UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE DE
LA MAÎTRISE EN MATHÉMATIQUES ET INFORMATIQUE
APPLIQUÉES

PAR
YOUSEF AICHOUR

DÉTECTION AUTOMATIQUE DES EXPRESSIONS FACIALES À PARTIR DE SÉQUENCES D'IMAGES POUR L'ÉVALUATION DE L'ÉTAT DES FACULTÉS D'UNE PERSONNE

OCTOBRE 2007

Résumé

Ces travaux de recherche présentent l'intégration complète d'un système de détection des expressions faciales à partir de séquences d'images vidéo par vision numérique. Conçu pour opérer dans un contexte d'évaluation de l'état des facultés d'une personne, le système a pour tâche de détecter automatiquement des expressions faciales associées principalement à la fatigue, et ce, à partir des séquences d'images contenues dans une base de données préalablement construite. Les techniques de reconnaissance sélectionnées utilisent uniquement les composantes anatomiques du visage comme caractéristique discriminante. Le système réalise par ailleurs la reconnaissance des expressions faciales en quatre phases principales, soient l'acquisition des images vidéo, la détection et la localisation du visage, l'extraction de paramètres découlant de la détection du mouvement à l'aide des modèles temporels et, finalement, la détection et la reconnaissance des expressions faciales. Les séquences d'images, obtenues sont pré-traitées par un système de détection et de localisation du visage, et ensuite, ces données seront traitées à l'aide des modèles temporels pour en extraire l'historique de mouvement. Les résultats obtenus sont ensuite acheminés à un module de reconnaissance utilisant des techniques basées sur le calcul des distances à l'aide de l'algorithme des k plus proches voisins (k-ppv). Finalement, la validation des techniques sélectionnées est réalisée à l'aide de la banque des séquences d'images MMI provenant du laboratoire de recherche de Maja Pantic (professeure au département d'informatique du collège royal de Londres) ainsi que

notre propre banque de séquences d'images utilisées pour l'expérimentation.

Remerciements

Je tiens à remercier les personnes suivantes:

Le professeur François Meunier d'avoir accepté d'être mon superviseur de maîtrise, et qui a toujours su me prodiguer des conseils judicieux tout au long de cette maîtrise. Ces études graduées m'ont permis d'acquérir de nombreuses connaissances et représentent sans aucun doute une expérience enrichissante.
Les professeurs Fathallah Nouboud et Ismail Biskri d'avoir accepté de juger ce travail.
Tous les membres du département mathématiques et informatique et plus particulièrement le chef du département Sylvain Delisle et le directeur du programme Mourad Badri.

J'aimerais également souligner le support de ma mère, mes frères, mes soeurs, ainsi que tous les membres de ma famille. Je tiens à remercier tout particulièrement mon frère Mustapha Aichour, qui restera toujours pour moi une très grande source de motivation.

Finalement, j'aimerais remercier les professeurs, le personnel de soutien ainsi que mes collègues de maîtrise, notamment Abdelkader Siddour et Alain Girard, pour ces belles années au laboratoire !

Table des matières

LISTE DES TABLES

LISTE DES FIGURES

Chapitre 1

Introduction

1.1 Contexte et objectif

Depuis quelques années, on observe un besoin croissant pour des systèmes automatiques de détection des expressions faciales. On n'a qu'à penser aux besoins relatifs au développement des systèmes actifs pour alerter un conducteur et surveiller son niveau de vigilance pour tenir les conducteurs réveillés et par conséquent réduire le nombre d'accidents de la route. La détection automatique des expressions faciales d'une personne peut être réalisée à partir des séquences d'images de l'individu, plus particulièrement de son visage. La vision numérique vise ainsi l'acquisition, le traitement et l'interprétation des images d'une séquence pour réaliser la reconnaissance des expressions faciales d'une personne.

1.2 Vision numérique, automatisation et reconnaissance

La vision numérique, ou vision artificielle, est un domaine visant la reproduction de la capacité de perception visuelle de l'oeil humain. Ainsi, des caméras sont utilisées pour observer des scènes qu'elles reproduisent sous la forme d'un signal vidéo ou d'une séquence d'images. Ce domaine inclut également le traitement et l'analyse qui est effectué sur les données brutes. Connue dans les milieux scientifiques depuis de nombreuses années, la vision numérique gagne maintenant le grand public. Pour ce faire, la plupart des ordinateurs vendus aujourd'hui sont équipés d'une caméra permettant l'acquisition d'images, de films ainsi que l'exécution d'applications interactives. Citons entre autres les programmes de communication et les jeux.

1.3 Description du projet

La détection automatique des expressions faciales étant un sujet d'actualité, jumelé avec les défis passionnants de la vision numérique, le développement d'un système complet de détection automatique des expressions faciales représente un projet très intéressant. L'objectif principal concerne la conception d'une application de détection des expressions faciales, en utilisant un montage simple et peu coûteux. Parmi les objectifs supplémentaires, il y a entre autres le choix des algorithmes et méthodes nécessaires pour effectuer les tâches de détection et de reconnaissance. De plus, différentes

expérimentations doivent présenter les niveaux de précision, de robustesse et d'efficacité des techniques sélectionnées.
En résumé, ce projet vise l'exploration des différentes facettes de la vision numérique à partir de l'acquisition des images et de leur traitement, jusqu'à l'interprétation et la reconnaissance des expressions faciales. Il s'agit donc en fait de la conception d'un système complet de détection automatique des expressions faciales.

1.4 Contraintes du projet

Pour réaliser ce projet, plusieurs contraintes ont été établies dès le départ afin de préciser les conditions réelles pour lesquelles l'application devait être conçue. Ces conditions ont permis le design d'un montage convenable ainsi que la sélection des différents algorithmes nécessaires au fonctionnement du système. Tout d'abord, le système de détection et de reconnaissance doit pouvoir observer une scène. Les conditions d'acquisition de chaque séquence d'images obtenue sont contrôlées. Habituellement, la capture des images du visage s'effectuera en vue frontale. Par conséquent, la présence d'un visage dans la scène est assurée, et le positionnement du visage dans la scène est connu a priori. Aucun mouvement principal rigide (mouvement de la caméra ou orientations du visage de la personne) ne doit être produit lors de la capture d'une séquence d'images. Comme contrainte supplémentaire, l'identification doit être réalisée seulement si le sujet observé est face à la caméra tout en se situant à une distance approximativement d'un mètre

devant elle. Concernant le montage physique, il se doit d'être simple et peu coûteux tout en offrant performance, robustesse et efficacité. Cette étape concerne donc le choix de l'équipement informatique et du capteur à utiliser. Aucune contrainte monétaire n'est réellement appliquée au système, mais il se doit d'utiliser des composantes standard et non spécialisées à des coûts raisonnables.

1.5 Description de la solution

Le système de détection automatique des expressions faciales développé comporte plusieurs modules effectuant chacun des tâches bien précises.
La figure 1.1 illustre un diagramme représentant ces différentes étapes.

Figure 1.1: Modules composant le système de détection automatique des expressions faciales.

Tout d'abord, l'acquisition est réalisée à l'aide d'une caméra dont les spécifications techniques sont illustrées dans la table 1.1.

Avec son coût approximatif de 2000$, cette caméra satisfait l'exigence du coût abordable associée au projet. Par la suite, les modèles de représentation de données (modèles temporels) basés sur le mouvement traitent les images brutes fournies au système. Ces modèles, utilisant une technique simple,

Caractéristiques	**Valeurs**
Modèle	Sony XC-EI50
Format du signal	EIA (RS-170)
Type de capteur	CCD
Pixels (H × V)	768 × 494
Débit d'images	jusqu'à 30 FPS
Prix	environ 2000$

Table 1.1: Spécifications techniques de la caméra utilisée pour le système de la détection automatique des expressions faciales.

mais efficace de représentation du mouvement en 2D dans un espace tridimensionnel (les deux dimensions spatiales et la dimension du temps) pour la création des images de l'historique du mouvement (Motion History Images, MHIs).

Finalement, ces images MHIs normalisées sont présentées au module de reconnaissance des expressions faciales. Ce module est basé sur le calcul des distances à l'aide de l'algorithme du k plus proches voisins (k-ppv) permettant de calculer la distance entre un vecteur de caractéristiques correspondant à une expression faciale inconnue avec un ensemble de vecteurs de caractéristiques de référence associée aux expressions faciales à reconnaître. Pour davantage de robustesse, ce module d'identification est basé sur une architecture de classification utilisant des méthodes de reconnaissance variées. Toutes les opérations et calculs nécessaires sont réalisés à partir d'un ordinateur standard munis de composantes courantes et non spécialisées.

Les spécifications de cet équipement sont résumées dans la table 1.2.

Composantes	**Valeurs**
Processeurs	Intel Pentium VI
Vitesse	1.70 GH
Mémoire vive	512 Mo
Carte graphique	ASUS A9250 TD Radeon 9250 128 Mo
Système d'exploitation	Microsoft windows xp pro

Table 1.2: Techniques de l'ordinateur utilisé pour les expérimentations.

1.6 Organisation du mémoire

Ce mémoire est donc organisé comme suit : Tout d'abord, le chapitre 2 portera sur la revue de littérature qui permet de décrire les principales approches de détection des expressions faciales retrouvées dans la littérature, alors que le chapitre 3 portera sur l'analyse et toutes les méthodes appliquées pour la réalisation des tâches de détection et de reconnaissance des expressions faciales. Ensuite, le chapitre 4 présentera le fruit de nos expérimentations dans lesquelles le modèle de détection et de reconnaissance des expressions faciales implémenté est testé avec des séquences provenant de la banque de séquences d'images MMI ainsi que de la banque des séquences d'images constituées dans le cadre de ces expérimentations. Finalement, on finira par une conclusion au chapitre 5.

Chapitre 2

Revue de la littérature

2.1 Introduction

Le visage humain est impliqué dans une variété impressionnante d'activités différentes. Il contient la majorité de notre appareil sensoriel : yeux, oreilles, bouche, et nez, nous permettant de voir, écouter, goûter et sentir. À part ces fonctions biologiques, le visage humain fournit un nombre de signaux essentiels pour la communication interpersonnelle dans notre vie sociale. Le visage comporte plusieurs systèmes qui coopèrent pour produire un ensemble de signaux de communication comme : la parole, le regard, le positionnement et les mouvements de la tête, les expressions faciales, etc.. Ces signaux étant produits essentiellement par le système de production de la parole et le système musculaire facial. Ils sont primordiaux pour déduire l'état affectif et les intentions d'une personne. D'autres informations comme, l'attraction, l'âge et le genre peuvent être aussi dérivées à partir du visage d'une personne. Automatiser les analyses des signaux faciaux, et en particulier les signaux faciaux rapides (actions des muscles faciaux), devrait être fortement bénéfique pour plusieurs champs d'intérêt comme la sécurité, la médecine, la communication, et l'éducation. Dans le contexte de la sécurité, les ex-

pressions faciales jouent un rôle crucial dans l'établissement ou l'évaluation de la crédibilité. En médecine, les expressions faciales fournissent une signification directe permettant d'identifier les processus mentaux spécifiques, par exemple, quand une personne est en état de sommeil. En éducation, les pupilles des auditeurs informent le professeur de la nécessité d'ajuster le message d'instruction. Les interfaces normales entre les êtres humains et les ordinateurs (ordinateurs personnels/ robots/ machines) sont aussi concernées, les expressions faciales fournissent une possibilité pour communiquer des informations de base sur des besoins et des demandes à une machine.

En fait, les analyses automatiques des signaux faciaux rapides semblent avoir une place naturelle dans divers sous-ensembles de systèmes de vision, incluant les outils automatisés pour le regard et reliés aussi à la lecture des lèvres, traitement de la parole bimodal, visage / synthèse de la parole visuelle, et le traitement facial. Certains signaux faciaux (clignement de l'oeil) peuvent être aussi associés à certaines commandes (click de souris) offrant une alternative à des commandes du clavier et souris traditionnelles. Les possibilités humaines à entendre dans les environnements bruyants au moyen de la lecture sur les lèvres sont la base du traitement de la parole bimodal qui peut mener à la réalisation des interfaces de discours robustes. La capacité humaine pour lire les émotions à partir des expressions faciales de quelqu'un est la base de l'analyse des messages faciaux pouvant mener au développement d'interfaces d'extension avec la communication émotive et, alternativement, à obtenir une interaction plus flexible, plus adaptable, et normale entre les hommes et les machines. Bien que les êtres humains soient parfaitement capables d'estimer l'état affectif d'une personne à partir d'une image statique, il y

a assurément plus d'informations sur le comportement facial contenu dans des observations dynamiques de la séquence vidéo d'un visage humain. Par conséquent, nous présentons plusieurs approches à travers ce chapitre qui visent à tirer bénéfice de cette information additionnelle.

2.2 Les expressions faciales

2.2.1 Définition

Tout d'abord, il est important de faire la distinction entre la reconnaissance d'expressions faciales et la reconnaissance d'émotions. Les émotions résultent de plusieurs facteurs et peuvent être révélées par la voix, la posture, les gestes, la direction de regard et les expressions faciales. Par contre, les émotions ne sont pas la seule origine des expressions faciales. En effet, celles-ci peuvent provenir de l'état d'esprit (ex: la réflexion), de l'activité physiologique (la douleur ou la fatigue) et de la communication non verbale (émotion simulée, clignotement de l'oeil, froncement des sourcils). Néanmoins, sept émotions de base correspondent chacune à une expression faciale unique, et ce, quelles que soient l'ethnicité et la culture du sujet, ces émotions sont : la colère, le dégoût, l'étonnement, la joie, le mépris, la peur, et la tristesse. La reconnaissance des expressions faciales consiste à classer les déformations des structures faciales et les mouvements faciaux uniquement à partir des informations visuelles. La reconnaissance des émotions, quant à elle, est une tentative d'interprétation qui requiert une information contextuelle plus complète.

2.2.2 Utilisation des expressions faciales pour la détection de la fatigue

A côté du mouvement de la tête et des yeux, les expressions faciales sont l'une des plus importantes sélections visuelles [1], parmi les sélections visuelles existantes, on trouve le mouvement de paupières, la détection de la direction du regard, le mouvement de la tête, et les expressions faciales. Ainsi les expressions faciales d'une personne en état de fatigue ou bien d'une personne au début de la fatigue sont souvent caractérisées par le traînement des muscles et le bâillement. Le développement des systèmes actifs pour alerter un conducteur et surveiller son niveau de vigilance est très important pour tenir les conducteurs réveillés et par conséquent réduire le nombre d'accidents. Certains efforts ont été rapportés dans la littérature sur le développement des systèmes actifs pour surveiller la fatigue en temps réel, mais la majorité de ces systèmes emploient une seule sélection visuelle ce qui est insuffisant. Le système proposé par Zhiwei Zhu et Qiang Ji [2] peut caractériser le niveau de vigilance d'un conducteur d'une façon non intrusive et en temps réel. L'armée de l'air des Etats-Unis a montré un intérêt en appliquant cette technologie pour surveiller la fatigue d'un pilote, l'administration fédérale des routes aux Etats-Unis a également montrée son intérêt.

- **Diagramme de fonctionnement**

La figure 2.1 présente une vue d'ensemble du système de surveillance de vigilance d'un conducteur dévellopé par Zhiwei et Qiang [1].

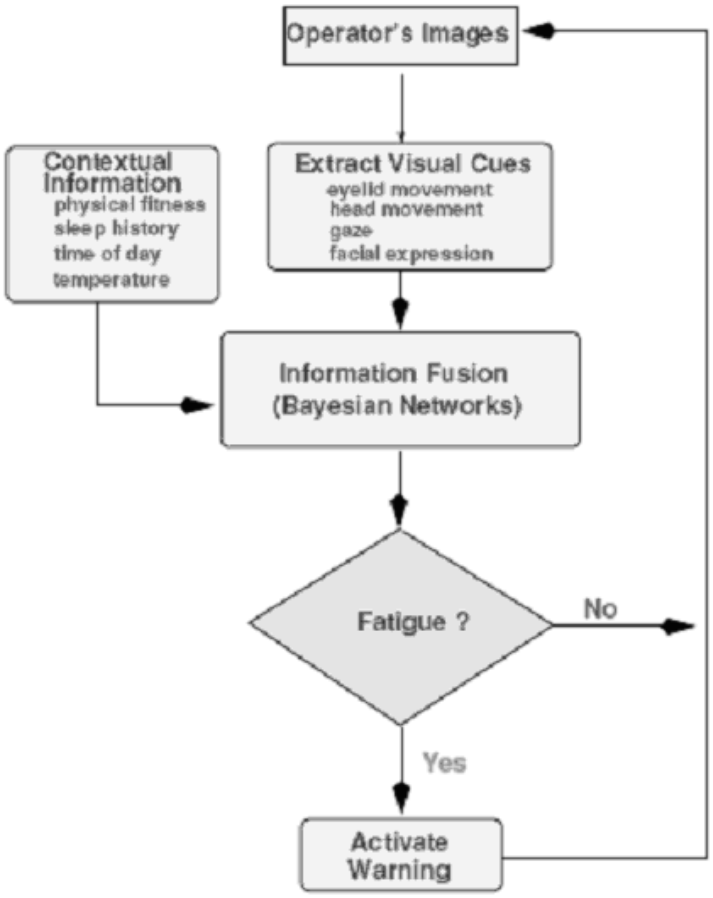


Figure 2.1: Diagramme de fonctionnement d'un système de surveillance de la vigilance d'un conducteur [1].

- **Le matériel nécessaire**

Le système consiste en deux caméras, avec une des caméras munie d'une lentille grand angle fixée en face du visage et une autre munie d'une lentille avec une ouverture étroite suivant les yeux. La caméra grand-angle surveille le mouvement de la tête ainsi que les expressions faciales tandis que l'autre

caméra surveille le regard et les mouvements de paupières.
La figure 2.2 présente une vue d'ensemble de système utilisé.

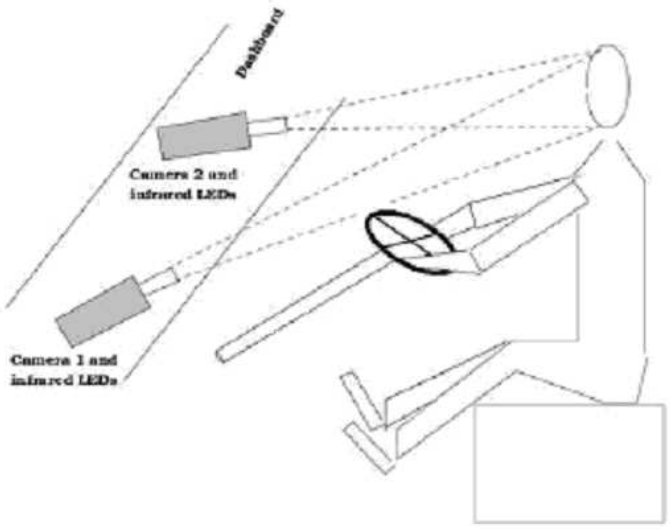


Figure 2.2: Vue d'ensemble d'un système de surveillance de la fatigue d'un conducteur [1].

La figure 2.3 représente les composantes matérielles utilisées.

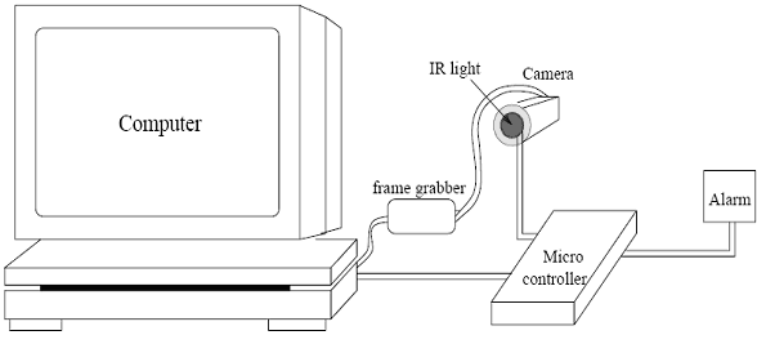


Figure 2.3: Les composantes matérielles utilisées [1].

La figure 2.4 quant à elle présente un essai de l'ensemble du système utilisé pour alerter un conducteur.

Figure 2.4: Essai de l'ensemble du système utilisé [1].

- **Traitement sur les expressions faciales**

Les caractéristiques visuelles utilisées dans ce système sont le mouvement de paupières, la direction du regard, le mouvement de la tête, et les expressions faciales. Les caractéristiques faciales autour des yeux et la bouche représentent les plus importants modèles composant les expressions faciales, ainsi 22 caractéristiques faciales sont détectées en temps réel par le système développé par Zhiwei Zhu et Qiang Ji [1].
La figure 2.5 présente les caractéristiques faciales suivies ainsi que les graphes locaux d'une personne en état de fatigue (baîllement).

(a) (b)

Figure 2.5: Les caractéristiques faciales suivies et les graphes locaux [1] appliqués sur deux images prisent à partir d'une séquence d'images montrant le baîllement. (a) Début de baîllement. (b) Baîllement (après deux secondes de temps).

À partir des points caractéristiques correspondant à chaque dispositif facial et leurs relations spatiales, on peut reconnaître différentes expressions faciales. Les auteurs [1, 3] se basent sur la forme de la bouche pour détecter le baîllement à l'aide du paramètre YAWNFREQ, le YawnFreq est un indicateur de baîllement qui correspond à la présence de baîllements. La fréquence d'ouverture et de fermeture de la bouche (baîllement) est schématisée par une courbe dans le temps. Par conséquent, le baîllement peut être détecté facilement par l'ouverture et la fermeture fréquente de la bouche pour une certaine période de temps.
La figure 2.6 schématise la courbe de la fonction YAWNFREQ durant une période de deux minutes et vingt secondes.

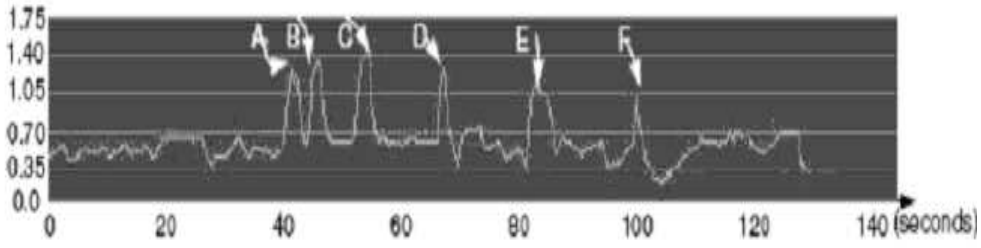


Figure 2.6: Courbe de la fonction YAWNFREQ d'ouverture de la bouche [1].

L'axe horizontal représente la durée de la séquence dans le temps (secondes) et l'axe vertical représente la fréquence d'ouverture de la bouche. Les marqueurs A, B, C, D, E et F indiquent qu'il y a une détection de baîllements. On voit clairement à partir de la figure 2.6 qu'il y a plusieurs détections de baîllements durant seulement soixante secondes entre la quarantième et la centième seconde du test et cela indique que la personne est vraiment en état

de fatigue.

La figure 2.7 montre les résultats obtenus par le système pour une personne en état de fatigue (baîllement).

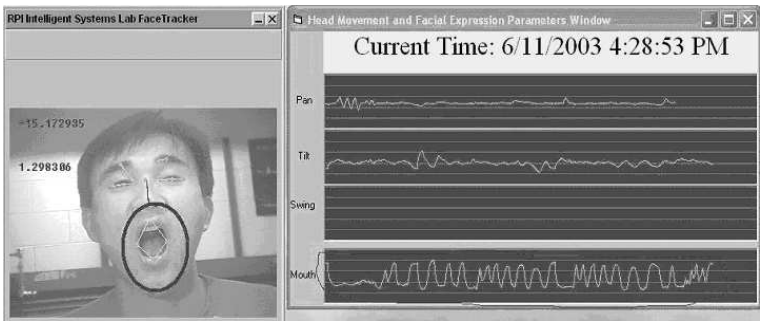


Figure 2.7: Résultats du mouvement de la tête et des expressions faciales obtenus d'une personne en état de fatigue [1].

On voit plusieurs détections de baîllement durant toute la période de test et cela indique encore une fois que la personne est en état de fatigue.

2.2.3 Analyse automatique des expressions faciales

Le but recherché ici est l'implémentation d'un système qui peut effectuer les analyses automatisées des expressions faciales. En général, trois étapes doivent être exécutées pour résoudre ce problème. Premièrement, avant qu'une expression faciale soit analysée, le visage doit d'abord être détecté dans une scène. Ensuite, les expressions faciales sont extraites des séquences d'images. Ce processus correspond à l'extraction du visage et de ses caractéristiques dans une scène. À ce point, une distinction claire sera faite entre

deux termes, nommés, caractéristiques faciales et modèle de caractéristiques faciales. Les caractéristiques faciales sont les sourcils, les yeux, le nez, la bouche, et le menton. Le modèle de caractéristiques faciales est présenté sous forme d'une combinaison de caractéristiques utilisées pour représenter le visage, qui peut se représenter de plusieurs façons, soient en unité totale (représentation holistique), sous forme d'un ensemble de caractéristiques (représentation analytique), ou en une combinaison de ces dernières (approche hybride). La dernière étape consiste à définir les catégories qu'on voudra utiliser pour la classification des expressions faciales et/ou l'interprétation des expressions faciales, et de diviser le mécanisme de catégorisation. Depuis le milieu des années 70, différentes approches ont été proposées pour l'analyse des expressions faciales des images faciales statiques ou des séquences d'images. Cette section discute les trois problèmes reliés au processus d'analyse des expressions faciales qui sont la détection du visage, l'extraction des expressions faciales, et la classification de ces dernières. Une exploration et une comparaison des approches d'analyse automatique des expressions faciales développées récemment sont aussi présentées (table 2.1).

La table 2.1 représente les caractéristiques d'un analyseur automatique des expressions faciales.

Caractéristiques générales	
1	**Acquisition automatique des images faciales**
2	**Sujets de tout âge et appartenance ethnique**
3	**Traitements des visages partiellement occlus**
4	**Aucuns marqueurs spéciaux / marquage requis**
5	**Traitements des mouvements rigides de la tête**
6	**Détection automatique du visage**
7	**Extraction automatique de données des expressions faciales**
8	**Traitement de données imprécis des expressions faciales**
9	**Classification automatique des expressions faciales**
10	**Distinction de toutes les expressions possibles**
Applications de recherche sur le comportement facial	
11	**Distinction des 44 actions faciales ([5])**
12	**Quantification des codes des actions faciales**
Applications multi-modales des médias des interfaces homme/machine (IHM)	
13	**Interprétation des catégories illimitées**
14	**Facilité d'apprentissage des caractéristiques adaptatives**
15	**Assignations des étiquettes d'interprétation quantifiées**
16	**Assignations des étiquettes d'interprétation multiples**
17	**Traitement temps réel des caractéristiques faciales**

Table 2.1: Propriétés d'un analyseur idéal [4].

