Université de Montréal

Analyse de mouvements faciaux à partir d'images vidéo

par

Mohamed Dahmane

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Thèse présentée à la Faculté des arts et des sciences
en vue de l'obtention du grade de
Philosophiæ Doctor (Ph.D.)
en informatique

Décembre 2011.

© Mohamed Dahmane, 2011

Université de Montréal

Faculté des arts et des sciences

Cette thèse intitulée :

Analyse de mouvements faciaux à partir d'images vidéo

présentée par

Mohamed Dahmane

a été évaluée par un jury composé des personnes suivantes:

Max Mignotte
(président-rapporteur)

Jean Meunier
(directeur de thèse)

Claude Frasson
(membre de jury)

Langis Gagnon
(examinateur externe)

James Turner
(représentant du doyen de la FAS)

Thèse acceptée le ____________________

SOMMAIRE

Lors d'une intervention conversationnelle, le langage est supporté par une communication non–verbale qui joue un rôle central dans le comportement social humain en permettant de la rétroaction et en gérant la synchronisation, appuyant ainsi le contenu et la signification du discours. En effet, 55% du message est véhiculé par les expressions faciales, alors que seulement 7% est dû au message linguistique et 38% au paralangage. L'information concernant l'état émotionnel d'une personne est généralement inférée par les attributs faciaux. Cependant, on ne dispose pas vraiment d'instruments de mesure spécifiquement dédiés à ce type de comportements.

En vision par ordinateur, on s'intéresse davantage au développement de systèmes d'analyse automatique des expressions faciales prototypiques pour les applications d'interaction homme–machine, d'analyse de vidéos de réunions, de sécurité, et même pour des applications cliniques. Dans la présente recherche, pour appréhender de tels indicateurs observables, nous essayons d'implanter un système capable de construire une source consistante et relativement exhaustive d'informations visuelles, lequel sera capable de distinguer sur un visage les traits et leurs déformations, permettant ainsi de reconnaître la présence ou absence d'une action faciale particulière.

Une réflexion sur les techniques recensées nous a amené à explorer deux différentes approches.

La première concerne l'aspect apparence dans lequel on se sert de l'orientation des gradients pour dégager une représentation dense des attributs faciaux. Hormis la représentation faciale, la principale difficulté d'un système, qui se veut être général, est la mise en œuvre d'un modèle générique indépendamment de l'identité de la personne, de la géométrie et de la taille des visages. La démarche qu'on propose repose sur l'élaboration d'un référentiel prototypique à partir d'un recalage par SIFT–flow dont on démontre, dans cette thèse, la supériorité par rapport à un alignement conventionnel utilisant la position des yeux.

Dans une deuxième approche, on fait appel à un modèle géométrique à travers lequel les primitives faciales sont représentées par un filtrage de Gabor. Motivé par le fait que les expressions faciales sont non seulement ambigües et incohérentes d'une personne à une autre mais aussi dépendantes du contexte lui–même, à travers cette approche, on présente un système personnalisé de reconnaissance d'expressions faciales, dont la performance globale dépend directement de la performance du suivi d'un ensemble de points caractéristiques du visage. Ce suivi est effectué par une forme modifiée d'une technique d'estimation de disparité faisant intervenir la phase de Gabor. Dans cette thèse, on propose une redéfinition de la mesure de confiance tout en introduisant une procédure itérative conditionnelle d'estimation du déplacement qui offrent un suivi plus robuste que les méthodes originales.

Mots clés: Vision par ordinateur, traitement d'images, reconnaissance d'expressions faciales, analyse d'émotions, analyse de visages, représentation des primitives de l'image, filtrage de Gabor, recalage, suivi de cibles.

ABSTRACT

In a face–to–face talk, language is supported by nonverbal communication, which plays a central role in human social behavior by adding cues to the meaning of speech, providing feedback, and managing synchronization. Information about the emotional state of a person is usually carried out by facial attributes. In fact, 55% of a message is communicated by facial expressions whereas only 7% is due to linguistic language and 38% to paralanguage. However, there are currently no established instruments to measure such behavior.

The computer vision community is therefore interested in the development of automated techniques for prototypic facial expression analysis, for human computer interaction applications, meeting video analysis, security and clinical applications.

For gathering observable cues, we try to design, in this research, a framework that can build a relatively comprehensive source of visual information, which will be able to distinguish the facial deformations, thus allowing to point out the presence or absence of a particular facial action.

A detailed review of identified techniques led us to explore two different approaches.

The first approach involves appearance modeling, in which we use the gradient orientations to generate a dense representation of facial attributes. Besides the facial representation problem, the main difficulty of a system, which is intended to be general, is the implementation of a generic model independent of individual identity, face geometry and size. We therefore introduce a concept of prototypic referential mapping through a SIFT–flow registration that demonstrates, in this thesis, its superiority to the conventional eyes–based alignment.

In a second approach, we use a geometric model through which the facial primitives are represented by Gabor filtering. Motivated by the fact that facial expressions are not only ambiguous and inconsistent across human but also dependent on the behavioral context; in this approach, we present a personalized facial expression recognition system whose overall performance is directly related to the localization performance of a set of facial fiducial points. These points are tracked through a sequence of video frames by a modification of a fast Gabor phase–based disparity estimation technique. In this thesis, we revisit the confidence measure, and introduce an iterative conditional procedure for displacement estimation that improve the robustness of the original methods.

Keywords: Computer vision, image processing, facial expression recognition, emotion analysis, face analysis, feature representation, Gabor filtering, registration, tracking.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLES

GLOSSAIRE

AFEA: Automatic facial Expression Analysis

AU: Action Unit

AWN: Active Wavelet Network

DFFS: Distance From Feature Space

DIFS: Distance In Feature Space

EBG: Elastic Bunch Graph

EF: Expression Faciale

FACS: Facial Action Coding System

FEP: Facial Expression Program

FER: Facial Expression Recogintion

FOA: Focus Of Attention

GWN: Gabor Wavelett Network

HMM: Hidden Markov Model

HOG: Histogram of Oriented Gradient

LBP: Local Binary Patterns

LDA: Linear Discriminant Analysis

LLE: Local Linear Embedding

PCA: Principal Component Analysis

PDM: Point Distribution Model

REF: Reconnaissance d'Expression Faciale

SIFT: Scale Invariant Feature Transform

SVD: Singular Value Decomposition

SVM: Support Vector Machine

À mon pépitou,

REMERCIEMENTS

Je souhaite adresser mes remerciements les plus sincères aux personnes qui m'ont apporté leur aide morale et physique pour la réalisation de cette thèse.

Mon directeur de recherche Jean Meunier, professeur titulaire et Directeur du laboratoire imagerie au DIRO, pour avoir accepté de diriger cette thèse. Sa disponibilité constante, son soutien moral et ses critiques associées à ses judicieux conseils ont largement contribué à l'aboutissement de ce travail. L'apport de ses orientations et ses remarques à la fois rigoureuses et objectives a été plus qu'indispensable à la concrétisation de cette recherche.

Docteur Sylvie Cossette de m'avoir accepté dans son unité de recherche à l'Institut de Cardiologie de Montréal et pour son apport financier durant le début de ces travaux.

Professeur Max Mignotte, pour l'honneur qu'il me fait en présidant le jury de cette thèse.

Monsieur Claude Frasson, professeur titulaire et Directeur du laboratoire HERON au DIRO, et Monsieur Langis Gagnon, chercheur principal au CRIM et Directeur de l'équipe vision et imagerie pour avoir accepté d'être membre de jury de cette thèse.

Je leur exprime mes remerciements, les plus vifs.

Un grand merci à tous mes collègues des laboratoires d'imagerie et de vision–3D, pour les échanges sympathiques, scientifiques ou non. Pour ne pas en oublier certains, je les laisserai se reconnaître.

Finalement, mes remerciements vont à toute ma famille particulièrement mes parents et spécialement à ma douce moitié envers qui je ne peux que être un sincère redevable.

Chapitre 1

L'ACTION FACIALE

1.1 Psychologie des mouvements faciaux

Dans le passé, Duchenne[1] avait identifié les muscles qui étaient à l'origine de différentes expressions faciales. Darwin comparait, plus tard, les expressions faciales de l'humain à celles des animaux [45].

Au cours des trente dernières années, les expressions faciales deviennent une partie intégrante des émotions universelles [63]. En 1971, Izard publia la première version de sa théorie des émotions différentielles. En 1972, Ekman établit les bases de la théorie neuroculturelle. Ce dernier mit en évidence des expressions de base universellement reconnaissables [52]. Ces expressions, dont on discute toujours la part de l'inné de celle de l'appris, correspondent aux sept émotions : la *neutralité*, la *joie*, la *tristesse*, la *surprise*, la *peur*, la *colère* et le *dégoût*.

En psychologie d'expression faciale, les éléments qui encadrent les émotions basiques et leur modélisation connues par le FEP[2] se basent sur les suppositions suivantes [153]:

1. Les expressions faciales constituent un petit ensemble fini d'expressions.

2. Elles peuvent être distinguées en utilisant des attributs spécifiques.

3. Se rapportent aux états internes, habituellement, les émotions de la personne.

4. L'universalité de la configuration et de la signification.

[1] À qui on associe le fameux *Duchenne smile*

[2] FEP: *Facial Expression Program*

Par ailleurs, lors d'une conversation libre, d'autres types de signes faciaux ont été considérés. Entre autres, les mouvements perçus sur le front du locuteur symbolisent des signes de ponctuation ou d'interrogation. Alors que chez l'auditeur, permettent d'appuyer la conversation en indiquant l'accord ou l'appel d'information [152].

1.2 Reconnaissance automatique des expressions faciales

La représentation du visage est un sujet d'intérêt pour les chercheurs dans le domaine de la reconnaissance des expressions faciales aussi bien dans la communauté des psychologues que celle des chercheurs en vision par ordinateur. Cet intérêt semble devenir majeur en identification de visages humains où beaucoup d'efforts ont été consacrés, particulièrement cette dernière décade. La double dissociation entre l'identification de visage et la reconnaissance d'expression faciale chez les patients avec dysfonctionnent au niveau du cerveau énonce l'évidence selon laquelle les deux tâches reposent sur des représentations et/ou des mécanismes de traitement différents au niveau du cerveau humain.

Alors que les méthodes globales (*holistic*) sont utilisées efficacement dans plusieurs techniques d'identification de visages, il a été démontré que les méthodes basées sur les attributs locaux (*non-holistic*) sont plus efficaces pour la reconnaissance des expressions. Ceci fournit une possible explication concernant la double dissociation [33], et nous amène à penser à la façon de mettre en pratique et de tirer profit des avancées enregistrées en représentation faciale dans le domaine de l'identification de visages.

Les concepts clés adressés, dans un système générique de reconnaissance d'expressions faciales, sont ceux de la **détection** du visage, l'**extraction** des caractéristiques faciales et enfin, basée sur une représentation particulière, la **reconnaissance**. Ceci est illustré par la figure (Fig. 1.1).

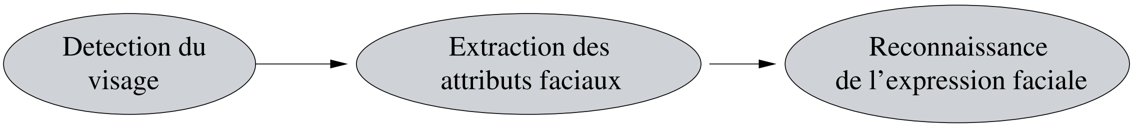

Detection du visage
Extraction des attributs faciaux
Reconnaissance de l'expression faciale

Figure 1.1. Structure générique d'un système automatique de reconnaissance d'expression faciale.

1.2.1 Détection et représentation du visage

Les premiers travaux de détection de visage et de reconnaissance d'expression faciale peuvent être trouvés dans [154]. Les techniques de détection et de suivi du visage, les plus récentes, peuvent être organisées en 4 catégories qui parfois peuvent avoir des aspects en commun. Elles sont rapportées dans ce qui suit.

1.2.1.1 Approches basées sur les composantes

Ce type d'approches se base sur la description individuelle des parties constituant le visage et leurs relations.

L'algorithme le plus populaire de localisation de visage est celui de Viola et Jones [182] qui se base sur l'*AdaBoost*, où l'on combine de nombreux descripteurs simples (fonctions de Haar), parmi lesquels un dopage (*boosting*) permet de retenir les plus discriminants [183] (voir Section 2.2). Heisele [83] a employé des machines à vecteurs de support (SVM) (voir Section 2.5 sur les SVM) comme classifieur de base et la LDA[3] pour combiner les résultats des classifieurs individuels relatifs aux différentes composantes du visage. Une nouvelle méthode, basée sur un arbre de décision et la LDA, a été développée afin d'améliorer le taux de détection en présence d'occultations [89]. On estime qu'*AdaBoost* performe mieux et qu'il est beaucoup plus rapide que le SVM.

L'analyse en composantes indépendantes sur un espace résiduel local [98], une

[3] LDA : *Linear Discriminant Analysis.*

technique supposée être robuste aux changements d'illumination et de pose, a été proposée pour l'identification de personnes à partir d'images faciales.
Dans [149], on estime qu'un classifieur basé sur des SNoW (Sparse Network of Winnows) est capable d'apprendre, à partir d'images naturelles, la frontière de discrimination entre les "objets" visages et non-visages.

1.2.1.2 Approches basées sur l'apparence

Classiquement, dans ces approches globales, il s'agit de projeter les patterns extraits (images) dans un espace de plus petite dimension, dans lequel on tente de trouver une fonction discriminante décrivant le modèle d'appartenance [141].

Sung et Poggio ont développé un système basé sur des fonctions de distributions pour la détection de visage [165]. Ils ont montré comment les distributions des patterns d'une classe donnée peuvent être apprises à partir d'exemples positifs (visages) et négatifs (non-visages). Le système utilise un perceptron multicouche comme classifieur.

La distance de l'espace des attributs (DFFS[4]), un concept basé sur l'analyse en composantes principales (ACP), a été utilisée pour la classification et la détection d'objets, en particulier pour la détection de visages [124]. La figure 1.2 donne une interprétation géométrique de la distance (DIFS[5]) relative au sous–espace propre F généré par les M premières valeurs propres, la DFFS correspond à l'espace résiduel orthogonal $\bar{F}$ généré par les $N - M$ composantes principales. Il est à noter que dans ce type de classifieurs, la répartition de l'information sur les espaces F et $\bar{F}$, et donc le choix de M, affecte directement leurs performances.

D'autres approches proposées par Rowley [151] et Kouzani [101] emploient des classifieurs basés sur des réseaux de neurones pour localiser des visages, dont l'apprentissage utilise un corpus d'images positives (visages) et négatives (non-visages).

[4] DFFS: *Distance From Feature Space.*

[5] DIFS: *Distance In Feature Space.*

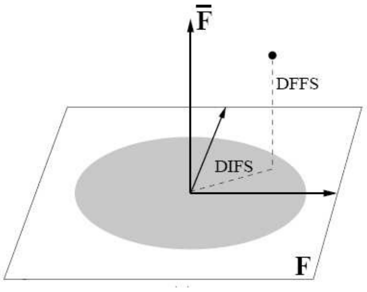


Figure 1.2. Décomposition de R^N en sous-espaces orthogonaux F et $\bar{F}$.

1.2.1.3 Appariement de gabarits

Les modèles actifs de forme (ASM) et les modèles actifs d'apparence (AAM) [30] ont souvent été employés en tant que techniques d'alignement en reconnaissance de visages [72].

Dans [77], on propose une technique basée sur les AAM pour le suivi de visage, on montre comment l'analyse en composantes principales est appliquée en cas de données manquantes pour modéliser les changements de forme et d'apparence. On présente aussi une nouvelle méthode de normalisation pour la construction du AAM, rendant le suivi de visages plus robuste aux occultations dont la présence altère sévèrement les performances des AMMs classiques.

Une nouvelle méthode d'alignement de visage, appelée AWN[6], qui s'annonce robuste à l'illumination et aux occultations partielles et plus efficace par ses capacités de généralisation, a été proposée dans [19]. L'idée principale consiste à remplacer le modèle de texture basé sur l'ACP par un réseau d'ondelettes.

D'autres approches employant l'information sur la structure en plus de la similarité locale permettent plus de flexibilité dans l'ajustement des positions des points d'intérêt et dans la recherche de ces positions selon une fonction d'optimisation donnée sous certaines contraintes de configuration [29].

[6] AWN : *Active Wavelet Network*

McKenna [121], utilise les ondelettes de Gabor (voir Section 4.3) pour extraire des points caractéristiques du visage, un PDM[7] est employé comme contrainte globale pour corriger les fausses positions et contraindre le décrochage de points. Similairement, un modèle Bayésien de formes (BSM), comme le montre la figure 1.3, peut aussi être utilisé pour l'extraction des attributs faciaux.

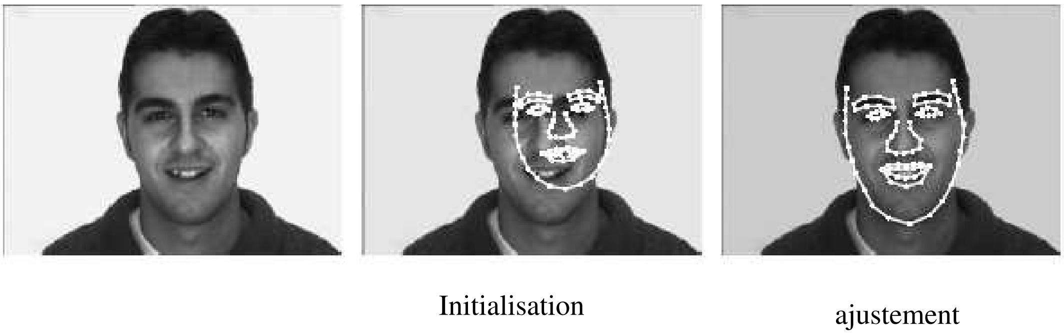


Figure 1.3. Extraction des attributs faciaux à l'aide de BSM [201].

1.2.1.4 Approches géométriques

Plusieurs méthodes géométriques [161, 163, 216] emploient des informations à priori au sujet du visage, et contraignent la recherche des attributs faciaux par des heuristiques ou en se servant de mesures anthropologiques impliquant des mesures d'angles ou de distances normalisées. Cette modélisation de l'information structurelle permet de valider les attributs faciaux localisés.

L'EBG[8] est un modèle typique des approches géométriques, c'est un graphe labélisé où les nœuds représentent les points d'intérêt (attributs faciaux) à l'aide de caractéristiques locales sous formes de jets, une représentation vectorielle de la transformée en ondelettes de Gabor (voir Section 4.3). L'information topologique est

[7] PDM: Point Distribution Model.

[8] EBG: Elastic Bunch Graph.

représentée par la structure du graphe, ceci permet de contraindre les jets sous certaines configurations géométriques. Basé sur une métrique particulière, un procédé d'appariement de graphes *(graph matching)* permet de localiser les attributs faciaux par une recherche exhaustive du graphe le plus similaire.

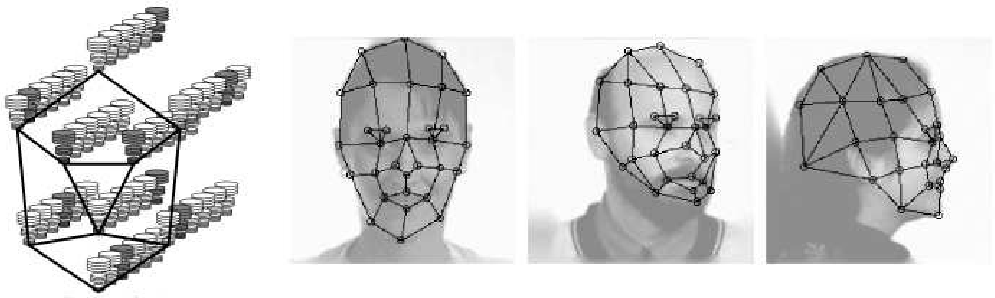


Figure 1.4. Représentation générique de visages par EBG [197].

D'une façon assez comparable aux EBGs, une architecture hiérarchique utilisant deux niveaux de réseaux d'ondelettes de Gabor GWN[9] a été proposée dans [62]. Le premier niveau représente un GWN correspondant au visage entier et sert à localiser, dans une image, une région susceptible de contenir un visage. Le deuxième niveau utilise d'autres GWNs qui servent à modéliser, plus localement, les attributs faciaux qui sont localisés sous une contrainte qu'impose le GWN du niveau supérieur.

En résumé, dans la littérature, les trois systèmes les plus réussis de détection de visage sont Rowley *et al.*, Roth *et al.* et Viola & Jones [58, 203], qui relèvent tous les trois des approches basées sur l'apparence.

[9] GWN: Gabor Wavelet Network.

1.2.2 Les mouvements faciaux

En sciences du comportement, le système le plus connu pour le codage des mouvements faciaux est le FACS[10] qui est un système plus complet et psychométrique rigoureux [28]. Chaque mouvement est décrit par une combinaison d'un ensemble d'actions unitaires (AUs)[11] qui repose sur des connaissances de l'anatomie du visage. L'apprentissage et le codage manuels des expressions humaines à partir de mesures psychophysiologiques sont deux tâches très ardues nécessitant une certaine spécialisation [104]. Torre *et al.* décrit beaucoup plus en détail l'apprentissage nécessaire et le processus d'encodage qui est laborieux [47].

En vision par ordinateur, plusieurs chercheurs tentent de résoudre le problème de la reconnaissance automatique des mouvements faciaux. Deux états de l'art [57, 139] dressent une collection de plusieurs recherches publiées ces dernières années. Pratiquement, toutes les méthodes développées se ressemblent du fait qu'elles extraient d'abord un ensemble d'attributs à partir d'images ou de séquences vidéo (voir Section 1.2.1), lequel constituera ensuite l'entrée d'un classifieur qui produira en sortie la catégorie de l'émotion émise. Ces techniques diffèrent principalement dans la sélection des éléments représentatifs du visage, considérés comme attributs faciaux à extraire, et aussi dans le type du classifieur employé. Une étude comparative ayant porté sur une variété de classifieurs montre que les performances sont relativement assez comparables. Cependant, on relève qu'un élément d'importance réside dans la précision et la spécificité des caractéristiques et des attributs visuels sélectionnés plutôt que dans le choix du type du classifieur [25].

Par ailleurs, notons qu'un grand nombre de techniques d'analyse automatique d'expressions faciales découlent des travaux antérieurs du psychologue P. Ekman. Les approches adoptées peuvent se classer en trois catégories principales. Il y a

[10] FACS : *Facial Action Coding System*[51].

[11] Une AU (*Action Unit*) exprime le changement le plus élémentaire visuellement discriminable dans une expression faciale.

celles utilisant les actions unitaires (AUs) [50] et celles qui se servent d'un ensemble d'expressions prototypiques [53]. Plus récemment, d'autres dimensions émotionnelles ont été considérées [67, 156].

1.2.2.1 Approches basées sur la décomposition en AUs

La reconnaissance par décomposition sous forme de (AUs) [5, 27] considère les contractions des différents muscles (Fig. 1.5) dans le système basique (FACS) d'encodage des expressions faciales, ce qui permet de:

- catégoriser toutes les déformations visuelles possibles sous forme de 44-AUs.
- offrir un support standard pour l'analyse d'expression faciale.

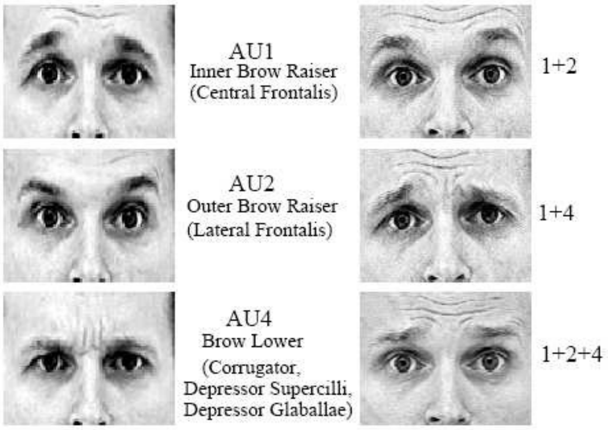


Figure 1.5. Décomposition d'une expression faciale en AUs.

En analysant l'effet sur la classification, le fait de considérer séparément ou conjointement les AUs, on parvient à établir que la corrélation naturelle entre les AUs est d'une grande importance car ces dernières seront correctement classifiées en considérant le visage tout entier. Le taux de classification diminue lorsqu'elles sont considérées séparément. Ce n'est pas acquis pour autant, car des actions adjacentes dans une combinaison d'AUs induisent une co–déformation, un effet semblable au phénomène de la coarticulation de la parole. On souligne en outre, l'importance d'établir une grande base de données, où les AUs seront accomplies individuellement, pour faciliter les futures recherches [195]. Certaines AUs sont réalistes mais d'autres impliquent des muscles dont le mouvement n'est pas volontaire donc la simulation de ceux-ci est faussée dans la tentative de les représenter individuellement. D'autre part, certaines AUs ne peuvent pas être individuelles comme AU26 (jaw drop) qui indique un mouvement d'abaissement du menton induit un abaissement de la lèvre inférieure impliquant AU25 (lips part).

1.2.2.2 Approches basées sur les EFs prototypes

Dans ce type d'approches [57, 139], on tente de trouver une représentation des déformations des attributs faciaux qui servirait de cadre au processus de reconnaissance.
En effectuant le suivi des points caractéristiques du visage, une évaluation de la quantité du mouvement facial perceptible permet de classer les différentes expressions dans l'une des classes des expressions de base. Les travaux récents sur l'analyse et l'identification d'expressions faciales ont employé un ensemble ou sous–ensemble d'expressions de base.

Dans [12] on récupère le mouvement non–rigide des attributs faciaux et on tente de dériver des prédicats à partir de paramètres locaux. Ces prédicats constitueront les entrées d'un système de classification à base de règles. Le système développé dans [54] utilise une technique de flux optique basée sur le calcul de l'énergie associée au mouvement spatio–temporel, pour l'identification d'expressions faciales. Une comparaison de différentes techniques de représentation de gestes faciaux, incluant le flux optique, est dressée dans [50]. Pour l'apprentissage des classes de variétés au sens topologique (*manifolds*) dans l'espace des attributs exprimé par un AWN[12], Hu [86] étudia deux

[12] AWN : *Active Wavelet Networks.*

types de méthodes non–linéaires de réduction de dimensions, il constata que la LLE[13] est plus appropriée à la visualisation des variétés des expressions faciales, par rapport à la méthode de Lipschitz [14, 94].

Plus récemment, un modèle probabiliste a été développé pour modéliser la variété des expressions faciales dans un espace multidimensionnel dans lequel, des expressions semblables se projettent en points concentrés autour d'une variété donnée représentant une expression particulière. Ainsi, une séquence d'expressions est représentée par un *patch* dans cet espace. Un modèle de transition probabiliste sert à déterminer la vraisemblance [20].

L'approche proposée dans [23] se sert d'un arbre augmenté de Bayes naïf (TANB)[14] comme classifieur d'expressions. Pour capturer l'aspect dynamique de celles–ci, une architecture multiniveau de modèles de Markov cachés (HMM) a été proposée. On estime qu'une segmentation automatique temporelle des vidéos a été possible.

Un réseau Bayésien dynamique peut être utilisé pour représenter l'expression faciale. Le modèle proposé dans [209] adopte une représentation par couches élémentaires. On distingue la couche de classification, la couche des actions unitaires (AUs), et la couche sensorielle qui agit comme récepteur visuel, où l'information correspond aux variables observables tels que les sourcils, les lèvres, etc.

La méthode proposée dans [68] présente un classifieur à deux étapes, premièrement des machines à vecteurs de support (SVMs) sont utilisées comme classifieurs par paires, chaque SVM est alors entraîné à distinguer entre deux émotions. Dans la deuxième étape, une régression logistique multinominale biaisée (MLR[15]) permet d'obtenir une représentation sous forme de distributions de probabilité sur l'ensemble des sept expressions de base, incluant l'expression neutre.

Dans [187], les auteurs emploient une forme modifiée de la décomposition en

[13] LLE: *Local Linear Embedding* [150].

[14] TANB: *Tree-Augmented Naïve Bayes.*

[15] MLR: *Multinomial Logistic Ridge Regression.*

valeurs singulières d'ordre supérieur (HOSVD) qui permet d'effectuer une décomposition de visages expressifs en deux sous–espaces, le sous–espace "expression" et le sous–espace "identité". L'isolation de l'identité, conduit simultanément, à une identification indépendamment des expressions et une reconnaissance d'expression indépendamment de l'identité de l'individu.

1.2.2.3 Approches basées sur les dimensions émotionnelles continues

Récemment, les efforts ont été orientés vers la façon de modéliser l'émotion en terme de dimensions affectives continues [156]. Fontaine *et al.* [67] soutiennent que la plupart des catégories affectives peuvent être exprimées par quatre dimensions émotionnelles à savoir, l'activation, l'anticipation, la dominance et la valence.

- *Activation*, indique l'intensité de l'émotion ressentie face au stimulus émotionnel.
- *Anticipation*, se réfère aux différents concepts d'anticipation, de passion, et d'être pris au dépourvu.
- *Dominance*, fait référence aux concepts de contrôle, d'influence, et de domination.
- *Valence*, cette dimension concerne le sentiment global de plaisance versus aversion.

