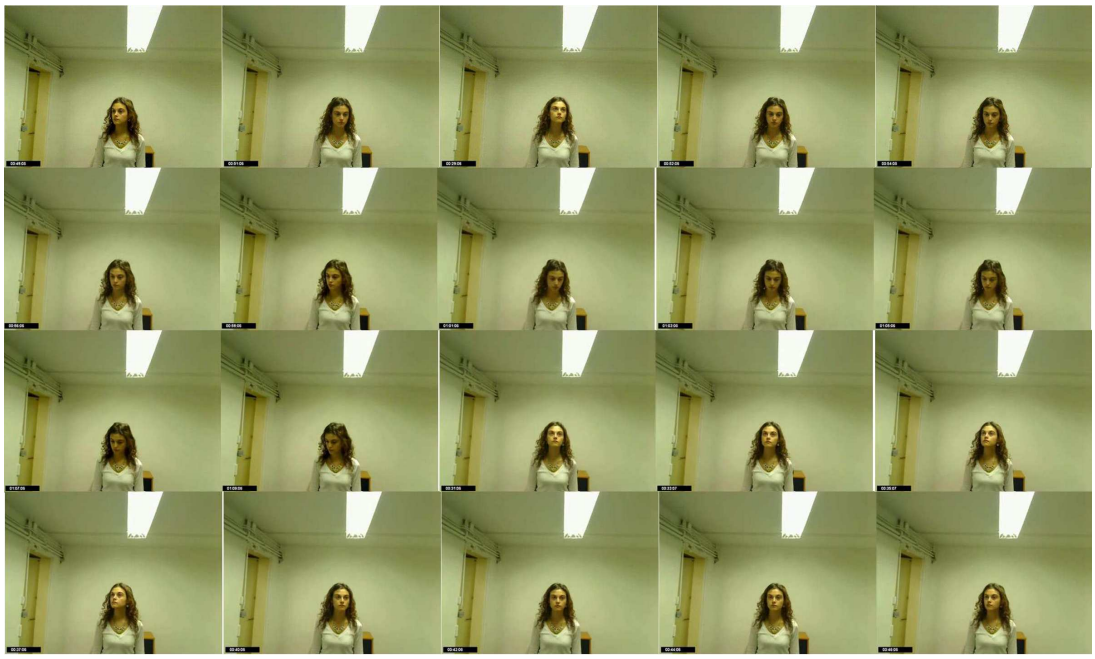


FIGURE B.3: Différents points de regard associés à une personne.

Annexe C

Calcul divers

C.1 Exemple Quaternion

En prenant à titre d'exemple le point A tel que le vecteur $\overrightarrow{OA}$ a pour coordonnées $(-1,1,1)^t$ (voir Figure 5.10) auquel on souhaite faire une rotation d'angle π autour de l'axe (Oy) de $\overrightarrow{OA}$. En appliquant la formule (5.13) les coordonnées de $\overrightarrow{OA'}$ l'image de $\overrightarrow{OA}$ par cette rotation sont données par :

$$\overrightarrow{OA'} = (0, Rot_{[\pi,\vec{j}]}(\overrightarrow{OA})) = (\cos\frac{\pi}{2}, \sin\frac{\pi}{2}.\vec{j}).(0,\overrightarrow{OA}).(\cos\frac{\pi}{2}, -\sin\frac{\pi}{2}.\vec{j})$$

On pose $Q_1 = (cos\frac{\pi}{2}, sin\frac{\pi}{2}.\vec{j})$, $Q_2 = (0,\overrightarrow{OA})$ et $Q_1^* = (cos\frac{\pi}{2}, -sin\frac{\pi}{2}.\vec{j})$

On a donc :

$$Q_1 = (0, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}), \quad Q_2 = (0, \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}) \quad \text{et} \quad Q_1^* = (0, \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix})$$

On est donc amené à calculer le double produit $Q_1.Q_2.Q_3$. Comme écrit précédemment, afin d'établir la correspondance entre quaternions et compositions de rotations vectorielles il faut que les quaternions soient tous normalisés. Q_1 et Q_1^* le sont mais pas Q_2. On calcule le versor $U_{Q_2} = Q_2/p_{Q_2}$ de Q_2 (avec $p_{Q_2} = \|Q_2\|$).

$$\text{On a } p_{Q_2} = \sqrt{3} \text{ et } U_{Q_2} = (0, \begin{pmatrix} -1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{pmatrix})$$

Le produit $Q_1.Q_2.Q_1^*$ est équivalent au produit $p_{Q_2}.Q_1.U_{Q_2}.Q_1^*$. On calcule le premier produit $Q_1.U_{Q_2}$ (on le nomme Q_3).

$$\begin{aligned} Q_3 = Q_1.U_{Q_2} &= (0, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}).(0, \begin{pmatrix} -1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{pmatrix}), \text{ on applique (5.10)} \\ &= (-\tfrac{1}{\sqrt{3}}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \wedge \begin{pmatrix} -1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{pmatrix}) = (-\tfrac{1}{\sqrt{3}}, \begin{pmatrix} 1/\sqrt{3} \\ 0 \\ 1/\sqrt{3} \end{pmatrix}) \end{aligned}$$

On fait le produit $Q_3.Q_1^*$ (on remarque que Q_3 est normalisé).

$$\begin{aligned} Q_3.Q_1^* &= (-\tfrac{1}{\sqrt{3}}, \begin{pmatrix} 1/\sqrt{3} \\ 0 \\ 1/\sqrt{3} \end{pmatrix}).(0, \begin{pmatrix} 0 \\ -1 \\ 0\sqrt{3} \end{pmatrix}), \text{ on applique (5.10)} \\ &= (0, \begin{pmatrix} 0 \\ -1/\sqrt{3} \\ 0 \end{pmatrix} + \begin{pmatrix} -1/\sqrt{3} \\ 0 \\ 1/\sqrt{3} \end{pmatrix} \wedge \begin{pmatrix} 0/\sqrt{3} \\ -1 \\ 0 \end{pmatrix}) \\ &= (0, \begin{pmatrix} 0 \\ 1/\sqrt{3} \\ 0 \end{pmatrix} + \begin{pmatrix} 1/\sqrt{3} \\ 0 \\ -1/\sqrt{3} \end{pmatrix}) = (0, \begin{pmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ -1/\sqrt{3} \end{pmatrix}) \end{aligned}$$

On multiplie le résultat par U_{Q_2}, on obtient :

$$Q_1.Q_2.Q_1^* = (0, \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix})$$

L'image de la rotation d'angle π dirigé par l'axe (Oy) du vecteur $(-1,1,1)^t$ est le vecteur $(1,1,-1)^t$

(voir Figure 5.10).