La table 2.2 représente une synthèse des plus récentes approches pour l'analyse automatique des expressions faciales des séquences d'images.

Références	**Caractéristiques d'un analyseur automatisé idéal des expressions faciales (Table 2.1)**														
	1	**2**	**3**	**5**	**6**	**7**	**8**	**9**	**10**	**14**	**16**	**17**	**18**	**19**	**20**
Analyses des séquences d'images faciales															
Black[6]	•	-	•	•	•	×	•	×	•	×	6	×	•	×	×
Cohn[7]	•	3	×	•	×	×	•	×	•	15	-	×	×	×	-
Essa[8]	•	•	•	•	-	•	•	•	•	2	4	×	×	×	•
Kimura[9]	•	×	•	•	×	•	•	•	•	×	3	×	•	×	-
Otsuka[10]	•	-	-	•	•	-	•	×	•	×	6	×	×	×	×
Wang[11]	•	1	×	•	×	×	•	-	•	×	3	×	•	×	×

Table 2.2: Méthodes récentes d'analyse des expressions faciales [4].

Légende: • = "oui", × = "non", - = entrée absente

2.2.4 Détection du visage

Dans la plupart des travaux en analyse des expressions faciales [8, 18, 19, 23], etc., les conditions d'acquisition de chaque séquence d'images sont contrôlées. Habituellement, le visage est capté sous forme d'une vue frontale. Par conséquent, la présence d'un visage dans la scène est assurée, et la position du visage dans la scène est aussi connue a priori. Cependant, la détermination de l'endroit exact du visage dans une image faciale digitalisée est un problème

plus complexe. D'abord, la dimension et l'orientation du visage peuvent changer d'une image à une autre. Si les images sont captées avec une caméra fixe, les visages peuvent être captés dans les images à des tailles diverses et des orientations différentes dues aux mouvements de la personne observée. Ainsi, il est difficile de rechercher un modèle fixe dans l'image. La présence du bruit et d'occlusions rend le problème bien plus difficile. On croit généralement que les images à deux niveaux de gris de 100 à 200 pixels forment une limite inférieure pour la détection d'un visage par un observateur humain. Une autre caractéristique du système visuel humain est qu'un visage est perçu dans son ensemble, pas comme une collection des caractéristiques faciales. La présence de ces caractéristiques et leur rapport géométrique réciproque semblent être plus importants que les détails de ces caractéristiques.

La table 2.3 nous fournit une classification des analyseurs d'expressions faciales selon le genre d'images d'entrée et la méthode appliquée.

Approche	Référence	Vue	Méthode	Description
Holistique	Huang [15]	Frontale	Le modèle PDM d'ajustement (détection des bords)	Rotations de tête non rigides
	Pantic [18]	Duelle	Analyses des histogrammes d'images	Caméra montée sur la tête du sujet
Analytique	Hara [13]	Frontale	Distribution de luminance	Pas de mouvements rigides de la tête (processus temps réel)
	Yoneyama [19]	Frontale	-	-
	Kimura [9]	Frontale	Projection intégrale [22]. (Ajustement de la fonction d'énergie)	Rotations de tête non rigides

Table 2.3: Résumé des méthodes pour la détection automatique de visage [4].

- Pour représenter le visage, C.L. Huang et Y.M. Huang [15] (table 2.3) appliquent un modèle statistique de distribution des points (Points Distribution Model, PDM). Afin de réaliser un placement correct d'un PDM initial dans une image d'entrée, Huang et Huang utilisent un détecteur d'arêtes pour obtenir une évaluation grossière de la position du visage dans l'image. La vallée dans la fonction de luminance située entre les lèvres et les deux bordures verticales symétriques représentant les frontières verticales externes du visage découle d'une évaluation grossière de l'endroit de cette dernière. Le visage ne doit pas être couvert de cheveux et de lunettes. La tête doit être statique et les variations d'illumination doivent être linéaires pour que le système fonctionne correctement.

- Pantic et Rothkrantz [18] (table 2.3) détectent aussi le visage comme une unité totale. Ils utilisent des paires d'images faciales en entrée : une de face et une de profil (Figure 2.9). Pour déterminer les frontières externes verticales et horizontales de la tête, ils analysent l'histogramme vertical et horizontal de l'image de vue frontale proposé dans [18]. Pour localiser le contour du visage, ils utilisent un algorithme de détection et de localisation du contour de visage basé sur le modèle de couleur HSV, qui est similaire à l'algorithme basé sur le modèle relatif RGB. Ainsi, ils utilisent des images en vue de profil pour faciliter la détection automatique des expressions faciales en cas de mouvement de la tête durant le traitement. Pour les images en vue de profil, ils appellent un algorithme de détection de profil, qui représente une approche spatiale pour prélever le contour de profil d'une image seuillée. Pour le seuillage

de l'image de profil en entrée, la valeur du seuil découlant du modèle de couleur HSV est exploitée. Les cheveux dans le visage et les lunettes ne sont pas permis non plus.

- Kobayashi et Hara [13] (table 2.3) appliquent une approche analytique pour la détection du visage. Ils ont utilisé une caméra CCD en mode monochrome pour obtenir une distribution en niveau de brillance du visage humain. Premièrement, la distribution des niveaux de brillance du visage de 10 sujets est obtenue. Ensuite, le système extrait la position des iris en utilisant une technique de corrélation, cette dernière tente de déterminer la position dans une image d'un visage à partir de modèles de base où la corrélation entre les modèles d'iris et des régions dans l'image du visage est maximale. Une fois que les iris sont identifiés, la position globale du visage est déterminée en employant la position relative des caractéristiques faciales dans le visage. Le sujet observé doit faire face à la caméra tout en se situant à la distance approximative d'un mètre devant elle.

- Yoneyama et al. [19] (table 2.3) emploient aussi une approche analytique pour la détection de visages. Les coins externes des yeux, la taille des yeux, et la taille de la bouche sont extraits d'une manière automatique. Une fois que ces caractéristiques sont identifiées, la taille du domaine facial examiné est normalisée et une grille rectangulaire de 8 par 10 est placée au-dessus de l'image [19]. Il n'est pas énoncé quelle méthode a été appliquée et aucune limitation de la méthode utilisée n'a été rapportée par Yoneyama et al.

- Kimmura et Yachida [9] (table 2.3) proposent des méthodes automatiques pour extraire des points de caractéristiques du visage à partir des images de couleurs normales. La méthode proposée inclut la localisation du visage, la position des caractéristiques du visage, les contours de ces caractéristiques, et les séquences de points de ces dernières. Pour extraire robustement le contour d'une caractéristique de visage, ils ont proposé des modèles de contour actif, qui emploient n contours pour résoudre le problème des méthodes de contours originales une fois appliquées au problème contenant les contours et ils proposent une nouvelle fonction d'énergie pour ces méthodes.

2.2.5 Extraction de données d'expressions faciales

Après avoir localisé un visage dans une image, la prochaine étape est d'extraire les informations sur l'expression faciale produite d'une manière automatique. Un analyseur d'expressions faciales entièrement automatique doit alors être développé. La représentation du visage et le genre d'images d'entrée affectent le choix de l'approche utilisée pour l'extraction des informations sur les expressions faciales. Un des objectifs fondamentaux de l'analyse des expressions faciales est la représentation de l'information visuelle qu'un visage testé peut contenir [24]. Les résultats de Johansson [26, 27] ont donné un indice de l'importance de ce problème. Les expériences de l'extraction des informations sur les expressions faciales suggèrent que les propriétés du visage, concernant les expressions faciales pourraient être obtenues en décrivant les mouvements des points de contrôle associés aux caractéristiques faciales

(sourcils, yeux, et bouche) et en analysant les rapports entre ces mouvements. Ceci a poussé les chercheurs sur l'analyse de visages à faire différentes tentatives pour définir les propriétés visuelles des ensembles de points faciaux permettant la modélisation des expressions faciales. Indépendamment du genre d'images d'entrée, les images faciales ou images arbitraires, la détection de la position du visage d'une image observée ou d'une séquence d'images a été approchée de deux manières. Dans l'approche analytique, le visage est localisé premièrement par la détection de certaines caractéristiques importantes du visage (ex : yeux, bouche, front). Dans l'approche holistique, le visage est considéré comme une unité totale. Diverses représentations analytiques du visage ont été rapportées, dans lesquelles le visage est modélisé comme un ensemble de points faciaux (par exemple : Figures 2.8 [13] et 2.9 [17]) ou comme un ensemble de gabarits adaptés aux caractéristiques faciales telles que les yeux et la bouche.

La figure 2.8 illustre les points caractéristiques du visage selon H. Kobayashi et F. Hara [13], (approche analytique).

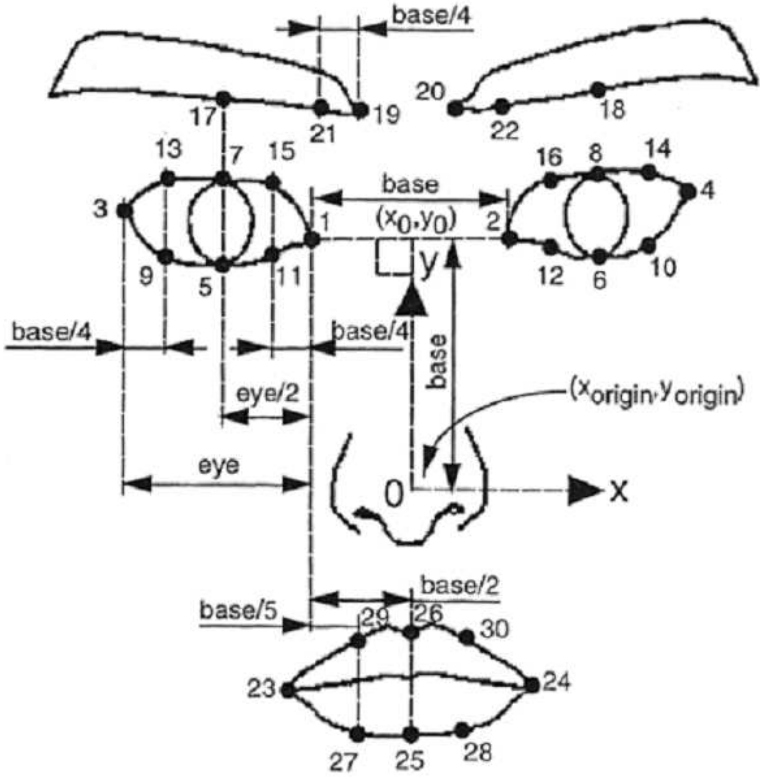
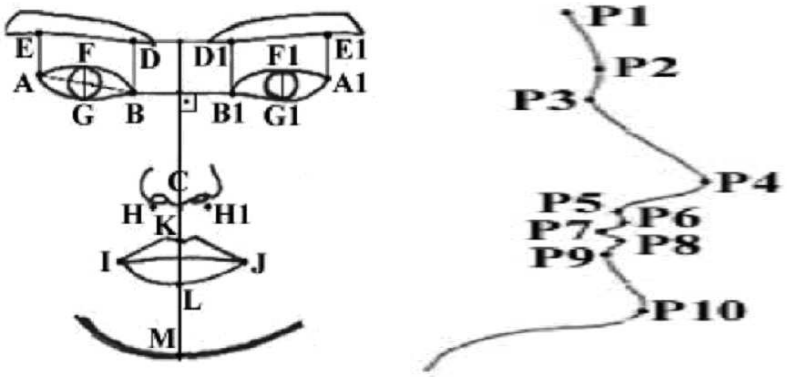


Figure 2.8: Points caractéristiques du visage [13].

La figure 2.9 représente une vue frontale et une vue de côté de l'ensemble des points du visage (approche analytique).

Une fenêtre 3D avec une texture tracée et un modèle spatio-temporel de mouvement du visage d'une image sont des exemples d'approches holistiques typiques pour la représentation de visage.



(a) (b)

Figure 2.9: Points faciaux [18]. (a) Modèle de visage à vue frontale. (b) Modèle de visage en vue de côté.

La figure 2.10 représente le modèle planaire pour représenter des mouvements rigides du visage et le modèle affine de courbure pour représenter des mouvements non rigide de visage (approche holistique).

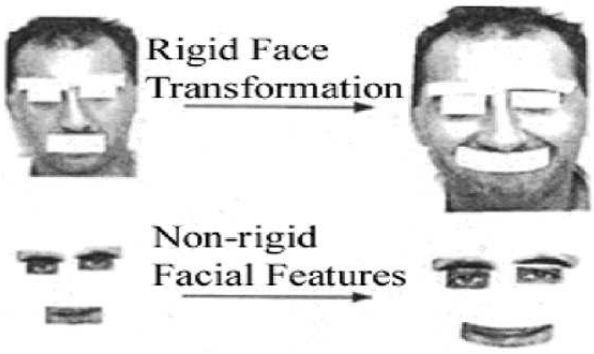


Figure 2.10: Modèle planaire pour représenter les mouvements faciaux rigides et le modèle affine de courbure pour représenter les mouvements faciaux non rigides [6].

Le visage peut aussi être modélisé en utilisant une approche hybride, ce qui caractérise une combinaison de l'approche analytique et holistique pour la représentation du visage. Dans cette approche, un ensemble de points faciaux sont habituellement utilisés pour déterminer la position initiale d'un modulateur du visage. Le système qui utilise cette approche est proposé par Kimura et Yachida [9]. Ils utilisent la fonction d'énergie pour adapter cette fonction à une image faciale normale. Ils calculent d'abord le contour de l'image en appliquant un filtre différentiel. Ensuite, afin d'extraire la force externe, qui correspond au gradient de contour dans l'image, ils appliquent un filtre gaussien. L'image filtrée est désignée sous le nom d'un champ énergétique.

La figure 2.11 représente la fonction d'énergie et le champ énergétique correspondant (approche hybride).

(a) (b)

Figure 2.11: La fonction d'énergie et son champ énergétique correspondant [9]. (a) La fonction d'énergie. (b) Le champ énergétique.

Indépendamment du genre de visage, le modèle est appliqué, des tentatives doivent être faites pour modeler et ensuite extraire les informations sur l'expression faciale montrée en perdant peu ou beaucoup de cette information. Plusieurs facteurs rendent cette tâche complexe. Le premier est la présence des cheveux, les lunettes, etc., ce qui cache les expressions faciales. Un autre problème est la variation au niveau de la taille et l'orientation du visage lors de la capture d'une séquence d'images. Ceci rend difficile la recherche des modèles fixes dans les images. Finalement, le bruit et l'occlusion sont toujours présents dans une certaine mesure.

2.2.6 Détection du mouvement

Afin de limiter le volume important de calculs et de traitements nécessaires à la segmentation, à la détection et à la reconnaissance des expressions faciales, il est avantageux d'utiliser des techniques de pré-filtrage qui restreindront l'espace de recherche. Parmi ces méthodes, la détection du mouvement demeure l'une des plus efficaces. La détection du mouvement, réalisée immédiatement après l'acquisition d'une image, représente une étape très avantageuse pour un système de vision numérique. En effet, un gain de performance considérable peut être réalisé lorsque des zones sans intérêt sont éliminées avant les phases d'analyse. Cette amélioration dépend cependant de la complexité des algorithmes de détection et de reconnaissance utilisés. Le problème de la détection du mouvement dans une séquence d'images consiste à séparer dans chaque image de la séquence les zones en mouvement des zones statiques. A chaque instant, chaque pixel doit ainsi être étiqueté par un identifiant binaire fixe/mobile. Lorsque la caméra est fixe, on peut effectuer une telle détection à partir des différences temporelles calculées pour chaque pixel dans une séquence d'images. Dans cette section, plusieurs méthodes de détection du mouvement par vision numérique seront présentées. Pour celles-ci, la performance varie, autant en ce qui à trait aux temps de traitement, qu'à la qualité des résultats produits.

Différences entre deux images consécutives

Étant peu complexe, la différence entre deux images consécutives [51] représente une solution très intéressante. Comme son nom l'indique, elle consiste à soustraire une image acquise au temps t_n d'une autre au temps t_{n+k}, où k est habituellement égal à 1. Ainsi, l'image résultante sera vide si aucun mouvement ne s'est produit pendant l'intervalle du temps observé car l'intensité et la couleur des pixels seront presque identiques. Par contre, si du mouvement a lieu dans le champ de vision, la différence d'images correspondant aux pixels frontières des objets en déplacement devraient changer drastiquement de valeurs, révélant alors la présence d'activité dans la scène. Cette technique nécessite très peu de ressources, car aucun modèle n'est nécessaire. Cela implique qu'il n'y a pas de phase d'initialisation obligatoire avec une scène statique, ce qui procure une très grande flexibilité d'utilisation. De plus, une opération de soustraction d'images requiert très peu de puissance de calcul, lui conférant un avantage supplémentaire. Par ailleurs, les résultats obtenus avec cette méthode ne sont pas aussi éloquents que ceux générés en utilisant un modèle statistique de l'arrière-plan. En effet, certains traitements supplémentaires sont nécessaires afin de déterminer la zone en mouvement, car l'information disponible ne concerne que les contours des régions en déplacement (ce qui inclus également les zones intérieures d'un objet).

Flux optique

Similaire à l'approche précédente, l'utilisation du flux optique procure une information du mouvement pour chaque pixel de l'image. Ainsi, il mesure les vecteurs du déplacement à partir de l'intensité des pixels de deux images consécutives ou temporellement rapprochées. Dans un contexte de détection du mouvement, les pixels inactifs posséderont une vélocité nulle contrairement aux pixels appartenant à des objets dynamiques. Une classification sous forme de regroupement est donc nécessaire afin d'isoler et de localiser les zones représentant du mouvement. Cette technique a notamment été utilisée pour la détection de piétons [48]. Il y a finalement plusieurs méthodes pour calculer le flux optique, mentionnons entre autres celle de Lucas et Kanade [49] ainsi que celle de Horn et Schunck [50]. L'inconvénient majeur de l'utilisation du flux optique est la somme importante de calculs à réaliser pour l'estimation du mouvement. Par ailleurs, une variante utilisant l'appariement par bloc (Block Matching) peut bénéficier de certaines instructions optimisées MMXTM, ce qui peut accélérer le traitement global. L'appariement par bloc est utilisé dans plusieurs algorithmes de compression vidéo pour la prédiction du mouvement. Il a donc fait l'objet de recherches intensives afin d'optimiser son exécution. Néanmoins, une tâche supplémentaire de classification et d'interprétation est nécessaire. De plus, si certaines parties d'un objet ne sont pas en mouvement, elles seront complètement ignorées par cette méthode. Ce pourrait être le cas par exemple d'une séquence vidéo contenant une personne assise par terre et agitant les bras. Dans cette situation bien précise, le corps de la personne ne serait pas détecté contrairement à ses bras.

Image de l'historique du mouvement

A.F. Bobick et J.W. Davis [51] utilisent la représentation MHI (Motion History Image) comme une base des histogrammes de mouvement. Ainsi, ils génèrent le mouvement entre fenêtres en faisant la différence successive des images binaires de la silhouette du visage de la personne (section 3.3.4). La raison de ceci est double. Premièrement, on croit que les méthodes à flux optique strict sont toujours trop fragiles pour l'imagerie réelle du mouvement des caractéristiques du visage (dû au bruit, textures, et le taux de mouvement). La différenciation entre images continue à être une méthode assez robuste pour localiser facilement la présence du mouvement. Un des problèmes principaux de la différentiation entre images est qu'on ne peut pas indiquer la magnitude ou la direction du mouvement mais seulement sa présence. Ainsi, il est difficile d'enlever le mouvement non désiré purement originaire du résultat de différentiation. Mais comme on montrera plus tard, l'accumulation des différences d'images peut rapporter une information directionnelle du mouvement. La deuxième raison de la différenciation des silhouettes correspond au fait que la texture fréquente au niveau du visage et le mouvement non rigide de la tête signalent un mouvement non désiré, ce qui peut causer des problèmes quand on utilise le mouvement pour la reconnaissance. Pour cette raison on choisit d'extraire la forme de la silhouette de la personne. Un effet secondaire d'utiliser des silhouettes est qu'aucun mouvement à l'intérieur de silhouette ne peut être vu. Donc, les différences d'images (l'union des différences aux résolutions normales et basses) montrent seulement le mouvement de la frontière des silhouettes, mais elles rapportent

toujours l'information de mouvement tout à fait utile pour divers types de mouvements.

2.2.7 Modèles de représentation de données

La représentation de données facilite la détection automatique des expressions faciales dans une séquence d'images représentant le visage humain. Les modèles temporels présentés par A.F. Bobick et J.W. Davis [51] sont une des méthodes de représentation de données. Ils servent à représenter le mouvement en 2D dans un espace tridimensionnel (les deux dimensions spatiales et la dimension du temps).
La figure 2.12 représente l'application des modèles temporels par M. Pantic, et I. Patras [53] pour la création des images de l'historique du mouvement. A.F. Bobick et J.W. Davis [51] proposent une représentation spatio-

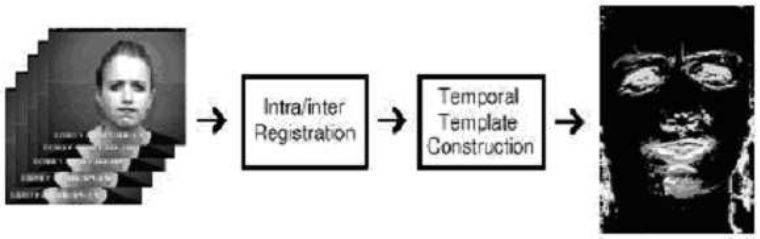


Figure 2.12: Création de l'image MHI à partir d'une séquence d'images [53].

temporelle d'activités humaines. Ils définissent les images d'énergie de mouvement et les images de l'historique du mouvement. Les premières, binaires, représentent la position du mouvement dans une séquence d'images. Dans les secondes, les pixels sont des valeurs correspondant à l'âge du mouvement.

Ils sont calculés par un simple remplacement et un opérateur de décalage. Si D(x,y) est un booléen indiquant si l'intensité lumineuse du point (x,y) à l'instant t a changé depuis l'instant t-1, alors, les pixels de l'image H de l'historique du mouvement à l'instant t sont définis par :

$$\mathbf{H}_\tau(\mathbf{x},\mathbf{y},\mathbf{t}) = \begin{cases} \tau & D(x,y,t)=1 \\ Max([H_\tau(x,y,t-1)-1],0) & Autrement \end{cases} \tag{2.1}$$

Dans cette équation, τ est l'âge considéré dans l'historique. L'historique au point (x,y) est affecté à la valeur τ si un mouvement est détecté en (x,y) au temps t, c'est-à-dire si D(x,y,t)=1. Sinon, l'âge du mouvement augmente et la valeur de H(x,y,t) est décrémentée. L'âge du mouvement en (x,y,t) est en fait déterminé par :

$$|H_\tau(x,y,t)-\tau| \tag{2.2}$$

L'image E d'énergie du mouvement est alors définie par le seuillage de l'image H à zéro, c'est-à-dire :

$$E_\tau(x,y,t) = \bigcup_{i=0}^{\tau} D(x,y,t-i) = \begin{cases} 1 \ si \ H_\tau(x,y,t)>0 \\ 0 \qquad\qquad sinon \end{cases} \tag{2.3}$$

Puisqu'on n'utilise pas les images d'énergie du mouvement (Motion Energy Images, MEIs) dans ce travail, on s'intéresse seulement à la construction des images d'historique du mouvement (Motion History Images, MHIs). M. Pantic, et I. Patras [53] ont fait des études sur les émotions produites par le visage humain. Dans la définition de leur problème il n'a pas été révélé quand le mouvement d'intérêt débute (début de l'émotion à détecter) ou se

termine (fin de l'émotion à détecter). Donc ils doivent varier la période τ observée et essayer de classifier toutes les MHIs résultantes. Dans le cadre de nos expérimentations, le début et la fin d'une expression faciale sont connus et coïncident avec la durée d'une séquence d'images, nous n'avons pas besoin de varier τ. Pour cette raison, ils sont capables de normaliser le comportement temporel en distribuant les valeurs de niveau de gris dans l'image MHI sur la gamme complète de leur image de sortie (0-255, supposons qu'ils utilisent les images de niveau de gris de 8 bits). Ainsi, les variations de la durée d'affichage d'une unité d'action sont annulées. Cela rend possible la comparaison des expressions faciales qui ont différentes périodes mais qui sont autrement identiques. Initialement les séquences d'images peuvent avoir différents nombres de fenêtres. Ainsi, quand les images d'historique de mouvement (MHIs) sont temporellement normalisées, le nombre de niveaux historiques dans ces images diffère d'une séquence d'images à une autre. Pour être capable de comparer les séquences correctement, on aura besoin de créer toutes les MHIs contenants le même nombre fixé de niveaux historiques n_h. Donc la séquence d'images est prélevée à n_{h+1} fenêtres. Le nombre des niveaux d'historiques est expérimentalement déterminé pour devenir le nombre permettant de construire les MHIs, donnant le taux de reconnaissance le plus élevé. En utilisant le paramètre connu n_h, on modifie l'opérateur MHI de la formule 3.1 comme suit:

$$\mathbf{H}(\mathbf{x},\mathbf{y},\mathbf{t}) = \begin{cases} s * t & D(x,y,t) = 1 \\ H(x,y,t-1) & Autrement \end{cases} \tag{2.4}$$

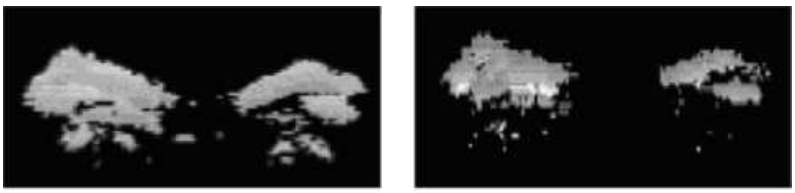
Où s=(255/n) est le saut d'intensité entre deux niveaux historiques et

H(x,y,t)=0 pour t≤0 .

Les figures 2.13 et 2.14 illustrent quelques résultats du modèle temporel.

(a) (b)

Figure 2.13: Image de sourire et son image MHI correspondante [53]. (a) Fenêtre montrant l'émotion sourire prise à partir de la séquence d'images de l'émotion sourire. (b) L'image MHI construite à partir de la séquence d'images de l'émotion sourire et contenant la fenêtre présentée en (a).



(a) (b)

Figure 2.14: Image MHI résultante d'une séquence vidéo représentant un mouvement des sourcils. (a) Relèvement des sourcils. (b) Abaissement des sourcils.