Ramirez *et al.* [144] présenta un modèle basé sur les champs aléatoires conditionnels (CRF, Conditional Random Fields) pour la classification des différentes dimensions émotionnelles. Ce système permet d'intégrer au mieux les motifs acoustiques, toutefois il fait usage d'un outil commercial pour extraire les motifs visuels. Il n'est pas clair si les performances sont dues à l'approche proposée ou aux caractéristiques visuelles de haut niveau que permet d'extraire cet outil, telles que la détection des coins des yeux, l'intensité du sourire etc.

La méthode décrite dans [36] s'inspire de la vision humaine et de la théorie de l'attention. Basé sur la fréquence de l'information visuelle, l'article adresse le problème de sous–échantillonnage de données de large dimensionnalité (très longues séquences) tout en maintenant une bonne représentativité, aussi minimale soit–elle au sein de ces données, sans pour autant sacrifier la performance.

Dans l'article présenté en annexe G, nous présentons une méthode pour la reconnaissance d'émotions en terme de dimensions latentes (activation, anticipation, dominance, valence). On se sert des filtres d'énergie de Gabor pour représenter l'image faciale. À partir d'une grille uniforme, des histogrammes unidimensionnels sont utilisés pour intégrer, sur toute la cellule, les réponses des différents filtres. La reconnaissance utilise une machine multiclasse à vecteurs de support comme apprenant de dimensions émotionnelles.

1.2.3 Défis et enjeux

Malgré ces avancées, on est encore confronté à un défi majeur pour établir des systèmes automatiques de reconnaissance des émotions humaines, en effet le comportement émotionnel humain est subtil et multimodal, auquel se greffent d'autres défis et enjeux non moins importants que nous présentons dans ce qui suit.

1.2.3.1 Relatifs à la géométrie et à l'apparence

La classification d'expression faciale naturelle est souvent pavée d'ambiguïtés, en effet les limites entre expressions ne sont pas toujours clairement définies. Dans plusieurs situations, même pour l'être humain, la notion de contexte est fondamentale en interprétation, pour pouvoir identifier distinctement ces expressions. Pour la machine, il est encore difficile de requérir à ce type de perception. Par ailleurs, la reconnaissance automatique est confrontée aux problèmes de changements physionomiques dus au genre, à l'âge et à l'appartenance ethnique, ce qui place un autre degré de difficulté.

La résolution des images et la taille de la tête nous informent sur les détails présents sur le visage. Ces paramètres sont capitaux, quant à l'adoption d'une démarche de reconnaissance plutôt qu'une autre.

D'autres facteurs qu'on peut identifier influencent la manière dont le visage apparaît dans une scène. Les conditions d'éclairage et l'angle d'incidence de la lumière peuvent changer dramatiquement l'apparence d'un visage (Fig. 1.6). Si ces conditions peuvent quand même être atténuées dans un environnement contrôlé, il n'est pas de même dans une interaction naturelle libre, où les changements de l'orientation de la tête sont difficilement compensables.

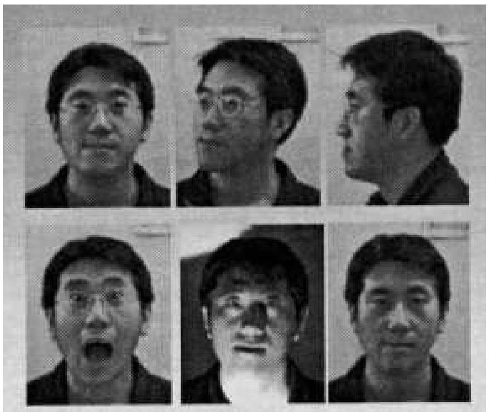


Figure 1.6. Exemple montrant la grande variabilité de l'apparence du visage [148].

1.2.3.2 Relatifs à un contexte réaliste

Alors qu'on assiste à des progrès notables en analyse automatique d'expressions faciales comme décrit un peu plus haut, beaucoup de questions demeurent encore ouvertes. L'un des problèmes majeurs réside dans la reconnaissance et l'interprétation des expressions faciales lors d'une interaction naturelle libre.

La basse résolution des images ainsi que la faible intensité des expressions sont parmi les facteurs qui compliquent l'analyse d'expression faciale [169]. On évoque aussi le problème de la reconnaissance dans les situations où l'on ne dispose pas de visage inexpressif (neutre), une représentation que construit mentalement l'être humain.

Par ailleurs, le mouvement global de la tête est une composante normale dans un contexte conversationnel. Celui–ci résulte d'un mouvement rigide associé à la pose et d'un mouvement non–rigide lequel serait dû à l'expression faciale elle–même. Il faut donc isoler adéquatement la composante non–rigide correspondant exclusivement à l'expression faciale [125], dans ce cas on parlera souvent d'un problème de factorisation. Un problème loin d'être trivial à cause de la forte non–linéarité du couplage [9].

1.2.3.3 Relatifs à la dynamique

Avec un ensemble d'attributs faciaux, les humains sont capables de mieux percevoir une émotion, à partir de points en mouvement, qu'à partir de points figés présentés sur une image fixe. Ce résultat peut expliquer comment les mouvements faciaux peuvent présenter une valeur ajoutée à la reconnaissance [193] et ceci même lorsque l'information texture est moins évidente [75].
Un classifieur hybride, conjuguant à la fois les avantages de l'aspect dynamique et statique, serait plus fiable [24].

Lorsqu'ils sont utilisés dans le cas d'un flux vidéo continu, les classifieurs statiques, plus faciles à entraîner et à implémenter, peuvent–être moins efficaces particulièrement lorsque l'expression faciale n'est pas à son extrema. Quant aux classifieurs dynamiques, en contrepartie, ils sont plus complexes et exigent plus de données d'apprentissage qui dépendent étroitement de certains paramètres temporels.

Le *timing*, cet aspect, assez élaboré en théorie, permet de décrire une expression à l'aide de trois paramètres temporels à savoir le début (attaque), le soutien (apex), et la fin (relaxation) [106, 209]. En pratique, peu d'études ont porté sur le marquage du début et de la fin d'une expression faciale, des paramètres dont l'information concernant la durée n'est fondée que sur l'intuition et l'approximatif [57].

1.3 Structure du manuscrit

1.3.1 Plan de la thèse

Cette thèse par articles est composée de deux articles de journaux. Dans le présent chapitre on a présenté une revue de littérature des techniques utilisées en reconnaissance faciale. Le chapitre 2 passe en revue les travaux connexes par rapport à notre première contribution présentée dans le chapitre 3. Dans ce chapitre, on présente notre premier article qui introduit un nouveau modèle d'apparence pour la reconnaissance des expressions faciales en se basant sur un référentiel prototypique. Au chapitre 4 on présente les travaux connexes relatifs aux modèles de représentation basés sur les attributs faciaux. Le chapitre 5 présente notre deuxième contribution avec un article portant sur la localisation et le suivi de points saillants du visage utilisant un modèle géométrique reflétant la topologie du visage. Dans cet article, on propose un système personnalisé de reconnaissance d'expressions faciales. Une discussion et conclusion sur nos travaux est présentée au chapitre 6. Les annexes (A, B et C) concernent les travaux portant sur la pose dont les articles ont été publiés dans différentes conférences internationales avec comité de lecture.

1.3.2 Publications

ARTICLES SOUMIS À DES REVUES AVEC COMITÉ DE LECTURE

1. M. Dahmane et J. Meunier. SIFT–flow Registration for Facial Expression Analysis using Prototypic Referential Models. *IEEE Transactions on Multimedia*, Février 2012.

2. M. Dahmane, S. Cossette et J. Meunier. Iterative Gabor Phase–based Disparity Estimation for "Personalized" Facial Action Recognition. *Signal processing: Image Communication*, Février 2012.

ARTICLES PUBLIÉS DANS DES COMPTE–RENDUS DE CONFÉRENCE AVEC COMITÉ DE LECTURE

1. M. Dahmane et J. Meunier. Oriented–Filters–based Head Pose Estimation. Dans *Fourth Canadian Conference on Computer and Robot Vision (CRV 2007)*, Montreal, QC, Canada, pages 418–425. IEEE Computer Society, May, 2007.

2. M. Dahmane et J. Meunier. Enhanced Phase–based Displacement Estimation – An Application to Facial Feature Extraction and Tracking. Dans *Proc. of Int. Conference on Computer Vision Theory and Applications (VISAPP'08)*, Funchal, Madeira, Portugal, pages 427–433, January 2008.

3. M. Dahmane et J. Meunier. An Efficient 3d Head Pose Inference from Videos. Dans *Image and Signal Processing, fourth International Conference (ICISP 2010)*, Trois-Rivières, QC, Canada, volume 6134 de *Lecture Notes in Computer Science*, pages 368–375, June-July 2010.

4. M. Dahmane et J. Meunier. Emotion Recognition using Dynamic Grid–based HOG Features. Dans *IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, Santa–Barbara, CA, USA, pages 884–888, March 2011.

5. M. Dahmane et J. Meunier. Individual Feature–Appearance for Facial Action Recognition. Dans *Proc. of Int. Conference on Image Analysis and Recognition*, Burnaby, BC, Canada, pages 233–242, June 2011.

6. M. Dahmane et J. Meunier. Continuous Emotion Recognition using Gabor Energy Filters. Dans *Proc. of the fourth international conference on Affective Computing and Intelligent Interaction - Volume Part*

II, Springer-Verlag (ACII'11), Memphis, TN, USA, pages 351–358, October 2011.

7. M. Dahmane et J. Meunier. Object Representation Based on Gabor Wave Vector Binning: An Application to Human Head Pose Detection. Dans *IEEE International Conference on Computer Vision Workshops, ICCV 2011 Workshops*, Barcelona, Spain, pages 2198–2204, November 2011.

1.3.3 Co–auteurs

Ce travail a été initié par un projet d'étude de l'interaction patient-infirmière à l'institut de Cardiologie de Montréal dirigé par la Dre Sylvie Cossette. Les co-auteurs des articles présentés dans cette thèse sont :

- Pr. Jean Meunier, directeur de thèse et directeur du laboratoire de traitement d'images et professeur titulaire au Département d'Informatique et de Recherche Opérationnelle de l'université de Montréal.

- Dre. Sylvie Cossette, chercheuse à l'Institut de Cardiologie de Montréal et professeure à la faculté des sciences infirmières de l'université de Montréal.

Chapitre 2

RECONNAISSANCE DES EXPRESSIONS FACIALES PAR RECALAGE SUR DES MODÈLES GÉNÉRIQUES

2.1 Avant–propos

Dans ce qui suit nous introduirons les notions utilisées pour la représentation et la reconnaissance de l'expression faciale présentées au chapitre 3. En premier, un détecteur de Viola et Jones est utilisé pour la localisation du visage [183]. La fenêtre englobant ce dernier servira à caractériser les traits faciaux à l'aide d'une approche basée sur l'apparence. On se sert de l'orientation des gradients dans cette partie de l'image pour permettre une représentation dense des attributs faciaux. Des histogrammes globaux sont obtenus en compilant les amplitudes de gradients pour un ensemble d'orientations dans des histogrammes 1-D plus locaux définis sur un ensemble de cellules d'une grille prédéfinie. Ce type de descripteurs appartient à la même classe de descripteurs que les motifs binaires locaux (LBP pour Local Binary Patterns), grandement utilisés, pour l'extraction des attributs faciaux afin de générer la signature de l'expression faciale à reconnaître.

Outre la caractérisation de l'expression faciale, la principale difficulté pour tout système de reconnaissance d'émotions faciales est de mettre en œuvre des modèles plus génériques. Étant donné que la taille et la géométrie du visage diffèrent d'un individu à un autre, un recalage s'avère nécessaire. Ce qui permettra, en plus, de dissocier le mouvement rigide (global) de la tête des déformations faciales non rigides (plus locales). Plusieurs algorithmes de recalage utilisant des modèles basés sur la forme du visage ont été proposés [30, 35, 78, 112, 155, 192, 214]. Dans ce travail, nous avons adopté l'algorithme SIFT–flow, une méthode dense regroupant à la fois la transformation de caractéristiques visuelles invariantes à l'échelle (communément appelée SIFT pour Scale-Invariant Feature Transform) et le flux optique. Nous comparerons cette approche à une méthode de normalisation plus classique reposant sur la position des yeux.

Les systèmes de reconnaissance automatique d'expressions faciales nécessitent de mettre en place une forme d'apprentissage. Plusieurs méthodes issues de reconnaissance de formes ont été utilisées telles que les Modèles de Markov cachés (HMM pour Hidden Markov Models) [136], réseaux bayésiens (BN pour Bayesian Networks) [210], réseaux de neurones (ANN pour Artificial Neural Networks) [105], etc. Comme apprenants de base d'émotions, nous avons adopté les machines à vecteurs de support (SVM pour Support Vector Machines) qui possèdent un faible nombre de paramètres à ajuster, et entre autres, sont appropriées aux données à grande dimensionnalité.

2.2 Localisation de visages

Le visage humain est un objet complexe à reconnaître, vu la variabilité des poses qu'il présente, essentiellement due à l'échelle, l'orientation et l'occultation (Fig. 1.6). Le modèle statistique tel que proposé par Viola et Jones permet de prendre en compte cette variabilité, à partir d'une base d'apprentissage comprenant des images positives (visages) et des images négatives (tout ce qui n'est pas un visage). Basé sur un ensemble de simples descripteurs (Fig. 2.1), le modèle essaie de trouver ceux qui sont capables de classifier correctement les visages (Fig. 2.2). La détection d'une instance visage est effectuée en parcourant l'image par une fenêtre glissante, laquelle est présentée en entrée à une chaîne de filtres. Des fenêtres susceptibles de ne pas contenir un visage seront éliminées le plus tôt possible dans une cascade de classifieurs.

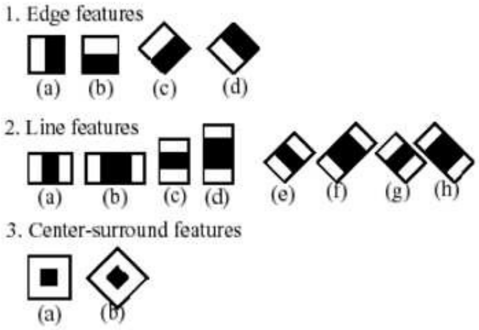


Figure 2.1. Différent types de filtres utilisés (fonctions de Harr) [15].

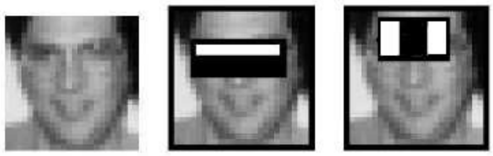


Figure 2.2. Exemple de descripteurs retenus (Fig. 2.1: 1-b,2-a).

Le calcul des descripteurs est basé sur la valeur de l'image intégrale (SAT[1]) :

$$SAT(X,Y) = \sum_{x<X, y<Y} I\left(x,y\right) \tag{2.1}$$

Cette représentation permet de calculer efficacement n'importe quelle somme à partir d'une région rectangulaire de l'image (Fig. 2.3).

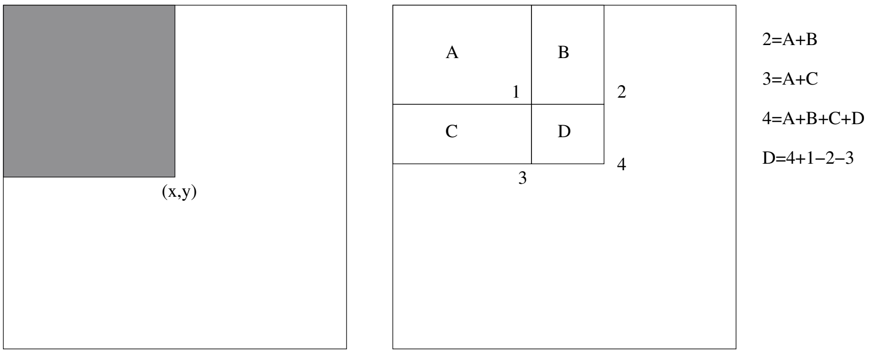


Figure 2.3. Somme intégrale de sous-fenêtre.

[1] SAT: Summed Area Table

Soit le classifieur faible F_k appliqué au descripteur D_k tel que :

$$F_k\left(I\right) = \begin{cases} +1 & \text{si } D_k(I_i) \geq T \\ -1 & \text{sinon} \end{cases}$$

La technique de *boosting* permet de construire un classifieur fort à partir d'une combinaison de classifieurs faibles (Eq. 2.2) en sélectionnant les descripteurs les plus discriminants. Les coefficients α_k dépendent de l'erreur de classification commise sur l'ensemble d'apprentissage.

$$H\left(I\right) = Signe\Big(\sum_{k=1}^{K} \alpha_k F_k(I)\Big) \tag{2.2}$$

Pour accélérer le processus de détection et pour de meilleures performances, Viola et Jones proposent un classifieur regroupant plusieurs classifieurs, disposés en cascade et par ordre croissant de complexité (nombre et simplicité des classifieurs faibles) (Fig. 2.4), afin de détecter de façon fiable les vraies instances de visage.

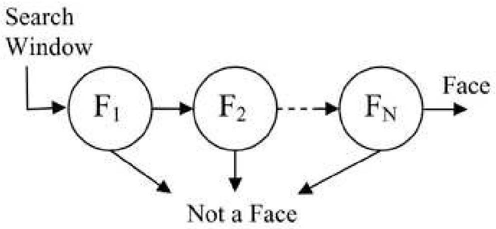


Figure 2.4. Représentation en cascade pour la détection de visages.

Une fois le visage localisé, on procède à la caractérisation de l'image faciale.

2.3 Caractérisation des images faciales

2.3.1 Motif binaire local

L'expression faciale peut être représentée à l'aide de motifs de texture et de structures locales [157] définis sur un voisinage donné en utilisant des motifs binaires locaux (LBP) [132].

Un voisinage local est défini comme étant un ensemble de points équidistants échantillonnés depuis un cercle centré sur le pixel à coder. Pour les points qui ne tombent pas à l'intérieur du pixel une interpolation est établie. La figure 2.5 montre quelques exemples de l'opérateur LBP, où P désigne le nombre de points échantillonnés et R le rayon du cercle.

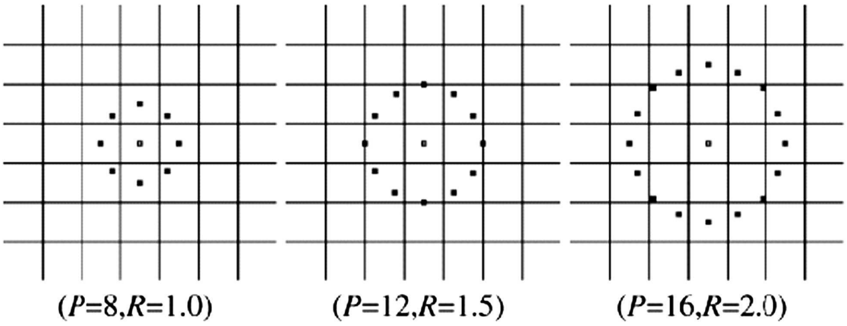


Figure 2.5. Exemples de voisinage utilisé pour définir la texture via l'opérateur LBP [133]. Voisinages circulaires (8,1), (11,1.5) et (16,2).

Formellement, le codage LBP du pixel (x_c, y_c) peut être exprimé comme suit.

$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} s(i_p - i_c)2^p$$

où i_c et i_p représentent, respectivement, le niveau de gris du pixel central et celui du $p^{\text{ème}}$ pixel dans le voisinage circulaire de rayon R, et où la fonction $s(x)$ est définie ainsi,

$$s(x) = \begin{cases} 1 & si \quad x \geq 0 \\ 0 & si \quad x < 0 \end{cases}$$

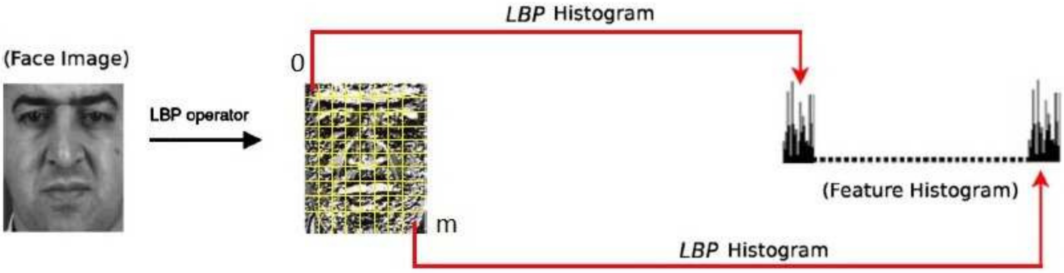


Figure 2.6. Caractérisation d'une image faciale par l'opérateur LBP [84].

Le codage LBP peut être utilisé directement, sinon pour réduire la taille de l'ensemble des descripteurs, une représentation par histogrammes peut être envisagée (Fig. 2.6). En pratique, ce type de codage s'est avéré sensible aux images bruitées, ceci est l'effet du seuillage adopté par l'opérateur LBP [143].

2.3.2 Histogrammes de gradients orientés

Le gradient orienté [103] est un descripteur utilisant l'information du contour (Fig 2.7). Dans ce cas, chaque fenêtre peut être décrite par la distribution locale des orientations et des amplitudes. Cette distribution est définie par le biais d'un histogramme local des gradients orientés (HOG pour Histogram of Oriented Gradients). Celui-ci est formé en divisant la fenêtre de détection en un ensemble de cellules sur lesquelles l'amplitude du gradient pour chaque intervalle (bin) d'orientation est intégrée sur tous les pixels de la cellule. Dans [44], on a montré la supériorité des HOG par rapport à quelques descripteurs connus tels que les PCA–SIFT et les contextes de contours (Shape Context).

Les gradients de l'image peuvent être calculés en chaque point (x, y) par un

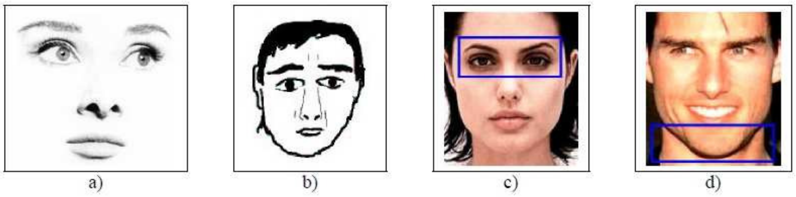


Figure 2.7. Exemple de caractéristiques globales vs. locales pouvant être collectées via des histogrammes d'orientations [103]. (a,b) montrent les caractéristiques globales d'un visage. (c,d) montrent d'importants attributs locaux ((c) contours horizontaux, (d) contours obliques des deux côtés).

opérateur de Sobel,

$$G(x,y) = \sqrt{G_x(x,y)^2 + G_y(x,y)^2}$$

où G_x et G_y représentent, respectivement, le gradient dans les directions x et y, l'orientation du contour en ce point est donc,

$$\theta(x,y) = \arctan\left(\frac{G_y\,(x,y)}{G_x\,(x,y)}\right)$$

Les contours sont, ensuite, classés en K intervalles (bins). La valeur du $k^{\text{ème}}$ intervalle est donnée par,

$$\psi_k(x,y) = \begin{cases} G(x,y) & \text{si } \theta(x,y) \in bin_k \\ 0 & \text{sinon} \end{cases}$$

Le calcul de $\psi_k(x,y)$ sur une région R de l'image peut se faire efficacement en utilisant le concept de l'image intégrale (Fig. 2.3).

$$SAT_k(R) = \sum_{(x,y)\in R} \psi_k(x,y)$$

Pour la génération du descripteur associé à une expression faciale, on divise la fenêtre de détection en 48 cellules, on utilisera $K = 9$ intervalles correspondant à des intervalles équirépartis sur $[0..\pi[$. Pour chaque cellule un histogramme local est généré. Ensuite, pour chaque bloc de 2×2 cellules, on concatène les 4 histogrammes associés en un seul histogramme qu'on normalise. L'histogramme global concaténant les histogrammes normalisés des 12 blocs définit alors la signature de l'expression faciale à reconnaître (Fig. 2.8).

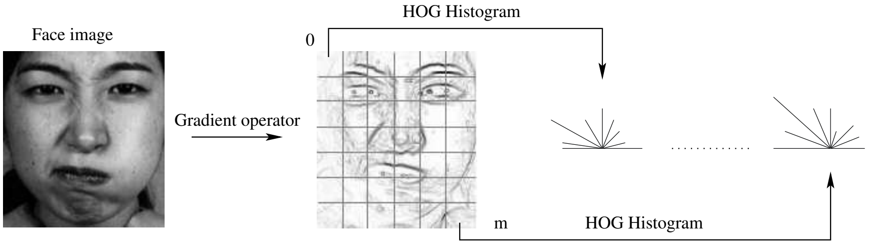


Figure 2.8. L'histogramme global d'orientation concaténant des HOGs locaux générés à partir des différentes cellules.

Étant donné que les histogrammes du gradient orienté sont des descripteurs à base de fenêtrage, le choix de la fenêtre de détection est crucial. Ceci nécessite un bon recalage, dont la tolérance dépend de la position de la fenêtre mais aussi de la taille des cellules.

2.4 Recalage des images faciales

Le recalage d'images faciales permet de normaliser la taille et la géométrie des différents visages, en outre, il permet de dissocier le mouvement rigide de la tête des déformations faciales pertinentes. Cette étape importante influe directement sur les performances globales du système de reconnaissance d'expressions faciales tout comme la reconnaissance de visages [145].

2.4.1 Normalisation par fenêtres adaptatives

Au lieu de se servir de fenêtres statiques telles que utilisées dans le cas des HOG classiques, on propose une normalisation de la fenêtre de détection elle–même.

Dans [41], on avait proposé des HOG adaptatives comme des descripteurs utilisant des fenêtres de détection dynamiques définies à partir de mesures anthropométriques du visage. La taille de la fenêtre de détection varie selon la taille de la cellule, celle-ci étant de forme carrée de côté variable dépendamment de la distance inter–pupilles de chaque visage.

2.4.2 Recalage par SIFT–flow

Au chapitre 3, on présente une méthode basée sur le SIFT–flow pour le recalage générique (i.e. personnes identiques ou pas) de visages.

Récemment proposé par Liu *et al.* [108], l'alignement par SIFT–flow a été conçu pour l'alignement de scènes. Il consiste à mettre en correspondance, entre deux images à aligner, des attributs SIFT [110] obtenus de part et d'autre par échantillonnage dense. Inspiré du concept du flux optique mettant en correspondance les intensités, le SIFT–flow met en correspondance une paire d'images SIFT (s_1, s_2) en utilisant une fonction objective (Eq. 2.3) similaire à celle du flux optique.

$$\begin{aligned} E(\mathbf{w}) = \; & \sum_{\mathbf{p}} \min\left(\left\| s_1(\mathbf{p}) - s_2(\mathbf{p} + \mathbf{w}(\mathbf{p}))\right\|_1, t\right) + \\ & \sum_{\mathbf{p}} \eta \left(|u(\mathbf{p})| + |v(\mathbf{p})|\right) + \\ & \sum_{(\mathbf{p},\mathbf{q}) \in \epsilon} \min\left(\alpha\, |u(\mathbf{p})| - |u(\mathbf{q})|, d\right) + \min\left(\alpha\, |v(\mathbf{p})| - |v(\mathbf{q})|, d\right) \end{aligned} \tag{2.3}$$

Le modèle d'optimisation (Fig. 2.9) utilise une méthode d'inférence dite la propagation de croyance en boucles (loopy belief propagation) [128] pour minimiser la fonction objective (Eq. 2.3), laquelle est formulée à partir des contraintes suivantes.

1. *Le terme de données*, celui-ci permet de contraindre la mise en correspondance des descripteurs SIFT via le vecteur de déplacement $\mathbf{w}(\mathbf{p}) = (u(\mathbf{p}), v(\mathbf{p}))$ au niveau du pixel $\mathbf{p} = (x, y)$.

2. *Le déplacement minimal*, ce terme pénalise les grands déplacements.

3. *Le terme de lissage*, lequel contraint les pixels voisins à avoir un mouvement cohérent.

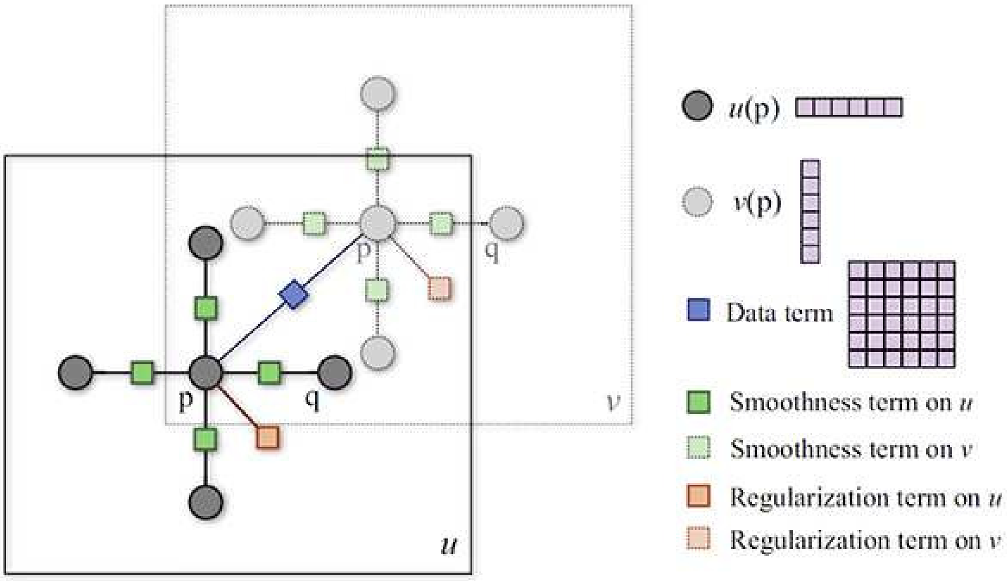


Figure 2.9. Propagation de croyance en boucles à double couches. La composante horizontale (u) et verticale (v) de la fonction objective du SIFT flow [108].

Les seuils t et d permettent de minimiser l'impact des erreurs de mise en correspondance (aberrations et discontinuités). Une implémentation pyramidale de la mise en correspondance par raffinement/propagation, à travers les niveaux hiérarchiques (du plus grossier au plus fin qui correspond à l'image originale), permet d'améliorer les performances du SIFT–flow.

La figure 2.10 montre un exemple de recalage facial par SIFT–flow. La construction de l'image de référence utilise une image arbitraire "neutre", sur laquelle toutes les images de la même catégorie d'expressions sont recalées puis moyennées. Bien que le procédé préserve la catégorie d'expression, la déformation semble être plus ou moins importante. Par exemple pour l'image du milieu "joie", la perte de détails importants était même prévisible au niveau de la bouche puisque la partie des dents est inexistante dans l'image référence.

Au chapitre 3, on propose d'utiliser une image de référence par catégorie d'expression afin de préserver, au maximum, les détails des déformations pertinentes. Ceci nous permettra d'exprimer toute expression faciale par rapport à un référentiel comprenant les sept catégories d'expressions prototypiques ("colère", "dégout", "peur", "joie", "neutre", "tristesse", "surprise") qu'on appellera *référentiel prototypique.*

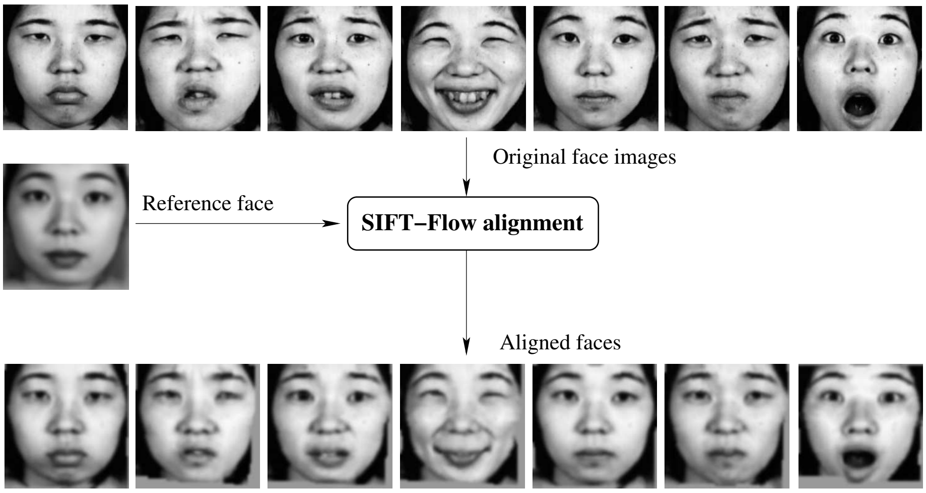


Figure 2.10. Recalage par SIFT–flow sur une image référence "neutre" moyennant toutes les images neutres recalées par rapport à une image arbitraire.

2.5 Machines à vecteurs de support

Les machines à vecteurs de support sont des classifieurs robustes, bien adaptés aux problèmes de discrimination de grande dimensionnalité [34, 180] et ce même lorsqu'on a relativement peu d'échantillons. Les SVMs sont donc particulièrement utiles en reconnaissance d'expressions faciales où le nombre d'attributs peut se chiffrer en milliers alors que les bases de données d'expressions faciales sont peut être limitées à des centaines d'échantillons. Basés sur le principe de minimisation du risque structurel[2] décrit par la théorie de l'apprentissage statistique de Vapnik [13, 31, 179], les SVMs établissent un mappage des données d'entrée vers un espace de plus grande dimensionnalité, dans lequel elles deviennent linéairement séparables.

Soit l'ensemble de données $x_i,\ i = 1, \ldots, l$ et les étiquettes correspondantes $y_i \in \{+1, -1\}$, la tâche principale pour entrainer un SVM consiste à résoudre le problème d'optimisation quadratique suivant [55],

$$\begin{aligned} \min_\alpha \qquad & \tfrac{1}{2}\alpha^T Q \alpha - \textstyle\sum \alpha_i \\ \text{tel que } 0 \leq \alpha_i \leq C\ ,\ & i = 1 \ldots l, \\ Y^T\alpha = 0 & \end{aligned} \tag{2.4}$$

La constante de régularisation C représente la borne maximale de toutes les variables, Q est une matrice symétrique $l \times l$, avec $Q_{ij} = y_i\, y_j\, K(x_i, x_j)$, où $K(x_i, x_j)$ représente la fonction noyau permettant d'introduire la non–linéarité. La fonction d'optimisation permet de trouver un plan optimal séparant les instances positives des instances négatives. L'ensemble de données est dit optimalement séparable, s'il est séparé sans erreurs et que la distance séparant le vecteur le plus proche et l'hyperplan est maximale. Dans le chapitre suivant, nous examinerons plusieurs seuils permettant de déplacer l'hyperplan des deux bords de sa position "optimale".

[2] Le risque structurel est le lien entre la vaste marge et la complexité (exprimée par la dimension de Vapnick-Chervonenkis) de la fonction de décision [181].

Pour un problème de discrimination à catégorisation multiple, tout comme la catégorisation des expressions faciales, “un contre tous” et “un contre un” sont les deux approches utilisées pour la construction de SVM multiclasse à partir d’un ensemble de SVM binaires. La première se base sur le principe du vote majoritaire (winner–take–all), dans lequel la classe est assignée par le SVM ayant le plus grand score en sortie. L’approche “un contre tous” consiste à utiliser un SVM par couple de catégories et à chaque fois que le SVM assigne une instance à l’une des deux classes, le vote de la classe gagnante sera bonifié de un et ainsi de suite. Finalement, la classe avec un maximum de votes détermine l’étiquette de l’instance.

L’approche “un contre tous” peut souffrir de la répartition déséquilibrée des exemples positifs par rapport aux exemples négatifs (ceux-ci étant plus nombreux). C’est la raison pour laquelle, dans ce qui suit, nous nous servirons de la stratégie “un contre un” pour résoudre le problème de catégorisation multiple de l’expression faciale.

Chapitre 3

(ARTICLE) SIFT–FLOW REGISTRATION FOR FACIAL EXPRESSION ANALYSIS USING PROTOTYPIC REFERENTIAL MODELS

Une première ébauche de cet article (voir annexe F) a été publiée comme l'indique la référence bibliographique [41]

> M. Dahmane et J. Meunier. Emotion recognition using dynamic grid–based HoG features. Dans *IEEE International Conference on Automatic Face Gesture Recognition and Workshops*, pages 884–888, 2011.

Comparée à la version originale, le présent article présente une méthodologie plus élaborée qui a été soumis pour publication dans le journal scientifique *IEEE Transactions on Multimedia*, par Mohamed Dahmane et Jean Meunier.

Abstract

Automatic facial expression analysis (AFEA) is the most commonly studied aspect of behavior understanding and human–computer interface. Aiming towards the application of computer interaction, human emotion analysis and even medical care, AFEA tries to build a mapping between the continuous emotion space and a set of discrete expression categories. The main difficulty with facial emotion recognition systems is the inherent problem of facial alignment due to person–specific appearance.

In the present paper, we investigate an appearance–based approach combining SIFT–flow registration and HOG features (SF/HOG), with a RBF–kernel SVM classifier. Experimental results show that this facial expression recognition strategy out-

performs other approaches with a recognition rate of up to 99.52% using a leave–one–out strategy and 98.73% using the cross–validation strategy. For the much more challenging person–independent evaluation, the SF/HOG still gave the best results (86.69%).

3.1 Introduction

In a face–to–face talk, language is supported by nonverbal communication, which plays a central role in human social behavior by adding cues to the meaning of speech, providing feedback, and managing synchronization [3]. Information about the emotional state of a person is usually inferred from facial expressions, which are primarily carried out by facial attributes. In fact, 55% of a message is communicated by facial expressions whereas only 7% is due to linguistic language and 38%, to paralanguage [123].

The computer vision community is therefore interested in the development of automated facial expression analysis (AFEA) techniques, for HCI (Human Computer Interaction) applications, meeting video analysis, security and clinical applications etc.

The overall performance of AFEA systems can severely be affected by different factors such as intra/inter individual variability, transitions among expressions, intensity of facial expression, deliberate versus spontaneous expressions, head orientation and scene complexity, image properties, reliability of ground truth, and databases. An ideal facial expression analysis system has to address all these dimensions [170]. With this in mind, a variety of systems have been developed and applied [7, 8, 11, 50, 91, 114, 199, 202] . These systems resemble each other by means of their processing.

First, designed approaches try to optimize the recognition performance on common facial expression databases such as Cohn-Kanade DFAT [95], CMU-PIE [162], MMI [140], UT Dallas [135], Ekman-Hager [50], FEED [186], JAFFE [116], GEMEP [90], and GENKI [126].

For instance, the authors in [194] used the GENKI dataset to recognize smiling faces. Valstar *et al.* [177] used the GEMEP–FERA dataset to recognize five different facial emotions (anger, fear, joy, relief, and sadness), they investigated both person–specific and person–independent partitions. The JAFFE database (described below), was used by Bucius and Pitas [17], Zheng *et al.* [212], Kotsia *et al.* [100], and Lyons *et al.* [116]. The Cohen dataset was used by Bartlett *et al.* [6], Pantic and Patras [138], Kotsai *et al.* [100], and Wang *et al.* [188] to recognize the six prototypic expressions (happy, angry, disgust, fear, sad and surprise).

Second, the approach commonly adopted in developing these systems is to first track the face and facial features, and apply a bank of linear filters (e.g. Harr-like features [182], Gabor filters [46], Scale Invariant Feature Transform (SIFT) [110], oriented gradient [44, 103], Local Binary Pattern (LBP) [133], and Local Phase Quantization (LPQ) [134]). Gabor filters [16, 50, 79, 116, 211, 212], PCA (Principal Component Analysis) [17], adaptive HOG (Adaptive Histogram of Oriented Gradients) [41], PHOG (pyramid of HOG) with LPQ [49], EOH (Edge Orientation Histograms) [103], LBP features [88, 91, 177], and optical flow [167] are used to characterize facial expressions. For high–dimensional derived features, dimension reduction techniques such as PCA, Linear Discriminant Analysis (LDA), and Locality Preserving Projections (LPP) [82] can be used.

Third, based on the derived feature vector, eigenspaces [130], LDA [116, 174], SVM (Support Vector Machines) [6, 16, 49, 100], k-Nearest Neighbors [17, 88], Artificial Neural Networks [100, 167], and Hidden Markov Models [2] are used for the expression classification task which automatically assigns each face image instance to the corresponding expression class.

Clinical studies of facial expressions has been also investigated. The authors in [191] automatically quantify emotional expression differences between patients with neuropsychiatric disorders. The authors in [99] examined facial expression differences based on duration and frequencies of evoked emotion expressions from a group of 12 persons with stable schizophrenia. McIntyre *et al.* [120] proposed an approach to measure and compare facial activity in depressed subjects, before and after treatment, of endogenous and neurotic depressives.

Among the existing facial expression recognition techniques, geometric–based techniques demonstrate high concurrent validity with manual FACS (Facial Action Coding System) coding [27, 32]. Furthermore, they share some common advantages such as explicit face structure, practical implementation, and collaborative feature–wide error elimination [87]. However, the precise localization of local facial features poses a significant challenge to automated facial expression analysis, since subtle changes in the facial expression could be missed due to localization errors [137]. Therefore, geometric approaches including only shape information may be rather irrelevant [113] by requiring accurate and stable localization of facial landmarks. These drawbacks can be naturally avoided by the appearance–based methods, which have attracted more and more attention, but in contrast, are in need of a well–designed feature set closely relevant to facial expression variations but at the same time reliably insensitive to facial variations that are irrelevant to facial expression.

In this paper, we investigate an appearance–based approach (SF/HOG) for the problem of automatic facial expression recognition. Histogram of oriented gradients (HOG) are used as features and irrelevant facial variations are managed with SIFT–flow registration procedure.

The structure of this paper is as follows. Section 3.2 describes the JAFFE database used for our tests and outlines the results published in the literature by other groups with it. Section 3.3 describes in details our SF/HOG approach. Section 3.4 describes the evaluation protocols, presents the experimental results, and provides a detailed comparison of various system performances. Finally, we draw our conclusion in section 3.5.

3.2 JAFFE database

The Japanese Female Facial Expression (JAFFE) database [116] is commonly used in measuring the performance of facial expression recognition systems [160]. We will use it as well to perform comparisons with other existing systems [16, 17, 21, 59–61, 70, 79, 116, 130, 211].

The JAFFE database consists of 213 images collected from ten female expressers, each one performing 2 to 4 examples for each of the seven prototypic expressions (happiness, sadness, surprise, anger, disgust, fear, and neutral). The grayscale images are of resolution 256 × 256 pixels. Figure 3.1 shows samples of three expressers, named "KM", "NM", and "UY", acting seven categories of facial expressions.

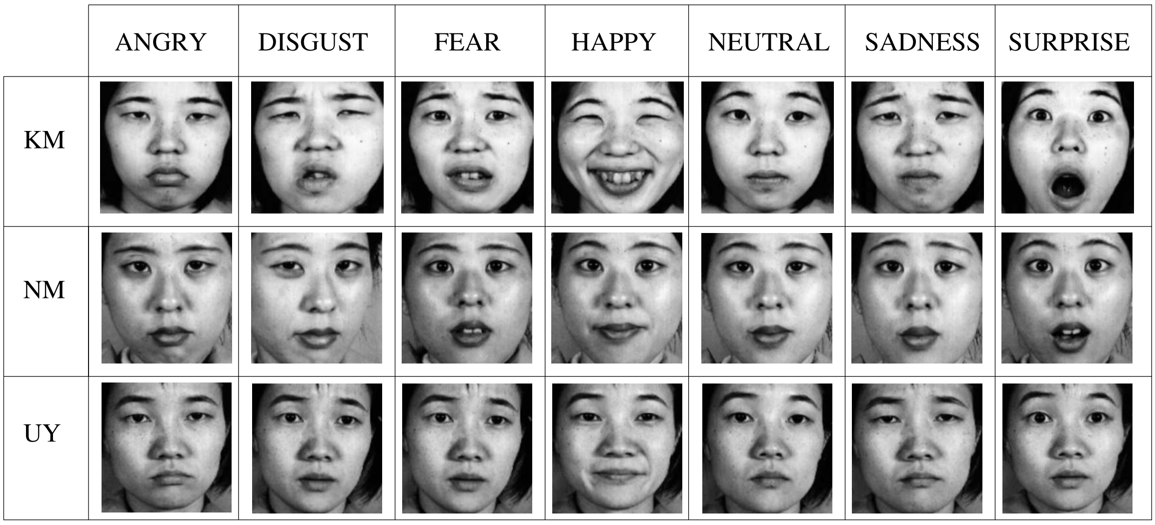


Figure 3.1. Examples from the JAFFE facial expressions database of three persons "KM","NM", and "UY" acting seven facial expressions.

In [211], facial expression images are coded using a multi–orientation, multi–resolution set of Gabor filters coefficients extracted from the face image at fiducial points, using a 34–node grid aligned manually with the face. Their facial expression recognition system that uses this input code and a two–layer perceptron classifier, achieved a generalized recognition rate of 90.1%. In [116] the fiducial grid was positioned by manually clicking on the same 34 easily identifiable points of each facial image. PCA was then used to reduce the dimensionality of the feature vectors of Gabor filter coefficients, which combined with LDA achieved a rate of 92% on the seven different facial expressions of the JAFFE dataset.

Buciu *et al.* [17] cropped and aligned all images, and made use of a nearest neighbor classifier with different similarity measures. PCA was employed to classify the seven facial expressions that were characterized by using two image representation approaches called non–negative matrix factorization and local non–negative matrix factorization. From the JAFFE database a higher classification accuracy of 81.42% was achieved when the maximum correlation classifier was applied. In [16], authors reported a performance rate of 90.34% when the feature vectors were obtained by Gabor representation with low frequency range and the classification was done using a quadratic–kernel SVM.

In [130], authors computed an eigenspace for each of the six facial expressions. The classification was based on measuring the similarity between a probe image and the reconstructed image of each class (i.e. each facial expression). An average performance rate of 77.5% was reported on the JAFFE dataset.

Authors in [60], extracted local texture features by applying LBP to facial feature points detected by an active appearance model, the direction between each pair of feature points was considered as geometrical (shape) features. In addition, they considered a global texture information by dividing the face image into small regions, and calculating the LBP histogram of each region. Local texture, shape and global texture features were combined and fed to a nearest neighbor classifier in which the weighted Chi–square statistic was used for classification. The average rate of recognition was 83%. In a combination of LBP features and linear programming technique, Feng *et al.* [61] reported a recognition rate of 93.8% on the JAFFE database. It should be noted that the two eye positions were manually selected in their approach to exclude non–face area from each image. Fu *et al.* [70] evaluated an operator called centralized binary pattern with image Euclidean distance on JAFFE database. The obtained recognition rate was 94.76%.

As in references [116] and [211], to solve the 7–expression recognition problem, Guo *et al.* [79] employed, as feature vectors, the amplitudes of Gabor–filter responses of 34 selected fiducial points. Their recognition accuracy was 91% with feature selection via a linear programming routine.

More recently, Chang *et al.* [21] used images from the JAFFE database that were manually cropped to extract the face region. They investigated a Gaussian process classification approach for facial expression recognition. The adopted leave–one–out cross–validation strategy gave a recognition rate of 93.43%.

3.3 Technical approach

In this section, we first present the face registration procedure of SIFT–flow to remove irrelevant facial variations. This registration will be integrated within an appearance–based method for facial expression recognition. This method will use oriented gradient histograms [44, 103] defined over a dense patchwork. Authors in [44] have shown the advantage of HOG (Histogram of Oriented Gradients) with respect to several other descriptors. The section is concluded with a brief description of the SVM used for classification.

3.3.1 SIFT–flow facial registration

By a facial registration stage, we want to align faces to normalize size and geometry across different persons, and remove irrelevant facial variations. This stage is a critical issue in feature–based approaches, since miss–located fiducial features will propagate the error through subsequent processing stages.

In this article, we use SIFT–flow to solve the person–dependent and person–independent registration problem.

Recently introduced by Liu *et al.* [108], the SIFT–flow alignment was designed for higher level image (scene) alignment. The SIFT flow algorithm consists of matching densely sampled SIFT features [110] between two images, while preserving spatial discontinuities. For every pixel in an image, the neighborhood (e.g. 16×16 pixels) is divided into a 4×4 cell array, to quantize the orientation into 8 bins for each cell. The per–pixel SIFT descriptor then consists of a $4 \times 4 \times 8 = 128$–dimensional vector corresponding to the SIFT image which has high spatial resolution that can preserve edge sharpness.

Inspired by optical flow, a pair of SIFT images (s_1, s_2) are matched via a dense correspondence using an objective function $E(\mathbf{w})$ on SIFT descriptors instead of intensity values (Eq. 3.1).

$$\begin{aligned} E(\mathbf{w}) = \; & \sum_{\mathbf{p}} \min\left(\left\|s_1(\mathbf{p}) - s_2(\mathbf{p} + \mathbf{w}(\mathbf{p}))\right\|_1, t\right) + \\ & \sum_{\mathbf{p}} \eta\left(|u(\mathbf{p})| + |v(\mathbf{p})|\right) + \\ & \sum_{(\mathbf{p},\mathbf{q})\in\epsilon} \min\left(\alpha\left|u(\mathbf{p})\right| - \left|u(\mathbf{q})\right|, d\right) + \min\left(\alpha\left|v(\mathbf{p})\right| - \left|v(\mathbf{q})\right|, d\right) \end{aligned} \tag{3.1}$$

$E(\mathbf{w})$ contains 3 terms, the *data term* which constrains the SIFT descriptors to be matched along with the flow vector $\mathbf{w}(\mathbf{p}) = (u(\mathbf{p}), v(\mathbf{p}))$ at the image position $\mathbf{p} = (x, y)$, the *small displacement term* constrains the flow vectors to be as small as possible when no other information is available, and the *smoothness term* which constrains the flow vectors of adjacent pixels to be similar. The set ϵ contains all the spatial neighborhoods (a four–neighbor system is used). The truncated L_1 norm is used in both the data term and the smoothness term to account for outliers and flow discontinuities, with the thresholds t and d, respectively.

A dual–layer loopy belief propagation is used as the base algorithm to optimize the objective function [108], and a coarse–to–fine SIFT flow matching scheme, that roughly estimates the flow at a coarse level of image grid then gradually propagates and refines the flow from coarse to fine, permits to significantly improve the matching performance.

The SIFT–flow facial alignment is applied in two different steps of the algorithm. First, we wish to construct seven reference facial expressions from the JAFFE database. This is simply done by aligning (with SIFT–flow) all images corresponding to a particular expression and then averaging them. For each expression, we select one arbitrary image for the alignment of all others. After this step, we get a set of seven generic reference images, one for each expression. Irrelevant individual variations are reduced by this averaging process. Figure 3.2 shows the seven generic images we obtained after this step.

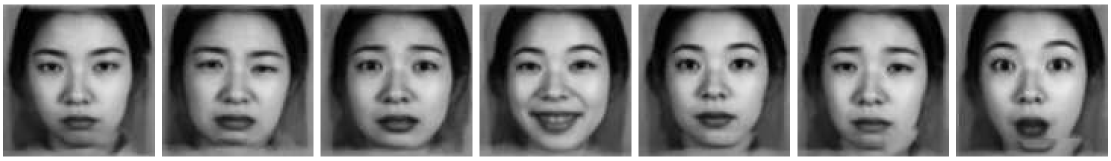


Figure 3.2. Prototypic facial expression referential models. Each reference image provides a generic representation of a prototypic facial expression.

The second step where SIFT–flow is applied is the recognition of an unknown expression. In this case, the unknown expression is registered with each generic reference expression in turn before any feature is computed. Only after this registration, HOG features (see below) are computed for classification. Figure 3.3 shows a diagram describing this process.

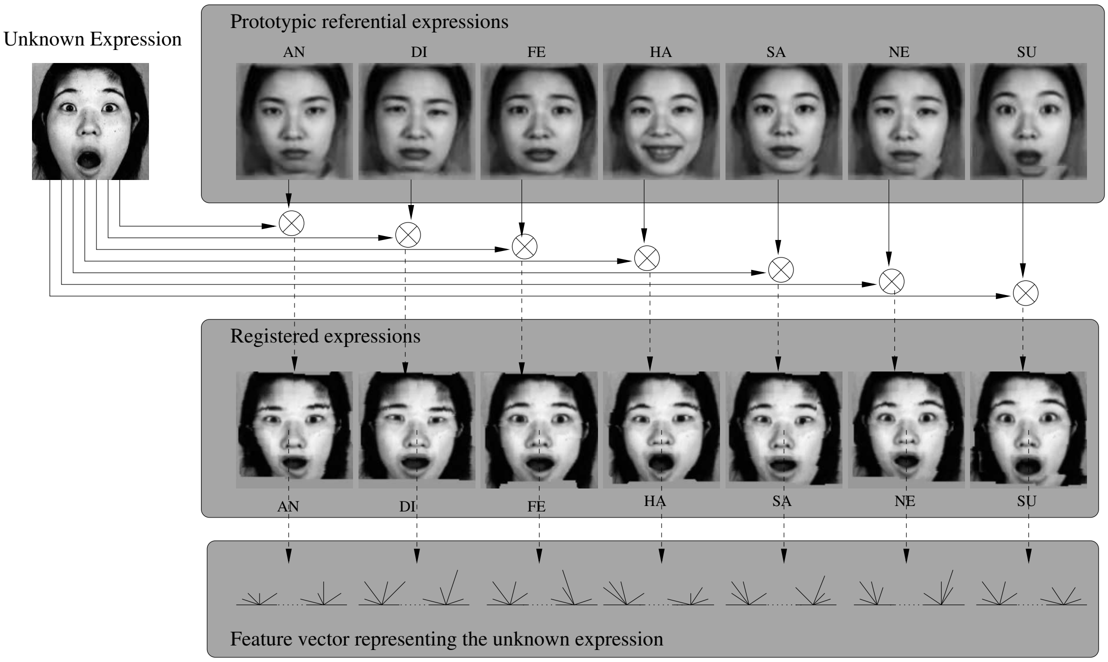

Unknown Expression
Prototypic referential expressions
AN
DI
FE
HA
SA
NE
SU
Registered expressions
AN
DI
FE
HA
SA
NE
SU
Feature vector representing the unknown expression

Figure 3.3. Diagram describing the SIFT–flow HOG process.

3.3.2 Facial expression characterization

3.3.2.1 Facial image pre–processing

Prior to performing image characterization, image pre–processing was performed by normalizing for scale. First, we used the standard Viola and Jones face detection to extract the face location in each image. The detected face, on which a contrast amelioration and brightness normalization have been performed, was scaled to be 200 by 200 pixels.

To characterize facial expressions from this face area, we then used an appearance approach with histogram of oriented gradients.

3.3.2.2 Histogram of Oriented Gradients

Histogram of oriented gradient (HOG) feature descriptors were proposed in [103] and [44]. The main idea behind the HOG descriptors is based on edge information.

That is each window region can be described by the local distribution of the edge orientations and the corresponding gradient magnitude. The local distribution is described by the local histogram of oriented gradients which is formed by dividing the detection window into a set of small regions called cells, where the magnitude of the edge gradient for each orientation bin is computed for all pixels. To provide better invariance, the local HOG is normalized over a block of neighboring cells. Dalal and Triggs [44] have shown that HOG outperforms wavelet, PCA-SIFT and Shape Context approaches.

To generate facial expression feature descriptors, the aligned 200 × 200 pixels face image is divided into a 8 by 6 cell array columns with a cell size side of 25 × 33 pixels. We used a $[-1, 0, 1]$ gradient filter, and linear gradient voting into 9 orientation histogram bins from 0° to 180°.

For each cell a local histogram is processed. Then for each block of 2 × 2 cells, the 4 local histograms are concatenated and the resulting histogram is normalized using the L_2-norm. The per–block normalized histograms are then concatenated into a global histogram giving a single 432 dimensional feature vector encoding the global facial image structure with local primitives (Fig. 3.4).

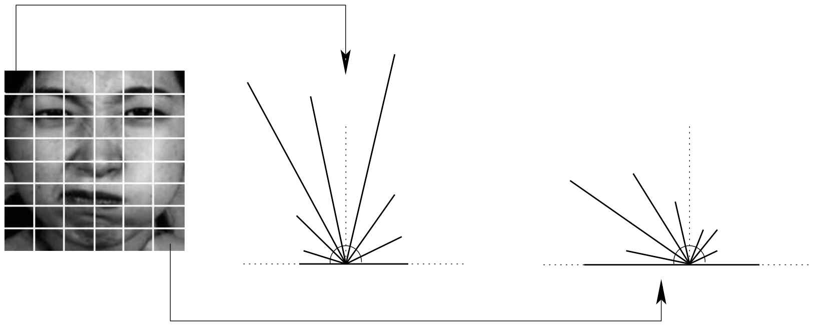


Figure 3.4. A Global HOG concatenating the local HOGs extracted from each cell.

Since HOGs are window–based descriptors, the choice of the detection window is therefore crucial. The SIFT–flow registration permitted to define from each generic image a single patchwork as the smallest window enclosing the visual features (e.g. eyes wide open and eyebrows raised high for the "surprise" facial expression) as shown in figure 3.5. The patchworks are superimposed on the corresponding registered face image for generating the histograms.

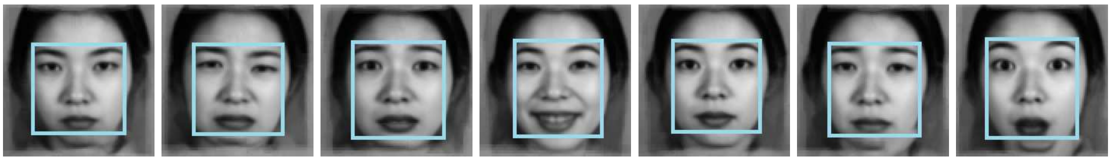


Figure 3.5. The variable patchworks independently defined on the seven generic images.

For an unknown facial expression, the image is registered with each prototypic facial expression. HOG descriptors are generated from the registered images and, then concatenated into a single 7 × 432 dimensional feature vector codifying the unknown facial expression (see figure 3.3).

3.3.3 Support vector machines (SVM)

Support vector machines are known as stronger classifiers for high dimensionality problems, and powerful algorithms for solving difficult classification problems [34, 180].

Different kernel functions (e.g. polynomial, sigmoid, and radial basis function) are used to non–linearly map the input data to a (linearly separable) high–dimensional space.

To handle the multiclass problem, the "one–versus–all" and the "one–versus–one" are the two most adopted strategies for building multiclass–SVM from multiple binary–SVM classifiers. The first paradigm is based on the "winner–take–all" rule, in which the class is assigned by the SVM with the highest output score. Whereas in the “one–versus–one” strategy, each time a pairwise SVM assigns the instance to one of the two classes, the vote of the winner class is increased by one. At last, the class with a maximum wins determines the label of the instance.