2.2.8 Classification des expressions faciales

Après que le visage et son aspect ont été perçus, la prochaine étape d'un analyseur automatisé d'expressions faciales est de classifier (identifier) cette expression donnée par le visage. Une problématique fondamentale de la classification des expressions faciales est de définir un ensemble de catégories que nous voulons traiter. Une autre problématique qui en découle est de concevoir des mécanismes de catégorisation. Des expressions faciales peuvent être classifiées de plusieurs façons en termes des actions faciales qui causent une expression, en termes de certaines expressions sans prototype (ex. : fronts augmentés) ou en termes de certaines expressions à prototype (expressions émotives). Il y a plusieurs approches pour la reconnaissance des changements du visage humain linguistiquement universelles basées sur l'activité musculaire du visage [6, 7, 8, 18]. Parmi ces approches, on trouve le système de codage d'actions faciales (Facial Action Coding System, FACS)

proposé par Ekman et al [5], qui est la meilleure approche connue et utilisée. C'est un système désigné pour les observateurs humains pour décrire les changements dans l'expression faciale en termes d'activations visuellement observables des muscles du visage humain.

Système de codage d'actions faciales

Les signaux rapides du visage humain sont les mouvements des muscles du visage qui tirent la peau, entraînant une déformation provisoire de la forme des caractéristiques faciales (yeux, bouche, nez, front) et de l'aspect des plis, des sillons, et des bombements de la peau. La terminologie commune pour décrire les signaux rapides du visage se réfère l'un ou l'autre aux limites linguistiques culturellement dépendantes indiquant un changement spécifique dans l'apparaissance d'une caractéristique faciale particulière (sourire, sourire affecté, froncement des sourcils, ricanement, etc.). Le système de codage d'actions faciales (FACS) est probablement l'étude la plus connue sur l'activité faciale. C'est un système qui a été développé pour faciliter la mesure objective de l'activité faciale pour des investigations comportementales de la science sur le visage. FACS est conçu pour les observateurs humains pour détecter les changements subtils indépendants de l'aspect facial provoqué par des contractions des muscles faciaux. Les changements au niveau de l'expression faciale sont décrits avec les FACS en termes de 44 unités d'action (Action Units, AU) différentes, qui sont anatomiquement liées à la contraction d'un muscle spécifique du visage ou d'un ensemble de muscles.
La figure 2.15 illustre quelques images des unités d'actions faciales ainsi que

la description de chaque unité d'action faciale présentées dans [23].

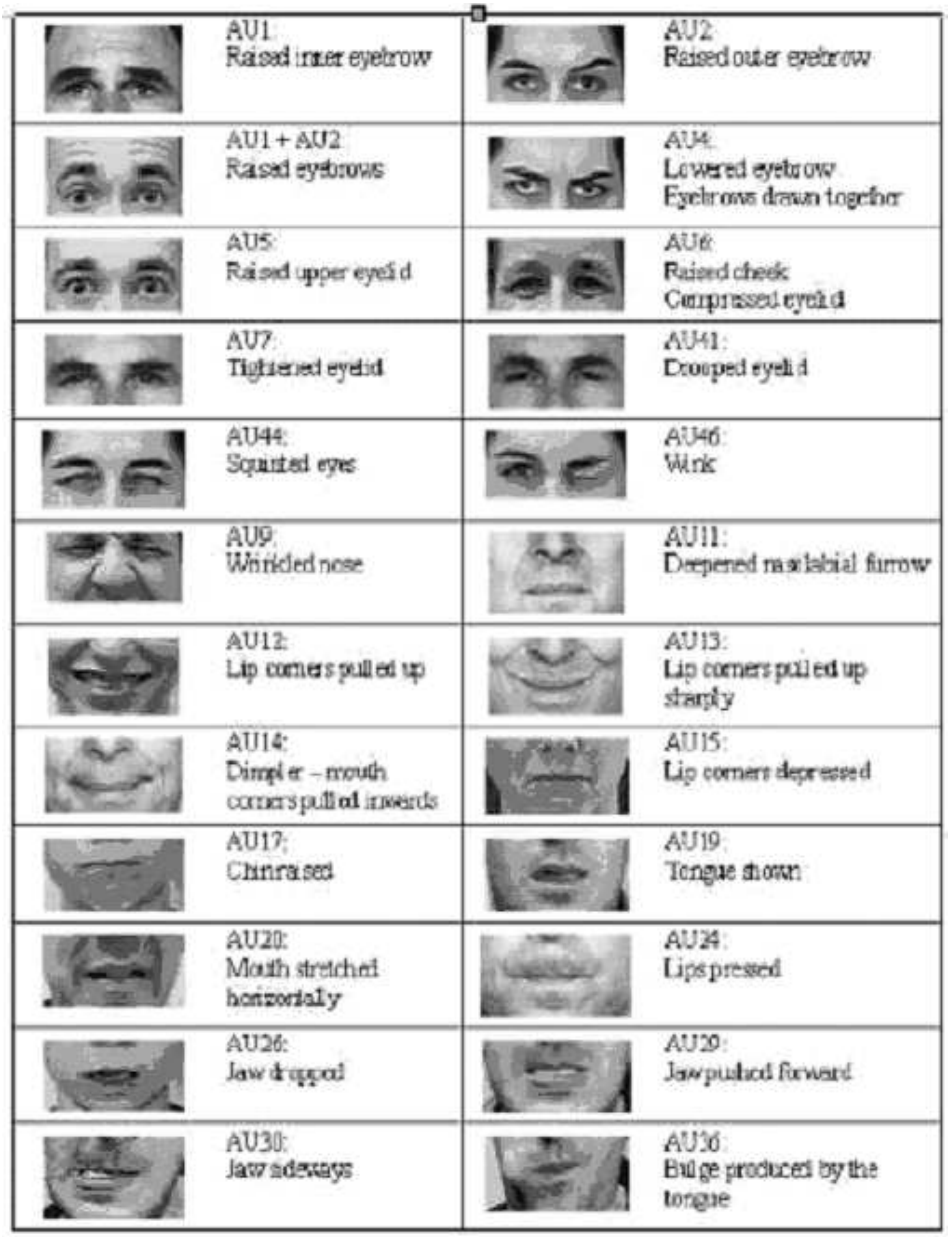


Figure 2.15: Quelques exemples des unités d'actions faciales [23].

Avec la définition des différentes unités d'action (AUs), les codeurs FACS (Facial Action Coding Systems) fournissent aussi les règles permettant la détection visuelle des unités d'action et leurs segments temporels (début, milieu, et fin de l'unité d'action) d'une image faciale. En utilisant ces règles, un codeur FACS décompose une expression faciale fournie en plusieurs unités d'action qui produisent l'expression faciale. Bien que les FACS fournissent une bonne fondation pour le codage des unités d'actions des images faciales par les observateurs humains, réaliser la reconnaissance des unités d'actions par un ordinateur n'est pas une tâche triviale. Le problème majeur de cette méthode est que les unités d'action peuvent se produire dans plus de 7000 combinaisons complexes différentes entraînant des bombements (bombement produit par la langue (AU36) comme la figure 2.15 le démontre) et divers mouvements d'entrée et de sortie des images planes de caractéristiques faciales permanents qui sont difficiles à détecter dans les images faciales à deux dimensions.

Approches utilisant FACS

Peu d'approches ont été rapportées pour l'identification automatique des unités d'actions dans les images représentant le visage. Quelques chercheurs ont décrit les modèles de mouvement facial qui correspondent à quelques unités d'actions spécifiques, mais n'ont pas rendu compte de l'identification actuelle de ces unités d'actions. Des exemples de tels travaux sont des études de Mase [28], Black et Yacoob [6], et Essa et Pentland [8]. Presque tous les autres efforts qui utilisent le codage de FACS ont adressé le problème de

reconnaissance automatique des unités d'actions dans une séquence vidéo représentant un visage humain.

Pour détecter six unités d'actions individuelles dans une séquence vidéo représentant le visage et sans mouvement de tête, Bartlett et al [29] ont utilisé un réseau de neurones de type 61×10×6 alimenté vers l'avant. Ils ont réalisé 91% de réussite en alimentant le réseau convenablement avec les résultats d'un système hybride combinant l'analyse holistique spatiale et le flux optique avec l'analyse des caractéristiques locales.

Pour reconnaître huit unités d'actions individuelles et quatre combinaisons des unités d'actions dans des séquences vidéo représentant le visage humain sans prendre en considération le mouvement de tête, Donato et al [30] ont utilisé une représentation d'ondelette de Gabor (Gabor wavelet) et la méthode d'analyse de composante indépendante (Independent Component Analysis, ICA). Ils ont rapporté 95.5% comme taux moyen de reconnaissance des unités d'actions.

Pour reconnaître huit unités d'actions individuelles et sept combinaisons des unités d'actions dans des séquences d'images vidéo représentant le visage et sans prendre en considération le mouvement de tête, Cohn et al [31] ont utilisé pour leur part le suivi des points de caractéristiques faciaux et l'analyse par fonction discriminante. Ils ont réalisé un taux moyen de reconnaissance de 85 %.

Tian et al [32] ont utilisé une approche pour le suivi des lèvres, un modèle de suivi (template matching), et des réseaux neutres (neutral networks) pour reconnaître 16 unités d'actions individuelles ou en combinaison à partir d'images extraites des séquences vidéo acquises en vue frontale. Ils ont rap-

porté 87.9 % comme taux moyen de reconnaissance.

Braathen et al [33] ont rapporté la reconnaissance automatique de trois unités d'actions utilisant le filtrage de particules appelé aussi les chaînes de Markov de Monte-Carlo [40, 41, 42] pour le suivi des séquences d'images en 3D, les filtres de Gabor [46] (Gabor Filters), les machines de vecteur de soutien (Support Vector Machines) [43, 44], et les modèles cachés de Markov [45] (Hidden Markov Models) pour analyser une séquence d'images d'entrée n'ayant aucune restriction placée sur la position de tête.

Ces approches pour la détection automatique des unités d'actions, traitent seulement des images de visage humain en vue frontale et ne peuvent pas manipuler la dynamique temporelle des unités d'actions. Pantic et Patras [34] ont adressé le problème de la détection automatique des unités d'actions et leur segments temporels (début, milieu, et fin de l'unité d'action) pour les séquences d'images à partir d'une vue de profil (profile-view). Ils ont utilisé un filtrage de particules pour le suivi de 15 points caractéristiques du visage provenant de séquences d'images vidéo représentant le visage en vue de profil et des règles temporelles pour effectuer la segmentation automatique et la reconnaissance des segments temporels de 23 unités d'actions apparaissant seules ou en combinaison dans la séquence vidéo. Ils ont réalisé 88 % de reconnaissance.

Le seul travail rapporté jusqu'à maintenant qui s'adresse au codage automatique des unités d'actions pour une séquence d'images représentant la tête statique est celui de Pantic et Rothkrantz [23]. Il s'agit d'un système automatisé pour la reconnaissance des unités d'actions dans les images couleur en vue frontale et/ou de profil représentant le visage sans mouvement de

tête. Une approche multi détecteurs [47] (approche qui effectue la détection au niveau d'un visage en vue frontale et en vue de profil) pour la localisation des caractéristiques du visage est utilisée spécialement pour prélever le contour d'un visage sous forme d'une vue de profil et les contours des composantes du visage comme les yeux et la bouche. Pour les contours des caractéristiques du visage, le contour de 10 points caractéristiques du visage sous forme d'une vue de profil et 19 points caractéristiques des contours des composantes du visage sont extraits. Basées sur ces derniers, 32 unités d'actions individuelles apparaissant seules ou en combinaison sont reconnues utilisant le raisonnement basé sur un ensemble de règles. Avec chaque unité d'actions marquée, l'algorithme utilisé associe un facteur dénotant une certitude avec laquelle l'unité d'actions pertinente est marquée. Un taux de reconnaissance de 86 % est réalisé avec cette méthode.

En résumé, quatre essais critiques dans l'automatisation du codage des FACS peuvent être distingués. La première concerne le traitement avec les variations du positionnement de la tête. La majorité des systèmes ont des capacités limitées pour surmonter les problèmes causés par les variations de la position de la tête. Xiao et al [35], ont signalé une méthode pour recouvrir le mouvement complet (trois rotations et trois translations) de la tête à partir d'une séquence d'images d'entrée utilisant un modèle de tête cylindrique. En utilisant les paramètres du mouvement recouverts, la région du visage peut être stabilisée et traitée par un analyseur d'expressions faciales. Le deuxième essai concerne les aspects temporels des expressions faciales. La synchronisation des différentes unités d'actions, la vitesse d'activation d'une unité d'action, et la détection du début, du milieu, et de la fin de l'émotion,

sont seulement trois exemples des aspects temporels de l'action faciale. Les chercheurs en codage automatique des FACS ont commencé à poursuivre ces essais. Troisièmement, l'occlusion du visage pose souvent un problème pour le codage automatique des FACS. Les barbes, moustaches et lunettes compliquent la détection des unités d'actions. L'importance de ces occlusions varie pour chaque système mais la plupart des publications ne mentionnent pas à quel point le système proposé fait face à ce problème particulier. Seulement Essa et Pentland [8] rapportent un système qui est capable de manipuler les occlusions causées par les cheveux dans le visage (barbe, moustache) et les lunettes.
La table 2.4 représente quelques méthodes sur la classification des expressions faciales en termes des actions faciales extraites de séquences d'images représentant le visage humain.

Comme indiqué par Fridlund et al. [37], l'étude la plus connue et généralement la plus utilisée sur la classification émotive des expressions faciales est l'étude sur l'existence des catégories universelles des expressions émotives. Ekman [5] a défini six catégories, désignées sous le nom des émotions de base : bonheur, tristesse, surprise, crainte, colère, et dégoût. Il a décrit chaque émotion de base en termes d'expressions faciales qui caractérisent uniquement cette émotion. Ces dernières années, beaucoup de questions ont été posées autour de cette étude. Pour un certain nombre de raisons il est difficile d'automatiser la classification émotionnelle des expressions faciales. Premièrement, la description d'Ekman pour les six modèles des expressions faciales de l'émotion est ambiguë. Il n'y a aucune description uniquement définie en termes d'actions faciales ou en termes de quelques autres codes faci-

Référence	Méthode	Nombre d'actions faciales	Étude de cas	Réussite
Méthodes basées sur les modèles				
Cohn [7]	Fonctions discriminantes	Combinaison de 15 unités d'actions	504 séquences et 100 sujets	88%
Essa[8]	Modèles spatiotemporels d'énergie de mouvement	Combinaison de 2 unités actions	22 séquences et 8 sujets	100%
Méthodes basées sur les règles				
Black [6]	Paramètres de mouvement seuillés	-	70 séquences et 8 sujets	88 %

Table 2.4: Classification des expressions faciales en termes d'actions faciales [4].

aux universellement reconnus. Par conséquent, la validation et la vérification du schéma de classification à employer sont des tâches difficiles et cruciales. Deuxièmement, la classification des expressions faciales dans de multiples catégories d'émotions devrait être faisable, par exemple les sourcils augmentés et le sourire est un mélange de surprise et de bonheur (Figure. 2.16).

A ce jour, il n'y a aucun examen psychologique minutieux en cette matière. Trois problématiques supplémentaires sont liées à la classification d'expressions faciales en général. Premièrement, le système devrait être capable d'analyser n'importe quel sujet, mâle ou femelle de n'importe quel âge et appartenance ethnique. En d'autres termes, le mécanisme de classification peut ne pas dépendre de la variabilité physionomique de la personne ob-

(a) (b)

Figure 2.16: Expressions d'un mélange d'émotions (sourire et surprise) de la même personne [4]. (a) Émotion surprise. (b) Émotion sourire.

servée. D'autre part, chaque personne a sa propre façon d'exprimer une expression faciale particulière. Par conséquent, si la classification obtenue doit être mesurée, les systèmes qui peuvent commencer par une classification générique d'expressions et ensuite s'adapter à un individu particulier ont un avantage certain. Par contre, l'information sur le contexte de chaque expression faciale est très difficile à obtenir de façon automatique. Finalement, il y a maintenant une croissance au niveau de la recherche psychologique qui avance le fait que la synchronisation des expressions faciales est un facteur critique dans l'interprétation des expressions faciales [26, 27, 39].

k plus proche voisin et métriques de distance

Le k plus proche voisin : Pour un bon nombre de techniques, l'algorithme du k plus proches voisins (k-ppv) [55, 56] est utilisé lors de la phase de reconnaissance. Le k-ppv se base tout simplement sur une liste ordonnée des voisins les plus près d'une image test. Pour ce faire, le vecteur test est comparé avec chacun des prototypes constituant la banque d'apprentissage. Pour chacune de ces comparaisons, une distance est calculée pour évaluer la ressemblance avec un certain prototype. Cette distance peut être mesurée par différentes métriques, dont voici une liste partielle que nous présenterons dans la section suivante.

Les métriques de distance : L'algorithme du k plus proches voisins (k-ppv) permet par un simple test utilisant une métrique de distance, [56] de compter quels k échantillons sont les plus proches de l'échantillon à classer. Il détermine ensuite l'étiquette majoritaire des voisins les plus proches pour décider la classe de l'échantillon de test. Les paramètres d'intérêt sont la métrique de distance étant utilisée et k, le nombre de voisins à considérer. Indépendamment de l'expertise du domaine, il n'y a aucune façon de déterminer quelle métrique de distance à utiliser ou quelle valeur de k on devrait prendre. Les métriques de distance les plus utilisées sont les suivantes :

- **Métrique de City-block (L1) :** La métrique de distance L1 consiste à calculer la somme des différences absolues entre les éléments des

vecteurs, soit la fonction suivante :

$$L_1 = |U - V| = \sum_{i=1}^{d} |U_i - V_i| \tag{2.5}$$

Où U et V représentent les vecteurs à comparer et d la taille des vecteurs.

- **Métrique Euclidienne (L2) :** La métrique de distance L2 ou métrique de distance euclidienne entre deux vecteurs consiste à calculer la racine de la somme des différences au carré, soit :

$$L_2 = \|U - V\| = \sqrt{\sum_{i=1}^{d} (U_i - V_i)^2} \tag{2.6}$$

Où U et V représentent les vecteurs à comparer et d la taille des vecteurs.

- **Métrique de Minkowski :**

$$L_m(a, b) = (\sum_{i=1}^{d} |a_i - b_i|^m) \tag{2.7}$$

Où la distance euclidienne connue (Norme L2, m=2) et la distance de Manhattan ou de bloc de ville (city block, m=1) (Norme L1) sont des cas spéciaux. La distance de Manhattan est la somme des distances de projections des points sur un ensemble des axes perpendiculaires prédéfinis.

- **Métrique de Tanimoto :** On trouve souvent cette métrique dans le domaine de la taxonomie et est définie comme suit :

$$D_{Tonimoto}(S_1, S_2) = \frac{n_1 + n_2 + 2n_{12}}{n_1 + n_2 - n_{12}} \tag{2.8}$$

Où n_1 et n_2 sont les nombres d'éléments dans les ensembles S_1 et S_2 respectivement, et n_{12} est le nombre d'éléments qui se trouvent dans les deux ensembles (n_{12}=n_1+n_2).

- **Métrique de Chamfer :** Cette métrique de distance permet d'évaluer la distance entre vecteurs de dimensions différentes ce qui est très utile avec la représentation de données d'images d'historique de mouvement à niveaux multiples MMHI. Cette métrique est définie comme suit :

$$D_{Chamfer}(a,b) = \sum_{i=1}^{l} min_k|a_i - b_k| + \sum_{i=1}^{m} min_k|b_i - a_k| \tag{2.9}$$

Où a et b sont des ensembles avec les cardinalités l et m, respectivement. Une fois appelées pour les images de l'historique de mouvement à niveaux multiples (Multilevel Motion History Images), les entités dans les ensembles a et b indiquent quels niveaux d'historiques sont actifs.

Une fois toutes les distances mesurées par rapport à chaque prototype d'expressions faciales, une liste ordonnée croissante est générée afin de départager les candidats. Habituellement, k se voit assigner une valeur de 1, ce qui signifie que le prototype de la banque d'apprentissage le plus proche est sélectionné. Si k est supérieur à 1, le prototype qui possède la majorité de votes parmi les k plus proches voisins sera choisi.

2.3 Conclusion

Cette revue de littérature nous a permis d'expliquer succinctement comment détecter automatiquement les expressions faciales d'une personne, plusieurs

approches ont été introduites, ces approches vont nous servir dans le cadre de ce mémoire à atteindre notre objectif qui sert à la détection automatique des expressions faciales à partir de séquences d'images pour l'évaluation de l'état des facultés d'une personne. L'analyse des expressions faciales est un problème intrigant que les humains résolvent avec une assez grande facilité. Nous avons identifié trois aspects importants de ce problème : la détection du visage, l'extraction des informations associées à chaque expression faciale, et la classification d'expressions faciales. La possibilité du système visuel humain pour la résolution de ces problèmes a été discutée. Elle devrait servir comme un point de référence à n'importe quel système de vision automatique essayant de réaliser les mêmes fonctionnalités. Parmi ces problèmes, la classification d'expressions faciales a été la plus étudiée, due à son utilité dans les domaines d'application de l'interprétation du comportement humain et les interfaces homme-machine. Cependant, la plupart des systèmes examinés sont basés sur des images représentant le visage sous forme de vues frontales et sans cheveux et lunettes, qui sont peu réalistes dans ces domaines d'application. En outre, toutes les approches proposées pour l'analyse automatique d'expressions effectuent seulement la classification d'expressions faciales dans les catégories de base d'émotion définies par Ekman et Friesen [5]. Néanmoins, ces approches sont peu réalistes puisqu'il n'est pas du tout certain que toutes les expressions faciales pouvant apparaître sur le visage puissent être classifiées dans les six catégories d'émotions de base (dégoût, colère, sommeil, sourire, surprise, tristesse).

Finalement, notre travail étudie la possibilité de détecter automatiquement quelques expressions faciales standards (Baîllement, Colère, Sommeil, Sourire,

Surprise) en utilisant les modèles de représentation de données (modèles temporels) présentés par A.F. Bobick, J.W. Davis [52], M. Pantic, et I. Patras [53].

Chapitre 3

Analyse et méthodes appliquées

3.1 Introduction

Le but de ce travail est d'investiguer de nouvelles combinaisons de représentation de données et de classification pour la recherche d'une solution qui facilite la détection de plusieurs unités d'actions (Action Units, AUs) menant à la détection des différentes expressions faciales. Pour atteindre ce but, nous appliquerons plusieurs méthodes, à savoir les méthodes de détection et localisation du visage [58, 59, 60, 63, 64], les modèles temporels d'A.F. Bobick et J.W Davis [51, 52] pour la détection du mouvement, et la méthode du k plus proches voisins (k-ppv) [55] pour la classification et la reconnaissance de ces expressions détectées. Les modèles temporels [51, 65] sont des images 2D construites à partir des séquences d'images, qui permettent de réduire effectivement un espace spatiotemporel 3D en une représentation 2D. L'idée est d'éliminer une dimension tout en préservant l'information temporelle. Les endroits où le mouvement s'est produit dans une séquence d'images d'entrée sont dépeints dans l'image 2D du mouvement. Pour être capable de construire les modèles temporels il faut que la caméra et le background soient statiques et/ou le mouvement de l'objet d'intérêt soit séparable du mouve-

ment introduit par la caméra et par des mouvements du fond. Pour la phase de classification de données nous appliquerons l'algorithme de classification du k plus proches voisins (k-ppv) [55]. C'est l'algorithme le plus connu et le plus utilisé comparativement à d'autres classifieurs et aussi il réalise une meilleure classification. Dans ce chapitre nous allons premièrement décrire tous les éléments nécessaires à la réalisation de notre projet. Ensuite nous effectuerons une analyse complète de toutes les méthodes retenues et appliquées et nous finirons par adapter ces méthodes pour obtenir une meilleure détection automatique des différentes expressions faciales en question.

3.2 Détection et normalisation du visage

3.2.1 Matériel nécessaire

Pour pouvoir réaliser ce projet un ensemble de pièces matérielles est nécessaire. Le système qui permet de préparer les séquences d'images consiste en une caméra vidéo munie d'une lentille grand angle fixée en face du visage. La caméra est reliée à un ordinateur qui permet d'effectuer l'enregistrement des séquences d'images. Les caractéristiques de ces pièces matérielles utilisées sont présentées à la section 1.5.

3.2.2 Séquences d'images

Bases de données des séquences d'images de visages

Dans un contexte de reconnaissance d'individu, la quantité d'informations requise peut être considérable. En effet, un système de détection automatique et de reconnaissance des expressions faciales doit en pratique identifier les émotions d'un grand nombre de visages différents, ce qui nécessite plusieurs séquences d'images pour chaque personne. Pour développer et évaluer les applications d'analyse de visages, des tests sur de grandes collections de séquences d'images sont requis. Puisque les enregistrements de séquences d'images du visage en mouvement sont nécessaires pour étudier la dynamique temporelle des expressions faciales, les images statiques sont quant a elles importantes pour obtenir de l'information sur la configuration des expressions faciales qui sont essentielles, donc les séquences vidéo du visage et aussi les images statiques sont importantes dans le processus de classification d'expressions faciales. Plusieurs chercheurs [68, 69, 70, 71] dans le domaine de l'analyse de machine d'effet facial se sont intéressés à des expressions faciales d'émotions. D'autres chercheurs [23, 34, 33, 32, 30], dans le développement de machine d'analyse sur les actions des muscles faciaux se sont intéressés à des expressions faciales produites par l'activation d'un seul muscle facial (unité d'action) ou par l'activation d'une combinaison des unités d'action. Donc les deux genres de formation et de test (activation d'un seul muscle et activation d'une combinaison des unités d'action) sont demandés. Dans une image faciale en vue frontale, les unités d'action faciales sont clairement observables comme les changements de l'aspect des yeux et

des sourcils. Par contre les actions faciales comme par exemple montrer la langue ou pousser la mâchoire représentent des mouvements non rigides qui sont difficiles à détecter. Des bases de données de visages sont rendues publiquement disponibles, mais pas encore utilisées par toutes les communautés de recherche sur les expressions faciales, se sont la base de données PIE [72] contenant les images de visages, la base de données AR [73], et la base de données JAFFE [74]. Ainsi, toutes ces bases de donnés contiennent seulement des images statiques et captées en vue frontale, aucune de ces bases de données ne contient des séquences vidéo en plus des images faciales. Nous présentons deux types de bases de données contenants en plus des images statiques, les séquences vidéo. De plus, la base de données construite dans le cadre de ce mémoire et basée sur ces deux bases de données est aussi présentée.

MMI La base de données MMI [75] contient les images faciales ainsi que les séquences vidéo montrant les visages humains sous forme de vue frontale. Elle est très connue et utilisée, elle consiste à plus de 1200 vidéos et 600 images statiques de 31 adultes âgés entre 18 et 35, 50 % femelles, 81 % sont des Caucasiens, 14 % Asiatiques et 5 % des africains. Presque toutes les vidéos sont enregistrées sous forme de vue frontale et sept seulement sont enregistrées sous forme d'une vue de profil en utilisant un miroir.

Cohn Kanade La base de données Cohn Kanade [36] contient plus de 2000 vidéos de 210 adultes âgés entre 18 et 50 ans, 69 % femelles, 81 % caucasiens, 13 % africains et 6 % des autres groupes ethniques. Seulement des expressions

réelles ont été enregistrées, ce qui signifie que jamais plusieurs unités d'actions n'apparaissent seules. Parmi toutes les bases de données citées auparavant, Cohn Kanade est la plus utilisée dans le domaine de recherche sur les analyses des expressions faciales automatisées.

Base de données utilisée dans l'expérimentation La base de données utilisée pour l'expérimentation contient plus de 150 échantillons des séquences d'images de visages en vue frontale et montrant les cinq expressions faciales citées auparavant. La base de données inclut huit visages différents des adultes (étudiants ainsi que le directeur du staff de recherche) âgés entre 25 et 45 ans. Parmi les huit adultes choisis, 25 % sont des femelles, 50 % sont des américains, 50 % africains.

Préparation des séquences d'images

Les séquences d'images utilisées pour le développement du programme proviennent de notre laboratoire de recherche. Chaque séquence représente une émotion à la fois. Ainsi, la durée de chaque séquence est la durée d'une émotion (Bâillement, Sourire, Sommeil, Colère, ou Surprise), en moyenne la durée d'une séquence d'images est de 3.5 secondes.

La table 3.1 illustre la durée moyenne de chaque émotion prise après plusieurs tests sur les séquences.

Émotion	Durée Moyenne (secondes)
Baîllement	4.3
Colère	3
Sommeil	4.5
Sourire	2.7
Surprise	2.8

Table 3.1: Durée moyenne de chaque émotion parmi les cinq émotions mise en test.

Une fois qu'une séquence est prête pour la mise en test, elle sera mise dans une base de données. La numérisation des séquences vidéo se fait à l'aide de notre programme. Ainsi, le taux de numérisation est de 30 images par seconde pour toutes les séquences.

Exigences sur les séquences

Pour la plupart des travaux en analyse des expressions faciales [23, 32], les conditions d'acquisition de chaque séquence d'images obtenue sont contrôlées. Habituellement, la capture des séquences d'images du visage s'effectuera en vue frontale. Par conséquent, la présence d'un visage dans la scène est assurée, et le positionnement du visage dans la scène est connu a priori. Pour que le système fonctionne correctement un ensemble de conditions doivent

être appliquées sur les séquences d'images.

- La séquence d'images doit être compressée à l'aide d'un logiciel de compression des séquences d'images.
- Le visage doit être sans cheveux (moustache, barbe) et lunettes.
- Le visage doit être en position de face par rapport à la caméra.
- Aucun mouvement principal rigide ne doit être produit.
- Les variations d'illumination doivent être linéaires.
- Le sujet observé devrait faire face à la caméra tout en se situant à une distance approximativement d'un mètre devant elle.
- Aucune variation au niveau de la taille et l'orientation de visage ne doit survenir lors de la capture d'une séquence d'images.

3.2.3 Protocole de découpage

Le découpage en plan d'une séquence vidéo consiste à donner le lieu et la nature de la transition entre deux plans. De nombreux travaux découpent automatiquement une vidéo en faisant appel à des algorithmes adaptés à la nature de la transition entre plans pour des séquences vidéo compressées [78] ou non [76, 77]. Chaque plan identifié est représenté par une image fixe caractéristique qui résume ce plan. Selon la durée d'une vidéo et le rythme du montage, le nombre d'images peut être élevé et un résumé vidéo exhaustif

serait de taille importante. On estime qu'il y a environ 30 images par seconde. D'où pour une séquence vidéo de 3.5 secondes nous nous retrouvons avec 105 images à traiter.

La figure 3.1 représente quelques images captées à partir d'une séquence vidéo.

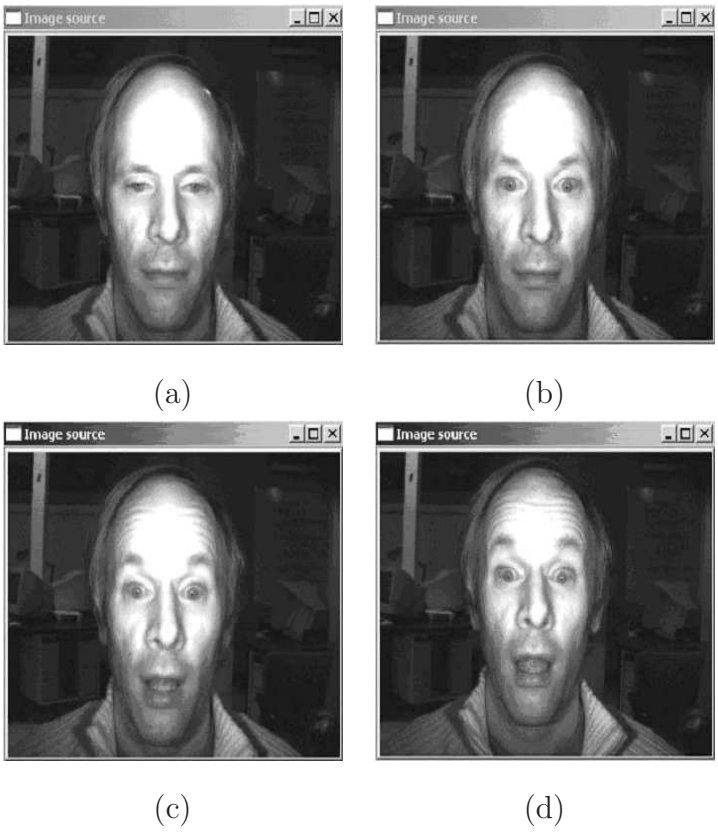


Figure 3.1: Quelques images prises à partir de la séquence représentant l'émotion surprise. (a) À l'instant t_0=0. (b) À l'instant t_1=0.6 seconde. (c) À l'instant t_2=1.2 seconde. (d) À l'instant t_3=2 secondes.

3.2.4 Structure du programme

Langage utilisé

L'architecture logicielle a été conçue avant tout afin de simplifier l'intégration des différentes méthodes de reconnaissance tout en demeurant la plus flexible et versatile possible. Le programme est développé en langage C. C'est un langage de programmation structurel impératif et aussi un des langages les plus utilisés dans le domaine de traitement d'images et de la vision par ordinateur pour plusieurs raisons :

- Il est très facile à porter d'une machine à une autre (le compilateur est écrit en C : il y a juste à changer les routines de génération de code et à le faire se compiler lui-même pour une machine cible).
- Il n'est pas très éloigné de la machine.
- Il est très puissant au niveau de l'analyse d'images et de vision par ordinateur.
- Il inclut la bibliothèque de traitement d'images et de vision par ordinateur OpenCV que nous allons aborder par la suite.

OpenCV

Définition OpenCV (Open Source Computer Library) est une bibliothèque gratuite d'analyse d'images et de vision par ordinateur, en langage C/C++,

proposée par Intel pour Windows et Linux. C'est une collection de fonctions en langage C et peu de classes de C++ qui mettent en application quelques algorithmes populaires du traitement d'images et de vision par ordinateur.

Opérateurs classiques OpenCV comprend un très grand nombre d'opérateurs, parmi lesquels :

- Création/libération d'images, macros d'accès rapides aux pixels.
- Opérateurs standards (morphologie, filtres dérivatifs, filtres de contours, suppression de fond, recherche de coins).
- Recherche, manipulation, traitement de contours.
- Pyramides d'images.
- Dessins de primitives géométriques (lignes, rectangles, ellipses, polygones... et même du texte).
- Création et utilisation d'histogrammes.
- Changement d'espaces de couleurs (RGB, HSV, L*a*b* et YCrCb).
- Interface Utilisateur (lecture/écriture d'images de type ≪ JPEG, PPM,-...≫, affichage à l'écran, gestion des signaux sur un clic de fenêtre, fermeture, ...).
- Lecture des séquences vidéo et découpage de celles-ci en plusieurs images (environ 30 par seconde).

Opérateurs complexes Comme pour les opérateurs classiques OpenCV comprend aussi un nombre élevé d'opérateurs complexes, parmi lesquels :

- Contours actifs: Modèle qu'on nomme aussi snakes (serpents) en raison des déformations subies pendant le processus d'adaptation qui s'apparentent au mouvement d'un serpent.
- Analyse de mouvement : Flux optique et MHI.
- Détection de visages.
- Calibrage d'une caméra (possible à partir d'un échiquier).
- Suivi d'objets 3D avec plusieurs caméras.
- Mise en correspondance entre deux images.
- Lecture d'images à la volée directement depuis une vidéo AVI ou une caméra (Windows seulement).

Fonctions OpenCV nécessaires Les fonctions OpenCV suivantes sont les plus intéressantes que notre programme utilise :

- **Modèles du mouvement**
  - **Mise à jour de l'historique du mouvement (cvUpdateMotionHistory)** Cette fonction permet la mise à jour de l'historique du mouvement en déplaçant la silhouette. Ainsi, un pixel brillant

dans l'image de l'historique du mouvement correspond à un mouvement récent et un pixel moins brillant correspond à un mouvement plus ancien.

- **Le gradient du mouvement (cvCalcMotionGradient)** Cette fonction permet le calcul du gradient d'orientation d'historique du mouvement de l'image MHI. Ainsi, elle calcule les dérivés D_x et D_y de l'historique du mouvement de l'image et ensuite l'orientation du gradient à l'aide de la formule suivante :

$$Orientation(x, y) = Arctangente(D_x(x, y)/D_y(x, y)) \qquad (3.1)$$

- **L'orientation globale (cvCalcGlobalOrientation)** Cette fonction calcule l'orientation globale du mouvement de quelques régions du visage sélectionnées (Figure 3.12) et retourne l'angle d'orientation entre 0° et 360° de chaque région d'intérêt. Ainsi, à chaque instant de la séquence vidéo et durant le traitement de chaque image, la fonction calcule l'orientation globale de chaque région des six régions d'intérêt choisies.

- **Le segment du mouvement (cvSegmentMotion)** Cette fonction trouve tous les segments du mouvement et les marques des six régions d'intérêt dans l'image où le masque trouvé devrait être stocké avec chacune des valeurs individuelles (1,2,...). Ensuite, la direction du mouvement pour chaque composant puisse être calculée avec la fonction de calcul de l'orientation globale du mouvement en utilisant le masque extrait du composant particulier.

- **Histogrammes** On peut définir des histogrammes standards. La création n'est pas directe : il faut auparavant désentrelacer les canaux de l'image. Ils sont utilisés pour évaluer la distribution du mouvement dans une région d'intérêt. Ainsi, la fonction de création d'histogramme d'une taille indiquée retourne le pointeur sur l'histogramme créé. Il existe plusieurs fonctions sur les histogrammes parmi lesquelles la copie d'un histogramme dans un autre, les fonctions de comparaison d'histogrammes, rétro projection sur une image, etc.

- **Fonctions d'entrée / sortie sur les séquences vidéo**

  - **Lecture d'une vidéo** Il est possible de lire une séquence d'images à partir d'un fichier vidéo ou directement à partir d'une caméra. Pour chaque étape, une image est automatiquement allouée (et ensuite libérée) en mémoire.

  - **Structure de capture d'une vidéo (CvCapture)** Cette structure n'a pas d'interface publique et elle est employée seulement comme paramètre pour des fonctions de capture de vidéo. Ainsi, il y a deux types d'initialisation de capture d'une vidéo. Le premier type est l'initialisation à partir d'un fichier AVI (cvCaptureFromFile("Fichier.avi")). Le deuxième type est l'initialisation en temps réel à partir d'une caméra (cvCaptureFromCAM (index de la caméra utilisée)).

  - **Saisie d'une image (cvQueryFrame(capture))** Cette fonction permet de saisir une image à partir d'une caméra ou un fichier AVI déjà capturé. L'image saisie est stockée en mémoire. Le but de

cette fonction est de saisir l'image rapidement ce qui est important pour la synchronisation dans les cas de lecture à plusieurs caméras simultanément. Les images saisies ne sont pas exposées parce qu'elles peuvent être stockées dans le format compressé (comme défini par la caméra/conducteur).

- **Obtention d'une image (cvSaveImage("Image.extension",image))** Cette fonction permet de sauvegarder l'image saisie. Ainsi elle retourne un pointeur sur l'image saisie par la fonction précédente. L'image retournée ne devrait pas être libérée ou modifiée par l'utilisateur.

3.2.5 Le diagramme de fonctionnement

La figure 3.2 illustre le diagramme de fonctionnement de notre programme. Il regroupe toutes les étapes nécessaires à la détection automatique d'une expression faciale à partir d'une séquence d'images.

3.3 Méthodes appliquées

3.3.1 Opérations Classiques

La reconnaissance de formes est un sujet traité de longue date en traitement d'images. Elle fait généralement appel à la détection des contours ainsi qu'à la morphologie mathématique pour établir des caractéristiques et des

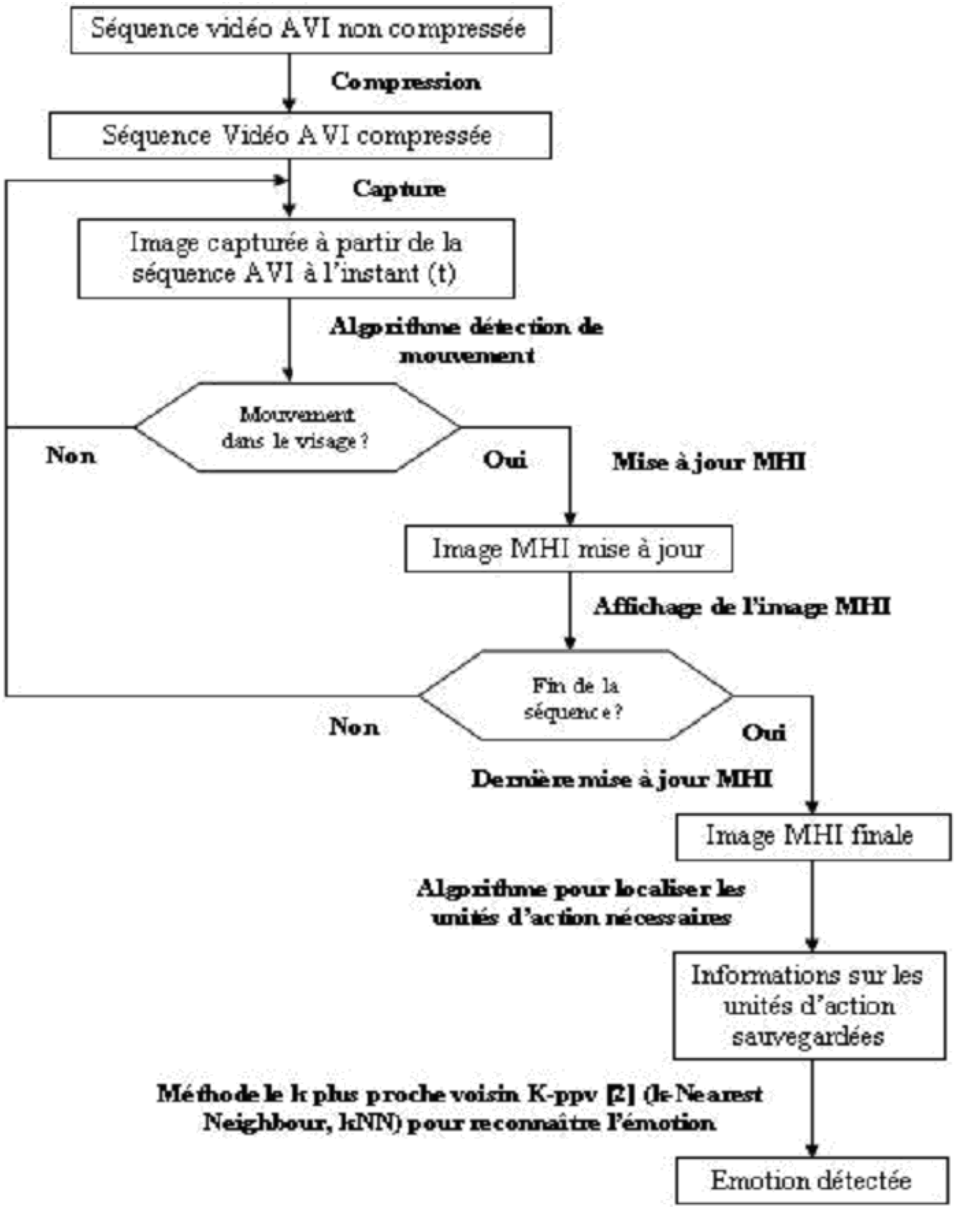


Figure 3.2: Le diagramme de fonctionnement regroupant toutes les étapes nécessaires à la détection automatique d'une émotion.

opérations sur les formes. Ainsi, l'approche retenue utilise des outils liés à la reconnaissance de formes ainsi que la géométrie spatiale et les probabilités pour parvenir à une solution efficace.

Détection de contours

Pour pouvoir détecter et localiser le visage dans une séquence d'images et par la suite l'approximer par une ellipse, on a appliqué des techniques de détection de contours [57, 64, 61, 79]. La détection de contours est très utilisée comme étape de prétraitement pour la détection d'objets et pour trouver les limites de régions. En effet, un visage peut être localisé à partir de l'ensemble des pixels de son contour. De plus, trouver cet ensemble permet d'obtenir une information sur la forme du visage. Du point de vue théorique, un contour est défini par un changement marqué de l'intensité d'un pixel à autre. En chaque point, le contour est considéré comme perpendiculaire à la direction du gradient de la fonction de luminance de l'image de la région entourant le visage, ce dernier est alors utilisé pour la détection. Selon les modèles classiques de traitement d'image, en tout point de l'image les dérivées partielles sont estimées. Ceux dont le gradient est le plus fort correspondent à des pixels de contour. En pratique, le calcul du gradient est effectué par la convolution avec des filtres linéaires. Ces filtres peuvent estimer les dérivées premières, les dérivées secondes, etc., et se focaliser sur des contours de directions différentes. Il existe ainsi de nombreux filtres, qui ont des fondements mathématiques différents. Cependant, quelle que soit leur justification théorique, leur objectif reste le même : mettre en évidence

les pixels qui ont une valeur très différente de leurs voisins. Le filtre de Sobel [57] est utilisé pour la détection des contours à transition élevée du niveau de gris permettant entre autres la localisation des contours du visage. Ce filtre à matrice de 3 x 3 a un seuil ajustable et accorde une pondération aux pixels environnants. Il permet après seuillage d'obtenir une image binaire où les pixels d'intensité maximum représentent les zones à fort contraste. C'est un filtre dont l'application est relativement rapide et qui donne de bons résultats.

La transformée de Hough [57] est une méthode utilisée pour la détection des droites et aussi adaptée à la détection des cercles. C'est une méthode d'ordre de complexité élevé et de temps d'exécution assez lent. Plusieurs chercheurs développent des optimisations de cette méthode mais un essai personnel n'ayant pas été satisfaisant dans le contexte de cette recherche, la méthode n'a pas été utilisée dans le développement. Elle est mentionnée ici puisqu'elle est citée dans certaines recherches en références [80, 81, 82, 83].

Les contours peuvent être également employés pour l'analyse de formes et la reconnaissance des objets. L'approche retenue pour la détection de contours fonctionne de la façon suivante : au début, elle effectue une recherche des contours de l'image binaire et renvoie le nombre de contours recherchés. Ensuite, un pointeur est retourné par la fonction. Il contiendra une référence sur le premier contour ou NULL si aucun contour n'est détecté (si l'image est complètement noire). D'autres contours peuvent être atteints en utilisant des liens de h_{next} ou v_{next} (fonction cvFindContour() d'OpenCV).

Les figures 3.3 et 3.4 représentent les résultats de détection de contours de deux personnes en état neutre (aucune émotion produite).

(a) (b)

Figure 3.3: Détection de contours d'une image prise à partir d'une séquence vidéo. (a) Image source. (b) Résultat de la détection de contours. Avec un seuil du gradient de 80.

(a) (b)

Figure 3.4: Détection de contours d'une image prise à partir d'une séquence vidéo. (a) Image source. (b) Résultat de détection de contours. Avec un seuil du gradient de 114.

Finalement, la détection de contours va nous servir pour la détection du

visage et ensuite permettre son approximation par une ellipse.

Morphologie mathématique

La morphologie mathématique [62] est une représentation ensembliste des éléments de l'image. Ainsi, les opérateurs morphologiques que nous avons appliqué servent à une détection, une localisation, et une approximation par une ellipse du visage plus précise et ce en éliminant le bruit et ensuite faciliter les prochaines tâches (figures 3.6 et 3.7). Les opérateurs morphologiques (l'érosion et la dilatation sont les plus connues) permettent de traiter efficacement des images binaires. Les pixels obéissant à une certaine connectivité forment des ensembles sur lesquels on peut appliquer des opérations morphologiques. La morphologie permet également de caractériser l'ensemble de pixels par l'utilisation des relations de dispersion ou de proximité. L'introduction de règles probabilistes assure une bonne certitude d'évaluation. Les opérations sur un ensemble restreint de pixels s'exécutent dans un temps très court et permettant une décision rapide.

Corrélation

La technique de corrélation [55] est utilisée pour faciliter la détection automatique des expressions faciales. Ainsi, elle est basée sur une comparaison simple entre une image d'un visage non classifié et les images des visages d'apprentissage. L'image du visage d'apprentissage se trouvant à la plus faible distance du visage non classifié sera sélectionné comme premier choix.

Plusieurs métriques peuvent être utilisées afin d'évaluer cette distance comme par exemple les distances L1 (city block) et L2 (euclidienne), la corrélation croisée, la distance de Mahalanobis, etc. Ce processus de décision est communément appelé algorithme du k plus proches voisins (k-ppv) [55] et est présenté avec davantage de détails à la section 2.2.8. De plus, toutes les séquences qui seront traitées représentent des visages clairs, sans mouvement non rigide, et sous forme d'une vue frontale. Dans ce contexte, cette méthode offre des avantages particulièrement intéressants. Par contre, la technique de corrélation offre peu de robustesse face aux expressions faciales, aux variations d'éclairage et aux changements physiques (ex. : barbe) ce qui n'est pas notre cas.

3.3.2 Détection et localisation du visage

De nombreux algorithmes de reconnaissance du visage ont été développés au cours des dernières années et plusieurs se révèlent très performants [59, 60, 63, 64, 80]. Cependant, le succès de ces méthodes dépend largement de la qualité des résultats de détection et de localisation des visages. En effet, plus la précision obtenue est élevée, plus les conditions se rapprocheront de celles de la phase d'apprentissage, ce qui augmentera les probabilités d'une identification efficace. La performance d'un système de reconnaissance de visages est évidemment tributaire de la qualité et de l'efficacité du module d'identification. Une mauvaise localisation du visage entraîne cependant une chute drastique du taux de reconnaissance. Dans ce cas bien précis, l'extraction de la représentation du visage sera erronée et difficilement com-

parable aux prototypes d'apprentissage. L'algorithme de détection et localisation s'effectue en trois phases :

- Détecter et affiner les contours dans l'image.
- Modéliser chaque contour avec une courbe elliptique.
- Raffiner la localisation des visages dans l'espace des paramètres.
  La figure 3.5 représente le système de segmentation de visages.

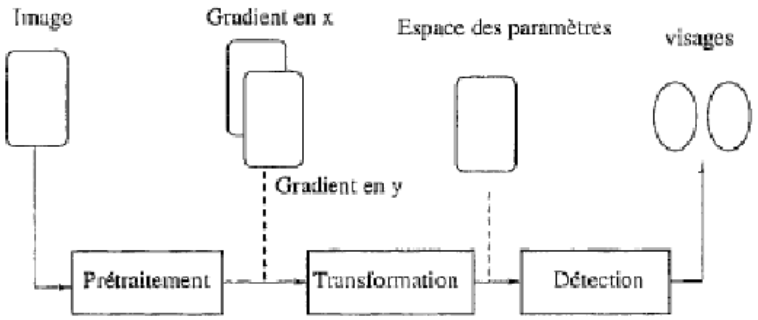


Figure 3.5: Le système de segmentation de visages.

Notre système de segmentation comprend trois phases majeures:

- Prétraitement : Cette phase requiert la présence obligatoire d'un visage dans l'image, changement des dimensions de l'image si différentes des images références, le visage doit être en position de face, et calcul des gradients en x et en y de l'image.
- Transformations : Création d'un espace de paramètres avant d'effectuer la détection des visages.

- Détection : Dernière étape de segmentation et qui permet la détection dans un espace des paramètres de tous les visages disponibles dans une image (dans notre cas un seul visage). Le résultat de détection est sous forme d'ellipses (chaque ellipse représente un visage).

Détection du visage

Une grande variété de méthodes de détection du visage ont été proposées dans les dernières années [59, 60, 80]. Nous allons se limiter à une seule méthode pour la détection de celui-ci. Ainsi, la méthode retenue et la plus adaptée à notre projet porte sur l'utilisation des arêtes grâce à son efficacité et sa robustesse par rapport aux autres approches. Les arêtes sont décrites comme étant formées des "points de discontinuités dans la fonction de luminance (intensité) de l'image" [61]. Ces informations utiles sont notamment employées pour l'interprétation de scène et la reconnaissance d'objets. Le principe de base consiste à reconnaître des objets (le visage pour notre cas) dans une image à partir des modèles de contours connus au préalable [61]. Pour réaliser cette tâche on appliquera la méthode de la transformée de Hough.