The one–over–all strategy may suffer from the problem of unbalanced positive, relative to negative data. Thus, every SVM–classification made along this work utilizes the one-versus-one strategy to solve for the *seven*–class problem.

3.4 Evaluation protocols and experimental results

3.4.1 Protocols

To perform comparison of our results with other existing methods, we studied three different division strategies of the JAFFE database.

In the first study, we applied a cross–validation procedure as in [116, 160, 211]. At each training cycle, the database was randomly divided into 10 roughly equal–sized segments in terms of different facial expressions, of which nine segments were used for training, and the remaining segment was used to test the generalization performance, with the results averaged over 30 distinct cycles. The random process was performed such that the seven facial expressions were roughly equally distributed over the 10 segments.

In the second study, we applied the leave–one–out strategy as in [16, 160] where, at each cycle, we used only one image to test the recognition performance and the remaining images were used for training and the results averaged over all cycles.

In the third study, to test the generalization ability across different individuals, we divide the database into 10 partitions, in which each partition corresponds to an expresser. The process is repeated 10 times, so that the class corresponding to each expresser is used once as the test set. In [116] authors evaluated their system on 9 expressers, since the subject with the initials “NM” was supposed to be an outlier expresser. In the present study, all the subjects present in the database have been considered.

3.4.2 Performance evaluation

Table 3.1 shows the performance comparisons between our proposed system and the other existing systems using the same JAFFE database. The experimental results show that our proposed method, using the SIFT–flow HOG, outperforms all other published works, whatever evaluation protocol was used, with a generalization rate of up to 99.52% ($s.d. = 2.56\%$) with the leave–one–out evaluation protocol, and 98.73% ($s.d. = 2.1\%$) with the cross–validation protocol.

Regarding the person–independent protocol, the SF/HOG feature set gave a superior performance of 86.69% ($s.d. = 15.85\%$).

The proposed method successfully provides both accuracy and computational efficiency. Currently, it takes the system, approximately, 0.09 sec to process one input image of size 256×256. All the experiments were conducted using the same SIFT–flow parameters. The data term and the smoothness term thresholds were, respectively, set to 10 and 40×255 in equation 3.1, with $\eta = 0.005 \times 255$ and $\alpha = 2 \times 255$.

We investigated the effects of two distinct kernels of SVM (linear and Radial Basis functions) among different strategies. We applied 10–fold cross–validation on the training partition to determine the best SVM kernel. At each fold, one subset is used for validation using the classifier trained on the nine subsets [85]. Overall, linear and RBF–kernel performed comparably. Linear separability of the SF/HOG image representation may have been responsible for the relatively high performance with the linear SVMs.

Our generalization rates for the more challenging person–independent facial expression recognition (86.69%) were superior to Lyons *et al.* [115], who reported a rate of 56.8% when all subjects were considered. Our SIFT–flow registration strategy allied with the HOG descriptors seems to better dissociate expression factors and identity.

Table 3.1. Comparison of the performances among different systems on JAFFE database.

	Strategy	Generalization Rate (%)
	Published works	
Lyons *et al.* [116]	cross–validation	92
Lyons *et al.* [115]	person–independent	56.8
Zhang *et al.* [211]	cross–validation	90.1
Buciu *et al.* [16]	leave–one–out	90.34
Shih *et al.* [160]	cross–validation	95.71
	leave–one–out	94.13
Feng *et al.* [61]	cross–validation	93.8
Fu *et al.* [70]	cross–validation	94.76
Guo *et al.* [79]	cross–validation	91
Cheng *et al.* [21]	cross–validation	86.89
	leave–one–out	93.43
	person–independent	55
	This work (SF/HOG)	
Linear/RBF SVM	cross–validation	98.73
Linear/RBF SVM	leave–one–out	99.52
RBF SVM	person–independent	86.69
Linear SVM		85.65

In comparison with a set of well–known classifiers (Tab. 3.2), our method obtained the highest overall accuracy (87%). The second best accuracy was achieved by the Gaussian process with only 55%. However, to be fair, we must notice that these classifiers were trained without any class–based feature selection/extraction, and observations were based solely on the image pixel intensities [21].

Further, in this study, we tested several values for the SVM–hyperplane parameter τ. The classification accuracies of the different facial expressions on the JAFFE database are shown in figure 3.6. The τ values indicate the different thresholds we used to translate the optimal separating hyperplane. The three strategies are represented, respectively, the leave–one–out, cross–validation, and person–independent.

Table 3.2. Performance comparison with reported results of some well–known learning algorithms using person–independent strategy

	“KA”	“KL”	“KM”	“KR”	“MK”	“NA”	“NM”	“TM	“UY”	“YM”	Overall
	Without any feature selection/extraction										
Gaussian process	56	50	63	55	76	61	40	33	47	68	55
SVM	26	13	13	15	14	14	15	14	14	18	16
3-NN	39	31	45	10	57	28	25	38	28	54	36
5-NN	30	45	63	30	66	23	20	38	19	50	39
Naive Bayes	34	63	22	10	42	33	35	47	42	68	40
Classification tree	17	36	36	35	23	19	35	09	42	18	27
C4.5 decision tree	17	36	50	30	42	42	25	28	28	27	33
	Our SF/HOG feature set										
SVM	87	95	100	90	76	48	85	100	86	100	87

The leave–one–out exhibited an accuracy gain at $\tau = -0.1$ with a score of 99.52% ($s.d. = 2.56\%$), relative to the 99.05% ($s.d. = 3.56\%$) obtained by the default hyperplane ($\tau = 0$) of the RBF–SVM. Similarly, the cross–validation strategy produced the highest score (98.73% $s.d. = 2.1\%$) at $\tau = -0.1$ against (98.25% $s.d. = 2.6\%$) at the default hyperplane. Likewise, the person–independent strategy gave unchanged highest score (86.69% $s.d. = 15.85\%$) at $\tau = 0$.

To compare our approach to a classical eyes–based alignment technique [41], we tested an adaptive window–based HOG that used a dynamic detection–window defined over eye position measurements as shown in figure 3.7. As we can see from figure 3.6, the SIFT–flow HOG performed substantially better than the eyes distance adaptive HOG. Especially, with the person–independent strategy, 86.69% ($s.d. = 15.85\%$) for the SF/HOG against only 62.59% ($s.d. = 23.73\%$) for the adaptive HOG. This may be attributable to the erroneous or noisy eye positions, compared to the densely matching process used by the SF/HOG.

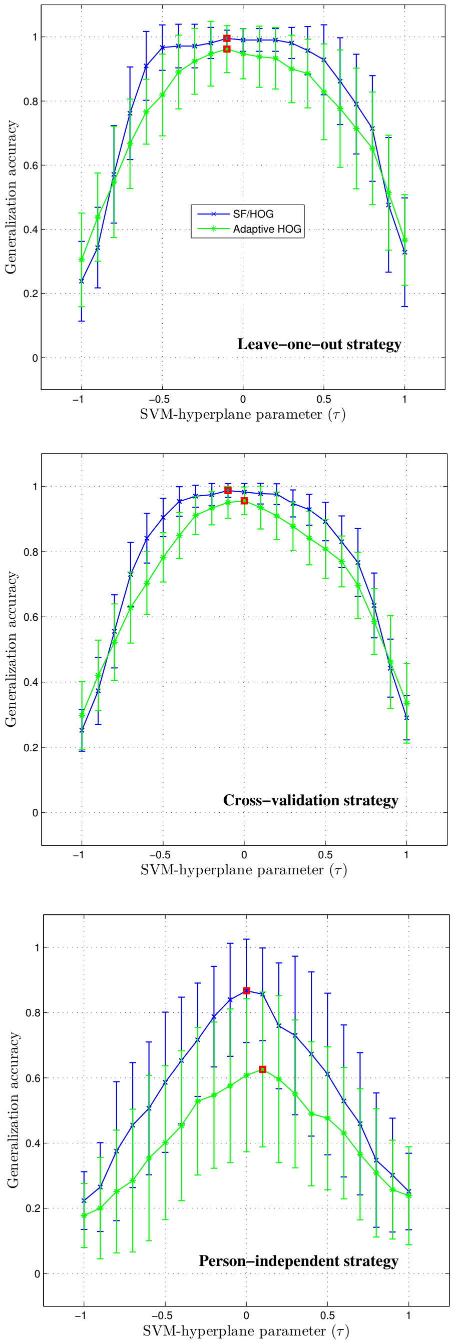
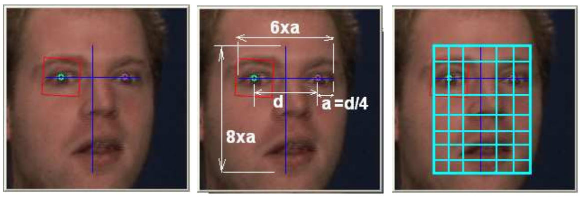


Figure 3.6. Relaxation of the optimal separating hyperplane parameter, SF/HOG vs. adaptive HOG.



(i) *(ii)* *(iii)*

Figure 3.7. Eye distance adaptive window–based HOG : (i). **In–plane rotated face,** (ii). **The baseline distance** a **extracted from physiognomic measurements of the rotated face,** (iii). **The cropped ROI based on** a.

3.5 Conclusion

In this paper, we presented a system for facial expression recognition. We compared various classification schemes with different feature representations. In particular, to deal with the inherent problem of facial alignment due to person–specific appearance, we explored a generic face alignment procedure based on the SIFT–flow registration technique.

The experiment evaluation on the JAFFE dataset, showed that the proposed system, using SIFT–flow registration with HOG descriptors and a RBF–kernel SVM, outperforms all other published methods including those with manual intervention.

We adopted a random validation strategy in which the 10–fold cross–validation was performed independently in many trials. The overall generalization rate reached 99.52% on average with a small standard deviation of 2.56%, by using the leave–one–out strategy, and 98.73% ($s.d. = 2.1\%$) with the cross–validation strategy.

For the much more challenging generic facial expression recognition, where expressions were considered independently of identity, our approach also yielded the best generalization performance with a recognition rate of 86.69%.

Finally, over the different evaluation protocols, our results have demonstrated that the SF/HOG may allow the construction of efficient generic facial expression recognition systems, that can meet the real–time requirement.

Acknowledgment

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Chapitre 4

SUIVI DE POINTS SAILLANTS POUR LA RECONNAISSANCE D'EXPRESSIONS FACIALES

4.1 Introduction

L'analyse automatique d'une séquence d'images faciales est très sensible au performance du suivi, une tâche qui se voit compliquée à cause des changements pouvant survenir dans l'environnement de prise de vue, et plus particulièrement à cause de la variabilité de l'apparence du visage dont la non–rigidité vient ajouter un autre degré de difficulté. Pour outrepasser tous ces problèmes, plusieurs techniques ont été développées qui peuvent être réparties en quatre différentes catégories à savoir celles qui se basent sur les composantes du visages, celles qui explorent l'apparence dites aussi holistiques, celles utilisant l'appariement de gabarits et les méthodes non–holistiques qui utilisent des modèles géométriques du visage mettant en évidence des attributs faciaux (voir Section 1.2.1).

L'analyse d'images faciales par les méthodes géométriques (Fig. 4.1) se distingue par rapport aux autres catégories en démontrant une validité concomitante au codage manuel par le système d'encodage[1] d'action faciale [27, 32]. En outre, lorsque les attributs faciaux sont correctement localisés, les méthodes géométriques se partagent plusieurs avantages qu'elles tirent de la structure explicite du visage et d'une mise en œuvre plus pratique. De plus, ces méthodes permettent une réduction/élimination des erreurs de localisation par inter–correction [87].

Le suivi des attributs faciaux s'inspire des techniques classiques de mise en correspondance qui permettent d'extraire deux ensembles de primitives à partir de

[1] FACS, voir Section 1.2.2.

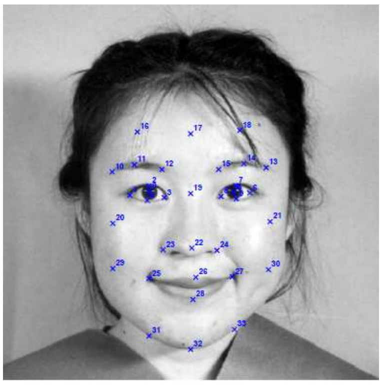


Figure 4.1. Exemple de méthode géométrique pour la reconnaissance d'expression faciale montrant les positions des points saillants d'où sont tirés les attributs faciaux [211].

deux images entre lesquels elles essaient d'établir une correspondance. Quant aux méthodes conventionnelles de corrélation, elles comparent deux fenêtres sur deux trames (frames) et la valeur maximale de la corrélation croisée détermine la meilleure position relative. Toutefois, ces techniques requièrent une recherche exhaustive dans un voisinage défini.

Des techniques plus récentes ont été mises en œuvre pour déterminer directement la position relative (disparité) sans qu'aucun procédé de recherche ne soit utilisé. Dans cette catégorie, les approches basées sur la réponse en phase des filtres de Gabor ont suscité un grand intérêt de par leur robustesse ainsi que leur motivation biologique [66, 166].

Pour extraire les primitives discriminatives de l'image, le filtrage de Gabor [158] a été adopté par plusieurs approches dont la plus grande partie se base sur la réponse en amplitude sous forme d'un ensemble de coefficients [102, 107, 168, 178]. Cependant, la réponse en phase de cette famille de filtres peut être considérée comme un bon critère pour mesurer le mouvement dû à sa propriété de variance au déplacement.

Au chapitre 5, on se sert de cette propriété intrinsèque de la phase pour permettre le suivi de points caractéristiques du visage le long de séquences vidéo. On y propose une modification d'une technique de suivi basée sur la phase de Gabor, ceci inclut une redéfinition de la mesure de confiance et introduit une procédure itérative d'estimation du déplacement.

Les positions des points caractéristiques, dont le choix tient au facteur de stabilité par rapport aux déformations, serviront de repère pour superposer sur le visage un graphe de 28 nœuds. La superposition utilise la transformation de Procrustes (translation, rotation et changement d'échelle) pour contraindre le graphe à s'ajuster à la topologie du visage via les points de repère.

Motivés par le fait que les expressions faciales sont ambigües et incohérentes d'une personne à une autre (Fig. 4.2) et souvent dépendantes du contexte, au chapitre 5, nous synthétiserons un système personnalisé de reconnaissance d'expressions faciales garantissant une performance optimale.

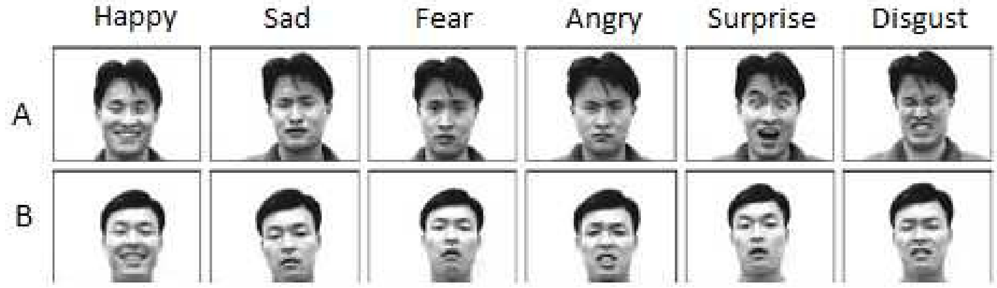


Figure 4.2. Deux ensembles d'expressions faciales de deux personnes différentes [97].

4.2 Transformation Procrustes généralisée

Soit $\mathbf{G}$ une configuration centrée dans $\mathcal{C}^k$ ($\mathbf{G}\,\mathbf{1}_k = 0$) d'un graphe G représenté sous forme d'un vecteur de k points 2d, en représentation complexe[2] $x + \imath y$. L'analyse Procrustes [96, 118] permet de modifier la forme de G_2 pour le faire correspondre à G_1 selon une transformation de similarité (Eq. 4.1) en minimisant la distance (Eq. 4.2) qui mesure le degré de dissemblance entre les formes G_1 et G_2.

$$\begin{cases} \mathbf{G}_1 = \alpha\,\mathbf{1}_k + \beta\,\mathbf{G}_2 & \alpha, \beta \in \mathcal{C} \\ \beta = |\beta|\; e^{\imath \angle \beta} \end{cases} \tag{4.1}$$

$$d_F(\mathbf{G}_1, \mathbf{G}_2) = 1 - \frac{|\mathbf{G}_1^*\,\mathbf{G}_2|^2}{\|\mathbf{G}_1\|^2\;\|\mathbf{G}_2\|^2} \tag{4.2}$$

On déduit la translation ($\alpha\,\mathbf{1}_k$), la mise en échelle $|\beta|$ et la rotation $\angle\beta$ à partir de,

$$\alpha = \bar{G_2} \qquad \beta = \mathbf{G}_1^*\mathbf{G}_2$$

4.3 Filtrage de Gabor

Pour extraire des caractéristiques discriminantes de l'image, les filtres de Gabor offrent un meilleur compromis entre la résolution spatiale et la résolution fréquentielle, et permettent de capturer des propriétés visuelles intéressantes dans l'image, telles que la localisation spatiale, la sélectivité angulaire (l'orientation) et la localisation spatio–fréquentielle. En outre, ces filtres fournissent un avantage considérable en traitement d'images parce qu'ils permettent notamment d'adresser le problème de la précision sous–pixel.

Le filtrage de Gabor permet de décrire, via ce qu'on appelle communément un jet, la répartition des niveaux de gris dans une image $I(\mathbf{x})$ au voisinage d'un pixel donné,

[2] La notation $\mathbf{G}^*$ dénote la transposée du conjugué complexe de $\mathbf{G}$.

sous forme d'un vecteur réarrangé de coefficients à partir d'une transformation définie par la convolution suivante,

$$J(\mathbf{x}) = \int I(\mathbf{x'})\Psi_j\left(\mathbf{x} - \mathbf{x'}\right) d^2\mathbf{x'} \tag{4.3}$$

avec une famille de noyaux de Gabor,

$$\Psi_j\left(\mathbf{x}\right) = \frac{\mathbf{k_j}\ \mathbf{k_j}^T}{\sigma^2}\ \exp(-\frac{\mathbf{k_j}\ \mathbf{k_j}^T\ \mathbf{x}\ \mathbf{x}^T}{2\sigma^2})\Big[\exp(i\mathbf{k_j}\ \mathbf{x}) - \exp(-\frac{\sigma^2}{2})\Big] \tag{4.4}$$

où $\mathbf{k_j}$ définit le vecteur d'onde comme suit,

$$\mathbf{k_j} = (k_\nu \cos(\phi_\mu), k_\nu \sin(\phi_\mu)) \quad \text{avec } k_\nu = 2^{-\frac{\nu+2}{2}}\ \pi \text{ et } \phi_\mu = \mu\frac{\pi}{8} \tag{4.5}$$

Les paramètres ν et μ correspondent, respectivement, à l'échelle et l'orientation des noyaux de Gabor, l'écart–type de la Gaussienne (σ/k_ν) est contrôlé par le paramètre σ dont la valeur est fixée à 2π.

Le premier facteur de l'équation de noyaux de Gabor (Eq. 4.4) représente l'enveloppe Gaussienne $\exp(-\frac{\mathbf{k_j}\ \mathbf{k_j}^T\ \mathbf{x}\ \mathbf{x}^T}{2\sigma^2})$, modulée par une fonction sinusoïdal complexe. Le terme $\exp(-\frac{\sigma^2}{2})$ compense pour la valeur moyenne non-nulle de la composante en cosinus. Un exemple de noyaux à différentes échelles est donné par la figure 4.3.

Le jet $J(\mathbf{x}) = \Big\{a_j\ \exp(i\phi_j)\Big\}$ est typiquement formé d'un ensemble de 40 coefficients complexes où $j = \mu + 8\nu$. Cet ensemble correspond à 5 fréquences $(\nu = 0, .., 4)$ et 8 orientations $(\mu = 0, ..7)$ différentes. Le terme a_j qui dénote l'amplitude de la réponse complexe pour une orientation et une fréquence particulière varie, graduellement, en fonction du vecteur d'onde.

4.3.1 Mesure de similarité

La comparaison de primitives (p.ex. pour la mise en correspondance des modèles géométriques) implique une comparaison de jets [102] par la fonction de similarité

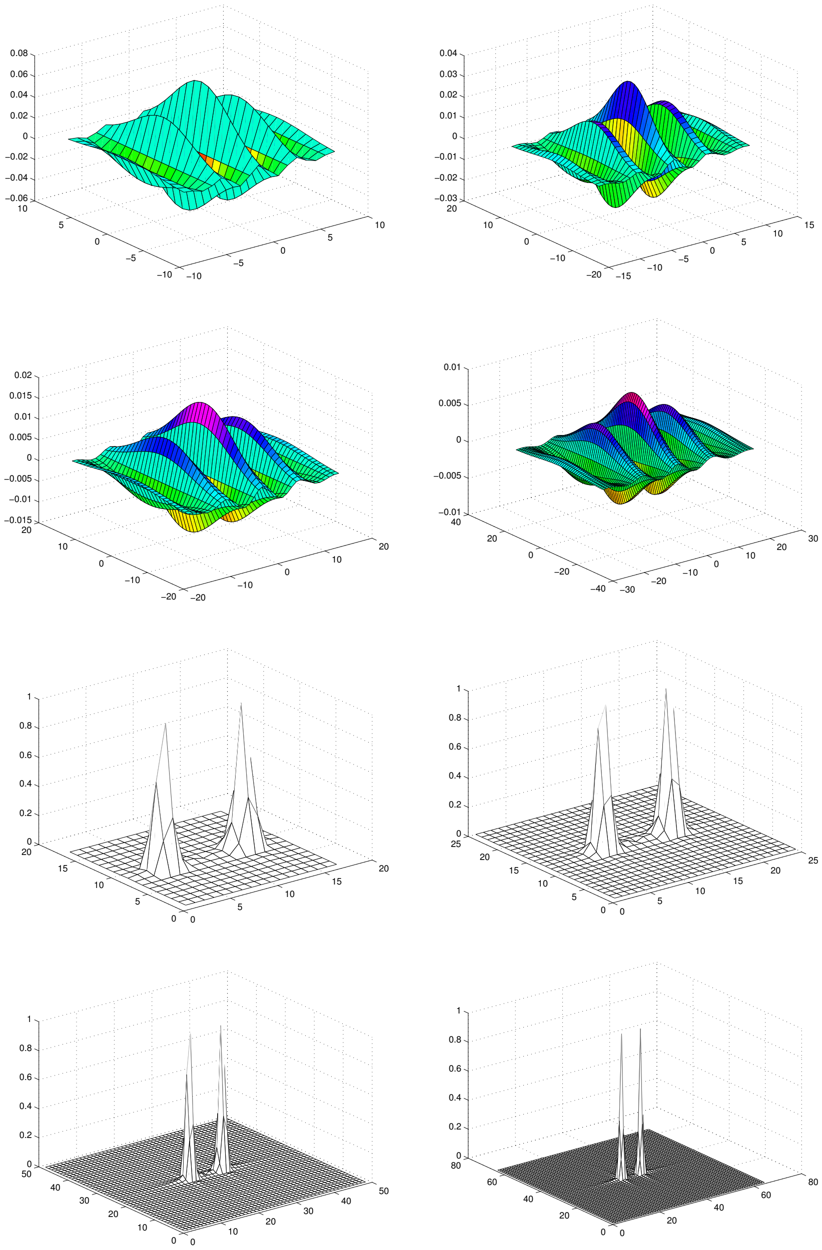


Figure 4.3. Exemple de filtres de Gabor dans le domaine spatial et les réponses fréquentielles correspondantes.

(Eq. 4.6).

$$S(J, J') = \frac{\sum_j a_j \ a'_j}{\sqrt{\sum_j a_j^2 \ \sum_j {a'_j}^2}} \tag{4.6}$$

Cette comparaison peut–être plus fiable en considérant la phase ϕ_j, cependant des problèmes de mise en correspondance peuvent surgir en raison des propriétés de variance au déplacement de la phase: deux pixels adjacents, n'auront pas nécessairement des jets similaires, bien qu'ils ont pratiquement une même répartition de niveaux de gris. La solution est de compenser les changements induits par la phase par le terme $\mathbf{d} \cdot \mathbf{k}$. On obtient, ainsi, une similarité sensible à la phase (Eq. 4.7).

$$S_\phi(J, J') = \frac{\sum_j a_j \ a'_j \cos(\phi_j - {\phi'}_j - \mathbf{d} \cdot \mathbf{k})}{\sqrt{\sum_j a_j^2 \ \sum_j {a'_j}^2}} \tag{4.7}$$

où $\mathbf{d}$, qui doit être estimé à partir de $J(\mathbf{x'})$, représente le vecteur de déplacement par rapport à la position prédite $\mathbf{x'}$.

4.3.2 Estimation du déplacement

Le vecteur de déplacement $\mathbf{d} = (d_x, d_y)$ est calculé par une technique d'estimation de disparité [65, 166]. Une maximisation de la forme réduite du développement de Taylor (Eq. 4.8) de S_ϕ (Eq. 4.7) donne le déplacement optimal $\mathbf{d}_{opt}$ (Eq. 4.9), qui permet d'associer la position du point facial considéré dans l'ancien trame à sa nouvelle position dans le nouveau trame.

$$S_\phi(J, J') \approx \frac{\sum_j a_j \ a'_j \left[1 - 0.5 \left(\phi_j - {\phi'}_j - \mathbf{d} \cdot \mathbf{k_j}\right)^2\right]}{\sqrt{\sum_j a_j^2 \ \sum_j {a'_j}^2}} \tag{4.8}$$

$$\mathbf{d}_{opt}(J, \acute{J}) = \frac{1}{\Gamma_{xx}\Gamma_{yy} - \Gamma_{xy}\Gamma_{yx}} \begin{pmatrix} \Gamma_{yy} & -\Gamma_{yx} \\ -\Gamma_{xy} & \Gamma_{xx} \end{pmatrix} \begin{pmatrix} \Phi_x \\ \Phi_y \end{pmatrix} \tag{4.9}$$

si $\Gamma_{xx}\Gamma_{yy} - \Gamma_{xy}\Gamma_{yx} \neq 0$, avec $\Phi_x = \sum_j a_j a'_j \ k_{jx}({\phi'}_j - \phi_j)$ et $\Gamma_{xy} = \sum_j \ a_j \, a'_j \, k_{jx} \, k_{jy}$

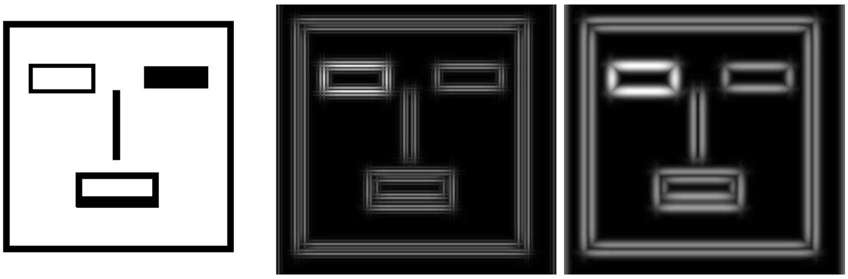


Figure 4.4. Réponse d'un filtre standard de Gabor (image du milieu), réponse d'un filtre d'énergie (image de droite).

Les écarts de phase sont corrigés par $\pm 2\pi$, pour les ramener dans l'intervalle $(-\pi, \pi]$. Pour accélérer le calcul des convolutions, essentiellement lors du suivi, une architecture multiéchelle peut être considérée (voir chapitre 5). Ainsi, le calcul des déplacements sur une image sous–échantillonnée utilisera une gamme réduite de fréquences. En plus, les mouvements seront plus locaux que dans l'image originale.

4.4 Filtres d'énergie de Gabor

Par rapport aux filtres classiques de Gabor, le filtre d'énergie de Gabor produit un résultat plus lisse en réponse à un contour ou à un trait d'une certaine largeur, avec un maximum local exactement au milieu du trait [76, 142] (voir figure 4.4). Ce filtre est obtenu par "superposition" de phases. Le filtre le plus utilisé combine par la norme L_2 les deux convolutions correspondant aux deux phases $(\varphi_0 = 0)$ et $(\varphi_1 = \pi/2)$.

Pour plus de détails concernant ce type de filtres voir l'article présenté en annexe G.

Chapitre 5

(ARTICLE) ITERATIVE GABOR PHASE–BASED DISPARITY ESTIMATION FOR "PERSONALIZED" FACIAL ACTION RECOGNITION

Ce chapitre présente une extension de deux articles ayant été publiés (voir annexes D et E) comme l'indique les références bibliographiques [38, 42]

> M. Dahmane et J. Meunier. Enhanced phase–based displacement estimation - An application to facial feature extraction and tracking. Dans *Proc. of Int. Conference on Computer Vision Theory and Applications*, pages 427–433, 2008.
>
> M. Dahmane et J. Meunier. Individual feature–appearance for facial action recognition. Dans *Proc. of Int. Conference on Image Analysis and Recognition*, pages 233–242, 2011.

La version étendue a été soumise pour publication dans la revue scientifique *Signal processing: Image Communication*, par Mohamed Dahmane, Jean Meunier et Sylvie Cossette.

Abstract

Within the affective computing research field, researchers are still facing a big challenge to establish automated systems to recognize human emotions from video sequences. Performances are quite dependent on facial feature localization and tracking.