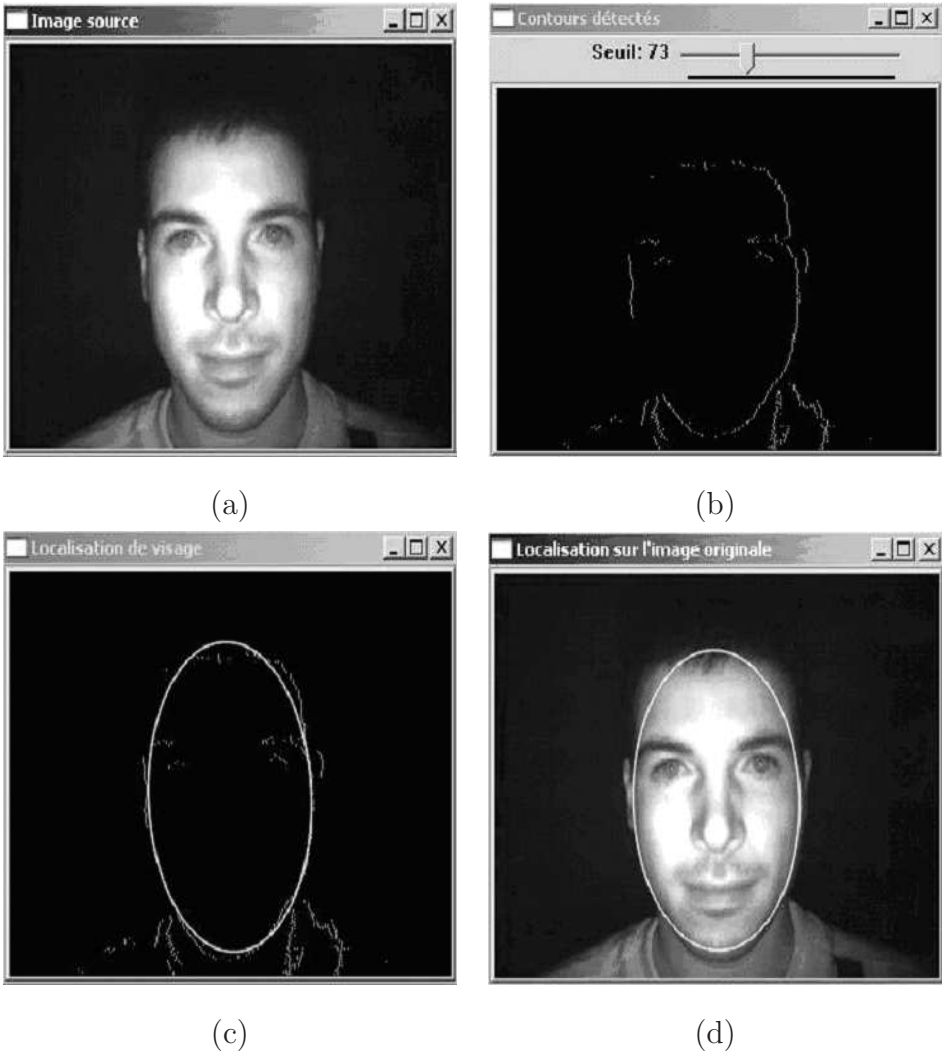
Transformée de Hough : La transformée de Hough est une méthode permettant d'extraire et de localiser des groupes de points respectant certaines caractéristiques. La transformée de Hough est largement utilisée en traitement d'images. Ainsi, nous avons appliquée la transformée de Hough dans nos expérimentations à l'aide de plusieurs fonctions openCV (cvFitEl-

lipse(), cvEllipse(), cvDrawEllipse()). Cette méthode fut introduite en 1962 par Paul Hough dans le but de détecter les trajectoires rectilignes de particules de haute énergie. Korosec et al. [80] sont les premiers à modéliser les contours extérieurs du visage par une ellipse, en appliquant une transformée de Hough standard. En outre, l'utilisation de la transformée de Hough autorise une vitesse de détection assez rapide pour envisager un traitement en temps réel. Nous avons donc décidé d'adapter la Transformée de Hough à une séquence vidéo en modélisant les contours extérieurs d'un visage par une courbe elliptique, et en recherchant ce type de courbe dans une scène. Par exemple, les particularités recherchées peuvent être des droites, des arcs de cercles, des formes quelconques, etc. Dans un contexte de détection du visage, ce dernier est représenté par une ellipse. L'application de la transformée de Hough circulaire produirait donc une liste de tous les candidats étant des cercles ou des dérivés [63]. Finalement, la transformée de Hough peut être utilisée pour détecter les yeux et les iris. Par contre, cette méthode échouera lorsque l'image est trop petite ou lorsque les yeux ne sont pas clairement visibles.

Localisation du visage Cette étape correspond à la recherche du centre d'ellipse dans l'espace des paramètres. Nous détectons ensuite dans l'espace des paramètres, les valeurs maximales locales qui correspondent dans l'espace image aux centres des l'ellipses d'approximation des visages. Si plusieurs visages sont détectés dans l'image, le système les considère les uns après les autres (les zones précédemment détectées sont remplacées par zéro) en utilisant un seuillage. Ainsi la vitesse d'exécution est d'autant moins rapide

que le nombre de personnes présentes dans l'image augmente. La distance entre la caméra et la scène filmée est fixe et la taille des visages varie peu dans l'image. La transformée de Hough exploite la valeur et la direction du gradient en chaque point de contour. Dans notre cas, la forme recherchée est une ellipse, donc le gradient en chaque point de contour est perpendiculaire à la tangente de la courbe générée par ce point et son voisinage [64]. Puisque les séquences d'images utilisées pour l'expérimentation contiennent un seul visage, cette approche est très bien adaptée pour notre projet et donne des meilleurs résultats de localisation du visage.

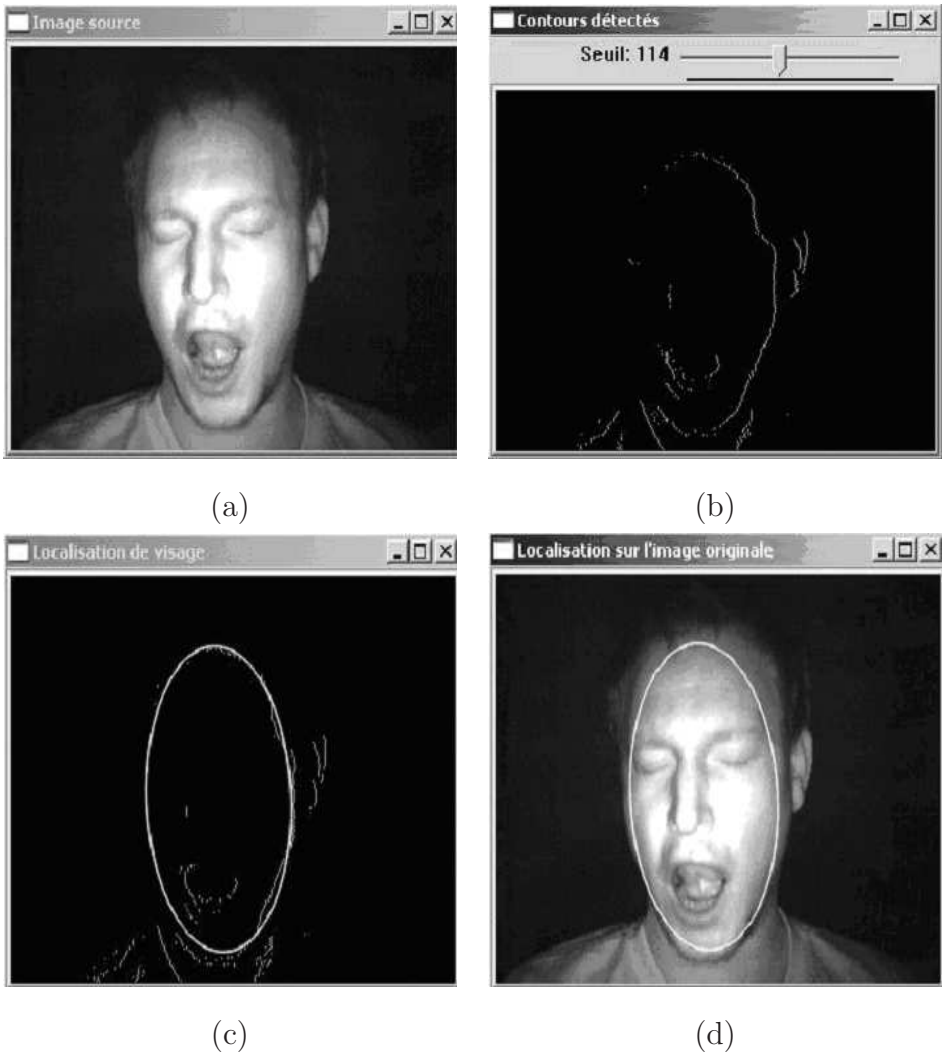
Les figures 3.6 et 3.7 illustrent les résultats de traitement de chaque étape en exploitant deux images contenants deux visages différents.



(a) (b)

(c) (d)

Figure 3.6: Résultats de détection et localisation du visage par une ellipse. (a) Image originale. (b) Contours affinés. (c) Ellipse d'approximation. (d) Localisation du visage par une ellipse.



(a) (b)

(c) (d)

Figure 3.7: Résultats de détection et localisation du visage par une ellipse d'une autre personne. (a) Image originale. (b) Contours affinés. (c) Ellipse d'approximation. (d) Localisation du visage par une ellipse.

Dans la majorité des cas, les visages intéressants à détecter sont positionnés selon une direction verticale (Vue frontale), donc il n'est pas indispensable

de définir l'orientation précise des visages détectés. Une ellipse verticale a pour expression :

$$\frac{X^2}{lh^2} + \frac{Y^2}{lv^2} = 1 \tag{3.2}$$

Avec : X=x-x_c et Y=y-y_c,

Où (x_c,y_c) est le centre de l'ellipse, (x,y) un point de contour de l'ellipse, la demi-hauteur lh et la demi-largeur lv de l'ellipse.

En effet, d'après nos mesures avec un objectif de 8 mm de distance focale et une distance du visage par rapport à la caméra de 70 cm (représentant la position d'un individu en face d'une caméra positionnée sur un PC), si la distance varie de (plus ou moins) 10 cm (soit 14 %), la variation de la taille du visage est de l'ordre de 5 %. Ainsi, les paramètres lh et lv peuvent être considérés comme des constantes pour notre application visée. Toutefois, pour augmenter la robustesse de notre système, nous tenons compte de cette faible variation de la taille des visages.

3.3.3 Enregistrement du visage dans une séquence d'images

Pour être capable de construire les modèles temporels il faut que la caméra et le background soient statiques, ou bien il faut que le mouvement de l'objet d'intérêt soit séparable du mouvement induit par la caméra et par des mouvements de fond. En outre, pour être capable de comparer les modèles temporels séparés d'une façon significative, le visage dans la séquence d'images doit avoir la même position et la même orientation. Par conséquent, pour construire les modèles temporels utiles et comparables, il faut que les séquences d'images d'entrée soient enregistrées de deux façons. Premièrement, tous les mouvements rigides de la tête dans la séquence d'images doivent être éliminés. Deuxièmement, toutes les séquences d'images utilisées doivent avoir le visage dans la même position, la même orientation, et la même échelle. Pour réaliser les deux modes d'enregistrement, on sélectionne manuellement à la main les neufs points faciaux de la première fenêtre capturée à partir de la séquence d'images à traiter (Figure 3.8).

Figure 3.8: Les neuf points faciaux utilisés pour l'enregistrement des images pour qu'elles soient utilisées pour la construction du modèle temporel.

Ces points seront alors suivis dans toutes les fenêtres suivantes (fenêtres capturées durant le traitement). Pour l'enregistrement de chaque fenêtre en respect de la première fenêtre on appelle une transformation affine en utilisant les points faciaux stables dont la position faciale demeure la même si une contraction faciale (activation d'une unité d'action AU) de certains muscles faciaux survient (figure 3.9).

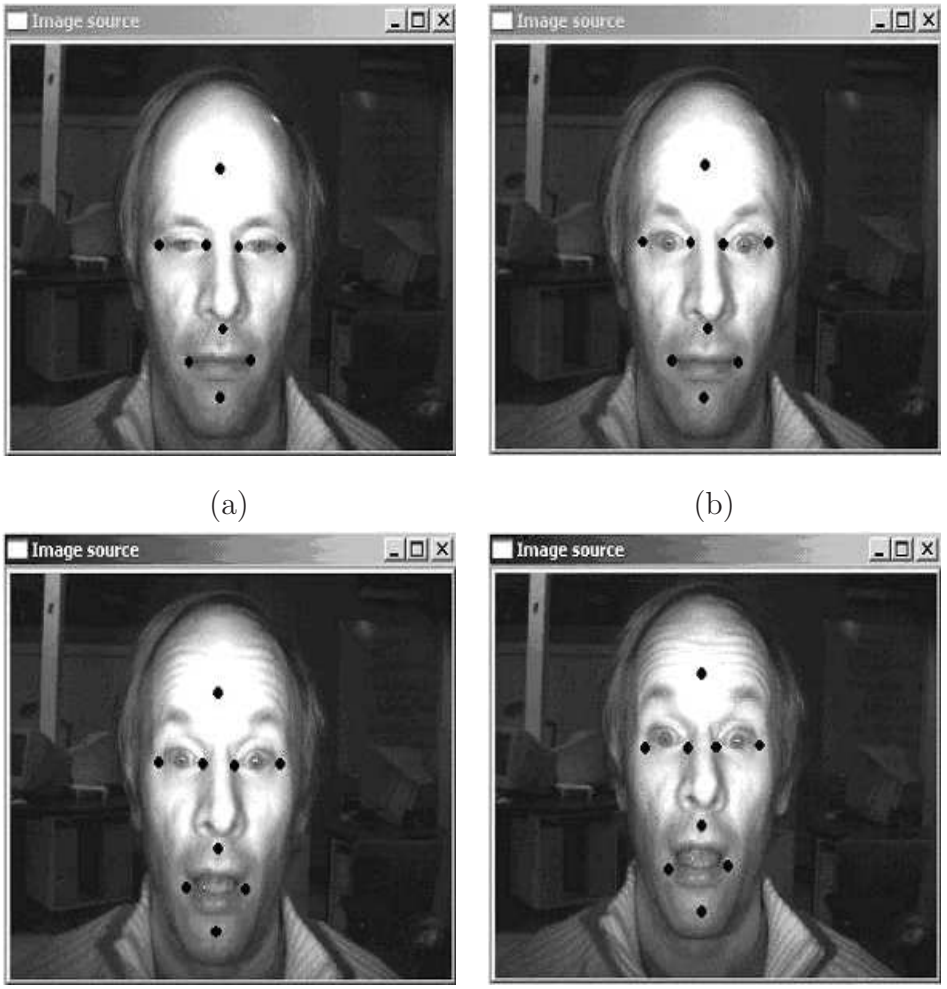


Figure 3.9: Exemple d'enregistrement de quatre fenêtres d'une séquence d'images représentant l'émotion surprise. (a) À l'instant t_0=0 (fenêtre de base). (b) À l'instant t_1=0.6 seconde. (c) À l'instant t_2=1.2 seconde. (d) À l'instant t_3=2 secondes.

En effet, si nous employons quelques points faciaux pouvant subir des déformations, nous ne pourrons pas être certain si le mouvement d'un point est

dû au mouvement de tête rigide non désiré ou dû à l'activation d'une ou de plusieurs unités d'actions (figure 3.10).

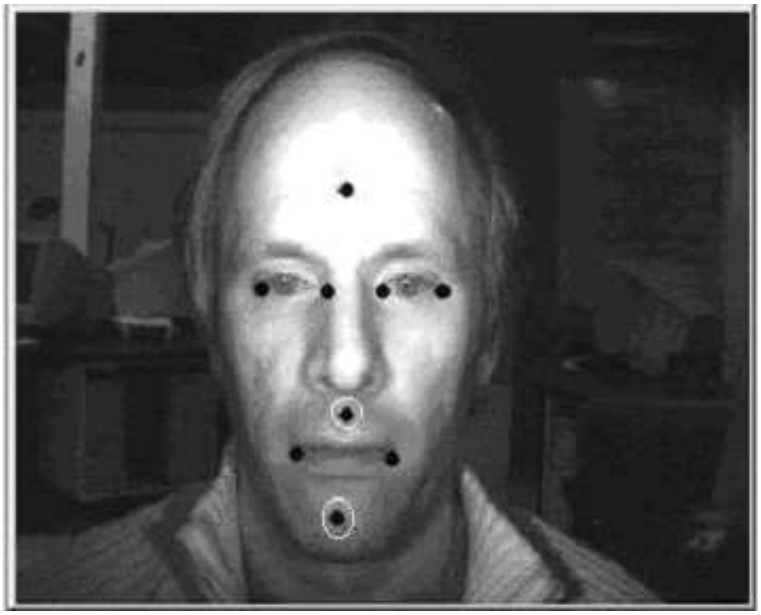


Figure 3.10: Les neuf points faciaux utilisés pour l'enregistrement, les deux cercles en blanc montrent les deux points non importants pour découvrir l'origine de mouvement (activation d'une unité d'action ou mouvement de tête non rigide).

3.3.4 Détection du mouvement

La détection du mouvement, réalisée immédiatement après l'acquisition d'une image, représente une étape non essentielle, mais très avantageuse pour un système de vision numérique. En effet, un gain de performance considérable peut être réalisé lorsque des zones sans intérêt sont éliminées avant les phases

d'analyse. Cette amélioration dépend cependant de la complexité des algorithmes de détection et de reconnaissance utilisés. Par ailleurs, il est important de bien définir le terme détection du mouvement afin d'éliminer les ambiguïtés qui pourraient survenir. La technique de différence d'images est bien connue en vision par ordinateur, elle consiste à effectuer une différence pixel à pixel entre deux images. Étant peu complexe, la différence entre deux images consécutives représente une solution très intéressante. Comme son nom l'indique, elle consiste à soustraire une image acquise au temps t_n d'une autre au temps t_{n+k}, où k est habituellement égal à 1. Supposons que notre séquence d'images se compose de l fenêtres. Posons H(x,y,t), t=1..k une séquence des intensités de pixels de la t^{ieme} fenêtre et posons D(x,y,t) l'image binaire indiquant les régions du mouvement qui résulte de la détection de changement d'intensité du pixel, cela s'effectue par le seuillage comme montré par la formule suivante,

$$|I(x, y, t) - I(x, y, t - 1)| > th_I \tag{3.3}$$

Où, x et y sont les coordonnées spatiales des éléments de l'image et th_I est le seuil de différence d'intensité entre deux images de détection du mouvement, un paramètre qui doit être déterminé expérimentalement. Ainsi, l'image résultante sera vide si aucun mouvement ne se produit pendant l'intervalle du temps observé car l'intensité et la couleur des pixels seront presque identiques. Par contre, si du mouvement a lieu dans le champ de vue, les pixels frontières des objets en déplacement devraient changer drastiquement de valeurs, révélant alors la présence d'activité dans la scène.

La figure 3.11 illustre certains résultats expérimentaux. Alors que les images des figures 3.11 (a) et (b) représentent les deux images consécutives utilisées,

les images des figures 3.11 (c) et (d) contiennent les résultats de la détection du mouvement. La différence entre celles-ci repose sur un seuillage appliqué aux données pour binariser les résultats et ainsi faciliter la visualisation.

3.3.5 Identification du mouvement dans le visage

L'estimation du mouvement a été un problème classique dans la vision par ordinateur. Plusieurs techniques pour décrire le mouvement dans une scène ont été proposées et sont utilisées [51, 52, 77]. La plupart tentent de calculer le flux optique à chaque pixel dans le domaine de l'image sans connaissance à priori au sujet de la scène. Souvent il est souhaitable d'employer une technique, qui peut ne pas être la solution du problème général mais devrait pouvoir travailler assez robustement pour un sous-ensemble d'applications. Notre système à été construit sur le concept des images d'historique du mouvement (Motion History Images, MHI) et des images du gradient de mouvement (Motion Gradient Images, MGI) développées par James Davis et Aaron Bobick [51, 52], et nous avons utilisé ces images pour déduire les histogrammes du mouvement (Motion Histograms) qui seraient indicatifs du genre de mouvement se produisant dans la scène.

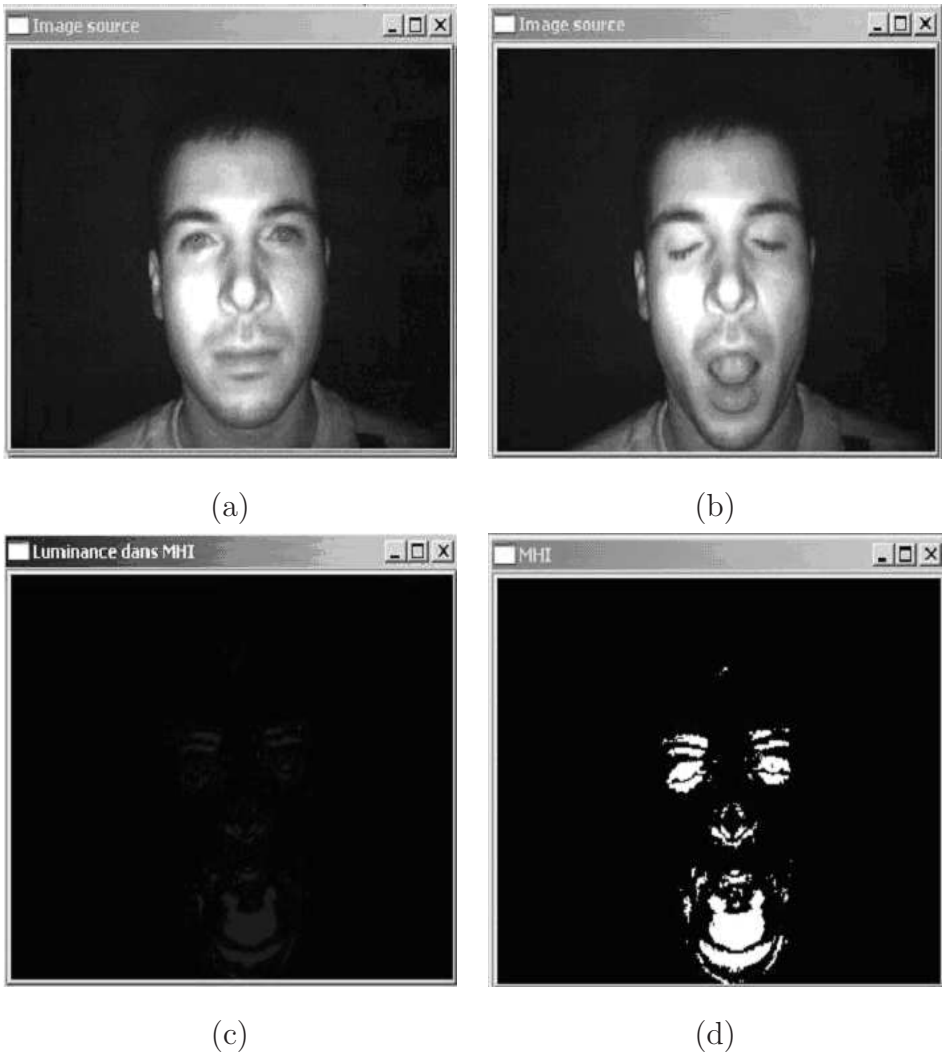


Figure 3.11: Exemple de détection du mouvement avec la méthode de soustraction d'images consécutives. (a) Image source à l'instant t_0. (b) Image source à l'instant t_1. (c) Détection du mouvement non seuillée. (d) Détection du mouvement.

Régions d'intérêt

La figure 3.12 représente les six régions d'intérêt (présence du mouvement durant la production d'une émotion) choisies manuellement pour faciliter la

détection des différentes expressions faciales en question (Baîllement, Colère, Sommeil, Sourire, Surprise).

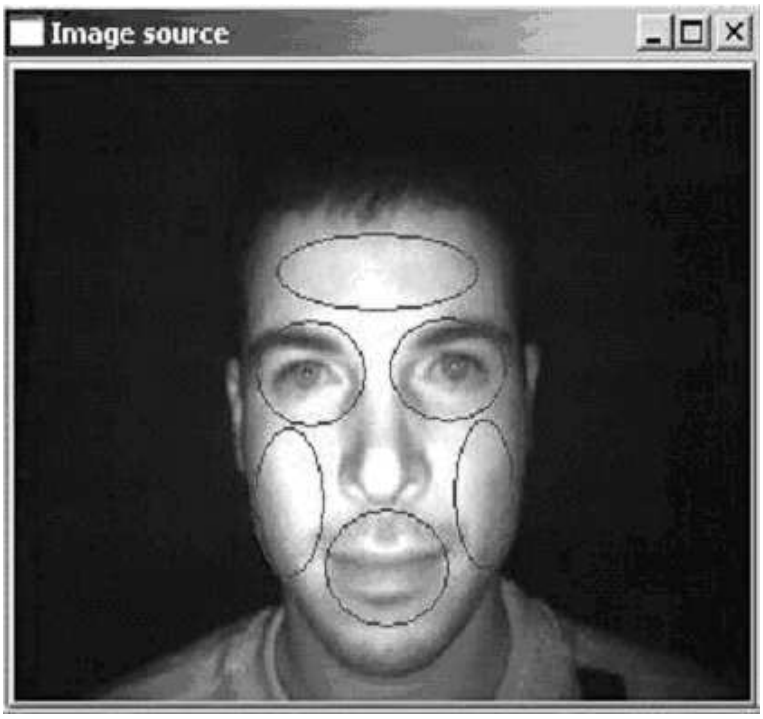


Figure 3.12: Les six régions d'intérêt choisies (front, oeil gauche, oeil droit, joue gauche, joue droite, et la région de la bouche).

Image de l'historique du mouvement

Pour générer l'image MHI (Motion History Image) du mouvement, nous pondérons et nous calculons les différences successives de l'image de la silhouette [51]. Dans l'image MHI, chaque valeur du pixel est une fonction de l'historique temporel du mouvement à ce point de toutes les fenêtres cap-

turées (30 fenêtres pas seconde) durant le traitement de la séquence d'images du mouvement. On utilise couramment un opérateur simple de rechange et d'affaiblissement basé sur l'emboutissage de temps.

$$\mathbf{MHI}(\mathbf{x},\mathbf{y}) = \begin{cases} \tau & Si\ le\ mouvement\ courant\ est\ en\ (x,y) \\ 0 & Sinon\ \ si\ \ MHI(x,y) < (\tau - \delta) \end{cases} \tag{3.4}$$

Où τ est l'horodateur courant, et δ est la constante maximale de la durée du temps. La fonction ci-dessus est appelée pour la mise à jour de l'image MHI à chaque fois qu'un résultat d'une nouvelle image de différence est calculé. En normalisant linéairement les horodateurs de l'image MHI à des valeurs entre 0 et 255, on voit que les pixels correspondant à un mouvement plus nouveau sont plus brillants que les pixels correspondant à un mouvement plus ancien. Le résultat du processus ci-dessus est montré dans la figure 3.13. Finalement, une MHI est une image obtenue en projetant le volume d'images dans le temps sur une image simple. Les valeurs d'intensité dans la MHI sont indicatives du temps à ce que ce pixel était témoin pour la dernière fois du mouvement ou de la présence de l'objet.

(a) (b) (c)

(d) (e)

Figure 3.13: Construction de l'image MHI d'historique de mouvement. (a) Image prise à partir d'une séquence d'images représentant une surprise. (b) et (d) Niveau de luminance (mouvement plus récent donne plus de luminance au niveau des pixels en mouvement). (c) Fin de l'émotion surprise. (e) Image MHI du mouvement.

Image du gradient du mouvement

Une MGI est l'image du gradient du mouvement de la MHI, où le gradient donne la direction du mouvement de chaque pixel dans l'image. Ainsi, l'image MHI présente les différences de silhouette de telle manière que le mouvement de la frontière de la silhouette puisse être perçu par le gradient déduit de

la fonction MHI(x,y). Ceci est très semblable au concept de flux optique [28]. On note que quand les sourcils sont montés vers le haut, l'effacement d'intensité (du foncé au plus clair) donne l'impression du mouvement dans la direction du mouvement des sourcils. Il s'est avéré que l'image MHI code visuellement quelques informations sur le mouvement de la silhouette. On voit clairement la direction du mouvement (Figure. 3.14), mais la magnitude est par contre inaccessible. Notre objectif est donc d'utiliser cette information directionnelle du mouvement pour la reconnaissance des expressions faciales. Les orientations locales du gradient de l'image MHI montrent directement le mouvement de la silhouette. Donc, on peut convoluer les masques classiques du gradient avec l'image MHI pour extraire l'information directionnelle du mouvement. Pour notre travail, on réuni la convolution à deux résolutions (l'originale et une résolution plus grossière) pour manipuler les gradients les plus répandus (en raison des vitesses différentes du mouvement). Les masques suivants de gradient de Sobel sont utilisés pour la convolution :

$$F(x) = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad , \quad F(y) = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{3.5}$$

Avec les vecteurs du gradient calculés, il est très simple d'avoir l'orientation du gradient d'un pixel comme suit :

$$\theta = \arctan \frac{F_y}{F_x} \tag{3.6}$$

On doit faire attention en calculant l'information sur le gradient parce qu'elle est valide seulement à des endroits particuliers dans l'image MHI. Les fron-

tières de l'image MHI ne devraient pas être utilisées car les pixels qui ne sont pas en mouvement (pixels en noir) seraient inclus dans le calcul du gradient, et cela donne souvent des résultats erronés. Seulement les pixels appartenant aux régions d'intérêt (présence du mouvement) dans une MHI devraient être examinés. De plus, on ne doit pas utiliser les gradients des pixels de la MHI qui ont un contraste trop bas ou trop élevé dans leur voisinage local. Un petit contraste ne donne pas une mesure fiable de la direction du gradient, et un grand contraste signifie une large disparité temporelle entre les pixels, ce qui rend l'information directionnelle polarisée et inutilisable. Les résultats du calcul de l'orientation du mouvement utilisant les masques du gradient sont présentés dans la figure 3.14.

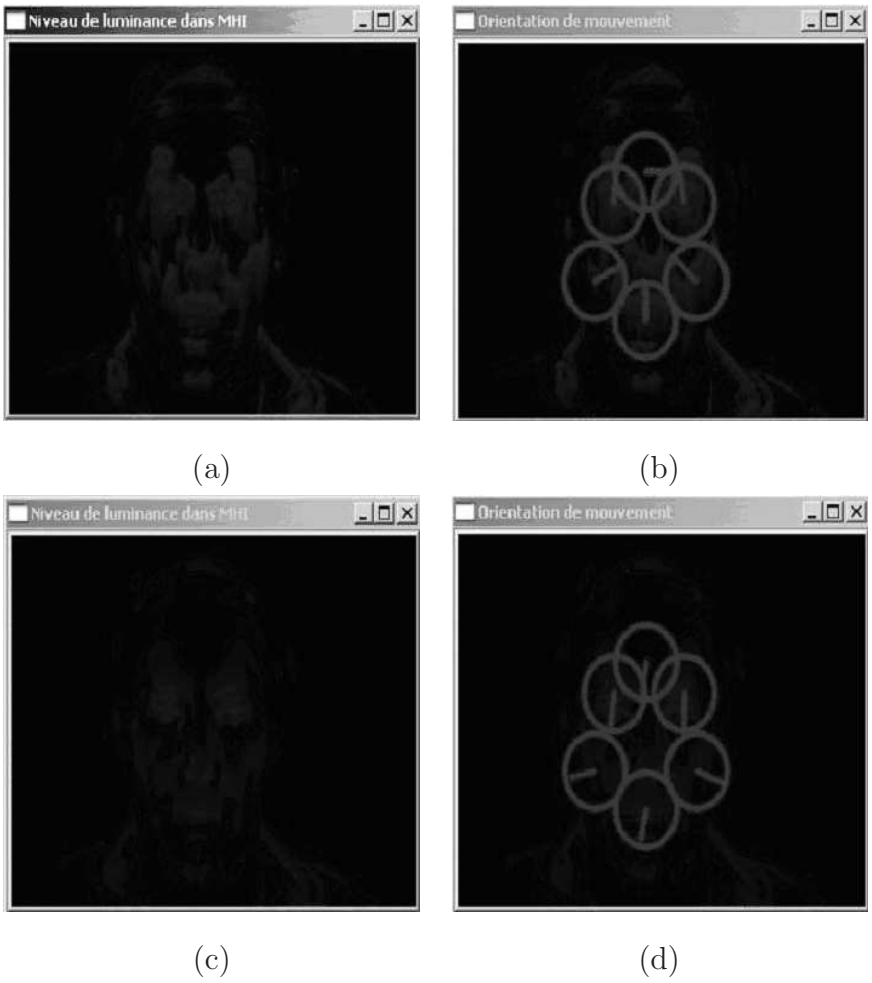


Figure 3.14: Directions du mouvement à partir des gradients d'une MHI. (a) et (c) MHIs résultantes du mouvement des caractéristiques faciales de l'émotion surprise ((a) au milieu de l'émotion, et (c) juste avant la fin de l'émotion). (b) et (d) Résultats de convolution des masques de gradient avec l'image MHI.

Ainsi, les directions du gradient montrent le mouvement approximatif de chaque région des six régions d'intérêt définies manuellement (front, yeux,

bouche, et joues).

La table 3.2 montre les résultats de mesure des orientations globales de mouvement des deux historiques de mouvement (a) et (b) de la figure 3.14 pour chaque région parmi les six régions (régions d'intérêt figure 3.12).

Fenêtre	**Orientation globale de la Région d'intérêt**					
	R1(F)	**R2**(OG)	**R3**(OD)	**R4**(B)	**R5**(JG)	**R6**(JD)
(a)	0	276 °	262 °	271 °	331 °	221 °
(b)	98 °	98 °	91 °	103 °	167 °	21 °

Table 3.2: Orientation globale de mouvement pour chaque région d'intérêt de visage d'une personne produisant l'émotion surprise selon les résultats de la figure 3.14.

NB : *F = Front, OG = Oeil Gauche, OD = Oeil Droit, JG = Joue Gauche, JD = Joue Droite, B = Bouche.*

Histogramme du mouvement

Un histogramme du mouvement décrit la distribution des angles (directions du mouvement) dans l'image MGI. La façon la plus simple de localiser le mouvement pour la reconnaissance des expressions faciales est de calculer un histogramme pour différentes régions autour du visage. Une solution pour cela est de diviser la MHI en diverses régions (ou fenêtres), et ensuite de caractériser chaque région par un histogramme. Une méthode de caractérisation est d'utiliser l'histogramme des orientations du mouvement

d'une région. Ainsi, on peut diviser le modèle du mouvement d'une MHI en régions, chaque région va être représentée par un histogramme d'orientation du mouvement local. La table 3.3 représente l'histogramme du mouvement de la séquence d'images représentant l'émotion surprise traitée dans la figure 3.13. Ainsi, elle montre la présence du mouvement ou non ainsi que l'orientation du mouvement en cas de présence du mouvement (angle Ri différent de zéro) pour chaque région à chaque instant t de cette séquence.

Durée(secondes)	**Orientation de mouvement**					
	R1(F)	**R2(OG)**	**R3(OD)**	**R4(B)**	**R5 (JG)**	**R6(JD)**
0.3	0	0	0	0	0	0
0.5	0	277 °	270 °	92 °	0	0
1	0	279 °	246 °	79 °	251 °	225 °
1.5	0	276 °	262 °	271 °	331 °	221 °
2	98 °	95 °	89 °	252 °	177 °	177 °
2.5	98 °	101 °	90 °	98 °	167 °	9 °
3	98 °	98 °	91 °	103 °	167 °	21 °

Table 3.3: Histogramme d'orientation du mouvement pour les six régions définies auparavant d'une séquence d'images d'une durée de trois secondes et représentant l'émotion surprise.

3.3.6 Classification et reconnaissance

Le résultat de production des histogrammes du mouvement du visage est la collection des histogrammes de toutes les régions localisées (table 3.2). Il y

a plusieurs façons possibles d'utilisation de ces données pour la reconnaissance. L'approche la plus simple est de concaténer les histogrammes en un seul vecteur et d'appeler l'algorithme de classification. Pour effectuer la classification, nous avons utilisé l'algorithme du k plus proches voisins (k-ppv) [55]. C'est un algorithme de classification très connu et très simple à utiliser. Parmi les autres algorithmes de classification, nous pouvons citer les machines de vecteur de soutien (Support Vector Machines, SVM) [66, 67]. C'est très complexe comme algorithme et ne pourrait pas être toujours adaptable, selon la représentation de données. Le calcul des distances peut s'effectuer par différentes métriques, les métriques qui sont adaptées à notre projet et qui sont retenues pour la classification des expressions faciales sont : La métrique de City-block (L1) , la métrique euclidienne (L2) ainsi que la métrique de Minkowski (Lm) (section 2.2.8).

Le modèle du mouvement

Dans cette experimentation, nous avons générer pour chaque émotion à traiter un modèle du mouvement qui sert de référence pour la classification. Pour générer un modèle du mouvement d'une émotion particulière, nous avons pris plusieurs exemples d'historiques et histogrammes de mouvement d'une personne (ou plusieurs personnes) produisant cette émotion particulière. De plus, nous avons répété cette approche pour chaque émotion traitée dans ce travail de recherche. Les histogrammes des orientations (table 3.3) du mouvement pour chaque émotion et pour chaque région des six régions d'intérêt du visage (figure 3.12) sont générés et stockés (sous forme

de vecteurs à six dimensions) une fois que chaque exemple est terminé (à partir d'une sélection manuelle durant le test). Pour obtenir un modèle simple de ce mouvement, un histogramme moyen des orientations du mouvement (table 3.2) est formé en calculant la moyenne de tous les histogrammes d'orientations (sous forme de vecteur) de toutes les régions d'intérêt du visage. Les exemples de test sont ensuite utilisés pour déterminer les voisins les plus proches entre ces vecteurs de test et le vecteur associé à une expression faciales inconnue et ce, à l'aide de l'algorithme de classification du k plus proches voisins (k-ppv). Ainsi, on peut choisir le seuil de reconnaissance basée sur la variabilité des mesures de données du test. Ce processus est alors répété pour chaque modèle du mouvement (séquence d'images représentant une émotion).

Les figures 3.15, 3.16, 3.17, 3.18 et 3.19 illustrent les modèles graphiques de mouvement des cinq séquences d'images utilisées comme références et montrant les cinq expressions faciales cherchées (baîllement, colère, sommeil, sourire et surprise).

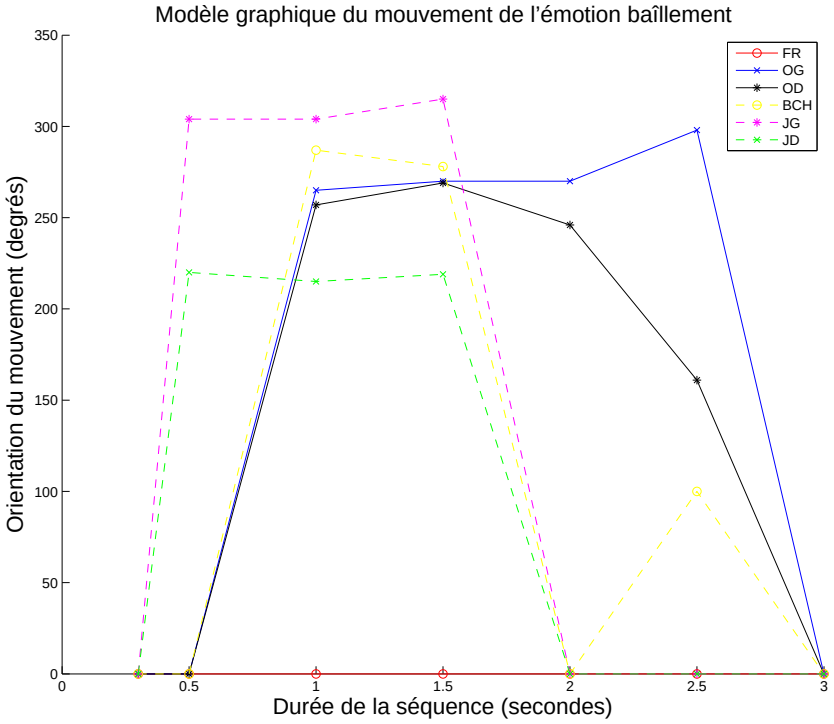


Figure 3.15: Modèle graphique de mouvement des six régions du visage de la séquence d'images représentant l'émotion baîllement et utilisée comme référence.

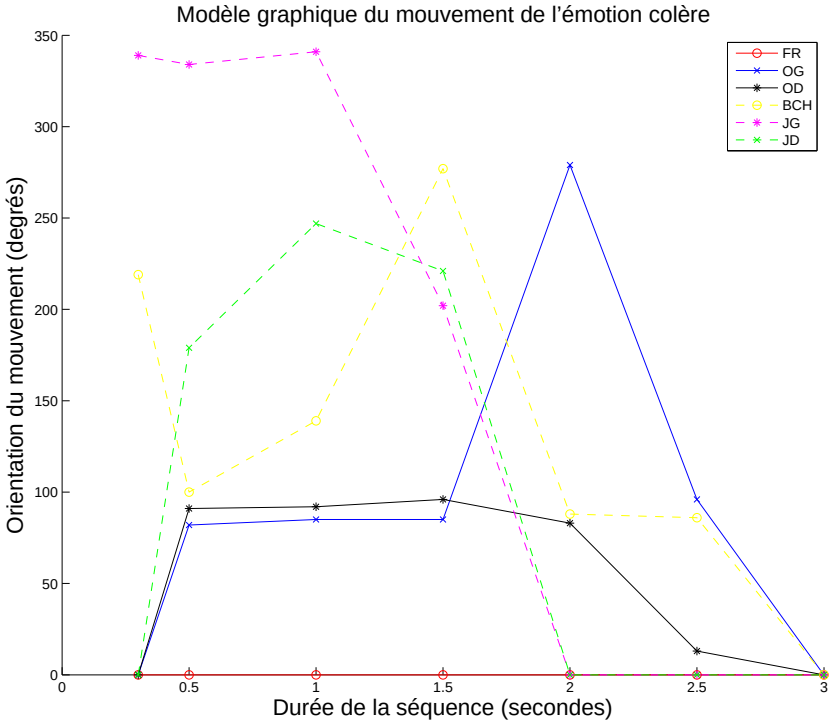


Figure 3.16: Modèle graphique de mouvement des six régions du visage de la séquence d'images représentant l'émotion colère et utilisée comme référence.

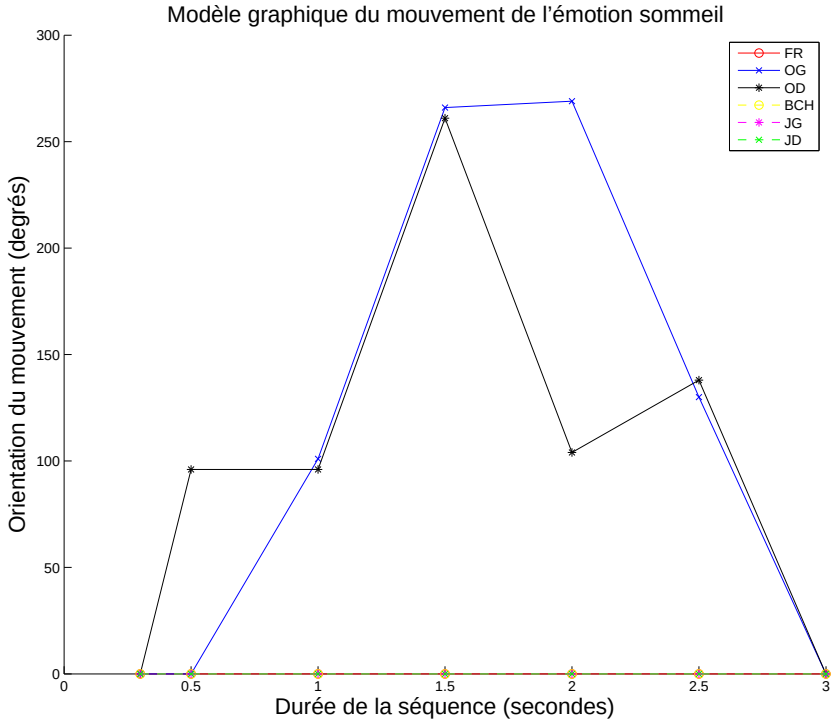


Figure 3.17: Modèle graphique de mouvement des six régions du visage de la séquence d'images représentant l'émotion sommeil et utilisée comme référence.

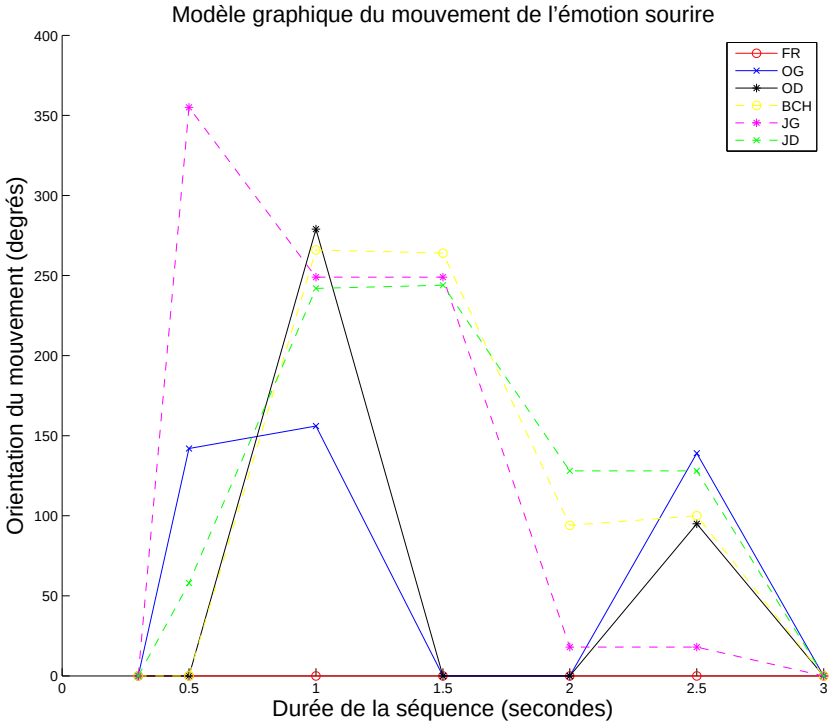


Figure 3.18: Modèle graphique de mouvement des six régions du visage de la séquence d'images représentant l'émotion sourire et utilisée comme référence.

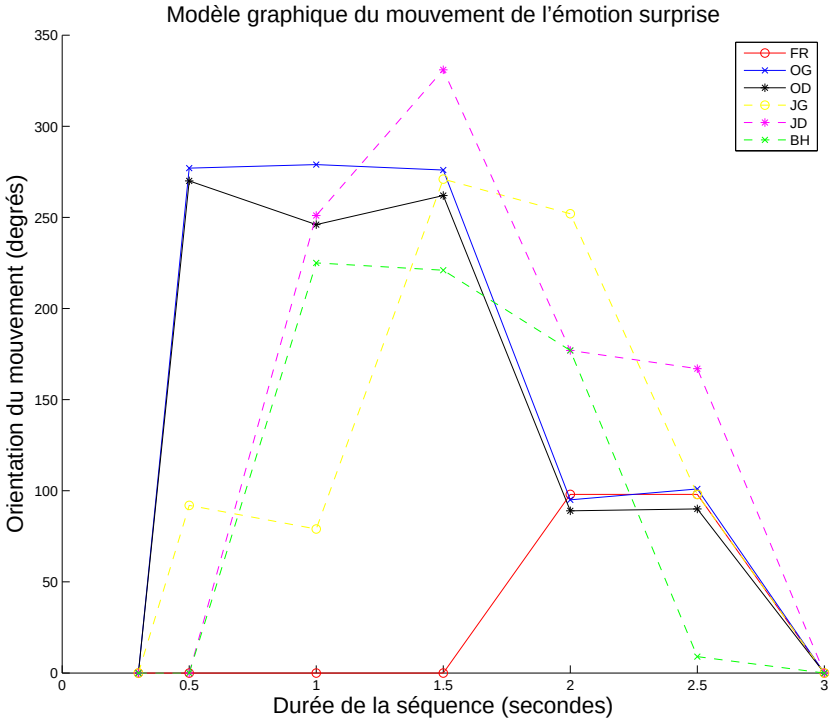


Figure 3.19: Modèle graphique de mouvement des six régions du visage de la séquence d'images représentant l'émotion surprise et utilisée comme référence.

Les données utilisées pour la construction de ce modèle sont celles de la table 3.3.

Reconnaissance

Pour la reconnaissance d'une nouvelle entrée de données (séquence d'images représentant une expression faciale inconnue), nous calculons tout simp-

lement la distance entre le vecteur d'entrée composé de l'histogramme des orientations du mouvement et chacun des vecteurs références composés aussi de l'histogramme des orientations prises à partir des séquences d'images références et représentants les différentes émotions (Baîllement, Colère, Sommeil, Sourire, et Surprise) et cela s'est fait à partir d'une sélection manuelle. En utilisant l'algorithme de classification du k plus proches voisins (k-ppv), on peut calculer les distances entre les vecteurs références (vecteurs représentants les séquences d'images utilisées comme références) représentants les différentes émotions et le vecteur représentant l'émotion à reconnaître. Donc, ce processus sera répété pour chaque entrée de données (séquences d'images des expressions faciales inconnues). Ainsi, la distance étant considérablement plus petite indique qu'il y a une ressemblance entre le vecteur représentant l'émotion utilisée comme référence et le vecteur représentant l'émotion utilisée comme entrée. Ces expérimentations nous montre que cette méthode de reconnaissance est clairement discriminante et que dans ce sens, l'utilisation de la localisation et la direction du mouvement à partir des histogrammes est un modèle de caractérisation d'expressions faciales efficace.

3.4 Conclusion

Plusieurs méthodes de reconnaissance ont été présentées tout au long de ce chapitre. Parmi celles-ci, certaines ne pouvaient être appliquées pour des raisons matérielles ou parce qu'elles transgressaient certaines conditions du projet. Notons tout particulièrement toutes les techniques intrusives ainsi que celles utilisant des informations tridimensionnelles. Les systèmes automatisés pour la détection des différentes émotions à partir des vidéos du visage qui ont été signalés jusqu'ici montrent principalement le problème de la modélisation spatiale des expressions faciales et, au mieux, peuvent détecter le mouvement dans les régions d'intérêt du visage. Nos méthodes développent d'autres aspects de la détection automatique des expressions faciales comparés aux travaux plus récents. La performance des méthodes proposées est invariable aux occlusions comme les lunettes et les cheveux de visage (barbe) tant que ces dernières n'occluent pas entièrement les points faciaux qui sont suivis. Afin d'implémenter efficacement le système, plusieurs fonctions OpenCV ont été utilisées. Cette bibliothèque permet notamment une gestion flexible des différents modules de reconnaissance. Finalement, cette architecture logicielle favorise les expérimentations de différentes configurations multi-classifieurs grâce à ses capacités de modification dynamique. Les résultats expérimentaux reliés à la reconnaissance des émotions seront présentés au chapitre suivant.

Chapitre 4

Résultats expérimentaux

4.1 Introduction

La détection automatique des expressions faciales par vision numérique est, comme démontrée précédemment, très complexe. Les différentes méthodes envisageables possèdent des avantages et des inconvénients qui doivent être considérés lors du design d'un système complet d'identification. Pour ce faire, il est primordial de valider les techniques choisies sur des ensembles de données relativement volumineux. Même si de telles séquences d'images ne représentent pas exactement les conditions réelles d'utilisation, elles procurent néanmoins une idée fiable du comportement des différents modules dans un environnement contrôlé. Ainsi, plusieurs expérimentations ont été réalisées afin d'évaluer la performance relative des différents algorithmes sélectionnés, tant sur le plan de la détection que de la reconnaissance. En effet, il est intéressant d'évaluer l'impact de la qualité de la détection du visage sur la performance globale du système. De plus, diverses configurations multiclassifieurs furent testées à l'aide de la banque des séquences d'images construite dans le cadre de ce mémoire et basée sur les bases de données Cohn Canade et MMI.

Cette dernière a été utilisée afin de comparer les différentes méthodes entre elles selon diverses conditions (c.-à-d. : éclairage, pose, occlusions, etc.). Ainsi, elle possède évidemment ses particularités spécifiques ainsi que ses qualités et défauts.

Ce chapitre exposera donc à la section 4.2 les différentes bases des séquences d'images retenues pour les expérimentations. Ensuite, la section 4.3 abordera le protocole expérimental utilisé lors des tests et finalement, la section 4.4 présentera de nombreux résultats expérimentaux, et à la section 4.5 nous finissons par une conclusion.

4.2 Banque de séquences d'images

Peu importe le problème de reconnaissance des expressions faciales, un point commun demeure toujours présent : la nécessité d'utiliser un ensemble de données volumineux, représentatif et standardisé. Cette particularité est effectivement primordiale pour la comparaison de techniques ou d'algorithmes, permettant ainsi une évaluation relative des performances. Cela étant dit, plusieurs points importants sont à considérer lors de la création ou de la sélection d'une banque de séquences d'images. Voici donc les particularités majeures à tenir en compte :

- **Nombre de personnes** La quantité d'individus dans une banque de séquences d'images est un des points les plus importants. En effet, ce nombre influence directement le niveau de difficulté de la banque : plus

la quantité est élevée, plus la tâche de reconnaissance sera difficile. De surcroît, la banque représentera davantage les tâches de reconnaissance en situations réelles, qui contiennent au minimum plusieurs milliers de personnes et aussi de multiples expressions à reconnaître.

- **Nombre de séquences d'images par individu** Une certaine quantité de séquences d'images est habituellement disponible pour chaque personne de la base de données. Un nombre élevé procure généralement un meilleur apprentissage du module de reconnaissance. Certaines banques d'images n'offrent cependant qu'une seule séquence d'images d'entraînement par individu, ce qui complique énormément le problème.