In this paper, we present a method based on a coarse–to–fine paradigm to characterize a set of facial fiducial points using a bank of Gabor filters. When the first face image is captured, the coarsest level is used to estimate a rough position for each facial feature. Afterward, a coarse–to–fine displacement refinement on an image pyramid is performed. The positions are then tracked over the subsequent frames using a modification of a fast Gabor–phase based technique. This includes a redefinition of the confidence measure and introduces an iterative conditional disparity estimation procedure.

We used the proposed tracking process to implement a "personalized" feature–based facial action recognition framework, motivated by the fact that the same facial expression may vary differently across humans.

Experimental results show that the facial feature points can be localized with high accuracy and tracked with sufficient precision leading to a better facial action recognition performance.

5.1 Introduction

The computer vision community is interested in the development of techniques, such as automated facial expression analysis (AFEA), to figure out the main element of facial human communication, in particular for human–computer interaction (HCI) applications or, with additional complexity, in meeting video analysis, and more recently in clinical research.

AFEA is highly sensitive to face tracking performance, a task which is rendered difficult owing principally to environment changes, and appearance variability under different head orientations, and non–rigidity of the face. To meet these challenges, various techniques have been developed and applied. Prior works have focused on both images and video sequences, and different approaches were investigated including feature–based and appearance–based techniques [57, 139]. Most of these techniques show satisfactory results using databases that were collected under non realistic conditions [194]. An advance emotion recognition system needs to deal with more natural behaviors in large volumes of un–segmented, un–prototypical, and natural data [156]. Moreover, these methods are still providing inaccurate results due to the variation of facial expression across different people and even for the same individual, since facial expression is context–dependent. It is noted that the notion of “universality”, as opposed to “personalization”, has been fashionable in the area of facial expression recognition [97].

However, it is advantageous to design a personalized model for facial emotion recognition, since facial physiognomy that characterizes each person leads to a personal facial action display [56] (see figure 5.1). This would explain why facial action units may be considered as “Facial behavior signatures” to recognize individuals [26].

A great number of results on facial expression recognition were reported in the literature of the last few decades. Nevertheless, the approach we report here, is reminiscent of only a few of them.

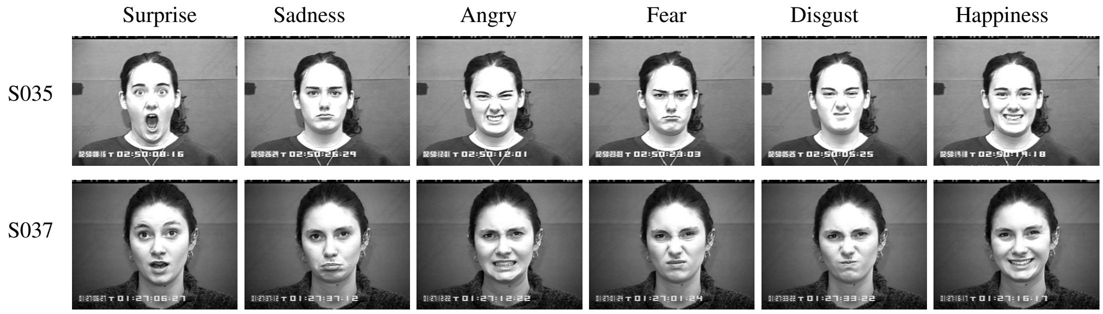


Figure 5.1. Two sets of facial displays from different persons.

In [113], authors employed for each individual in the datasets a person–dependent active appearance model (AAM), that was created for Action Units (AUs) recognition by using similarity normalized shape and similarity normalized appearance. By integrating user identification, the person–dependent method proposed in [22] performs better than conventional expression recognition systems, with high recognition rate reaching 98.9%. In reference [56], authors obtained the best facial expression recognition results by fusing facial expression and face identity recognition. Their personalized facial expression recognition setup combines outputs of convolutional neural networks that were either trained for facial expression recognition or face identity recognition. In [80], the authors conclude that the recognition rates for familiar faces reached 85%. In contrast, for unfamiliar faces, the performance score does not exceed 65%. Their system utilizes elastic graph matching and a personalized gallery to recognize expression on identified face. More recently, authors in [97], designed a “personalized” classifier using a neurofuzzy approach for “personalized” facial expression recognition. They reported a recognition rate of 91.8% on the Cohn–Kanade database (described below).

For both person–independent and person–dependent approaches, facial expression recognition (FER) rate is highly dependent on facial tracking performance. This task is rendered difficult due to environment changes, appearance variability under different head orientations, and the face non–rigidity. Several approaches have been suggested to alleviate these problems, which can be divided into knowledge–, feature–, template–, and appearance–based approaches. The feature–based techniques demonstrate high concurrent validity with manual FACS[1] coding [27, 32]. Furthermore, they have some common advantages such as explicit face structure, practical implementation, and collaborative feature–wide error elimination [87]. However, the tracking performance depends on the precise configuration of the local facial features, which poses a significant challenge to the geometric–based facial expression analysis, since subtle changes in the facial expression could be missed due to errors in facial point localization [137]. Though an effective scheme for the facial feature points tracking can compensate for this drawbacks, it could be possibly not sufficient. Feature–based

[1] Facial Action Coding System.

approaches including only shape information may be rather irrelevant [113]. Figure 5.2 shows an example of two different facial expressions (*fear* vs. *happy*) where the respective appearances were significantly different, while the two expressions have a high degree of shape similarity. Therefore, including appearance matching should improve the recognition rate, which can be done by including the local appearance around each facial feature [102].

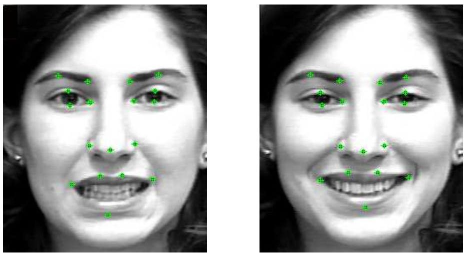


Figure 5.2. Facial point position may not be sufficient to achieve reliable FER (e.g. fear vs. happy).

In this paper, we propose a modified Gabor phase–based tracking approach that we used to track a set of facial key points using an iterative conditional displacement estimation algorithm (for simplicity, along this work, we use the initials `ICDE`). These points are then used to perform a feature–based "personalized" facial action recognition system using a prestored facial action graphs. At the first frame of the video, four facial key points are automatically found and tracked over time. Then, at each frame, the most similar graph, through the prestored ones, is chosen based on these key points. The selection utilizes Procrustes transformation and a set of Gabor jets that are stored at each node of the graph.

In what follows, we will describe the modified Gabor phase–based tracking approach, and give details about the proposed `ICDE` algorithm in section 5.2. In section 5.3, we describe the facial action recognition process. Performance evaluation is presented in section 5.4. Finally, in section 5.5 we draw some conclusions.

5.2 Iterative Gabor–phase–based displacement estimation

In both detection and tracking processes, the classical matching techniques extract features from two frames and tries to establish a correspondence, whereas correlation–based techniques compare windowed areas in two frames, and the maximum cross correlation value provides the new relative position. Recent techniques have been developed to determine the correct relative position (disparity[2]) without any searching process as it is required by the conventional ones. In this category, Gabor phase–based approaches have attracted attention because of their biological plausibility and robustness [64, 66, 166].

In the literature, one can find several attempts at designing feature–based methods using Gabor wavelets [158], due to their interesting and desirable properties including spatial locality, self-similar hierarchical representation, optimal joint uncertainty in space and frequency. However, most of them are based on the magnitude part of the filter response [102, 107, 168, 178]. In fact, under special consideration, particularly because of shift–variant property, the Gabor phase can be a very discriminative information source [208]. In this paper, we use this property of Gabor phase for facial feature tracking.

5.2.1 Gabor wavelets

A Gabor jet $J(\mathbf{x})$ describes via a set of filtering operation (Eq. 5.1), the spatial frequency structure around the $N \times N$ neighborhood of pixel $\mathbf{x}$, as a set of complex coefficients.

$$J_j(\mathbf{x}) = \int_{N^2} I(\mathbf{x}')\Psi_j\left(\mathbf{x} - \mathbf{x}'\right) d\mathbf{x}' \tag{5.1}$$

[2] Along this work, we use, interchangeably, the words "disparity" and "displacement"

A Gabor wavelet is a complex plane wave modulated by a Gaussian envelope:

$$\Psi_j\left(\mathbf{x}\right) = \eta_j \, e^{-\frac{\|\mathbf{k}_j\|^2\,\|\mathbf{x}\|^2}{2\sigma^2}} \left[e^{\imath\,\mathbf{k}_j \cdot \mathbf{x}} - e^{-\frac{\sigma^2}{2}} \right] \tag{5.2}$$

where $\sigma = 2\pi$, and $\mathbf{k}_j = (k_{jx}, k_{jy}) = (k_\nu \cos(\phi_\mu), k_\nu \sin(\phi_\mu))$ defines the wave vector, with

$$k_\nu = 2^{-\frac{\nu+2}{2}}\,\pi \qquad \text{and} \qquad \phi_\mu = \mu\frac{\pi}{8} \tag{5.3}$$

Notice that the last term of equation 5.2 compensates for the non–null average value of the cosine component. We choose the term η_j so that the energy of the wavelet Ψ_j is unity (Eq. 5.4).

$$\int_{N^2} \left|\Psi_j\left(\mathbf{x}\right) d\mathbf{x}\right|^2 \;=\; 1 \tag{5.4}$$

A jet $J(\mathbf{x}) = \{a_j\, e^{\imath\,\phi_j} \;/\; j = \mu{+}8\nu\}$, is commonly defined as a set of $5{\times}8 = 40$ complex coefficients (a_j : amplitude, ϕ_j : phase) constructed from different filters (Fig. 5.3) spanning different orientations ($\mu \in [0, 7]$) under different scales ($\nu \in [0, 4]$). We use for each filter a variable window size that depends on the ratio σ/k_ν, with $\sigma = 2\pi$. Gabor filter bank responses to an expressive (angry) face is depicted in figure 5.4.

5.2.2 Facial key point detection

When the first face image is captured, a pyramidal image representation is created, where the coarsest level is used to find near optimal starting points for the subsequent facial feature localization and refinement stage. Each trained graph (Fig. 5.5) from a set of prestored face graphs is displaced as a rigid object over the coarsest image. The graph position that maximizes the weighted magnitude–based similarity function (Eq. 5.5) provides the best fitting node positions.

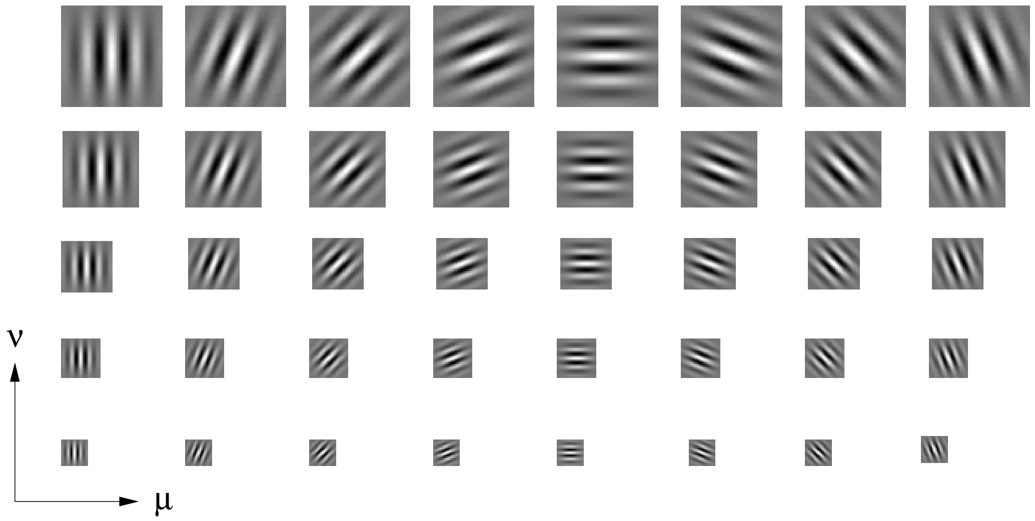


Figure 5.3. Real part of the 5×8 Gabor filters.

$$Sim(\mathrm{I}, \mathrm{G}) = \frac{1}{L} \sum_{l}^{L} S(J_l, J_l') \tag{5.5}$$

$S(J, J')$ refers to the similarity between the jets of the corresponding nodes (Eq. 5.6), $L = 28$ stands for the total number of nodes.

$$S(J, J') = \sum_j c_j \frac{a_j \, a_j'}{\sqrt{\sum a_j{}^2 \sum a_j'^2}} \quad \text{with} \quad c_j = \left(1 - \frac{\left|a_j - a_j'\right|}{a_j + a_j'}\right)^2 \tag{5.6}$$

5.2.3 Facial key point position refinement and tracking

The rough position of the facial key points are refined by estimating the displacement using the `ICDE` procedure (sec. 5.2.3.2) so as to handle the non–rigid facial deformation. The procedure is also used to track the key points over time. The difference is that, in the refinement stage, the two jets are calculated in the same frame, in this case the disparity represents the amount of position correction. In the case of tracking, the two jets are processed from two consecutive frames and the disparity represents the displacement amount.

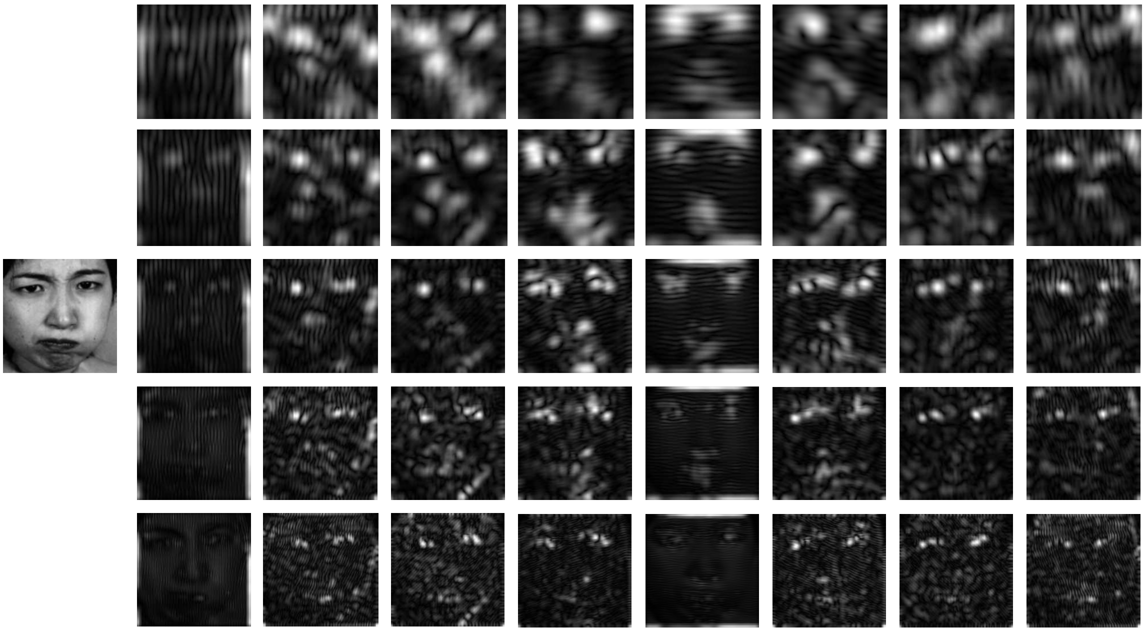


Figure 5.4. Gabor filter bank responses to an angry face (left image).

5.2.3.1 Disparity estimation

The displacement estimation technique [166] exploits the strong variation of the phases of the complex filter response [119]. Later adopted by [215] and investigated in [119] and [197], the technique is based on the maximization of a phase–based similarity function which is equivalent to minimize the squared error within each fre-

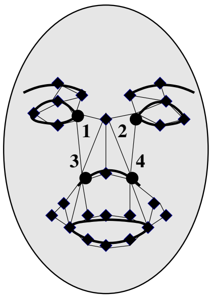


Figure 5.5. The tracked key points (circle) and the adjusted points (diamond).

quency scale ν given two jets J and J' (Eq. 5.7), as it has been proposed in [166].

$$e_\nu^2 = \sum_\mu c_{\nu,\mu} (\Delta\phi_{\nu,\mu} - \mathbf{k}_{\nu,\mu} \cdot \mathbf{d}_\nu)^2 \tag{5.7}$$

The optimization function integrates a saliency term (Eq. 5.8) as weighting factor $c_{\nu,\mu}$, privileging displacement $\mathbf{d}_\nu$ estimation from filters with higher amplitude response. Also, for such response it seems that the phase is more stable [121]. $\Delta\phi_{\nu,\mu}$ denotes the principal part of the phase difference within the interval $[-\pi, \pi)$.

$$c_j \; = \; a_j \, a_j' \tag{5.8}$$

Authors in [166] defined another weighting factor c_j as a confidence value (Eq. 5.9), that assesses the relevance of a single disparity estimate, and tends to reduce the influence of erroneous filter responses.

$$c_j \; = \; 1 - \frac{\left| a_j - a_j' \right|}{a_j + a_j'} \tag{5.9}$$

Both saliency term and normalized confidence ignore the phase of the filter response. In the present work, we propose to penalize the response of the erroneous filters by using a new confidence measure that combines both amplitude and phase (Eq. 5.10).

$$c_j \; = \; {a_j}^2 \left(1 - \frac{\left| a_j - a_j' \right|}{a_j + a_j'} \right)^2 \; \frac{\pi - |\Delta\phi_j|}{\pi} \tag{5.10}$$

The first term in this formulation represents the saliency term that is incorporated as a squared value of only the amplitude of the *reference* jet J which, contrary to the *probe* jet J', necessarily ensures high confidence. The reference jet consists of the jet calculated from the previous frame or a prestored jet. The second bracket squared–term holds the normalized magnitude confidence. The last term gives more weight to filters where the phase difference (computed within $[-\pi, \pi)$) has a favorable convergence and limits the influence of outlier filters.

The displacement $\mathbf{d}$ can then be estimated with sufficient accuracy by minimizing (Eq. 5.7) which leads to a set of linear equations, that can be directly resolved from (Eq. 5.11).

$$\mathbf{d}(J, J') = \begin{pmatrix} \sum_j \; c_j \, {k_{jx}}^2 & -\sum_j \; c_j \, k_{jx} \, k_{jy} \\ -\sum_j \; c_j \, k_{jx} \, k_{jy} & \sum_j \; c_j \, {k_{jy}}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_j \; c_j \, k_{jx} \Delta\phi_j \\ \sum_j \; c_j \, k_{jy} \Delta\phi_j \end{pmatrix} \tag{5.11}$$

5.2.3.2 Iterative displacement estimation

Disparity estimates $\mathbf{d}_\nu$ can be resolved for the different orientations ϕ_μ relative to each scale ν using the least squared error criterion (Eq. 5.7). The optimal disparity can then be calculated by averaging over all scales with an appropriate weighting coefficients (Eq. 5.9). Some approaches use a least squared solution in one pass over all considered frequencies and orientations [197], others propose, at first, to use a higher frequencies subset (e.g. $\nu \in [0, 2]$), and then to resolve for a lower frequencies subset (e.g. $\nu \in [2, 4]$).

These solutions may carry risk of unfavorable results since at each scale, there exists a displacement value, above which, the displacement estimate would not be reliable, due to the lack of a large overlap of the Gabor kernels. Obviously, this value depends on the radius (σ/k_ν) of the Gaussian envelope.

As the power spectrum of the Gabor signal (Eq. 5.2) is concentrated in the interval $[-\sigma/(2k_\nu), \sigma/(2k_\nu)]$, we can compute the maximum disparity $\mathbf{d}_\nu^{max}$ that can be estimated within one scale as in equation 5.12.

$$d_\nu^{max} = \frac{\sigma}{2\,k_\nu} = \frac{\pi}{k_\nu} \tag{5.12}$$

If, for instance, the true displacement is $d = 7$ pixels, then according to the

Gabor–kernel family we used (Section 5.2.1), only the lowest frequency filter gives a reliable estimation of the disparity.

Therefore, we propose to estimate the disparity iteratively, from the lowest frequency to a highest critical frequency, depending on a stopping criterion involving the maximum allowed disparity. Some values of d_ν^{max} are shown in table 5.1 as a function of scale.

ν	0	1	2	3	4
$d_\nu^{\max}(pixel)$	2	≈ 3	4	≈ 6	8

Table 5.1. Critical displacement for each frequency.

Given a reference jet $J(\mathbf{x}) = \{a_j\, e^{\imath\phi_j}\}$ calculated at the position x_1 and $J'(\mathbf{x}+\mathbf{d})$=-$\{a'_j\, e^{\imath\phi'_j}\}$ calculated at any peripheral position x_2 the iterative disparity estimation procedure IterativeDisparityEstimation (Fig. 5.6) gives the optimal displacement $\mathbf{d}_{\texttt{opt}} \approx x_2 - x_1$, that makes the two jets as much as possible similar. The first step consists to set ν with the lowest frequency index, then calculate the jet for only the components that refer to the frequency ν at different orientations. After that, we estimate the disparity $\delta\mathbf{d}$ using equation 5.11, by considering at different orientations all the currently processed frequencies. Step 4 compensate for the phase. The process cumulates the disparity in step 5, and performs the convergence test in step 6. If the stopping criterion is not met, i.e. the overall displacement is less than the critical displacement value, we set the current frequency to the next higher one, and repeat from step 2. Iteratively, the algorithm will unroll on the novel position $\mathbf{x}^{\texttt{new}} \leftarrow \mathbf{x} + \mathbf{d}_{\texttt{opt}}$ until a convergence criterion is achieved (i.e. $\mathbf{d}_{\texttt{opt}}$ tends to $\mathbf{0}$) or the maximum number of iterations is reached.

We recall that the ICDE algorithm will be used (1) to improve the facial key point detection (position refinement stage) and (2) to track these points over time.

Procedure ITERATIVEDISPARITYESTIMATION ($\mathbf{x}$)

```
1 $\nu_c$ = lowest frequency index $\nu_c = 4$;

2 Calculate $J'_{\nu_c}(\mathbf{x})$ within the current
    scale $\nu_c$ ;

3 Estimate $\delta\mathbf{d}$ as in (Eq. 5.11) using
    the current processed scale range ie. $\nu_c \leq \nu \leq 4$;

4 $\phi'_j = \left\lfloor \phi'_j - \mathbf{k}_j \cdot \delta\mathbf{d} \right\rfloor_{2\pi}$;

5 $\mathbf{d} = \mathbf{d} + \delta\mathbf{d}$;

6 if $\delta\mathbf{d} > T$ goto (3);

7 Check the stopping criterion:
    if $\|\mathbf{d}\| < d^{max}_{\nu_c}$ put $\nu_c = \nu_c - 1$ and goto (2);

8 $\mathbf{d}_{opt} = \mathbf{d}$;
```

Figure 5.6. Iterative conditional disparity estimation (ICDE) process.

5.3 Facial action recognition

For the personalized facial expression recognition, a set of prestored graph models is constructed as shown in figure 5.5. In this study, the prestored graphs will consist of only one graph per person displaying each expression at its peak. Each graph represents a node configuration that characterizes the appearance of a facial action to recognize. A set of jets, describing the appearance around each point of interest, is generated and attached to the corresponding node. In our implementation, a set of 28 nodes defines the facial graph of a given expression.

5.3.1 Action recognition

From frame to frame, the four facial key points indicated by circles in figure 5.5, which are known to have relatively stable local appearance and to be less sensitive to facial deformations, are tracked and refined. The new positions are used to deform each graph from the prestored graph set using translation, rotation and scaling transformations in order to consider appearance changes (Fig. 5.7).

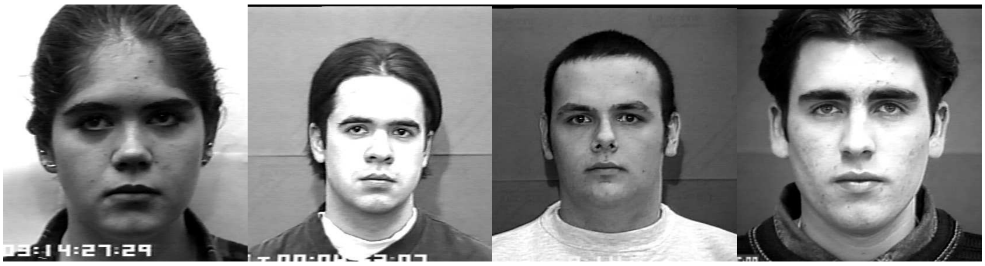


Figure 5.7. Examples of different faces.

These linear transformations are used to adequately deform each prestored graph G_r. Given the positions of the 4 reference key points (Fig. 5.5), the transformation that best wraps these points, is used to adjust the positions of the twenty–four remaining points of G_r . Then for each key point node position a refinement stage is performed to obtain the final position by estimating the optimal displacement of each point using the iterative displacement estimation algorithm (sec. 5.2.3.2). At each refined node position the corresponding jet is recalculated and updated. The final graph with adjusted positions is named distorted graph G_d. The graph G_r defines the facial action that closely corresponds to the displayed facial action, and $sim(G_r, G_d)$ (Eq. 5.5) the scoring value indicating its intensity.

The entire facial action recognition process is presented in the flow diagram of figure 5.8.

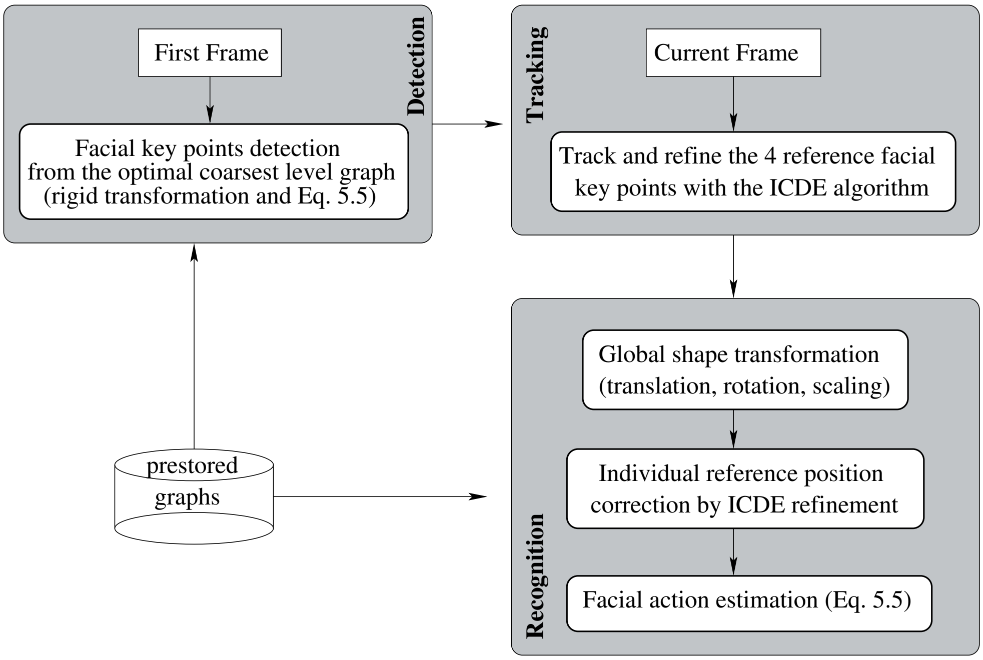

Detection
First Frame
Facial key points detection
from the optimal coarsest level graph
(rigid transformation and Eq. 5.5)
Tracking
Current Frame
Track and refine the 4 reference facial
key points with the ICDE algorithm
Recognition
Global shape transformation
(translation, rotation, scaling)
Individual reference position
correction by ICDE refinement
Facial action estimation (Eq. 5.5)
prestored
graphs

Figure 5.8. Flow diagram: Facial key points tracking and refinement using `ICDE` **algorithm for facial action recognition.**

5.4 Performance evaluation

The videos we used in this work for testing are from the Cohn–Kanade database [95]. The sequential images consist of a set of prototypic facial expressions (happiness, anger, fear, disgust, sadness, and surprise) that were collected from a group of psychology students of different races with ages ranging from 18 to 30 years. Each sequence starts from a neutral expression, and terminates at the peak of a given expression. We selected 15 subjects out of 97 that have complete set of non–ambiguous facial expressions. This was necessary because only combinations of action units (AU) are given in this database, instead of clearly defined facial expressions.

5.4.1 Facial feature tracking

5.4.1.1 Facial localization

As described in section 5.2.2, to initialize the four positions of the facial reference points we used the prestored facial action graphs as trained graphs. Even if a facial action graph was generated for a person displaying a given facial expression at its peak, we used it, successfully, to localize the four reference facial points, since these points were assumed to have relatively stable local appearance and to be less sensitive to facial deformations. This rough localization process was performed via an exhaustive search through the coarsest face image level in the first frame of the video. The optimal subgraph was that which maximizes the weighted magnitude similarity function (Eq. 5.5).

The detection process was performed via a three–level pyramid image representation (Fig. 5.9) to decrease the inherent average latency of the graph search operation by reducing the image search area and the Gabor–jet sizes. For images at the finest level (640 × 480 pixels resolution), jets are defined as sets of 40 complex coefficients constructed from a set of Gabor filters spanning 8 orientations and 5 scales. Whereas those for images at the coarsest level (320×240) are formed by 16 coefficients obtained from filters spanning 8 orientations under 2 scales. The intermediate level images use jets of (8×3) coefficients.

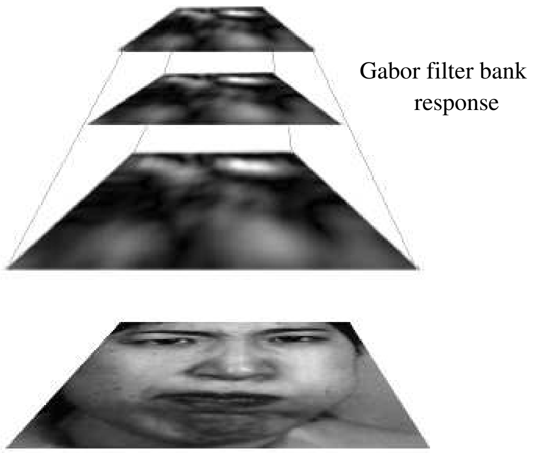


Figure 5.9. Pyramidal image representation.

In the first frame, the rough positions were refined through the three image levels using the `ICDE` procedure (Fig. 5.6) at each image level. The threshold T, in step 6, was set experimentally to 0.07.

5.4.1.2 Facial tracking

The tracking of the reference points was performed by applying the `ICDE` procedure again, this time from frame to frame between each pair of reference points. The displacement was roughly estimated at the coarsest level of the pyramid face image, then gradually propagated and refined from coarse to fine as illustrated in figure 5.10.

To avoid drifting during tracking, for each reference point, a local search was performed through the trained graphs for the subgraph maximizing the weighted magnitude similarity (Eq. 5.5). The rough positions of the four feature points were given by the optimal subgraph node positions, which were then adjusted using once more the `ICDE` procedure.

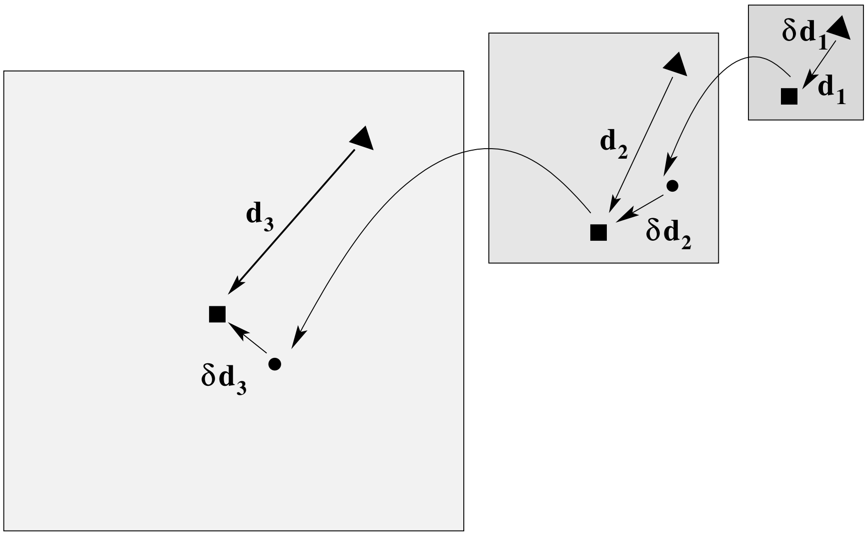


Figure 5.10. An illustration of coarse–to–fine disparity estimation on pyramid. The triangle–dot is the initial position. The circle–dot is the propagated position and the square–dot represents the refined position.

We conducted a series of tests using different confidence term in equation 5.11. Figure 5.11 shows snapshots of the convergence of the displacement through the 3 levels of hierarchy, using saliency (Eq. 5.8), normalized (Eq. 5.9) and phase difference (Eq. 5.10) as confidence term, the latter showing better iterative convergence over the pyramid image.

In figure 5.12, we present a quantitative comparison of the global tracking error of the x-y-positions of each reference point with `ICDE`. The phase difference confidence term (Eq. 5.10) achieves lower error rates (1.4 pixels, s.d. = 10.2), relative to the saliency (Eq. 5.8) with (3.0 pixels, s.d. = 16.4) and the normalized confidence (Eq. 5.9) with (2.1 pixels, s.d. = 13.8).

A better tracking performance rates of the x-y-positions of the reference points was achieved by the `ICDE` process in comparison to the conventional non–iterative displacement estimation procedure (2.1 pixels, s.d. = 10.6), as it is shown in figure 5.13.

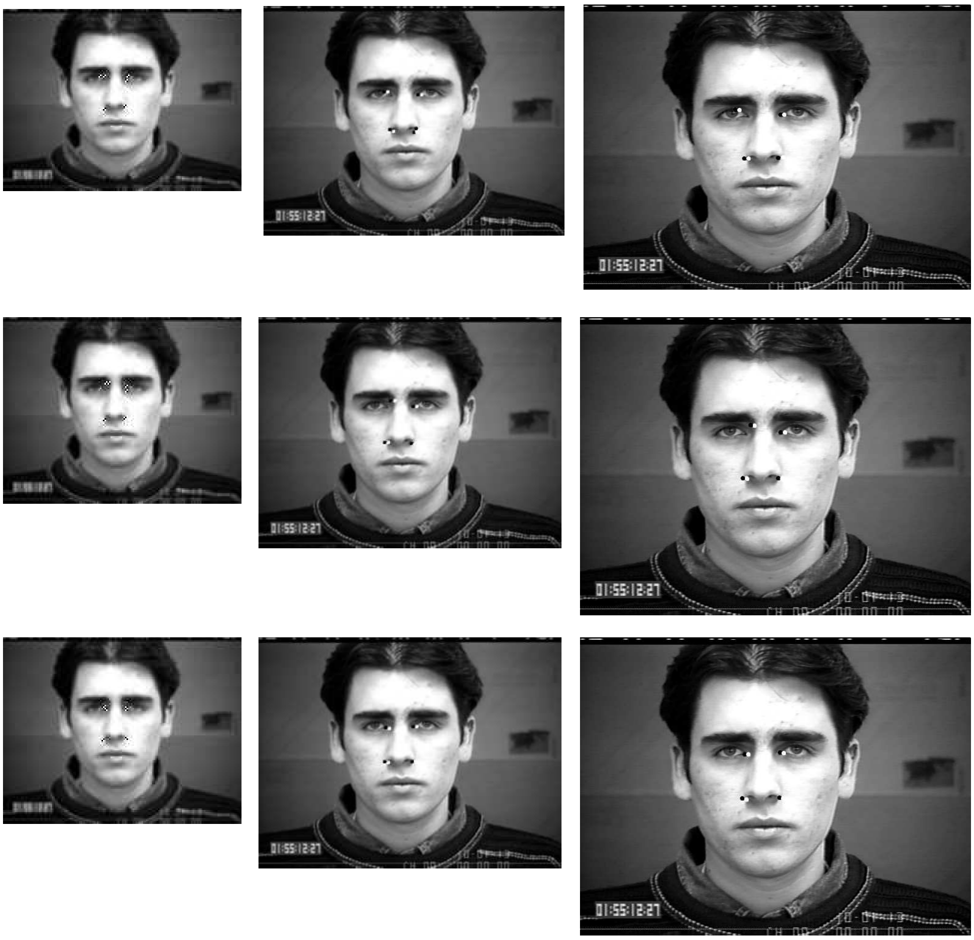


Figure 5.11. Snapshots of the displacement estimation algorithm through the 3 levels of hierarchy, using different confidence terms (saliency (Eq. 5.8), normalized (Eq. 5.9) and phase difference (Eq. 5.10)) from top to bottom in this order.

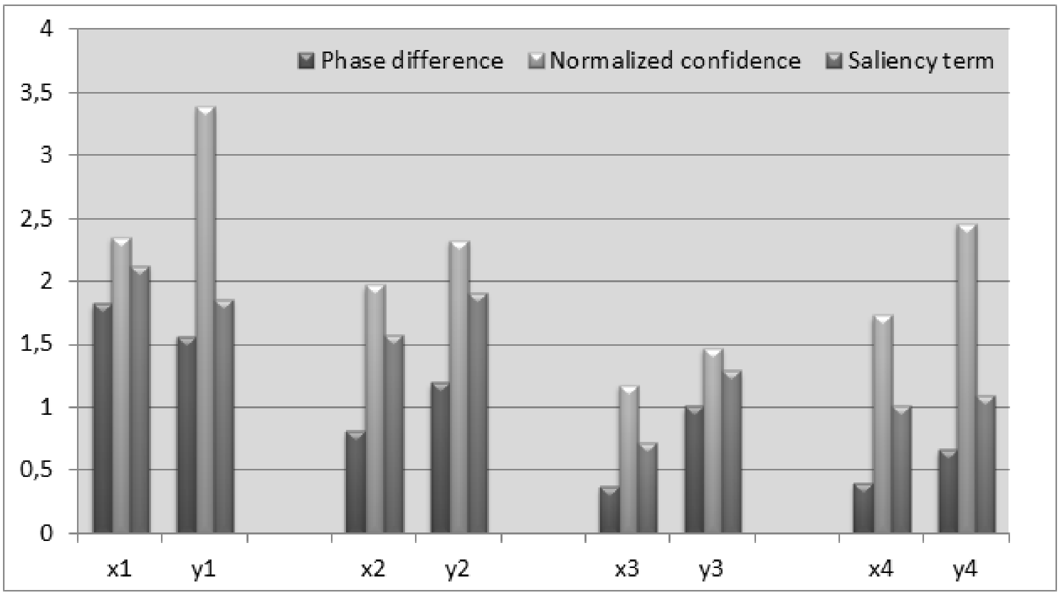


Figure 5.12. Mean tracking error of the 4 reference key point x-y-positions using different confidence terms (saliency (Eq. 5.8), normalized (Eq. 5.9) and phase difference (Eq. 5.10))

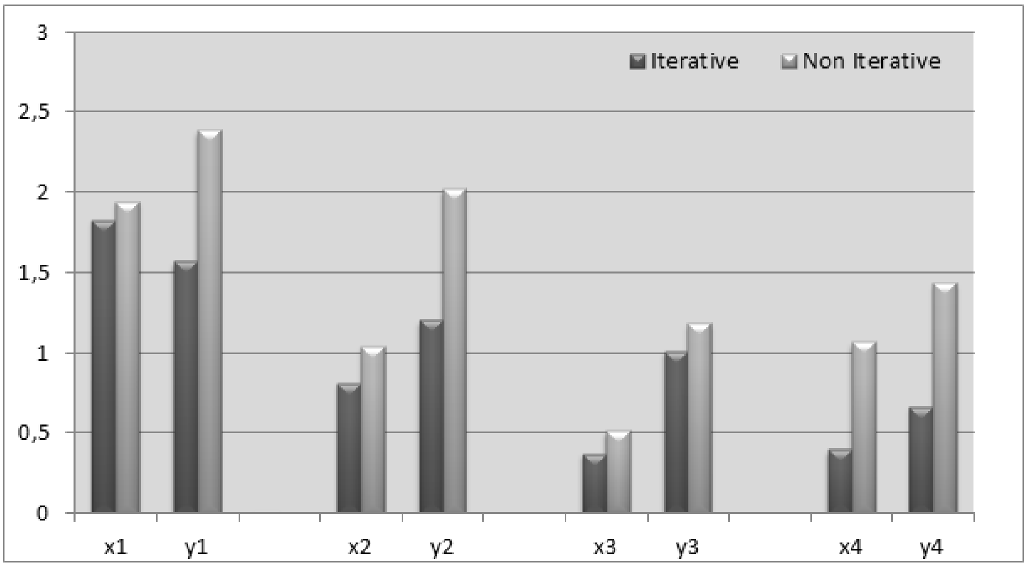


Figure 5.13. Overall tracking error of the 4 reference key point x-y-positions: Conventional vs. iterative conditional displacement estimation (ICDE) procedure.

5.4.2 Facial action recognition

As described in section 5.3.1, the twenty–eight facial point positions are obtained from the linear (translation, rotation, scaling) transformed positions of the reference subgraph by warping the positions of the key points from the selected graph to fit the four reference tracked positions.

After `ICDE` position refinement, the reference graph that maximizes the weighted magnitude–based similarity over the prestored graphs defines the final facial action that corresponds to the displayed expression. In this study, the prestored graphs consist only of one graph per person displaying a given expression at its peak.

Figure 5.14 shows, for each expression, an example of the intensity profile evolving in time as expected from a neutral display to its peak.

The overall performance on the six prototypic facial expressions reached 98.7% (Tab. 5.2). The most difficult expression to recognize was *sadness* with a rate of 92.3%. The ambiguous *sadness* sequence frames were classified as *fear* with only a small margin above the correct expression *sadness* (see figure 5.15), this is due to their very similar appearance particularly around the mouth region (see person "S035" figure 5.1). These facial expressions are distinguished by subtle changes in facial appearance.

Table 5.2. The overall recognition rates of different facial expressions considering up to 15% of ending frames.

Angry	Disgust	Fear	Happiness	Sadness	Surprise	Overall
100%	100%	100%	100%	92.31%	100%	98.72%

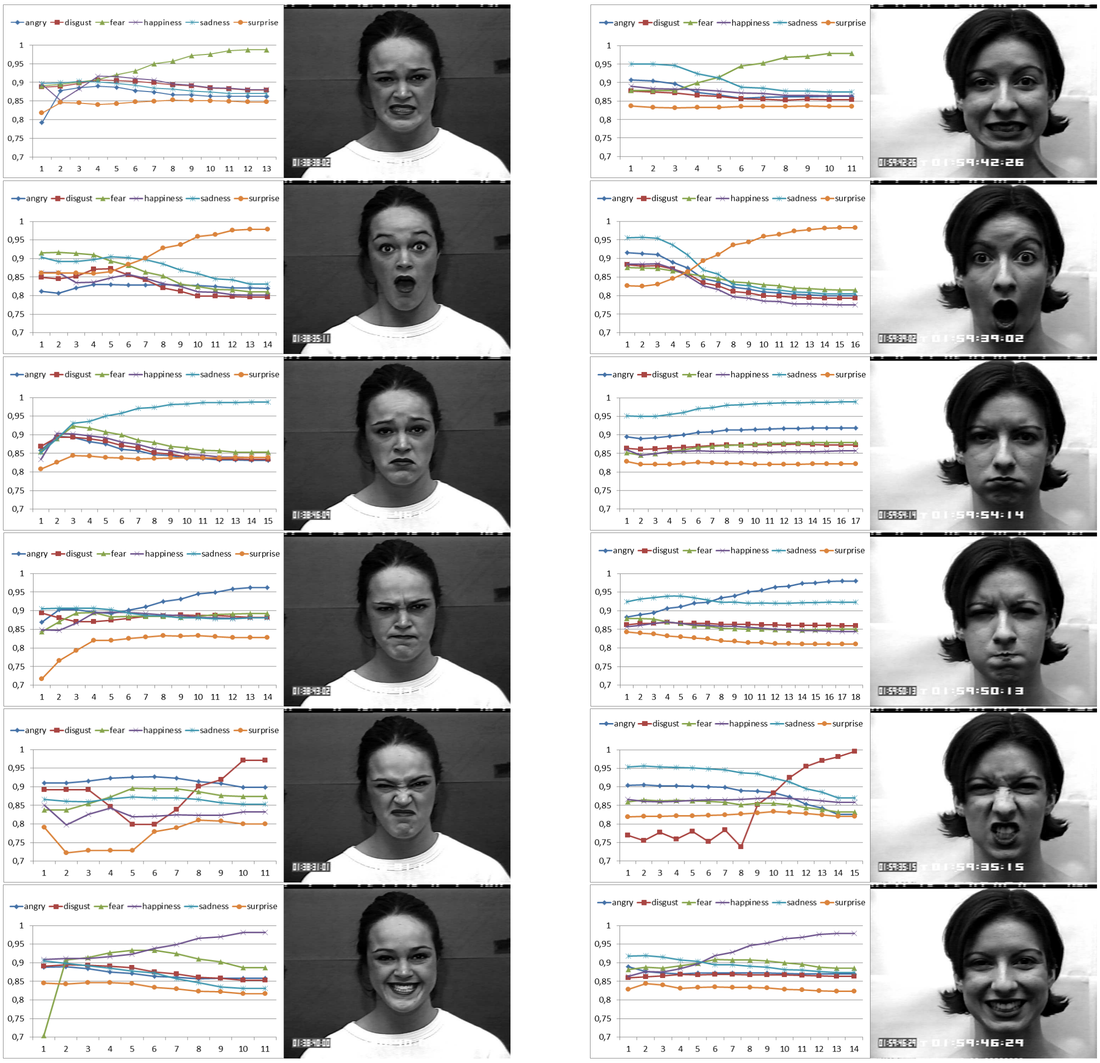


Figure 5.14. Facial action (FA) recognition performance. The similarity curve reflects its intensity.

5.5 Conclusion

In this work, we investigated a feature–based tracking algorithm that permits to eliminate accumulation of tracking errors, offering a good facial landmark localization, which is a crucial task in a feature–based facial expression recognition system. The proposed algorithm was implemented with a "personalized" facial action recognition

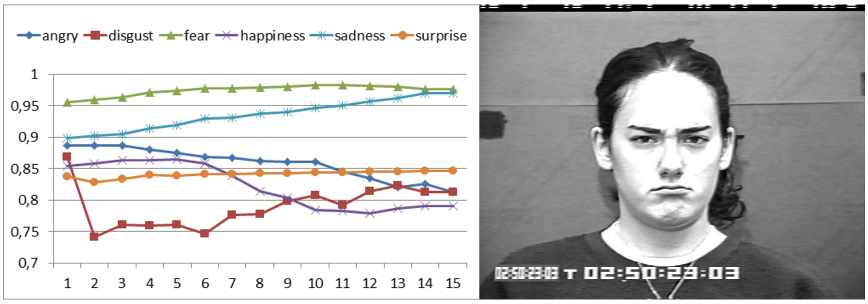


Figure 5.15. An example of ambiguous *sadness* expression of the person named "S035" classified as *fear* with a small margin above the true expression.

approach that utilizes local appearance provided by Gabor jets and global geometric configuration.

The novelty of the approach lies in the facial feature tracking process. We introduced a modified phase–based displacement estimation procedure that includes a new confidence measure and an iterative conditional disparity estimation. With a global constrained tracking on shape using linear transformations, the proposed facial tracking scheme, naturally enforces the optimal performance due to the localized appearance–based feature representation.

Some future developments of our approach are possible, particularly those related to the dynamic aspect. The emotional information is conveyed not only by the nature of facial movements, but also by their temporal evolution. The challenging task will be then how to determine the best way to set the timing parameters (e.g. limits between facial actions, transition, duration, speed). Also, we intend to investigate the performance of our method with respect to the problem of posed (deliberate) vs. spontaneous facial expression recognition.

Finally, we guess that the proposed approach is particularly suited for recognizing any facial action other than the prototypic facial expressions of the six basic emotions.

Acknowledgment

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Chapitre 6

DISCUSSION ET CONCLUSION

Ce chapitre présente les contributions apportées dans cette thèse portant sur la reconnaissance automatique des mouvements faciaux. Elles concernent principalement trois points: une méthode générique de reconnaissance d'expressions faciales, la localisation et le suivi des points saillants du visage et une approche personnalisée pour la reconnaissance d'actions faciales.

Les applications aussi bien que les extensions possibles des solutions proposées y sont discutées, ainsi que les travaux futurs potentiels qui pourraient en découler.

Modèle référentiel prototypique

Cette thèse a présenté un nouveau modèle pour la reconnaissance des expressions faciales. Nous appelons ce paradigme *référentiel prototypique* car il définit une base d'expressions prototypiques à partir de laquelle on peut exprimer n'importe quelle action faciale. Cette base est représentée par sept images références générées par un recalage SIFT–flow. Des histogrammes de gradients orientés sont ensuite calculés sur les images recalées par rapport à ce référentiel, produisant ce que nous désignons par SF–HOG. Ces descripteurs ont prouvé leur nette supériorité, en effet, le SIFT–flow par sa qualité de recalage confère aux SF–HOG une propriété supplémentaire d'alignement "garanti", un critère crucial pour les HOG qui se basent sur le fenêtrage.

Le modèle proposé a démontré sa capacité de généralisation, indépendamment de l'identité de la personne. Une évaluation quantitative montre que nous obtenons les meilleurs taux de reconnaissance comparés aux résultats expérimentaux obtenus par d'autres groupes de recherche, peu importe le procédé d'évaluation utilisé.

Certains développements futurs de notre approche sont possibles, particulièrement, ceux concernant l'aspect dynamique.

L'information émotionnelle est véhiculée non seulement par la nature des mouvements faciaux, mais également par leur synchronisation et leur évolution temporelle. Il s'agit de trouver la meilleure façon d'inclure la composante temps, ceci passe par la mise en pratique de toute la théorie relative au "*timing*" des expressions faciales.

D'autres perspectives concerneront la manière avec laquelle les paramètres optimaux de la fonction noyau du SVM sont trouvés durant l'apprentissage. À date, le modèle de sélection le plus adopté est la validation croisée, un modèle direct mais basé sur une simple recherche naïve de paramètres sur des intervalles choisis.

Algorithme de suivi de points fiduciels

Nous avons présenté au chapitre 5 un modèle géométrique reflétant la topologie du visage. La contribution majeure dans cette partie est la présentation d'un nouvel algorithme de suivi basé sur une modification d'une technique d'estimation de disparité faisant intervenir la phase de Gabor. Cette reformulation inclut une redéfinition de la mesure de confiance et introduit une procédure itérative et conditionnelle d'estimation du déplacement. Une analyse comparative par rapport aux méthodes originales a permis d'en affirmer la robustesse.

Motivé par le fait que les expressions faciales sont non seulement ambigües et incohérentes d'une personne à une autre, mais aussi dépendantes du contexte lui–même, la bonne performance du suivi et de la localisation des points saillants nous a permis de synthétiser un système personnalisé de reconnaissance d'expressions faciales garantissant une performance optimale.

On pense que plusieurs aspects seront sujets à amélioration dans le système de reconnaissance d'expressions faciales tel que proposé. Parmi tant d'autres, on suppose que, à travers les bases de données faciales existantes, des modèles statistiques sophistiqués de données peuvent être obtenus en utilisant des algorithmes d'apprentissage comme l'espérance–maximisation (EM, Expectation–Maximisation), pour représenter l'ensemble des graphes similaires (i.e. correspondant à des actions faciales similaires mais ne concerne pas nécessairement des visages familiers) dans des regroupements.

Estimation de la pose

En ce qui concerne la pose et la direction du regard plusieurs recherches ont porté sur leurs fonctions communicatives. Dans une conversation naturelle, la direction du regard permet d'évaluer l'autre : une personne qui regarde son interlocuteur durant de courtes durées est jugée comme *défensif* ou *évasif* alors que quelqu'un qui regarde pendant de longues périodes est jugé *amical*, *mature* et *sincère*.

La pose peut représenter un indicateur du degré d'intérêt et l'engagement que porte l'allocutaire lors d'une conversation. Elle renseigne le locuteur sur le degré de perception du message. Comme elle peut révéler un sentiment d'*ennui* chez un auditeur qui regarde ailleurs [196].

Par ailleurs, le mouvement global de la tête est une composante normale dans un contexte conversationnel. De ce fait, la nouvelle génération d'analyse automatique d'expressions faciales, devrait prendre en compte les mouvements et rotations de la tête [125]. Ceci revient à élaborer une solution qui tient compte de la posture de la tête.

À travers cette thèse nous nous sommes intéressés au problème d'estimation de la pose en optant pour des filtres dérivatifs Gaussiens (voir Annexe A).

Nous avons aussi proposé une approche itérative basée sur un algorithme d'estimation de la posture de Lowe, lequel utilise un ensemble de points de l'image et leurs positions relatives correspondantes dans un model `3d` (voir Annexe B).

Dans une autre méthode, on propose un nouveau type de descripteurs utilisant des histogrammes de vecteurs d'onde considérant à la fois la fréquence et l'orientation des différentes réponses en amplitude des filtres de Gabor. Nous montrons que ces descripteurs sont sensibles aux variations de la pose et insensibles aux déformations non pertinentes à la pose. Le taux de détection dépasse même le taux de reconnaissance atteint par l'humain (voir Annexe C).

Apparemment, ce type de résolutions séquentielles cause un problème puisque les deux sous–problèmes (expressions faciales vs. pose) sont mutuellement liés. C'est pourquoi une solution qui résout globalement le problème en prenant en charge à la fois les deux aspects du mouvement serait peut–être à considérer. Une résolution qui est loin d'être triviale à cause de la forte non–linéarité du couplage.

Cadre applicatif

Notre approche peut s'insérer dans un projet de grande envergure, particulièrement en intervention infirmière[1]. À travers ce projet, on vise à voir en quoi le *caring*[2] peut favoriser et promouvoir la santé d'un patient cardiaque. Les différents aspects de communication entre l'infirmière et le patient impliquent des éléments de **contenu** et des éléments de **relation**, les premiers représentent l'action sur laquelle porte l'interaction (Ex. rassurer une personne, ou l'assister dans la réponse à un besoin particulier), quant aux deuxièmes éléments, ils évoquent des indicateurs observables dites de processus tels que le toucher, le contact visuel, etc.

Notre but serait d'examiner et d'explorer le problème de détection automatique des interactions dites observables, plus particulièrement les trois types de comportements.

1. l'expressivité faciale : sourire, froncement de sourcils, etc.;

2. l'affection positive : indicatrice de hochement de tête, le sourire;

3. les retraits : indicateurs de non–sourire et de regarder ailleurs.

[1] Un projet d'étude de l'interaction patient–infirmière dirigé à l'institut de Cardiologie de Montréal par la Dre Sylvie Cossette a été initiateur de ce projet de thèse.

[2] Le *caring* est défini comme étant la finalité des soins infirmiers qui consiste à aider la personne à atteindre un plus haut niveau d'harmonie entre son corps, son âme et son esprit.

Par ailleurs, les études portant sur l'aspect acoustique, une autre modalité de l'émotion humaine, montrent l'importance de considérer l'intensité vocale présente lors de l'interaction. En effet, dans le cadre de ce même projet, les comportements non–verbaux audibles ont mis en évidence quatre caractéristiques importantes à considérer: le pourcentage d'interactions comportant un seul mot, le nombre moyen de mots par interaction, le débit, l'intensité vocale et le silence. Une extension de notre système qui intégrerait cette modalité pourra supporter l'élaboration de meilleures pratiques possibles pour le *caring* ainsi que pour diverses autres applications (Ex. conception de jeux vidéo).

RÉFÉRENCES

[1] A. Albiol, D. Monzo, A. Martin, et J. Sastre. Face recognition using HOG-EBGM. *Pattern Recognition Letters*, 29(10):1537–1543, 2008.

[2] P. S. Aleksic et A. K. Katsaggelos. Automatic Facial Expression Recognition Using Facial Animation Parameters and MultiStream HMMs. *IEEE Trans. Information Forensics and Security*, 1(1):3–11, 2006.

[3] M. Argyle. *Bodily Communication.* Routledge, London, second édition, 1990.

[4] M. Bart. Computing and visualization in science. *Springer-Verlag*, 11(5):53–63, 2002.

[5] M. S. Bartlett, J. C. Hager, et T. J. Sejnowski. Measuring facial expressions by computer image analysis. *Psychophisiology*, 36(2):253–263, 1999.

[6] M.S. Bartlett, G. Littlewort, I. Fasel, et R. Movellan. Real Time Face Detection and Facial Expression Recognition: Development and Application to Human Computer Interaction. Dans *Proc. CVPR Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction*, Cambridge, U.K., 2003.

[7] M.S. Bartlett, G.C. Littlewort, M.G. Frank, C. Lainscsek, I. Fasel, et J.R. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1(6):22–35, 2006.

[8] M.S. Bartlett, G.C. Littlewort-Ford, et J. Movellan. Computer expression recognition toolbox. Dans *Proc. of IEEE Int. Conference on Automatic Face and Gesture Recognition*, pages 1–2, 2008.

[9] B. Bascle et A. Blake. Separability of Pose and Expression in Facial Tracking and Animation. Dans *Proc. Int. Conference on Computer Vision*, pages 323–328, 1998.

[10] S. Belongie, J. Malik, et J. Puzicha. Matching shapes. Dans *Proc. Int. Conference on Computer Vision*, pages 454–461, 2001.

[11] M. J. Black, D. J. Fleet, et Y. Yacoob. A framework for modeling appearance change in image sequences. Dans *Proc. Int. Conference on Computer Vision*, pages 660–667, 1998.

[12] M. J. Black et Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *Int. Journal of Computer Vision*, 25(1):23–48, 1997.

[13] E. Boser, I. Guyon, et V. N. Vapnik. A training algorithm for optimal margin classiers. Dans *Proc. of the Fifth Annual Workshop on Computational Learning Theory, ACM Press*, pages 144–152, 1992.

[14] J. Bourgain. On Lipschitz Embedding of Finite Metric Spaces in Hilbert Space. *J. Math.*, 52:46–52, 1985.

[15] G. Bradski, A. Kaehler, et V. Pisarevsky. Learning-based computer vision with Intel's open source computer vision library. *Intel Technology Journal*, 9(2):119–130, 2005.

[16] I. Buciu, C. Kotropoulos, et I. Pitas. ICA and Gabor representation for facial expression recognition. Dans *Proc. of IEEE Int. Conference on Image Processing*, pages 855–858, 2003.

[17] I. Buciu et I. Pitas. Application of Non-Negative and Local Non Negative Matrix Factorization to Facial Expression Recognition. Dans *Proc. of Int. Conf. on Pattern Recognition*, pages 23–26, Cambridge, U.K., 2004.

[18] M. L. Cascia, S. Sclaroff, et V. Athitsos. Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. *IEEE Tran. on Pattern Analyssi and Machine Intelligence*, 4(22):322–336, 2000.