- **Hommes/femmes** Le ratio d'hommes et de femmes représente une particularité intéressante. Étant donné que certaines différences relatives au genre peuvent être modélisées efficacement, une banque ne contenant que des hommes ne pourra être de difficulté égale à une autre contenant 50% de femmes. Finalement, il y a habituellement un plus grand nombre de femmes portant les cheveux longs, ce qui peut influencer certains algorithmes de détection et de localisation du visage.

- **Arrière-plan** La plupart des banques de séquences d'images contiennent des séquences avec un arrière-plan neutre ou de couleur blanche. Les conditions d'acquisition ne sont par contre pas toujours idéales, occasionnant parfois la présence d'objets nuisibles ou d'arrière-plans complexes.

- **Dimension des séquences** La taille en pixels des séquences d'images a généralement beaucoup d'influence sur les algorithmes de reconnaissance. Les séquences d'images utilisées pour le test doivent avoir les mêmes dimensions que celles utilisées comme références. Dans le cas ou la dimension d'une séquence d'images est différente, on l'ajuste à l'aide d'un logiciel afin qu'elle garde les mêmes dimensions que les séquences d'images utilisées comme références.

- **Couleurs/tons de gris** L'utilisation de couleurs dans les techniques de reconnaissance est peu répandue. Celle-ci peut par contre s'avérer fort utile pour une détection des pixels représentant la peau ou pour la pré-classification d'individus de races différentes (reconnaissance des individus et leurs races). Les techniques utilisées dans ce mémoire n'utilisent pas de couleurs pour la reconnaissance des expressions faciales.

- **Coordonnées cartésiennes des composantes du visage** Ces informations supplémentaires s'avèrent particulièrement pratiques pour la comparaison de méthodes de reconnaissance. En effet, les résultats obtenus ne dépendant pas de la qualité de la détection du visage, des analyses plus robustes et plus représentatives peuvent être réalisées.

- **Cas particuliers ou difficiles** Des conditions spéciales peuvent également être présentes dans les bases de séquences d'images. Citons notamment les cas d'occlusions (exemple : lunettes fumées, chapeau, bandeau, cigares, etc.), de changements corporels (exemple : barbe, moustache, maquillage, verres de contact de couleurs, couleurs de cheveux,

cheveux détachés, etc.) et d'éclairage (exemple : incandescent, directionnel, etc.).

- **Pose** La pose de la tête de l'individu représente finalement un autre point important. En effet, la reconnaissance d'une expression faciale d'un visage de profil sera différente d'un visage face à la caméra.

Il y a donc plusieurs propriétés importantes à vérifier lors de la sélection d'une banque de séquences d'images pour fins d'expérimentations. Ces particularités s'appliquent également lors de la création d'une banque de séquences d'images.

Dans le cadre de ce mémoire nous avons utilisé notre propre banque de données en se basant sur les caractéristiques des bases de données Cohn Canade et MMI définies dans la section 3.2.2. Cette banque a été conçue spécialement pour le projet et était destinée aux expérimentations de l'application. Les facteurs ayant favorisés sa sélection résident entre autres dans la grande quantité des séquences d'images disponibles, toutes les séquences d'images qu'elle contient sont sous forme d'une vue frontale, et contenant cinq expressions (baîllement, colère, sommeil, sourire, surprise (section 3.2.2)). Afin de valider notre modèle de reconnaissance, nous avons aussi traité plusieurs séquences d'images provenant de la banque des séquences d'images MMI.

4.3 Protocole expérimental

Un protocole expérimental rigoureux permet avant tout la structuration des étapes nécessaires à l'exécution d'une certaine tâche. Dans le cas présent, l'objectif est de préparer les séquences d'images brutes pour les expérimentations.
Les étapes suivantes forment donc le protocole expérimental utilisé :

1. Compression des séquences d'images AVI utilisées pour l'expérimentation.

2. Numérisation des séquences d'images à l'aide de notre application avec un taux de numérisation de 30 images par seconde pour toutes les séquences d'images utilisées.

3. Vérification des coordonnées des caractéristiques du visage fournies avec la banque des séquences d'images.

4. Représentation des données à l'aide des modèles temporels.

5. Mise à jour des images de l'historique du mouvement (Motion History Images, MHIs) construites à l'aide des modèles temporels.

6. Localisation et identification des différentes régions (régions d'intérêt) du mouvement dans le visage (front, oeil gauche, oeil droit, joue gauche, joue droite, et la bouche) nécessaires pour la détection d'une expression faciale.

7. Apprentissage des différentes méthodes de reconnaissance.

8. Réalisation des expérimentations.

Un protocole rigoureux permet alors la répétabilité des expérimentations. Ainsi, quiconque désirant reproduire les résultats pour fins de vérification ou de comparaison pourra alors suivre la recette fournie par le protocole. De cette façon, la procédure utilisée est standard et assure qu'aucune modification ou altération n'est réalisée dans le processus.

4.4 Résultats expérimentaux

La présente section regroupe le fruit des expérimentations réalisées dans le cadre du projet de reconnaissance des expressions faciales par vision numérique. Les sections suivantes présenteront donc différents aspects ayant été évalués. Tout d'abord, la section 4.4.1 traitera de l'impact des métriques sur les algorithmes de reconnaissance. La section 4.4.2 présentera des expérimentations sur les modèles temporels.

4.4.1 Impact des métriques utilisées

La prochaine série d'expérimentations porte sur les différentes métriques utilisables lors de l'application de l'algorithme k-ppv. Ce dernier utilise en effet une métrique particulière pour déterminer l'ordre de proximité de la représentation test avec les différents prototypes d'apprentissage. Certaines des métriques envisageables ont été présentées au chapitre 2 (section 2.2.8),

citons notamment la métrique de city-block (L1), la métrique euclidienne (L2) ainsi que la métrique de Minkowski (Lm). Les autres métriques ne sont pas envisagées car elles ne peuvent pas être appliquées à notre méthode de classification.

Discussion Les figures 4.1, 4.2, 4.3, 4.4 et 4.5 illustrent les différents taux de reconnaissance obtenus sur la banque des séquences d'images utilisées pour les expérimentations en appliquant les métriques L1 (métrique de city-block), L2 (métrique euclidienne) et L3 (métrique de Minkowski avec m=3).

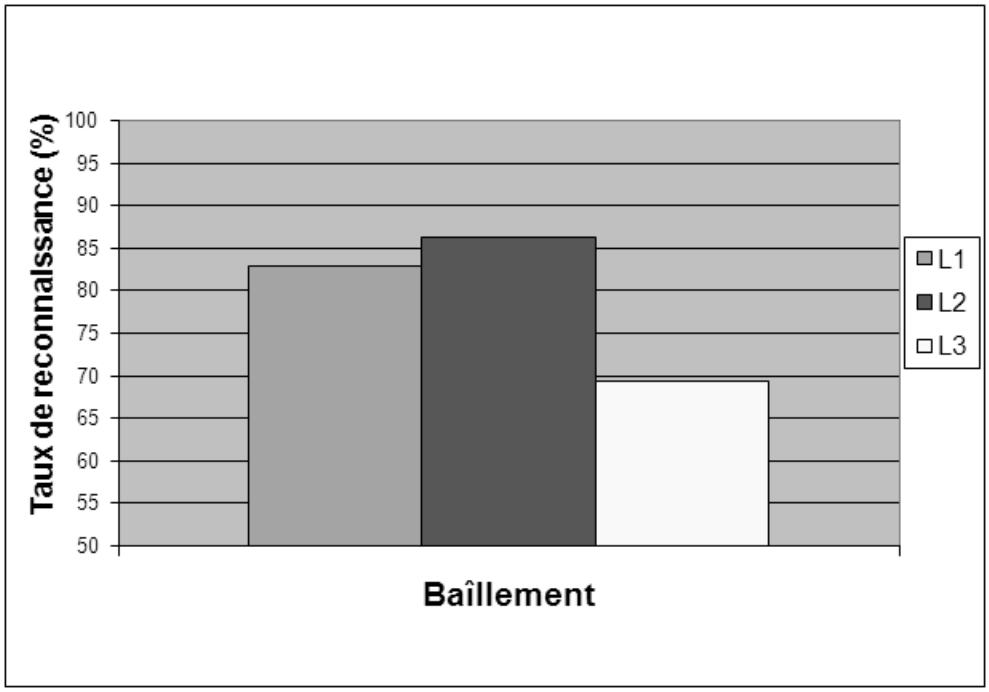


Figure 4.1: Impact des métriques utilisées (L1, L2 et L3) sur le taux de reconnaissance de l'émotion baîllement.

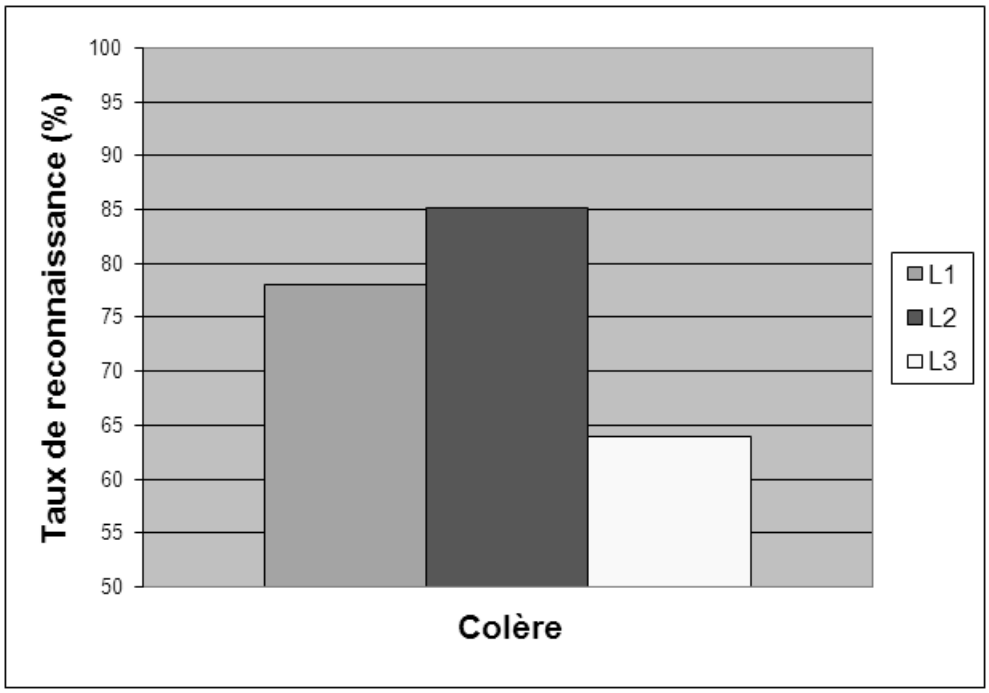


Figure 4.2: Impact des métriques utilisées (L1, L2 et L3) sur le taux de reconnaissance de l'émotion colère.

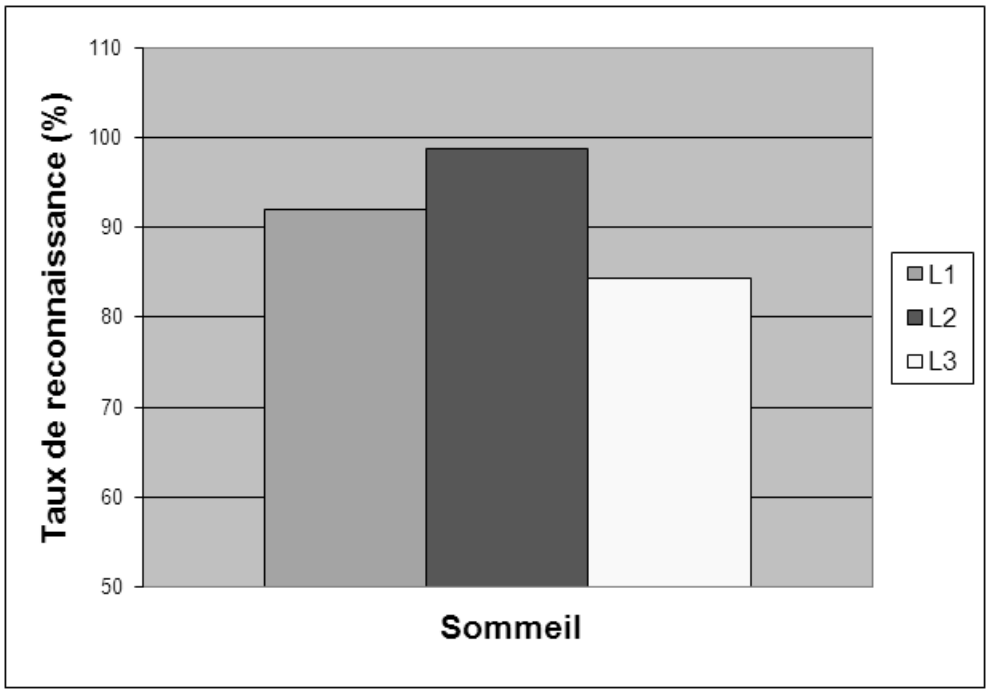


Figure 4.3: Impact des métriques utilisées (L1, L2 et L3) sur le taux de reconnaissance de l'émotion sommeil.

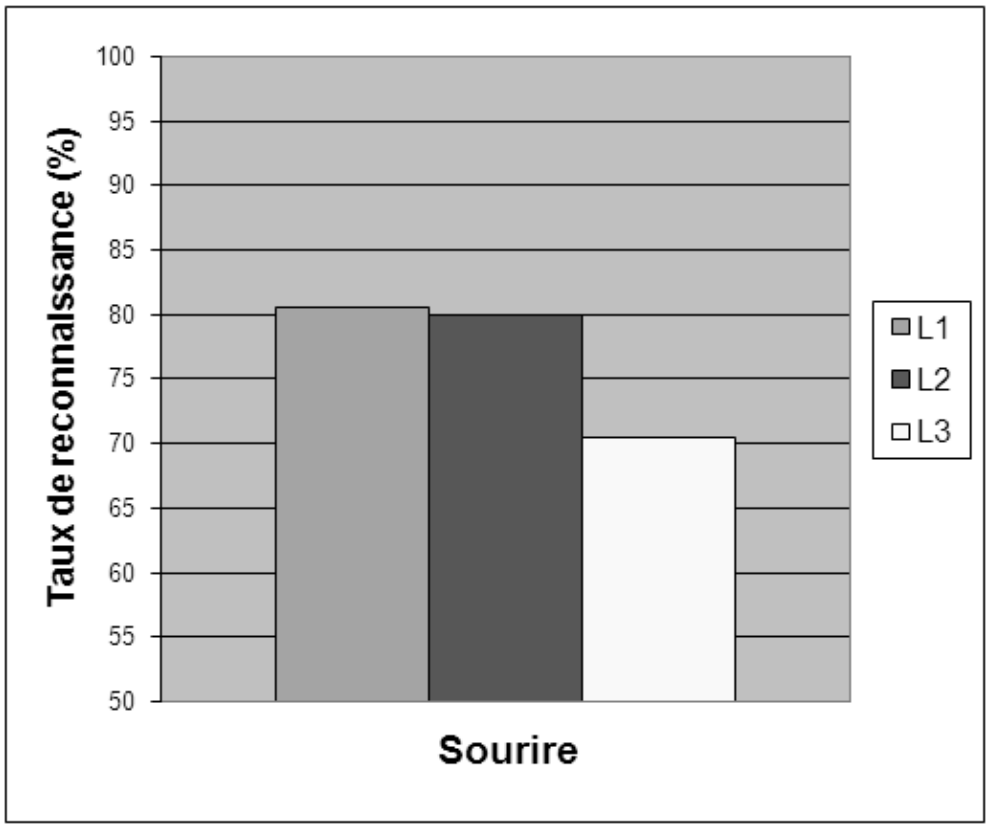


Figure 4.4: Impact des métriques utilisées (L1, L2 et L3) sur le taux de reconnaissance de l'émotion sourire.

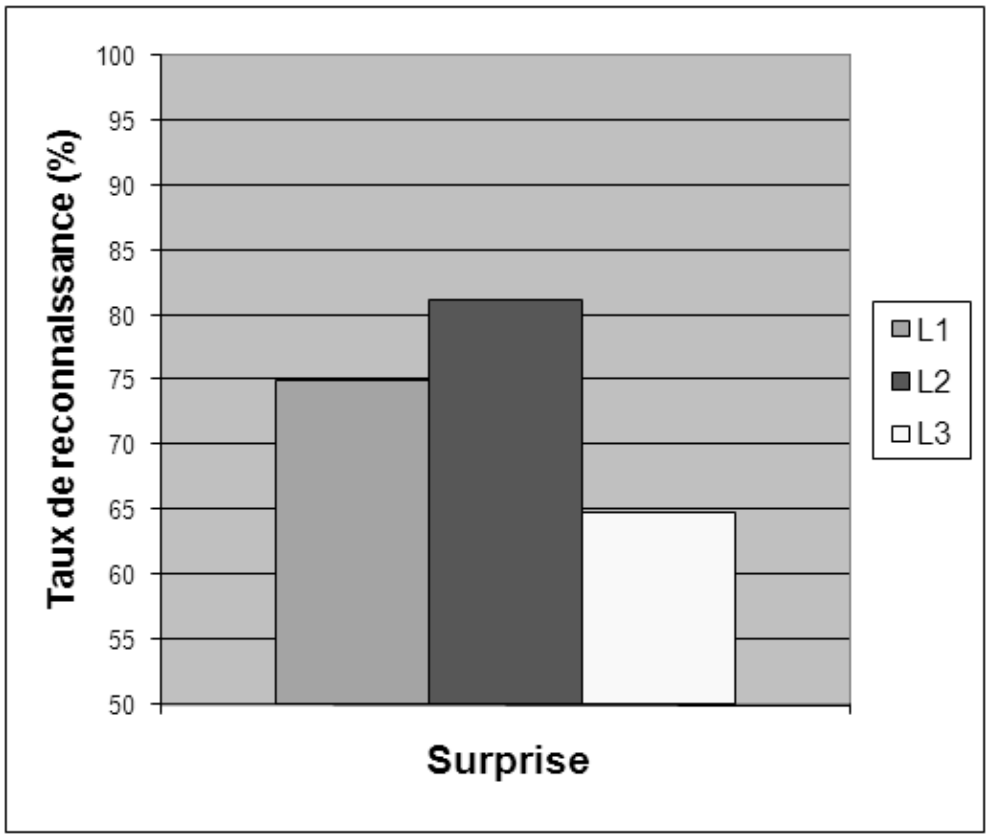


Figure 4.5: Impact des métriques utilisées (L1, L2 et L3) sur le taux de reconnaissance de l'émotion surprise.

Voici donc certains points d'analyse :

- La distance L2, ou distance euclidienne, procure les meilleurs taux de reconnaissance pour toutes les expressions faciales à reconnaître (baîllement, colère, sommeil, sourire, surprise).
- La distance L2 est préférable à la L1, identique à ce qui est utilisé par certains auteurs [84].
- Les taux de reconnaissance de la distance L1 et L3 ne permettent pas d'établir clairement la métrique la plus performante. La distance de city-block (L1) semble cependant être en moyenne légèrement plus appropriée par rapport à L3.

Suite à ces conclusions, toutes les expérimentations ont été réalisées uniquement à l'aide de la distance euclidienne (L2). Il serait cependant intéressant de vérifier l'impact de l'utilisation de différentes métriques dans un système de classification. Qui plus est, les erreurs de classifications commises en utilisant une certaine métrique ne sont peut-être pas identiques à celles réalisées par les autres classifieurs.

4.4.2 Validation de notre modèle

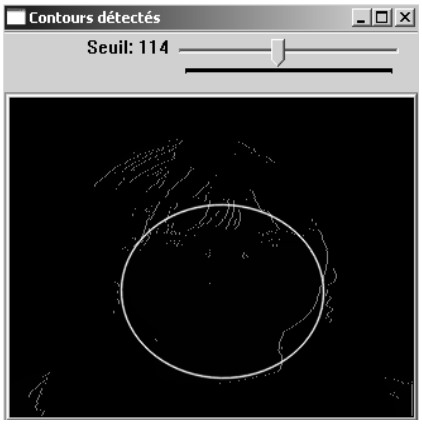
Afin de valider notre approche de détection automatique des expressions faciales, nous avons effectué plusieurs tests sur la banque de séquences d'images MMI (section 3.2.2). Ainsi, le nombre de vidéos dans la base de données

MMI est de plus de 1200 avec 41 sujets. La plupart des sujets dans cette base représentent les cinq émotions (baîllement, colère, sommeil, sourire, surprise) que notre propre base de données contient. Chaque séquence d'images représente une seule émotion. Cette séquence commence toujours par une expression neutre et finit par une expression neutre juste après la fin de l'expression qu'elle représente (3.5 secondes en moyenne). Pour nos expérimentations, nous nous sommes intéressés à des séquences d'images sous forme de vue frontale. Ainsi, la validation de notre modèle nécessitait plusieurs tests sur des séquences d'images provenant de la banque des séquences d'images MMI (400 séquences d'images testées) ainsi que notre banque de séquences d'images (50 séquences d'images testées). Les résultats des expérimentations sont présentés à la section 4.4.3.

Détection et localisation du visage

Les figures 4.6, 4.7 et 4.8 illustrent les résultats de détection et de localisation du visage par une ellipse en exploitant trois séquences d'images de trois visages différents de la base de données MMI.

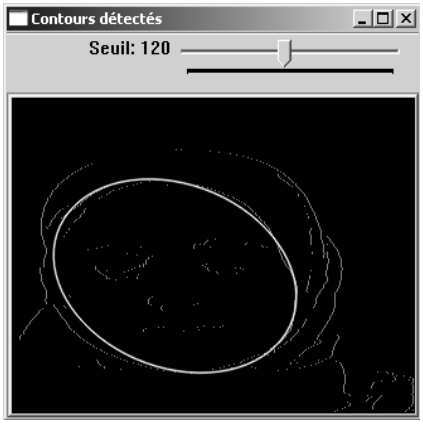
(a) (b)



(c)

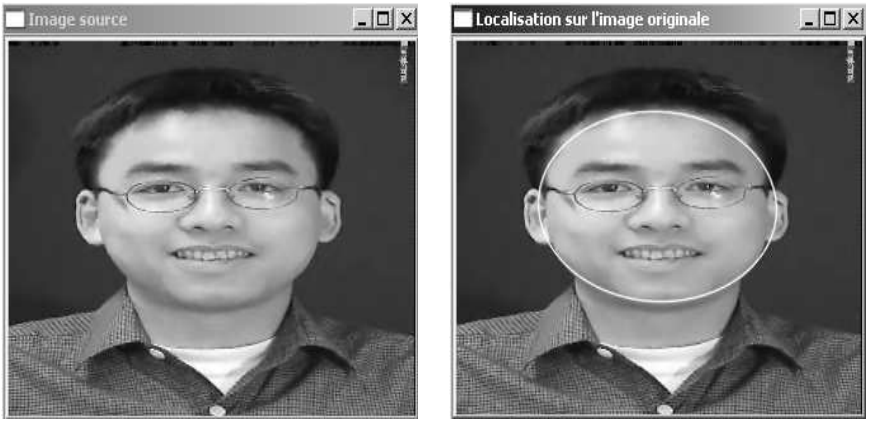
Figure 4.6: Résultat de détection du visage. (a) Image source. (b) Détection de contours et localisation du visage par une ellipse. (c) Localisation du visage sur l'image originale.

(a) (b)

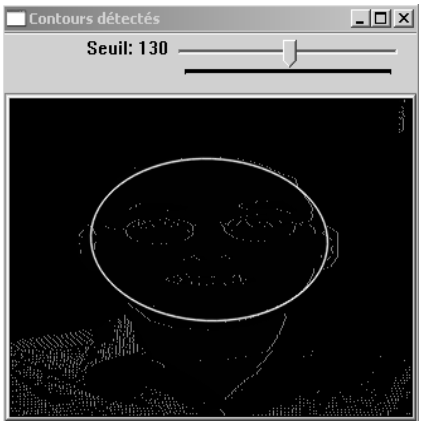


(c)

Figure 4.7: Résultat de détection du visage. (a) Image source. (b) Détection de contours et localisation du visage par une ellipse. (c) Localisation du visage sur l'image originale.



(a) (b)



(c)

Figure 4.8: Résultat de détection du visage. (a) Image source. (b) Détection de contours et localisation du visage par une ellipse. (c) Localisation du visage sur l'image originale.

Modèles temporels

- **Images de l'historique du mouvement**

Les figures 4.9 et 4.10 illustrent l'historique du mouvement de deux séquences d'images différentes de la même base de données MMI.

(a) (b)

Figure 4.9: Résultat de calcul de l'historique du mouvement. (a) Image source après un instant t (instant après le début de l'émotion). (b) Image de l'historique du mouvement correspondante.

L'historique du mouvement de l'émotion colère (figure 4.9) contient un mouvement important au niveau des régions des yeux et de la bouche.

(a) (b)

Figure 4.10: Résultat de calcul de l'historique du mouvement. (a) Image source après un instant t (instant après le début de l'émotion). (b) Image de l'historique du mouvement correspondante.

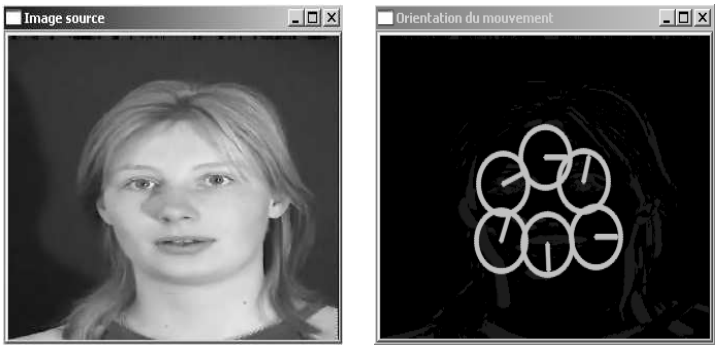
L'historique du mouvement de l'émotion surprise (figure 4.10) contient un mouvement moins important au niveau des régions des yeux, de la bouche, et le front.

- **Orientation du mouvement**

Les figures 4.11 et 4.12 illustrent les résultats de l'orientation du mouvement des deux séquences d'images traitées dans les figures 4.9 et 4.10. Ainsi, les flèches montrent la direction du mouvement de chaque région parmi les six régions d'intérêt.

(a) (b)

Figure 4.11: Résultat de calcul de l'orientation du mouvement de chaque région des régions d'interêt du visage. (a) Image source après un instant t (instant après le début de l'émotion). (b) Orientation du mouvement pour chaque région d'interêt.