[19] Y. Chang, C. Hu, et M. Turk. Manifold of facial expression. Dans *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pages 28–35, 2003.

[20] Y. Chang, C. Hu, et M. Turk. Probabilistic expression analysis on manifolds. Dans *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 520–527, 2004.

[21] F. Cheng, J. Yu, et H. Xiong. Facial expression recognition in JAFEE dataset based on Gaussian Process classification. *IEEE Transactions on Neural Networks*, 21(10):1685–1690, Oct 2010.

[22] C. Chuan-Yu, H. Yan-Chiang, , et Y. Chi-Lu. Personalized Facial Expression Recognition in Color Image. Dans *Proc. of Int. Conf. on Innovative Computing, Information and Control*, pages 1164–1167, 2009.

[23] I. Cohen. Facial expression recognition from video sequences: Temporal and static modelling. *Computer Vision and Image Understanding*, 91(1-2):160–187, 2003.

[24] I. Cohen, N. Sebe, Y. Sun, M. S. Lew, et T. S. Huang. Evaluation of Expression

Recognition Techniques. Dans *Proc. of International Conference on Image and Video Retrieval*, pages 184–195, 2003.

[25] J. F. Cohn et T. Kanade. Use of automated facial image analysis for measurement of emotion expression. Dans J. A. Coan and J. B. Allen, editeur, *The handbook of emotion elicitation and assessment. Oxford University Press Series in Affective Science*. 2006.

[26] J. F Cohn, K. Schmidt, R. Gross, et P. Ekman. Individual differences in facial expression: Stability over time, relation to self–reported emotion, and ability to inform person identification. Dans *Proc. of IEEE Int. Conference on Multimodal Interfaces*, pages 491–496, 2001.

[27] J. F. Cohn, A. J. Zlochower, J. Lien, et T. Kanade. Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding. *Journal of Psychophysiology*, 36(1):35–43, 1999.

[28] J.F. Cohn et M.A Sayette. Spontaneous facial expression in a small group can be automatically measured: an initial demonstration. *Behavior Research Methods*, 42(4):1079–1086, 2010.

[29] D. Colbry, G. Stockman, et A. K. Jain. Detection of Anchor Points for 3D Face Verification. Dans *IEEE Workshop on Advanced 3D Imaging for Safety and Security*, 2005.

[30] T. F. Cootes et C. J. Taylor. Statistical models of appearance for computer vision. Rapport technique, http://www.isbe.man.ac.uk/~bim/refs.html, 2001.

[31] C. Cortes et V. N. Vapnik. Support-vector network. *Machine Learning*, 20:273–297, 1995.

[32] G. W. Cottrell, M. N. Dailey, et C. Padgett. *Is All Faces Processing Holistic? The view from UCSD*. M. Wenger, J Twnsend (Eds), Computational, Geometric and Process Perspectives on Facial Recognition, Contexts and Challenges, 2003.

[33] Garrison W. Cottrell, Matthew N. Dailey, Curtis Padgett, et Ralph Adolphs. *Is all face processing holistic? The view from UCSD*. Eds., Computational, Geometric, and Process Perspectives on Facial Cognition: Contexts and Challenges. Erlbaum, 2003.

[34] N. Cristianini et J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2012.

[35] D. Cristinacce et T.F. Cootes. Feature Detection and Tracking with Constrained Local Models. Dans *British Machine Vision Conference*, pages 929–938, 2006.

[36] A. Cruz, B. Bir, et Y. Songfan. A psychologically-inspired match-score fusion mode for video-based facial expression recognition. Dans *Proceedings of the 4th international conference on Affective computing and intelligent interaction - Volume Part II*, ACII'11, pages 341–350, Berlin, Heidelberg, 2011. Springer-Verlag.

[37] M. Dahmane et J. Meunier. Oriented-filters based head pose estimation. Dans *Fourth Canadian Conference on Computer and Robot Vision (CRV 2007), 28-30 May 2007, Montreal, Quebec, Canada*, pages 418–425. IEEE Computer Society, 2007.

[38] M. Dahmane et J. Meunier. Enhanced Phase–Based Displacement Estimation - An Application to Facial Feature Extraction and Tracking. Dans *Proc. of*

Int. Conference on Computer Vision Theory and Applications, pages 427–433, Madeira, Portugal, 2008.

[39] M. Dahmane et J. Meunier. An efficient 3d head pose inference from videos. Dans *Image and Signal Processing, 4th International Conference, ICISP 2010, Trois-Rivières, QC, Canada, June 30-July 2, 2010.*, volume 6134 de *Lecture Notes in Computer Science*, pages 368–375. Springer, 2010.

[40] M. Dahmane et J. Meunier. Continuous Emotion Recognition using Gabor Energy Filters. Dans *Proceedings of the 4th international conference on Affective computing and intelligent interaction - Volume Part II*, ACII'11, pages 351–358, Berlin, Heidelberg, 2011. Springer-Verlag.

[41] M. Dahmane et J. Meunier. Emotion recognition using dynamic grid-based HoG features. Dans *IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, pages 884–888, 2011.

[42] M. Dahmane et J. Meunier. Individual Feature–Appearance for Facial Action Recognition. Dans *Proc. of Int. Conference on Image Analysis and Recognition*, pages 233–242, 2011.

[43] M. Dahmane et J. Meunier. Object Representation based on Gabor Wave Vector Binning: An application to human head pose detection. Dans *IEEE International Conference on Computer Vision Workshops, ICCV 2011 Workshops, Barcelona, Spain*, pages 2198–2204, 2011.

[44] N. Dalal et B. Triggs. Histograms of oriented gradients for human detection. Dans *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[45] C. Darwin. *The Expression of the Emotions in Man and Animals.* J. Murray, London, 1872.

[46] J. G. Daugman. Uncertainty relations for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America*, 2(7):1160–1169, 1985.

[47] F. De la Torre, T. Simon, Z. Ambadar, et J. F. Cohn. Fast-facs: A computer-assisted system to increase speed and reliability of manual facs coding. Dans *Affective Computing and Intelligent Interaction (ACII)*, 2011.

[48] Daniel F. DeMenthon et Larry S. Davis. Model-based object pose in 25 lines of code. *International Journal of Computer Vision*, 15:123–141, 1995.

[49] A. Dhall, A. Asthana, R. Goecke, et T. Gedeon. Emotion Recognition Using PHOG and LPQ features. Dans *IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, 2011.

[50] G. Donato, M. Bartlett, J. Hager, P. Ekman, et T. Sejnowski. Classifying facial actions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21(10):97–4989, 1999.

[51] P. Ekman et W. V. Frieesen. Facial Action Coding System. *Investigator's Guide, Consulting Psychologists Press*, 1978.

[52] P. Ekman et W. V. Friesen. Constants Across Cultures in the Face and Emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971.

[53] P. Ekman, W. V. Friesen, et Phoebe E. *Emotion in the Human Face.* Oxford University Press, 1982.

[54] I. Essa et A. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):757–763, 1997.

[55] R. E. Fan, P. H. Chen, et C. J. Lin. Working Set Selection using Second Order Information for Training Support Vector Machines. *Journal of Machine Learning Research*, 6:1889–1918, 2005.

[56] B. Fasel. Robust face analysis using convolutional neural networks. volume 2, pages 40–43, 2002.

[57] B. Fasel et J. Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36(1):259–275, 2003.

[58] I.R. Fasel et J.R. Movellan. A Comparison of Face Detection Algorithms. *Lecture Notes in Computer Science*, 2415:1325–1330, 2002.

[59] X. Feng, A. Hadid, et M. Pietik̈inen. A Coarse–to–Fine Classification Scheme for Facial Expression Recognition. Dans *Proc. of Int. Conference on Image Analysis and Recognition*, 2004.

[60] X. Feng, B. Lv, Z. Li, et J. Zhang. A novel feature extraction method for facial expression recognition. Dans *Proc. Joint Conf. Inform. Sci. Issue Adv. Intell. Syst. Res.*, pages 371–375, Kaohsiung, Taiwan, 2006.

[61] X. Feng, M. Pietikäinen, et T. Hadid. Facial expression recognition with local binary patterns and linear programming. *Pattern Recognition and Image Analysis*, 15(2):546–548, 2005.

[62] R.S. Feris et al. Hierarchical wavelet networks for facial feature localization.

Dans *Proc. of IEEE Int. Conference on Automatic Face and Gesture Recognition*, pages 118–123, 2002.

[63] J. M. Fernandez-Dols, P. Carrera, et C. Casado. *The Meaning of Expression: Views Say not to Say: New perspectives on miscommunication.* IOS Press, 2001.

[64] K. Flaton et S. Toborg. An approach to image recognition using sparse filter graphs. Dans *Proc. of Int. Joint Conference on Neural Networks (1)*, pages 313–320, 1989.

[65] D. Fleet et A. Jepson. Computation of component image velocity from local phase information. *IEEE Trans. on Computers*, 5(1):77–104, 1990.

[66] D. Fleet et A. Jepson. Stability of phase information. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(12):1253–1268, 1993.

[67] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, et P. C. Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(2):1050–1057, 2007.

[68] G. L. Ford. Fully automatic coding of basic expressions from video. Rapport technique, Technical Report INC-MPLab-TR-2002.03, Machine Perception Lab, Institute for Neural Computation, University of California, San Diego, 2002.

[69] W. T. Freeman et E. H. Adelson. The design and use of steerable filters. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 3(19):891–906, 1991.

[70] X. Fu et W. Wei. Centralized Binary Patterns Embedded with Image Euclidean Distance for Facial Expression Recognition. Dans *Fourth International Conference on Natural Computation*, pages 115–119, 2008.

[71] C.A. Gee et R. Cipolla. Non-intrusive gaze tracking for human-computer interaction. *IEEE Proc. of Mechatronics and Machine Vision in Practice*, pages 112–117, 1994.

[72] D. Gorodnichy. Seeing faces in video by computers. *Editorial for Special Issue on Face Processing in Video Sequences, Elsevier*, 24(6):551–556, 2006.

[73] N. Gourier, D. Hall, et J. L. Crowley. Estimating face orientation from robust detection of salient facial features. Dans *in Proceedings of POINTING'04 International Workshop on Visual Observation of Deictic Gestures*, 2004.

[74] N. Gourier, J. Maisonnasse, D. Hall, et J. L. Crowley. Head pose estimation on low resolution images. Dans *Multimodal Technologies for Perception of Humans: Proc. First Int'l Workshop Classification of Events, Activities and Relationships*, pages 270–280, 2007.

[75] L. Gralewski et al. Using a Tensor Framework for the Analysis of Facial Dynamics. Dans *Proc. of IEEE Int. Conference on Automatic Face and Gesture Recognition*, pages 217–222, 2006.

[76] C. Grigorescu, N. Petkov, et M. A. Westenberg. Contour detection based on nonclassical receptive field inhibition. *IEEE Transactions on Image Processing*, 12(7):729–739, 2003.

[77] R. Gross et al. Constructing and Fitting Active Appearance Models With Occlusion. Dans *IEEE Workshop on Face Processing in Video (FPIV)*, 2004.

[78] L. Gu et T. Kanade. A Generative Shape Regularization Model for Robust Face Alignment. Dans *Proc. European Conference on Computer Vision*, Wisconsin, 2008.

[79] G. Guo et C. R. Dyer. Simultaneous Feature Selection and Classifier Training via Linear Programming: A Case Study for Face Expression Recognition. Dans *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 346–352, 2003.

[80] H. Hai, H. Neven, et C. von der Malsburg. Online facial expression recognition based on personalized galleries. Dans *Proc. of IEEE Int. Conference on Automatic Face and Gesture Recognition*, pages 354–359, Apr 1998.

[81] Z. Hammal, L. Couvreur, A. Caplier, et M. Rombaut. Facial expression classification: An approach based on the fusion of facial deformation unsing the transferable belief model. Dans *Int. Jour. of Approximate Reasonning*, 2007.

[82] X. He, S. Yan, Y. Hu, et H. Zhang. Learning a Locality Preserving Subspace for Visual Recognition. Dans *Proc. Int. Conference on Computer Vision*, 2003.

[83] B. Heisele et al. Component-based face detection. *the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 657–662, 2001.

[84] G. Heusch, Y. Rodriguez, et S. Marcel. Local Binary Patterns as an Image Preprocessing for Face Authentication. Dans *Proc. of IEEE Int. Conference on Automatic Face and Gesture Recognition*, FGR '06, pages 9–14, Washington, DC, USA, 2006. IEEE Computer Society.

[85] C. W. Hsu, C.-C. Chang, et C. J. Lin. A Practical Guide to Support Vector Classification. Rapport technique, Department of Computer Science and Information Engineering, National Taiwan University, 2003.

[86] C. Hu, R. Feris, et M. Turk. Active wavelet networks for face alignment. Dans *In British Machine Vision Conference*, pages 757–763, 2003.

[87] Y. Hu, L. Chen, Y. Zhou, et H. Zhang. Estimating face pose by facial asymmetry and geometry. Dans *Proc. of IEEE Int. Conference on Automatic Face and Gesture Recognition*, 2004.

[88] Y. Hu, Z. Zeng, L. Yin, X. Wei, J. Tu, et T. Huang. Multi-view facial expression recognition. Dans *Proc. of IEEE Int. Conference on Automatic Face and Gesture Recognition*, pages 1–6, 2008.

[89] K. Ichikawa, T. Mita, et O. Hori. Component-based robust face detection using AdaBoost and decision tree. Dans *Automatic Face and Gesture Recognition*, pages 413–420, 2006.

[90] T. JBaenziger et K.R. Scherer. *Introducing the geneva multimodal emotion portrayal (GEMEP) corpus.* Blueprint for affective computing: A sourcebook, T. B. K. R. Scherer and E. Roesch, Eds., England: Oxford University Press, 2010.

[91] B. Jiang, M. Valstar, et M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. Dans *IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, 2011.

[92] F. Jiao, S. Li, H-Y. Shum, et D. Schuurmans. Face alignment using statistical models and wavelet features. Dans *Computer Vision and Pattern Recognition (1)*, pages 321–327, 2003.

[93] T. Jilin, T. Huang, et T. Hai. Accurate head pose tracking in low resolution video. Dans *7th Int. Conf. on Automatic Face and Gesture Recognition*, pages 573–578, 2006.

[94] W. Johnson et J. Lindenstrauss. Extension of Lipschitz Mapping into a Hilbert Space. *Contemporary Math*, 26:189–206, 1984.

[95] T. Kanade, J. Cohn, , et Y.L. Tian. Comprehensive database for facial expression analysis. Dans *Proc. of IEEE Int. Conference on Automatic Face and Gesture Recognition*, pages 46–53, March 2000.

[96] J. T. Kent. *New Directions in Shape Analysis.* the Art of Statistical Science, Wiley, 1992.

[97] D. J. Kim et Z. Bien. Design of "Personalized" Classifier Using Soft Computing Techniques for "Personalized" Facial Expression Recognition . *IEEE Transactions on Fuzzy Systems*, 16(4):874–885, 2008.

[98] T. K. Kim, H. Kim, W. Hwang, S.C. Kee, et J. Kittler. Independent Component Analysis in a Facial Local Residue Space. Dans *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 413–420, 2003.

[99] CG. Kohler, EA. Martin, M. Milonova, P. Wang, R. Verma, C.M. Brensinger, W. Bilker, R.E. Gur, et R.C. Gur. Dynamic evoked facial expressions of emotions in schizophrenia. *Schizophrenia Research*, 105(1):30–39, 2008.

[100] I. Kotsia, I. Buciu, et I. Pitas. An Analysis of Facial Expression Recognition under Partial Facial Image Occlusion. *Image and Vision Computing*, 26(7):1052–1067, 2008.

[101] A. Z. Kouzani. Locating human faces within images. *Computer Vision and Image Understanding*, 91(3):247–279, 2003.

[102] M. Lades, J. Vorbruuggen, J. Buhmann, J. Lange, W. Konen, C. von der Malsburg, et R. Wurtz. Distortion Invariant Object Recognition in the Dynamic Link Architecture. *IEEE Trans. Computers*, 42(3):300–311, 1993.

[103] K. Levi et Y. Weiss. Learning Object Detection from a Small Number of Examples: the Importance of Good Features. Dans *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 53–60, 2004.

[104] W. K. Liao et I. Cohen. Classifying Facial Gestures in Presence of Head Motion. Dans *IEEE Conference on Computer Vision and Pattern Recognition CVPRW*, page 77, 2005.

[105] C. Lisetti et D. Schiano. Automatic Facial Expression Interpretation: Where Human-Computer Interaction, Artificial Intelligence and Cognitive Science Intersect. Dans *Pragmatics and Cognition*, pages 185–235, 2000.

[106] G. Littlewort, M. Stewart Bartlett, I. Fasel, J. Susskind, et J. Movellan. Dynamics of facial expression extracted automatically from video. pages 615–625, 2004.

[107] C. Liu et H. Wechsler. Independent component analysis of Gabor features for face recognition. *IEEE Trans. on Neural Networks*, 14(4):919–928, 2003.

[108] C. Liu, J. Yuen, et A. Torralba. SIFT Flow: Dense Correspondence across Scenes and its Applications. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 33(5):978–994, 2010.

[109] D. G. Lowe. Fitting parametrized three-dimensional models to images. 13(5):441–450, 1991.

[110] D. G. Lowe. Object recognition from local scale–invariant features. Dans *Proc. Int. Conference on Computer Vision*, pages 1150–1157, 1999.

[111] D. G. Lowe. Distinctive image features from scale–invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[112] P. Lucey, S. Lucey, et J. Cohn. Registration invariant representations for expression detection. Dans *International Conference on Digital Image Computing: Techniques and Applications*, pages 255–261, Amsterdam, 2010.

[113] S. Lucey, A. B. Ashraf, et J. Cohn. *Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face.* Face Recognition Book, Pro Literatur Verlag, 2007.

[114] S. Lucey, I. Matthews, C. Hu, Z. Ambadar, F. de la Torre, et J. Cohn. AAM derived face representations for robust facial action recognition. Dans *Proc. of IEEE Int. Conference on Automatic Face and Gesture Recognition*, 2006.

[115] M. J. Lyons, S. Akamatsu, M. Kamachi, et J. Gyoba. Coding Facial Expressions with Gabor Wavelets. Dans *Proc. of IEEE Int. Conference on Automatic Face and Gesture Recognition*, pages 200–205, 1998.

[116] M. J. Lyons, J. Budynek, et S. Akamatsu. Automatic classification of single facial images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21(12):1357–1362, Dec 1999.

[117] B. Ma, S. Shan, X. Chen, et W. Gao. Head yaw estimation from asymmetry of facial appearance. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 38(6):1501–1512, 2008.

[118] K. Mardia et P. Jupp. *Directional Statistics.* New York, Wiley, 2000.

[119] T. Maurer et C. von der Malsburg. Tracking and Learning Graphs and Pose on Image Sequences of Faces. Dans *Proc. of IEEE Int. Conference on Automatic Face and Gesture Recognition*, page 76, 1996.

[120] G. McIntyre, R. Gocke, M. Hyett, M. Green, et M. Breakspear. An approach for automatically measuring facial activity in depressed subjects. Dans *Affective Computing and Intelligent Interaction and Workshops*, pages 1–8, 2009.

[121] S.J. McKenna, S. Gong, R. P. Würtz, J. Tanner, et D. Banin. Tracking facial feature points with Gabor wavelets and shape models. Dans *International conference on audio and video-based biometric person authentication*, pages 35–42, 1997.

[122] G. McKeown, M. Valstar, M. Pantic, et R. Cowie. The semaine corpus of emotionnally coloured character interactions. Dans *Proc . Int'l Conf. Mutlimedia & Expo*, pages 1–6, 2010.

[123] A. Mehrabian. Communication without Words. *Psychology Today*, 2(4):53–56, 1968.

[124] B. Moghaddam, W. Wahid, et A. Pentland. Beyond Eigenfaces: Probabilistic Matching for Face Recognition. Dans *Proc. of IEEE Int. Conference on Automatic Face and Gesture Recognition*, pages 30–35, 1998.

[125] J. R. Movellan et M. S. Bartlett. *The next generation of automatic facial expression measurement*, pages 393–426. In P. Ekman (Ed.), What the Face Reveals, 2nd Edition, Oxford University Press, 2005.

[126] MPLab. The MPLab GENKI Database. http://mplab.ucsd.edu.

[127] D.P. Mukherjee, A. Zisserman, et M. Brady. Shape from symmetry – detecting and exploiting symmetry in affine images. *Phil. Trans. R. Soc. Lond. A*, 351:77–106, 1995.

[128] K. P. Murphy, Y. Weiss, et M. Jordan. Loopy Belief Propagation for Approximate Inference: An Empirical Study. Dans *Uncertainty in Artificial Intelligence*, pages 467–475, 1999.

[129] E. Murphy-Chutorian et M. M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626, 2009.

[130] G.R.S. Murthy et R.S. Jadon. Effectiveness of Eigenspaces for Facial Expressions Recognition. *International Journal of Computer Theory and Engineering*, 1(5):1793–8201, 2009.

[131] S. Niyogi et W. Freeman. Example-based head tracking. Dans *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, pages 374–378, 1996.

[132] T. Ojala, M. Pietikainen, et D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[133] T. Ojala, M. Pietikainen, et I. Maenpaa. Multi-resolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[134] V. Ojansivu et J. Heikkil. Blur insensitive texture classification using local phase quantization. Dans *International Conference on Image and Signal Processing*, 2008.

[135] A. OToole, J. Harms, S. Snow, D. Hurst, M. Pappas, et H. Abdi. A video database of moving faces and people. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(5):812–816, 2005.

[136] T. Otsuka et J. Ohya. Recognizing Multiple Persons' Facial Expressions Using HMM Based on Automatic Extraction of Significant Frames from Image Sequences. Dans *Proc. Int. Conf. on Image Processing*, pages 546–549, 1997.

[137] M. Pantic et M. S. Bartlett. Face Recognition, Machine Analysis of Facial Expressions. I-Tech Education and Publishing, 2007.

[138] M. Pantic et I. Patras. Detecting facial actions and their temporal segments in nearly frontal-view face image sequences. Dans *Proc. IEEE conf. Systems, Man and Cybernetics,*, pages 3358–3363, 2005.

[139] M. Pantic et L. Rothkrantz. Automatic Analysis of Facial Expressions: The State of the Art. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, December 2000.

[140] M. Pantic, M. Valstar, R. Rademaker, et L. Maat. Web-based database for facial expression analysis. Dans *in Proc. IEEE Int'l Conf. Multimedia and Expo (ICME'05)*, Amsterdam, The Netherlands, 2005.

[141] A. Pentland, B. Moghaddam, et T. Starner. View-based and modular eigenspaces for face recognition. Dans *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 84–91, jun 1994.

[142] N. Petkov et M. A. Westenberg. Suppression of contour perception by band-limited noise and its relation to non-classical receptive field inhibition. *Biological Cybernetics*, 88(10):236–246, 2003.

[143] M. Pietikainen, A. Hadid, G. Zhao, et T. Ahonen. *Computer Vision Using Local Binary Patterns.* Computational Imaging and Vision. Springer, Dordrecht, 2011.

[144] G. Ramirez, T. Baltrusaitis, et L. P. Morency. Modeling latent discriminative dynamic of multi-dimensional affective signals. Dans *Proceedings of the 4th international conference on Affective computing and intelligent interaction - Volume Part II*, ACII'11, pages 396–406. Springer-Verlag, 2011.

[145] E. Rentzeperis, A. Stergiou, A. Pnevmatikakis, et L. Polymenakos. Impact of Face Registration Errors on Recognition. Dans *AIAI*, pages 187–194, 2006.

[146] E. Ricci et J. M. Odobez. Learning large margin likelihoods for realtime head pose tracking. Dans *In Proceedings of the IEEE International Conference of Image Processing*, 2009.

[147] T. Riemersma. Colour metric. http://www.compuphase.com/cmetric.htm, 2007.

[148] A. Ross. Multibiometric Systems: An overview of Information Fusion in Biometrics. Dans *International Conference on Biometrics*, pages 159–196, 2006.

[149] D. Roth et al. A SNoW-based Face Detector. *Advances in Neural Information Processing Systems*, 12:855–861, 2000.

[150] Sam T. Roweis et L. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290:2323–2326, 2000.

[151] H. A. Rowley, S. Baluja, et T. Kanade. Neural network based face detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.

[152] J. A. Russell et J. M. Fernandez-Dols. Facial displays. *in The psychology of facial expression, Cambridge University Press*, 1997.

[153] J. A. Russell et J. M. Fernandez-Dols. What does a facial expression mean ? *in The psychology of facial expression, Cambridge University Press*, pages 3–30, 1997.

[154] A. Samal et P. A. Iyengar. Automatic recognition and analysis of human faces and facial expression: a survey. *Pattern Recognition*, 25(1):66–77, 1992.

[155] J. Saragih, S. Lucey, et J. Cohn. Face alignment through subspace constrained mean-shifts. Dans *Proc. Int. Conference on Computer Vision*, 2009.

[156] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, et M. Pantic. Avec 2011 - the first international audio/visual emotion challenge. LNCS, 2011.

[157] C. Shan, S. Gong, et P. W. McOwan. Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.

[158] L. Shen et L. Bai. A review on Gabor wavelets for face recognition. *Pattern Analysis and Applications*, 9(2):273–292, 2006.

[159] J. Shi et C. Tomasi. Good features to track. *IEEE Conf. on Computer Vision and Pattern Recognition*, 1994.

[160] F. Y. Shih. *Image processing and pattern recognition : fundamentals and techniques: Facial expression recognition in JAFFE database.* Piscataway, NJ : Hoboken, NJ : IEEE Press ; Wiley, 2010.

[161] F.Y. Shih et C. Chuang. Automatic Extraction of Head and Face Boundaries and Facial Features. *Information Sciences*, 158:117–130, 2004.

[162] T. Sim, S. Baker, et M. Bsat. The CMU pose, illumination, and expression database. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(12):1615–1618, 2003.

[163] K. Sobottka et I. Pitas. A Fully Automatic Approach to Facial Feature Detection and Tracking. Dans *Audio- and Video-based Biometric Person Authentication, LNCS*, pages 77–84, 1997.

[164] R. Stiefelhagen, J. Yang, et A. Waibel. Towards tracking interaction between people. *Proc. of the AAAI Spring Symposium on Intelligent Environments, AAAI Press*, pages 123–127, 1998.

[165] K. K. Sung et T. Poggio. Example-Based Learning for View-Based Human Face Detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1998.

[166] W. Theimer et V. Mallot. Phase–based binocular vergence control and depth reconstruction using active vision. *CVGIP: Image Understanding*, 60(3):343–358, 1994.

[167] Y. Tian, T. Kanade, et J. Cohn. Recognizing Action Units for Facial Expression Analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.

[168] Y. Tian, T. Kanade, et J. Cohn. Evaluation of Gabor wavelet–based facial action unit recognition in image sequences of increasing complexity. Dans *Proc. of IEEE Int. Conference on Automatic Face and Gesture Recognition*, pages 229–234, 2002.

[169] Ying-Li Tian. Evaluation of Face Resolution for Expression Analysis. Dans *Proc. of CVPR Workshop on Face Processing in Video*, 2004.

[170] Y.L. Tian, T. Kanade, et J.F. Cohn. *Chapter 11. Facial Expression Analysis.* New York ; Toronto : Wiley and Sons, ISBN 0-471-37739-2, 2008.

[171] Carlo Tomasi et Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *Int. J. Comput. Vision*, 9(2):137–154, 1992.

[172] E. Trucco et A. Verri. *Introductory Techniques for 3-D Computer Vision.* Prentice-Hall, NJ, englewood cliffs édition, 1998.

[173] J. Tu, Y. Fu, Y. Hu, et T. Huang. Evaluation of head pose estimation for studio data. Dans *Multimodal Technologies for Perception of Humans: Proc. First Int'l Workshop Classification of Events, Activities and Relationships*, pages 281–290, 2007.

[174] M.Z. Uddin, J.J. Lee, et T.S. Kim. An enhanced independent component-based human facial expression recognition from video. *IEEE Transactions on Consumer Electronics*, 55(4):2216–2224, November 2009.

[175] S. Ullman et R. Basri. Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):992–1006, Oct 1991.

[176] L. Vacchetti, V. Lepetit, et P. Fua. Fusing online and offline information for stable 3d tracking in real-time. Dans *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–241–8 vol.2, June 2003.

[177] M. Valstar, B. Jiang, M. Méhu, M. Pantic, et K. Scherer. The First Facial Expression Recognition and Analysis Challenge. Dans *IEEE International Confer-*

ence on Automatic Face Gesture Recognition and Workshops (FG 2011), Santa Barbara, USA, 2011.

[178] M. Valstar et M. Pantic. Fully Automatic Facial Action Unit Detection and Temporal Analysis. Dans *Computer Vision and Pattern Recognition Workshop*, page 149, 2006.

[179] V. N. Vapnik. *The nature of statistical learning theory.* Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[180] V. N. Vapnik. *Statistical Learning Theory.* Wiley, New York, NY (USA), 1998.

[181] V. N. Vapnik et S. Kotz. *Estimation of dependences based on empirical data.* New York : Springer, 2nd ed, New York, NY, USA, 2006.

[182] P. Viola et M. Jones. Rapid object detection using a boosted cascade of simple features. Dans *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 511–518, 2001.

[183] P. Viola et M. Jones. Robust real–time object detection. Dans *Second International Workshop on Statistical and Computational Theories of Vision - Modeling, Learning, Computing, and Sampling*, 2001.