(a) (b)

Figure 4.12: Résultat de calcul de l'orientation du mouvement de chaque région des régions d'interêt du visage. (a) Image source après un instant t (instant après le début de l'émotion). (b) Orientation du mouvement pour chaque région d'interêt.

4.4.3 Expérimentations sur les modèles temporels

Pour le but de détecter des expressions faciales à partir des images de l'historique du mouvement (MHIs) construites en appliquant la méthode des modèles temporels sur des séquences d'images représentant le visage humain sous forme de vue frontale, nous avons appliqué la technique de classification basée sur la notion de distances (distance euclidienne). Pour la reconnaissance d'une nouvelle entrée de données (séquence d'images représentant une expression faciale inconnue), nous calculons tout simplement la distance entre le vecteur d'entrée correspondant à l'histogramme des orientations du mouvement (table 3.3 chapitre 3) et chacun des vecteurs de l'histogramme des orientations prises à partir des séquences d'images références et représentant les cinq émotions traitées dans ce mémoire (bâillement, colère, sommeil, sourire, et surprise) dont leurs représentations graphiques sont présentées aux figures 3.1, 3.2, 3.3, 3.4 et 3.5. En appliquant l'algorithme de classification du k plus proches voisins (k-ppv), on peut calculer les distances entre les vecteurs références représentants les différentes émotions et le vecteur représentant l'émotion à reconnaître. Donc, ce processus sera répété pour chaque entrée de données (séquence d'images représentant une émotion inconnue). Ainsi, la distance étant considérablement plus petite indique qu'il y a une ressemblance entre le vecteur représentant l'émotion utilisée comme référence et le vecteur d'entrée de données (table 4.1). Ainsi, cette méthode de reconnaissance est clairement satisfaisante d'après cet ensemble de tests utilisant la localisation et la direction du mouvement à partir de son histogramme.

Émotion	Baîllement	Colère	Sommeil	Sourire	Surprise
T_1	**32,59**	105,19	180,11	197,42	208,06
T_2	78,98	**55,82**	280,71	229,20	228,18
T_3	240,36	304,58	**8,22**	448,74	350,47
T_4	153,40	230,24	311,82	**43,39**	124,98
T_5	189,79	324,40	92,43	89,38	**88,06**

Table 4.1: Résultats de reconnaissance en appliquant la méthode des distances euclidiennes.

NB : T_1 = *Baîllement.* T_2 = *Colère.* T_3 = *Sommeil.* T_4 = *Sourire.* T_5 = *Surprise.*

La table 4.1 présente les résultats de distances moyennes d'un test sur plusieurs séquences d'images de la base de données MMI utilisée pour les expérimentations. Les chiffres représentent les distances euclidiennes entre le vecteur test d'entrée T_i et les cinq vecteurs utilisés comme références (baîllement, colère, sommeil, sourire, surprise). Les chiffres en gras représentent les distances minimales pour chaque entrée de données (plus proches voisins).

La table 4.2 montre la performance de notre système de reconnaissance des expressions faciales pour chaque région choisie (front, yeux, joues, bouche) et le classifieur basé sur la combinaison de ces régions. Cette table indique que les régions choisies donnent une information valable pour la classification des émotions. Elle montre aussi que les joues qui sont couramment utilisées dans le domaine de la reconnaissance des expressions faciales donnent la performance la moins discriminante car d'après plusieurs tests, les résultats de calcul des orientations du mouvement de ces régions sont plus proches pour

Région	Baîllement	Colère	Sommeil	Sourire	Surprise
Front	83,5	84	75,22	84,54	90
Yeux	79,92	81	98,22	82,77	84,11
Joues	76,63	79,90	72,35	80	77,9
Bouche	153,40	230,24	311,82	43,39	124,98
Combinaison	**86,22**	**85,05**	**98,78**	**79,9**	**81,23**

Table 4.2: Performance de notre classifieur des expressions faciales.

toutes les types des émotions abordées (baîllement, colère, sommeil, sourire, et surprise). La table 4.2 montre aussi que le classifieur des expressions faciales combiné a un taux de reconnaissance moyen de 86,43, qui est plus haut que la plupart des classifieurs pour ces mêmes quatre régions.
La Table 4.3 montre la matrice de confusion de cette méthode de classification pour analyser en détail la limite de ce système de reconnaissance d'émotions. Les colonnes représentent l'expression faciale à reconnaître et

Émotion	Baîllement	Colère	Sommeil	Sourire	Surprise
Baîllement	**86,22**	13,9	0,46	4,18	3,12
Colère	11,2	**85,05**	0,2	2,34	2,95
Sommeil	0,5	1,7	**98,78**	0,85	1,03
Sourire	4,9	3,06	1	**79,9**	21,77
Surprise	3,79	1,5	1,3	16,43	**81,23**

Table 4.3: Matrice de confusion de la méthode de distance euclidienne.

les lignes représentent l'expression sélectionnée par cette méthode de classi-

fication.

Cette table indique que l'émotion sommeil est reconnue avec un taux de reconnaissance très élevé (98,78 %). Les autres émotions sont classifiées avec un taux de reconnaissance moyen de 83,35 %. La table 4.3 nous montre aussi que dans le domaine des expressions faciales, la surprise et le sourire sont souvent confondus (Figure. 2.16 du chapitre 2), ainsi que le baîllement et la colère. Pour rendre nos résultats moins confondus, on peut ajouter d'autres régions d'intérêt(région entre les yeux et le nez par exemple).

Il y a plusieurs raisons possibles qui donnent une mauvaise classification. La première est la présence des cheveux dans le visage et les lunettes, ce qui rend difficile la detection du mouvement présent dans quelques régions d'intérêt et aussi cache les expressions faciales. La deuxième raison est la variation au niveau de la taille et l'orientation du visage lors de la capture d'une séquence d'images, ce qui rend la taille du visage à traiter différente par rapport à la taille autorisée pour permettre une meilleure classification, et cela donne une mauvaise classification. Finalement, le bruit et l'occlusion sont toujours présents dans une certaine mesure.

4.5 Conclusion

Notre étude portant sur la méthode des distances a démontré la robustesse à certaines transformations pouvant survenir lors de la reconnaissance des expressions faciales. Cette technique peut même tolérer la présence de plusieurs effets négatifs simultanément, en autant que des limites spécifiques ne sont pas transgressées. Après avoir présenté les performances individuelles des modules de reconnaissance, l'impact des métriques utilisées lors de l'identification a été abordé. En effet, le choix de cette métrique est crucial et agit directement sur le taux de reconnaissance de notre méthode de classification. Parmi les distances testées, la distance L2 a permis l'obtention des meilleurs résultats. L'algorithme k-ppv donne des meilleurs résultats de classification mais il a ses propres désavantages : il est lent en temps d'exécution et il occupe beaucoup d'espace mémoire au moment de l'exécution. Pour analyser la confusion entre les différentes émotions, nous avons présenté la matrice de confusion de notre classifieur (table 4.3). Parmi les raisons possibles de la confusion : la production d'une émotion par plusieurs personnes (l'émotion surprise par exemple est différente d'une personne très surprit par rapport à une autre personne moins surprit, le baîllement d'une personne fatiguée par rapport à une autre personne moins fatiguée), similarité au niveau des résultats de test sur deux séquences d'images représentants deux émotions différentes. Notons que la majorité des expressions sont détectées avec un taux de reconnaissance plus élevé. Afin d'étendre les conclusions obtenues, des expérimentations supplémentaires pourraient être réalisées sur la banque des séquences d'images choisie. En effet, pour valider l'efficacité de notre ap-

proche, de nouvelles sections devraient être créées pour obtenir une banque d'apprentissage contenant plusieurs séquences images par individu.

Chapitre 5

Conclusion

Malgré tous les travaux réalisés au cours des dernières années, la détection automatique des expressions faciales demeure un problème complexe et non parfaitement résolu. Plusieurs sous problèmes incombent à cette tâche de détection et de reconnaissance et chacun d'eux n'est pas trivial. Il y a également de nombreuses conditions réelles influençant la performance d'un système. Cela étant dit, le système complet de détection et de reconnaissance proposé dans ce mémoire n'a pas la prétention d'être le meilleur de tous ou de résoudre toutes les situations problématiques. Il représente néanmoins une solution efficace respectant les contraintes initiales et accomplissant les différentes tâches demandées. De nombreuses techniques ont été présentées tout au long de ce mémoire, certaines furent adaptées ou tout simplement rejetées. Les prochains paragraphes résument brièvement les méthodes retenues ainsi que les principales conclusions qui ressortent de ce travail.

Résumé des modules spécifiques Le système de reconnaissance développé contient quatre phases principales accomplissant les tâches d'acquisition des images, de détection et de localisation du visage, de représentation de données à l'aide des modèles temporels et, finalement, de détection et de reconnaissance des expressions faciales. Ces différentes étapes sont accomplies

par des modules spécifiques. L'acquisition des images à partir des séquences vidéo est réalisée à l'aide de notre programme et le taux de numérisation est de 30 images par seconde pour toutes les séquences d'images. La détection et la localisation du visage sont réalisées ensuite par un module utilisant des arêtes grâce à son efficacité et sa robustesse par rapport aux autres modules de détection et de localisation. Le principe de base consiste à reconnaître des objets dans une image en appliquant la méthode de la transformée de Hough. Les particularités recherchées peuvent être des droites, des arcs de cercle, des formes quelconques, etc. Dans un contexte de détection et de localisation du visage, ce dernier est représenté par une ellipse. La représentation de données est réalisée à l'aide des modèles temporels présentés par A.F. Bobick et J.W. Davis [51] qui servent à représenter le mouvement en 2D dans un espace tridimensionnel (les deux dimensions spatiales et la dimension du temps). Le dernier module du système réalise la tâche de détection et de reconnaissance proprement dite. Ainsi, le système de classification des expressions faciales a été développé permettant l'utilisation des techniques de reconnaissance basées sur le calcul des distances. Pour le calcul des distances entre les vecteurs références représentant les différentes émotions et le vecteur représentant l'émotion à reconnaître, nous avons appliqué l'algorithme de classification du k plus proches voisins (k-ppv).

Atteintes des objectifs et respect des contraintes Le système proposé dans ce mémoire satisfait toutes les contraintes établies au début du projet. Il réalise en effet, toutes les opérations nécessaires à la détection automatique des expressions faciales. Le montage proposé, qui est composé

d'un ordinateur standard et d'un capteur, satisfait également les contraintes du coût associées au projet.

Travaux futurs La conception de ce projet a permis l'apprentissage d'une grande quantité d'informations. Qui plus est, plusieurs améliorations potentielles ont été observées à l'usage et pourraient faire l'objet de travaux futurs. Tout d'abord, de nombreuses expérimentations supplémentaires pourraient être réalisées, notamment à propos de notre système de classification. Les métriques du k-ppv, la complémentarité des méthodes et la sélection dynamique du classifieur représentent des sujets intéressants à investiguer. Des techniques d'intelligence artificielle comme les réseaux de neurones pourraient également être expérimentées comme fonction de décision.

Références

[1] Qiang Ji et Zhiwei Zhu, "Real Time and Non-intrusive Driver Fatigue Monitoring," 2004 IEEE Intelligent Transportation Systems Conference Washington, D.C., USA, Octobre 3-6.2004.

[2] Qiang Ji et Zhiwei Zhu, "Eye and gaze tracking for interactive graphic display," dans 2nd Inrernational Symposium on Smart Graphics, Hawthorne, NY, USA. 2002.

[3] Haisong Gu, Qiang Ji, et Zhiwei Zhu, "Active facial tracking for fatigue detection," IEEE Workshop on Applications of Computer Vision, Orlando, Florida, 2002.

[4] M. Pantic et L. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," IEEE Trans. Pattern Anal. Machine Intell., vol. 22, no. 12, pp. 1424-1445, 2000.

[5] P. Ekman et W.V. Friesen, Facial Action Coding System (FACS): Manual. Palo Alto: Consulting Psychologists Press, 1978.

[6] M.J. Black et Y. Yacoob, "Recognizing Facial Expressions in Image

Sequences Using Local Parameterized Models of Image Motion," lnt'l J. Computer Vision, vol. 25, no. 1, pp. 23-48, 1997.

[7] J.F. Cohn, A.J. Zlochower, J.J. Lien, et T. Kanade, "Feature-Point Tracking by Optical Flow Discriminates Subtle Differences in Facial Expression," Proc. Int'l Conf. Automatic Face and Gesture Recognition, pp. 396-401, 1998.

[8] I. Essa et A. Pentland, "Coding, Analysis Interpretation, Recognition of Facial Expressions," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 757-763, July 1997.

[9] S. Kimura et M. Yachida, "Facial Expression Recognition and Its Degree Estimation," Proc. Computer Vision and Pattern Recognition, pp. 295-300, 1997.

[10] T. Otsuka et J. Ohya, "Spotting Segments Displaying Facial Expression from Image Sequences Using HMM," Proc. Int'l Conf. Automatic Face and Gesture Recognition, pp. 442-447, 1998.

[11] M. Wang, Y. Iwai, et M. Yachida, "Expression Recognition from Time-Sequential Facial Images by Use of Expression Change Model," Proc. Int'l Conf. Automatic Face and Gesture Recognition, pp. 324-329, 1998.

[12] G.J. Edwards, T.F. Cootes, and C.J. Taylor, "Face Recognition Using Active Appearance Models," Proc. European Conf. Computer Vision, vol. 2, pp. 581-695, 1998.

[13] H. Kobayashi et F. Hara, "Facial Interaction between Animated 3D Face

Robot and Human Beings," Proc. Int'l Conf. Systems, Man, Cybernetics, pp. 3,732-3,737, 1997.

[14] H. Hong, H. Neven, et C. von der Malsburg, "Online Facial Expression Recognition Based on Personalized Galleries," Proc. Int'l Conf. Automatic Face and Gesture Recognition, pp. 354-359, 1998.

[15] C.L. Huang et Y.M. Huang, "Facial Expression Recognition Using Model-Based Feature Extraction and Action Parameters Classification," J. Visual Comm. and Image Representation, vol. 8, no. 3, pp. 278-290, 1997.

[16] M.J. Lyons, J. Budynek, et S. Akamatsu, "Automatic Classification of Single Facial Images," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21, no. 12, pp. 1,357-1,362, 1999.

[17] C. Padgett et G.W. Cottrell, "Representing Face Images for Emotion Classification," Proc. Conf. Advances in Neural Information Processing Systems, pp. 894-900, 1996.

[18] M. Pantic et L.J.M. Rothkrantz, "Expert System for Automatic Analysis of Facial Expression," Image and Vision Computing J., vol. 18, no. 11, pp. 881-905, 2000.

[19] M. Yoneyama, Y. Iwano, A. Ohtake, et K. Shirai, "Facial Expressions Recognition Using Discrete Hopfield Neural Networks," Proc. Int'l Conf. Information Processing, vol. 3, pp. 117-120, 1997.

[20] Z. Zhang, M. Lyons, M. Schuster, et S. Akamatsu, "Comparison between Geometry-Based and Gabor Wavelets-Based Facial Expression

Recognition Using Multi-Layer Perceptron," Proc. Int'l Conf. Automatic Face and Gesture Recognition, pp. 454-459, 1998.

[21] J. Zhao et G. Kearney, "Classifying Facial Emotions by Backpropagation Neural Networks with Fuzzy Inputs," Proc. Conf. Neural Information Processing, vol. 1, pp. 454-457, 1996.

[22] L. Wiskott, "Labelled Graphs et Dynamic Link Matching for Face Recognition and Scene Analysis," Reihe Physik, vol. 53, Frankfurt a.m. Main: Verlag Harri Deutsch, 1995.

[23] M. Pantic et L.J.M. Rothkrantz, "Facial Action Recognition for Facial Expression Analysis from Static Face Images," IEEE ¿¡ DEFANGED.2911 Trans. Systems, Man, and Cybernetics, Part B, vol. 34, no. 3, pp. 1449-1461, 2004.

[24] H. Wu, T. Yokoyama, D. Pramadihanto, et M. Yachida, "Face and Facial Feature Extraction from Color Image," Proc. Int'l Conf. Automatic Face and Gesture Recognition, pp. 345-350, 1996.

[25] H. Yamada, "Visual Information for Categorizing Facial Expressions of Emotions," Applied Cognitive Psychology, vol. 7, pp. 257-270, 1993.

[26] J.N. Bassili, "Facial Motion in the Perception of Faces and of Emotional Expression," J. Experimental Psychology 4, pp. 373-379, 1978.

[27] V. Bruce, Recognizing Faces. Hove, East Sussex: Lawrence Erlbaum Assoc., 1986.

[28] K. Mase, "Recognition of facial expression from optical flow", IEICE Trans., vol. E74, no. 10, pp. 3474-3484, 1991.

[29] M.S. Bartlett, J.C. Harger, P. Ekman, et T.J. Sejnowski, "Measuring facial expressions by computer image analysis", Psychophysiology, 36, 253-263, 1999.

[30] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, et T.J. Sejnowski, "Classifying Facial Actions", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21 no. 10, pp. 974-989, 1999.

[31] J.F. Cohn, A.J. Zlochower, J. Lien, et T. Kanade, "Automated face analysis by feature point tracking has high concurrent validity with manual faces coding", Psychophysiology, vol. 36, pp. 35-43, 1999.

[32] Y. Tian, T. Kanade et J.F. Cohn, "Recognizing action units for facial expression analysis", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 2, pp. 97-115, 2001.

[33] B. Braathen, M.S. Bartlett, G. Littlewort, E. Smith et J.R. Movellan, "An Approach to Automatic Recognition of Spontaneous Facial Actions", Proc. IEEE Int'l Conf. Face and Gesture Recognition, pp 345-350, 2002.

[34] M. Pantic, et I. Patras, "Temporal modeling of facial actions from face profile image sequences", Proc. Int'l Conf. Multimedia and Expo, 2004.

[35] J. Xiao, T. Moriyama, T. Kanade et J.F. Cohn. "Robust Full Motion Recovery of Head by Dynamic Templates and Re-registration Techniques" Int'l Journal of Imaging Systems and Technology, vol. 13,pp. 85-84, 2003.

[36] T. Kanade, J. Cohn et Y. Tian, "Comprehensive database for facial expression analysis", Proc. IEEE Int.'l Conf. Face and Gesture Recognition, pp. 46-53, 2000.

[37] A.J. Fridlund, P. Ekman, et H. Oster, "Facial Expressions of Emotion: Review Literature 1970-1983," Nonverbal Behavior and Communication, A.W. Siegman and S. Feldstein, eds., pp. 143-224. Hillsdale NJ: Lawrence Erlbaum Assoc., 1987.

[38] J.A. Russell et J.M. Fernandez-Dols,"The Psychology of Facial Expression", eds. Cambridge: Cambridge Univ. Press, 1997.

[39] C.E. Izard, "Facial Expressions and the Regulation of Emotions," J. Personality and Social Psychology, vol. 58, no. 3, pp. 487-498, 1990.

[40] A. Doucet, N. de Freitas, et N. Gordon. An introduction to sequential Monte Carlo methods. In A. Doucet, N. de Freitas, and N. Gordon, editors, Sequential Monte Carlo Methods in Practice. Springer-Verlag, 2001.

[41] M. Isard et A. Blake. Condensation: conditional density propagation for visual tracking. Int. J. Computer Vision, 29(1):5-28, 1998.

[42] G. Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. Journal of Computational and Graphical Statistics, 5(1):1-25, 1996.

[43] Y. Lee, Y. Lin, et G.Wahba. Multicategory support vector machines. Technical Report TR 1040, U. Wisconsin, Madison, Dept. of Statistics, 2001.

[44] G. Littlewort-Ford, M. Bartlett, et J. Movellan. Are your eyes smiling? detecting genuine smiles with support vector machines and Gabor wavelets. In Proceedings of the 8th Joint Symposium on Neural Computation, 2001

[45] P. Baldi et Y. Chauvin. Smooth on-line learning algorithms for hidden Markov models. Neural Computation, 6(2), 1993.

[46] B. Yegnanarayana, G. PavanKumar et SukhenduDas; " One-Dimensional GaborFiltering for Texture Edge Detection"; Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP-98), 21-23 December 1998, New Delhi, INDIA, pp 231-237.

[47] J. M. Vincent, D. J. Myers, et R. A. Hutchinson, "Image feature location in multi-resolution images using multi-layer perceptrons," in Neural Networks for Vision, Speech and Natural Language, R. Lingard, D. J. Myers, and C. Nightingale, Eds. London, U.K.: Chapman and Hall, 1992, pp. 13-29.

[48] M. Kunert D. M. Gavrila et U. Lages : A multi-sensor approach for the protection of vulnerable traffic participants - the protector project. Dans Proc. of the IEEE Instrumentation and Measurement Technology Conference, volume 3, pages 2044-2048, 2001.

[49] B.D. Lucas et T. Kanade : An iterative image registration technique with an application to stereo vision. Dans IJCAI81, pages 674-679, 1981.

[50] B. K. P. Horn et B. G. Schunck : Determining optical flow. Dans Artificial Inteligence, volume 17, pages 185-203, 1981.

[51] Davis, J. et A.Bobick. The representation and recognition of human movement using temporal templates. In Proc. Comp. Vis. And Pattern Rec., pages 928-934, June 1997.

[52] A.F. Bobick et J.W. Davis, "the Recognition of Human Movement using Temporal Templates", IEEE trans. Pattern Analysis and Machine Intelligence, vol 23, pp. 257-267, 2001.

[53] M. Pantic, et I. Patras, "Temporal modeling of facial actions from face profile image sequences", Proc. Int'l Conf. Multimedia and Expo, 2004.

[54] M. Isard et A. Blake, "Condensation - Conditional Density Propagation for Visual Trackint", Int'l. J. Computer Vision, pp. 5-28, 1998.

[55] J. S. Pan, Y. L. Qiao, et S. H. Sun, "A fast K nearest neighbors classification algorithm." lEICE Trans. Fundamentals. E87-A, no. 4,2004 (accepted)

[56] C. Domeniconi, D. Gunopulos, et J. Peng. Large margin nearest neighbour classifiers. IEEE Transactions on Neural Networks, 16(4):899-909, 2005.

[57] Castleman, K.R. (1996). Digital Image Processing. New Jersey : Prentice Hall.

[58] Jingdong Wang, Changshui Zhang, et Heung-Yeung Shum, "FACE IMAGE RESOLUTION VERSUS FACE RECOGNITION PERFORMANCE BASED ON TWO GLOBAL METHODS" www.cs.ust.hk/ welleast/tsinghua/publishpapers/ACCV04.pdf

[59] Erik Hjelmas et Boon Kee Low: Face detection: A survey. Computer Vision and Image Understanding, 83(3):236-274, September 2001.

[60] Ming-Hsuan Yang, David J. Kriegman et Narendra Ahuja, Detecting faces in images: A survey IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 24(1), pages 34-58, 2002.

[61] Robert Bergevin : Vision numérique : aspects cognitifs (notes de cours GEL-64793). Université Laval, Automne 2000.

[62] Schmitt, M. et Mattioli, J. (1993). Morphologie mathématique. Paris : Masson.

[63] D. Maio et D. Maltoni : Real-time face location on gray scale static images. Pattern Recognition, 33(9):1525-1539, September 2000.

[64] R. Seguier, Détection et localisation de visages dans des séquences d'images vidéo, Thèse soutenue à l'université de Rennes 1, 1995.

[65] Davis, J. et A. bobick. SIDEshow: A silhouette-based interactive dual-screen environment. MIT Media Lab Perceptual Computing Group Technical Report No. 457, MIT, 1998.

[66] N. Cristianini et J. Shawe-Taylor, Support Vector Machines, Cambridge, England, Camridge University Press 2000.

[67] P. Mitra, C.A. Murthy et S.K. Pal, "A probabilistic active learning support vector learning algorithm", IEEE trans. On Pattern Analysis and Machine Intelligence, pp. 413-418 2004.

[68] L.S. Chen, T.S. Huang, T. Miyasato, et R. Nakatsu, "Multimodal Human Emotion/Expression Recognition," Proc. Int'l Conf. Automatic Face and Gesture Recognition, pp. 366-371, 1998.

[69] G.W. Cottrell et J. Metcalfe, "EMPATH: Fface, Emotion, Gender Recognition Using Holons," Advances in Neural Information Processing Systems 3, R.P. Lippman, ed., pp. 564-571, 1991.

[70] L.C. De Silva, T. Miyasato, et R. Nakatsu, "Facial Emotion Recognition Using Multimodal Information," Proc. Information, Comm., and Signal Processing Conf., pp. 397-401, 1997.

[71] P. Ekman, Emotion in the Human Face. Cambridge Univ. Press, 1982.

[72] R. Gross et J. Shi, "The CMU Motion of Body (Mobo) Database", Tech. report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, Juin, 2001.

[73] A.M Martinez and R. Banavente, "The AR face database", CVC Tech. Report N° 24, Juin 1998.

[74] M.J. Lyons et J. Budynek and S. Akamatsu, "Automatic Classification of Single Facial Images", IEEE trans. On Pattern Analysis and Machine Intelligence, vol. 21(12), pp. 1357-1362, 1999.

[75] L. Maat, R. Sondak et P. Maja, "MMI Face Database", technical report MMI-BS-2004-02, Delft, 2004.

[76] DEMARTY (C.H), BEUCHER (S.). Efficient morphological algorithms

for video indexing. Content-Based and Multimedia Indexing, CBMI'99, Octobre 1999.

[77] LAWRENCE (S.), AUCLAIR-FORTIER (M-F), ZIOU (D.), BEGHADI (A.) Détection insensible au mouvement des frontières de plans franches et graduelles dans les séquences vidéo numériques. CORESA'2001, Dijon, 12-13 novembre 2001.

[78] WANG (H.L.) et CHANG (S.F). A Highly Efficient System for automatic Face Region Detection in MPEG Video. CirSys Video, 7(4), pp 615-628, août 1997.

[79] A.S. Piquemal et O. Le Cadet, On the Application of Edge Detection using Wavelets to Watermark Images: Definition of the Algorithms, Rapport interne de l'IMATI-CNR, Avril 2003.

[80] J. Korosec, L.Gyergyek et al., Face contour detection based on the Hough transformation, Automatika, Vol. 31, 1990.

[81] S. Carlsson, Multiresolution hough transform an eficient method of detecting pattern in images, Real Time Imaging., (1992), pp. 1090-1095.

[82] R. O. Duda et P. E. Hart, Use of the Hough Transformation to Detect Lines and Curves in Pictures, CACM, (15) (1972), pp. 11-15.

[83] W. Eric et L. Grimson and D.P. Huttenlocher, On the Sensitivity of the Hough Transform for Object Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence., (12) (1990), pp. 255-274.

[84] Ziad M. Hafed et Martin D. Levine : Face recognition using discrete cosine transform. International Journal of Computer Vision, 43(3):167-188, Juillet - Août 2001.