[184] P. Viola et M. Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137–154–154, May 2004.

[185] M. Voit, K. Nickel, et R. Stiefelhagen. Neural network-based head pose estimation and multi-view fusion. Dans *Multimodal Technologies for Perception of Humans: Proc. First Int'l Workshop Classification of Events, Activities and Relationships*, pages 291–298, 2007.

[186] F. Wallhoff. Facial expressions and emotion database. http://www.mmk.ei.tum.de/waf/fgnet/feedtum.html., 2006.

[187] H. Wang et N. Ahuja. Facial expression decomposition. Dans *Proc. Int. Conference on Computer Vision*, pages 958–965, 2003.

[188] J. Wang et L. Yin. Static Topographic Modeling for Facial Expression Recognition and Analysis. *Computer Vision and Image Understanding*, 108(1-2):19–34, 2007.

[189] J. G. Wang et E. Sung. EM Enhancement of 3D Head Pose Estimated by Point at Infinity. *Image and Vision Computing*, 25(12):1864–1874, 2007.

[190] J.J. Wang et S. Singh. Video analysis of human dynamics–a survey. *Real-Time Imaging*, 5(9):321–346, 1999.

[191] P. Wang, F. Barrett, E. Martin, M. Milonova, R. E. Gur, R. C. Gur, C. Kohler, et R. Verma. Automated Video Based Facial Expression Analysis of Neuropsychiatric Disorders. *Journal of Neuroscience Methods*, 168(1):224–238, 2008.

[192] Y. Wang, S. Lucey, et J.F. Cohn. Enforcing Convexity for Improved Alignment with Constrained Local Models. Dans *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[193] T. Wehrle et al. Studying the Dynamics of Emotional Expression Using Synthesized Facial Muscle Movements. *Journal of Personality and Social Psychology*, 78(1):105–119, 2000.

[194] J. Whitehill, G. Littlewort, I.R. Fasel, M.S. Bartlett, et J.R. Movellan. Toward practical smile detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(11):2106–2111, 2009.

[195] J. Whitehill et C. W. Omlin. Local versus Global Segmentation for Facial Expression Recognition. Dans *Proc. of IEEE Int. Conference on Automatic Face and Gesture Recognition*, pages 357–362, 2006.

[196] S. Whittaker et B. Oi'Connaill. *The Role of Vision in Face-to-Face and Mediated Communication*, pages 23–49. Video-Mediated Communication, Eds. Finn, K., Sellen, A., Wilbur, S. Lawerance Erlbaum Associates, New Jersey., 1997.

[197] L. Wiskott et al. Face Recognition by Elastic Bunch Graph Matching. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.

[198] J. Wu, J. Pedersen, D. Putthividhya, D. Norgaard, et M. M. Trivedi. A two-level pose estimation framework using majority voting of gabor wavelets and bunch graph analysis. Dans *Proc. ICPR Workshop Visual Observation of Deictic Gestures*, 2004.

[199] T. Wu, M.S. Bartlett, et J. Movellan. Facial expression recognition using gabor motion energy filters. Dans *IEEE CVPR workshop on Computer Vision and Pattern Recognition for Human Communicative Behavior Analysis*, Atlantic City, New Jersey, USA, 2010.

[200] J. Xiao, S. Baker, I. Matthews, et T. Kanade. Real-time combined 2d+3d active appearance models. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2:535–542, 2004.

[201] Z. Xue, S. Z. Lib, et E. K. Teoh. Bayesian Shape Model for Facial Feature Extraction and Recognition. *Pattern Recognition*, 36:2819–2833, 2003.

[202] Y. Yacoob et L. Davis. Computing spatio-temporal representations of human faces. Dans *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 70–75, Atlantic City, New Jersey, USA, 1994.

[203] M.H. Yang. Recent Advances in Face Detection. Dans *Tutorial of IEEE Conferece on Pattern Recognition*, pages 23–26, 2004.

[204] Wu Ying et K. Toyama. Wide-range, person- and illumination-insensitive head orientation estimation. Dans *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 183–188, 2000.

[205] J.J. Yokono et T. Poggio. Oriented filters for object recognition: an empirical study. *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 755–760, 2004.

[206] Mian-Shui Yu et Shao-Fa Li. Tracking facial feature points with statistical models and gabor wavelet. Dans *Fifth Mexican International Conference on Artificial Intelligence*, pages 61–67, Nov. 2006.

[207] F. Yun et T.S. Huang. Graph embedded analysis for head pose estimation. Dans *7th International Conference on Automatic Face and Gesture Recognition*, April 2006.

[208] B. Zhang, S. Shan, X. Chen, et W. Gao. Histogram of gabor phase patterns (HGPP): A novel object representation approach for face recognition. *IEEE Tran. on Image Processing*, 16(1):57–68, 2007.

[209] Y. Zhang et Q. Ji. Facial expression understanding in image sequences using dynamic and active visual information fusion. Dans *Proc. Int. Conference on Computer Vision*, pages 113–118, 2003.

[210] Y. Zhang et Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(5):699–714, may 2005.

[211] Z. Zhang, M. Lyons, M. Schuster, et S. Akamatsu. Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron. Dans *Proc. of IEEE Int. Conference on Automatic Face and Gesture Recognition*, pages 454–459, 1998.

[212] W. Zheng, X. Zhou, C. Zou, et L. Zhao. Facial Expression Recognition Using Kernel Canonical Correlation Analysis. *IEEE Trans. Neural Networks*, 17(1):233–238, 2006.

[213] Zhu Zhiwei et Ji Qiang. Robust pose invariant facial feature detection and tracking in real-time. Dans *18th International Conference on Pattern Recognition.*, pages 1092–1095, 2006.

[214] Y. Zhou, L. Gu, et H. Zhang. Bayesian Tangent Shape Model: Estimating Shape and Pose Parameters via Bayesian Inference. Dans *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–116, Wisconsin, 2003.

[215] Z. Zhu et Q. Ji. Robust pose invariant facial feature detection and tracking in real-time. Dans *Proc. of Int. Conf. on Pattern Recognition*, pages 1092–1095, 2006.

[216] M. Zobel et al. Robust Facial Feature Localization by Coupled Features. Dans *Proc. of IEEE Int. Conference on Automatic Face and Gesture Recognition*, 2000.

Annexe A

(ARTICLE) ORIENTED–FILTERS BASED HEAD POSE ESTIMATION

Cet article [37] a été publié comme l'indique la référence bibliographique

> M. Dahmane et J. Meunier. Oriented-filters based head pose estimation. Dans *Fourth Canadian Conference on Computer and Robot Vision (CRV 2007), 28-30 May 2007, Montreal, Quebec, Canada*, pages 418–425. IEEE Computer Society, 2007.

Abstract

The aim of this study is to elaborate and validate a methodology to assess automatically the head orientation with respect to a camera in a video sequence. The proposed method uses relatively stable facial features (upper points of the eyebrows, upper nasolabial-furrow corners and nasal root) that have symmetric properties to recover the face slant and tilt angles. These fiducial points are characterized by a bank of steerable filters. Using the frequency domain, we present an elegant formulation to linearly decompose a Gaussian steerable filter into a set of x, y–separable basis Gaussian kernels. A practical scheme to estimate the position of the occasionally occluded nasolabial-furrow facial feature is also proposed. Results show that the head motion can be estimated with sufficient precision to get the gaze direction without camera calibration or any other particular settings are required for this purpose.

Introduction

In a face–to–face talk, language is supported by non-verbal communication, which plays a central role in human social behaviour [3] by adding cues to the meaning of speech, providing feedback and managing synchronisation. Unlike verbal communication which is generally explicit, information about visible behaviours usually is inferred from: (*i*) *Gaze*, that provides spatial information about the focus of attention, (*ii*) *Facial expressions*, which are primarily carried out by facial attributes and allow to infer the emotional state of a person, (*iii*) *Gestures* (hands, arms and head movements), that have usually conventionalized meaning such as the yes/no head motions [196].

Processing such information by "hand" from audio-video recordings, will not be practical in the context of large scale studies. In a larger project, we are interested in a computer vision system to investigate nurse/patient interaction patterns. For this purpose, a first step is the assessment of the head orientation or pose to get the patient's or nurse's gaze direction but also to normalize and warp the face for automatic facial expressions analysis.

The work presented here uses robust feature descriptors to track reference points on the face and infer its orientation without camera calibration or any other settings.

Previous work

To estimate the pose from a continuous video stream, first we have to correctly track the face, a task which is rendered difficult due principally to environment changes and the face appearance variability under different head orientations. Its non–rigidity adds yet another degree of difficulty. To overcome these problems, a great number of techniques have been developed using color information, facial features, templates, optical flow, contour analysis and a combination of these methods [164]. One can group these methods into two main categories: motion-based and model-based approaches. The first one depends on visual motions whereas the latter imposes high-level semantic knowledge.

For videos in which the face occupies less than 30 by 30 pixels (very low resolution), 3D model based approaches are inefficient [93]. The pose estimation problem is rather converted into a classical classification problem. For somewhat higher resolution videos, since facial features remain hard to detect and track, adapted techniques were developed. For instance, Cascia [18] proposed to model the frame differences around face area as a linear combination of some template difference images caused by small perturbation of the pose parameters. The head pose variation can be obtained according to the estimated linear coefficients. In the case of high resolution face images, geometric-based approaches typically share some advantages with the appearance-based approaches, when the facial features can be accurately recovered [87]. Using for example Harris detector, the head pose could be estimated by DeMenthon algorithm which uses, in turn, successive scaled orthographic approximation in a Ransac framework [176]. In [204], a training phase makes it possible to characterize each point of the ellipsoidal model of the head by a probability density function. A maximum a posteriori classification on the extracted edge-density features provides an estimate of the pose.

More investigations are conducted on head tracking using facial features like eye corners and mouth corners which if combined with other features can help achieve better tracking accuracy [190].

Feature-based face tracking

It is well know that localisation of facial features is often hard to achieve due to several possible corruptions (illumination, noise, occlusion) or the presence of a complex background. However, the motivation behind using feature-based techniques is their near invariance under pose and orientation changes [203]. Therefore a feature–based pose tracker seems to be advantageous to explore since a similar geometric-based approach has been exploited successfully in [71].

The purpose of this paper and the major contributions of this work consist first in using a bank of steerable filters as strong descriptors of a set of facial attributes, because of their robustness with respect to occlusions and to global geometrical deformations [164]. We present, in the Fourier domain, an elegant formulation to find a set of Cartesian x, y–separable basis Gaussian kernels and the corresponding coefficients to construct a Gaussian steerable filter. Furthermore, since the mouth and lip corners are more involved in facial actions they represent less stable features. Therefore, using them as facial attributes like in [71] will result in an inaccurate pose estimation when some facial expression instances are performed at the same time. Improving performance, by allowing facial expressions during pose recovering, requires more stable features. The facial features used here are the two upper points of the eyebrows, the two upper nasolabial–furrow corners and the nasal root that are more stable. When one of the two nasolabial-furrow corners is occluded, in the case of a rotated face, we propose a geometric scheme to recover it. Finaly, we try to derive the slant and tilt angles by solving an eigenvalue problem.

Local characteristic descriptors

Steerable filters [69] are good candidates for tracking task. They are based on Gaussian derivatives, which have both spatial and frequential orientation selectivity, and have been shown to be robust to view point changes [205]. Such representation is crucial for an accurate tracking since no feature–based vision system can work unless good features can be identified and tracked from frame to frame [159]. The local information is characterized by applying a bank of filters (jet) which extracts the grey-level distribution of pixels in the neighbourhood of each facial fiducial point. In practice, good performance under both criteria of selectivity and invariance are obtained by jets with derivatives up to third order, four orientations and three widths.

Local structure, may be kept in an extra–jet by combining the center jet descriptor with surrounding jets [205].

Steerable filter framework

Steering with self-similar functions

Freeman [69] synthesized steerable filters by linearly combining a set of basis filters using gain maps $k_{iN}(\theta)$, and denoted filter responses on N^{th} order Gaussian derivatives G_N^θ to an arbitrary orientation θ by:

$$G_N^\theta = \sum_{i=1}^{N+1} k_{iN}(\theta)\ G_N^{\theta_i} \tag{A.1}$$

with

$$\theta_i = \frac{(i-1)\pi}{N+1}$$

$$k_{iN}(\theta) = \frac{2}{N+1} \sum_{r=0}^{(N-1)/2} \cos(2r+1)(\theta - \theta_i)\ \text{, odd } N$$

$$k_{iN}(\theta) = \frac{1}{N+1} \left(1 + 2\sum_{r=1}^{N/2} \cos 2r(\theta - \theta_i)\right)\ \text{, even } N$$

Steering with Cartesian partial derivatives

Cartesian x, y-derivatives can be used to find a set of basis function needed for the construction of $G_{N,M}^\theta$, a steered Gaussian derivative kernel [4]. In the frequency domain, the N^{th} order $x-$partial derivative of a filter $f(x,y)$ streered to an arbitrary orientation is obtained by multiplying the Fourier transform by an imaginary oriented ramp raised to the N^{th} power $(-j\omega_\theta)^N$:

$$f^{\theta}_{N,0}(x,y) = \sum_{\omega_x=-\pi}^{\pi} \sum_{w_y=-\pi}^{\pi} (-j\omega_\theta)^N \; F(\omega_x,\omega_y) \\ \exp(-j(\omega_x x + \omega_y y)) \tag{A.2}$$

Expressing this oriented ramp in terms of the horizontal and vertical ramps provides the basis and coefficients needed to steer the N^{th} order $x-$derivative to an arbitrary orientation θ:

$$(\omega_\theta)^N = (\cos\theta\,\omega_x + \sin\theta\,\omega_y)^N$$

In the same way, we compute the basis and the coefficients needed to steer $y-$derivative up to the arbitrary order M, but this time by expressing the 90° counter-clockwise rotated oriented–ramp $(\omega_{\theta_\perp})$ in term of its horizontal and vertical ramps :

$$(\omega_{\theta_\perp})^M = (\sin\theta\,\omega_x - \cos\theta\,\omega_y)^M \tag{A.3}$$

Thus, the product $(\omega_\theta)^N\,(\omega_{\theta_\perp})^M$ provides the basis and coefficients needed to express the (N,M) order (x,y)-derivative steered filter $(f^{\theta}_{N,M})$:

$$(\omega_\theta)^N\,(\omega_{\theta_\perp})^M = (\cos\theta\,\omega_x + \sin\theta\,\omega_y)^N\,(\sin\theta\,\omega_x - \cos\theta\,\omega_y)^M$$

As an illustration, the $(x,y)-$separable basis $G^0_{1,1}, G^0_{2,0}, G^0_{0,2}$ and the corresponding coefficients c_0, c_1, c_2 needed to express the $(x,y)-$derivative steered filter $G^{\theta}_{1,1}$ (Fig. A.1) are straightforwardly given by :

$$(\cos\theta\ \omega_x + \sin\theta\ \omega_y)\ (\sin\theta\ \omega_x - \cos\theta\ \omega_y)\ = \\ \cos(2\theta)\ \omega_x\omega_y + \frac{1}{2}\sin(2\theta)\ {\omega_x}^2 - \frac{1}{2}\sin(2\theta){\omega_y}^2$$

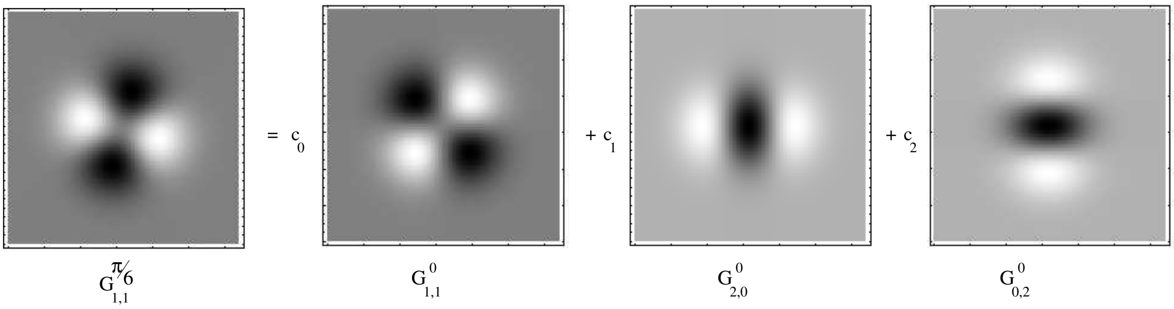


Figure A.1. Decomposition of a partial derivative steerable kernel $(\theta = \frac{\pi}{6})$**.**

Hence, we can write :

$$G^{\theta}_{1,1} = \cos(2\theta)\ G^{0}_{1,1} + \frac{1}{2}\sin(2\theta)\ G^{0}_{2,0} - \frac{1}{2}\sin(2\theta)\ G^{0}_{0,2}$$

We find that the proposed decomposition scheme which uses the Fourier domain is more convenient and more straightforward than the one proposed by Bart in [4].

One–feature correspondence

The RGB pixels are transformed to single-value attributes representing the *subjective color* according to [147]. Explicitly, the subjective–like color measures the weighted distance to the black color. The red and the blue channels weighting factors depend on how large is the red component as described in [147].

Tracking the facial features from frame to frame is performed using a jet similarity measure based on the cosine correlation coefficient:

$$S(m, m') = \frac{\sum_n m_n\ m'_n}{\sqrt{\sum_n {m_n}^2\ \sum_n {m'_n}^2}}$$

The best feature match within a search window around the previous position corresponds to the highest S–value.

Estimating the facial orientation

If we suppose that the face plane is determined by the two upper points of the eyebrows and the two upper nasolabial-furrow corners, then a $3D$ facial orientation refers to two parameters (σ, τ). The slant (σ) which is the angle between the optic axis and the face plane normal vector, and the tilt (τ) which represents the angle between the parallel projection of the face plane normal and the x axis (Fig. A.2).

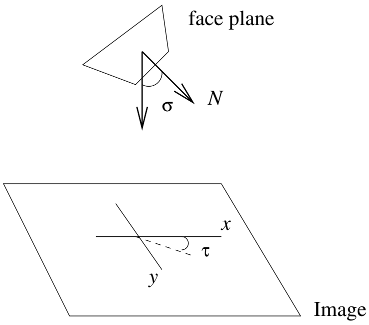


Figure A.2. 3d **surface orientation.**

Estimating the occluded point

In order to track a significantly rotated face we have to estimate an occluded facial feature (one of the upper nasolabial-furrow corners occluded by the tip of the nose) from the visible ones[1]: If $l_p > l_q$ (the distances from the nasal root point s to the upper points of the eyebrows, respectively p_1 and q_1) (Fig. A.3) we expect that the right part of the face is visible. Then q_2 could be occluded by the nose. So we have to track p_1, q_1, p_2 and to estimate q_2.

If we suppose that r is the reflection of p_1 through the point s' the projection of the

[1] For simplicity, we consider the right side as visible. The same reasoning can be done for the left one.

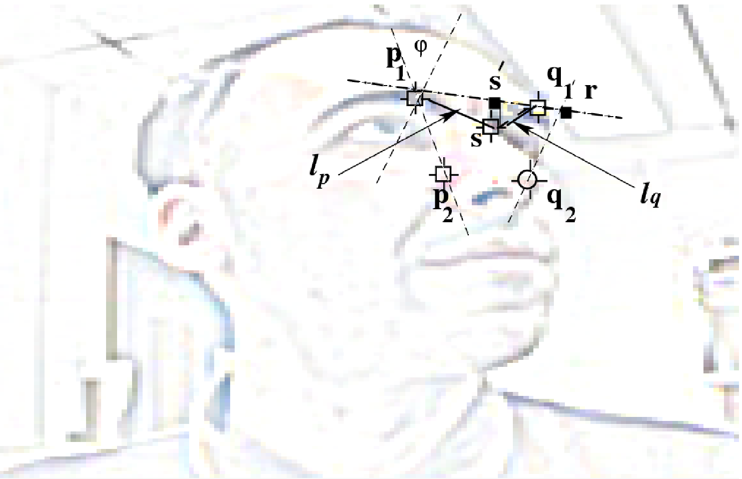


Figure A.3. The estimation of the occluded point (q_2) position.

nasal root point s on the line (p_1q_1), the estimated point q_2 is obtained from:

$$\overrightarrow{r\,q_2} = Rot\,(-\varphi) \cdot \overrightarrow{p_1\,p_2}$$

where Rot refers to a $2D$ rotation matrix, and

$$\varphi = \pi \,-\, 2\,\angle(p_1\,p_2\,,\,p_1\,q_1)$$

Slant and tilt estimation

Unlike [71], we do not force a parallel configuration between the line defined by the two upper eyebrows features (p_1q_1) and that defined by the upper nasolabial–furrow corners (p_2q_2) (Fig. A.3) in the image plane. We calculate the "mapping vectors" (μ_1, μ_2) as eigenvectors of the matrix A which forms with a vector b the affine transformation $\{A, b\}$ that backprojects the line (q_1q_2) to (p_1p_2) (Fig. A.4).This backprojection is parameterized by $\{a, b_x, b_y\}$ [127]:

$$A = \begin{pmatrix} a & -b_x \frac{1+a}{b_y} \\ b_y \frac{1-a}{b_x} & -a \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} b_x \\ b_y \end{pmatrix}$$

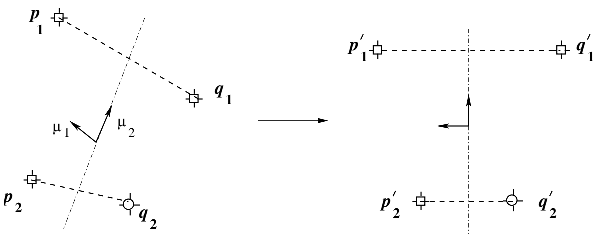


Figure A.4. Mapping vectors backprojection.

Then, the slant (σ) and tilt (τ) are recovered from the vectors μ_1 and μ_2 (Fig. A.4). If we put the equation of the image skewed symmetry line in the following form :

$$\alpha\ x + \beta\ y = 1$$

and from its equation [127]:

$$(1-a)b_y x + (1+a)b_x y - b_x b_y = 0$$

we can write :

$$\alpha = \frac{1-a}{b_x} \quad \text{and} \quad \beta = \frac{1+a}{b_y}$$

After appropriate manipulations, we can form and solve :

$$W.X = Z$$

where

$$X = \begin{pmatrix} a & b_x & b_y \end{pmatrix}^T \quad ;$$

$$Z = \begin{pmatrix} p_{1x} & p_{1y} & p_{2x} & p_{2y} & 1 & -1 & 2 \end{pmatrix}^T$$

$$\text{and} \quad W = \begin{pmatrix} q_{1x} & 1 - \beta\, q_{1y} & 0 \\ -q_{1y} & 0 & 1 - \alpha\, q_{1x} \\ q_{2x} & 1 - \beta\, q_{2y} & 0 \\ -q_{2y} & 0 & 1 - \alpha\, q_{2x} \\ 1 & \alpha & 0 \\ 1 & 0 & -\beta \\ 0 & \alpha & \beta \end{pmatrix}$$

Given the parameters a, b_x and b_y, we can now form the matrix A and find its eigenvectors (μ_1, μ_2). From figure (A.4), the skewed frame vectors μ_1 and μ_2 are respectively backprojected to the unskewed frame vectors[2] $(-1, 0)^T$ and $(0, \nu)^T$, by a transformation U, so we can write :

$$U = \begin{pmatrix} 0 & \nu \\ -1 & 0 \end{pmatrix} * \begin{pmatrix} \mu_1 & \mu_2 \end{pmatrix}^{-1}$$

Let $\{e_1, e_2\}$ be the eigenvalues of the matrix $U^T\, U$ and $\mathbf{v}$ the eigenvector corresponding to the highest eigenvalue e_1, the tilt τ will be given by :

$$\tau = \arctan\left(\frac{v_x}{v_y}\right)$$

and the slant σ by :

$$\sigma = \arccos\left(\lambda\right)$$

where λ^2 stands for the ratio of the eigenvalues $\frac{e_2}{e_1}$ [127] .

[2] The object (face) plane aspect ratio ν is set to 0.56

Results

For evaluating the system, video sequences with a 320 × 240 pixels resolution were taken where different head movements and orientations were performed. First, the facial features are interactively initialized in the first frame, however, assuming a pre-training phase, automatic initialization could be possible. After that, for each visual facial feature an extra-jet which combines the center pixel jet vector with four neighboring jets localised at a distance of five pixels along diagonals, is computed through a partial derivative operator convolution as a linear combination of convolutions with a set of Cartesian x, y-separable Gaussian derivative kernels. Considering derivative up to the third order, four orientations and three widths, the feature vectors are 180–dimension jets. The tracking is performed as a one-feature correspondence within a search window around the previous position.

Figure (A.5) shows that facial attributes were tracked fairly well despite the strong tilt (2^{nd} dial) that occurs during the motion. After checking the visibility of each of the two upper nasolabial-furrow facial features, the transfer process allows to trigger the tracking of the newly visible feature (green square) and at the same time the estimation of the position of the newly occluded one (red square). As we can see, from the same figure, the switching from the visible-feature tracking mode to the occluded-feature position estimation mode was sufficiently well performed to allow the head-pose estimation procedure to provide full effectiveness. It is difficult to obtain ground truth data for assessing exactly the head-pose parameters, but we can see that the estimation results are visually satisfactory (Fig. A.5).

We should note, however, that the feature locations drift somewhat, particularly in a brightly illuminated scene as it is shown in the two last frames from figure (A.6) which result in a lost of tracking, the problem is principally due to the similarity function which is based on one-feature correspondence. This difficulty can be avoided using a similarity function that employs all of the five facial fiducial points , which

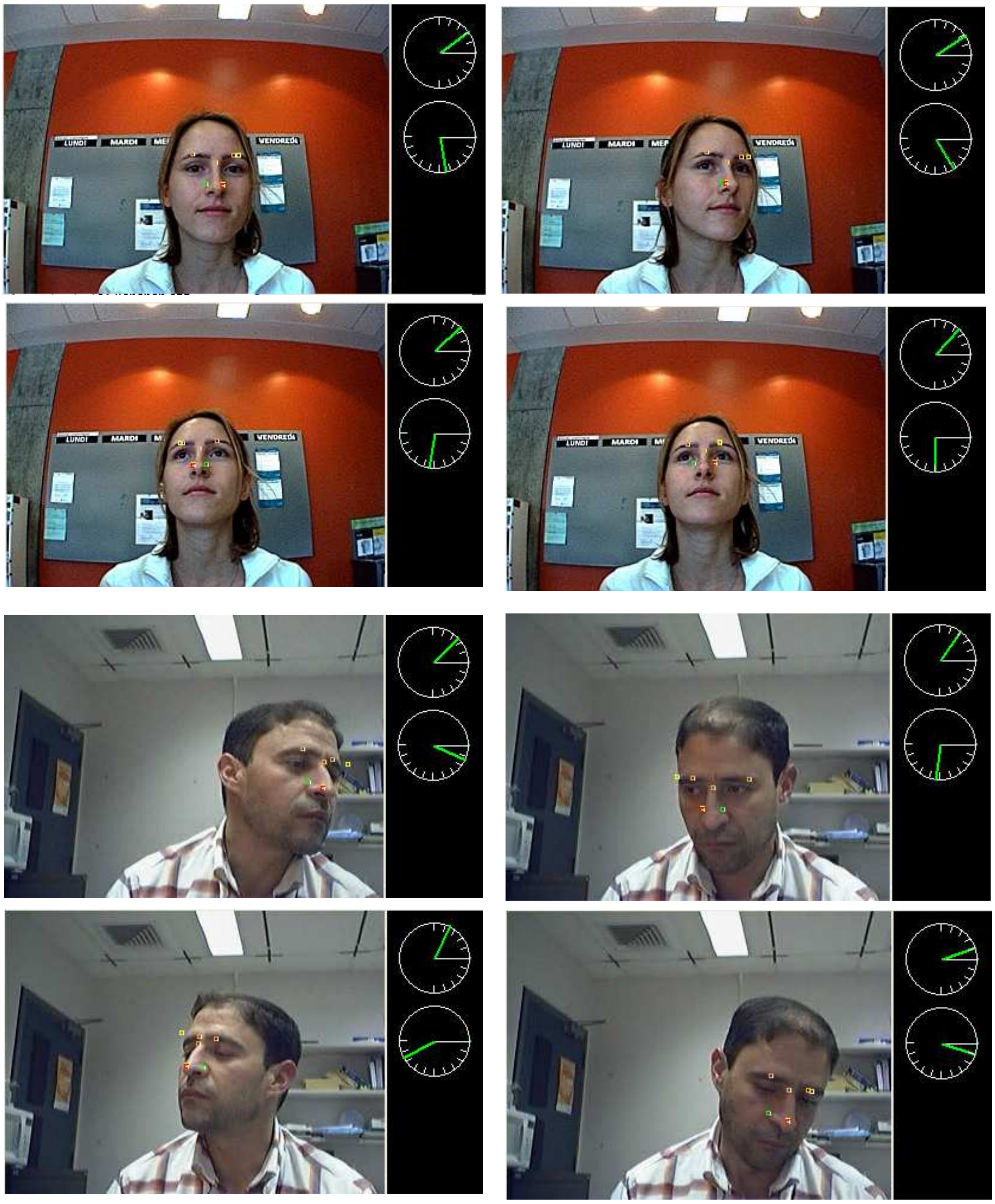


Figure A.5. Tracking and pose estimation results from a uniformly illuminated scene (1^{st} dial: slant angle, 2^{nd} dial: tilt angle)**.**

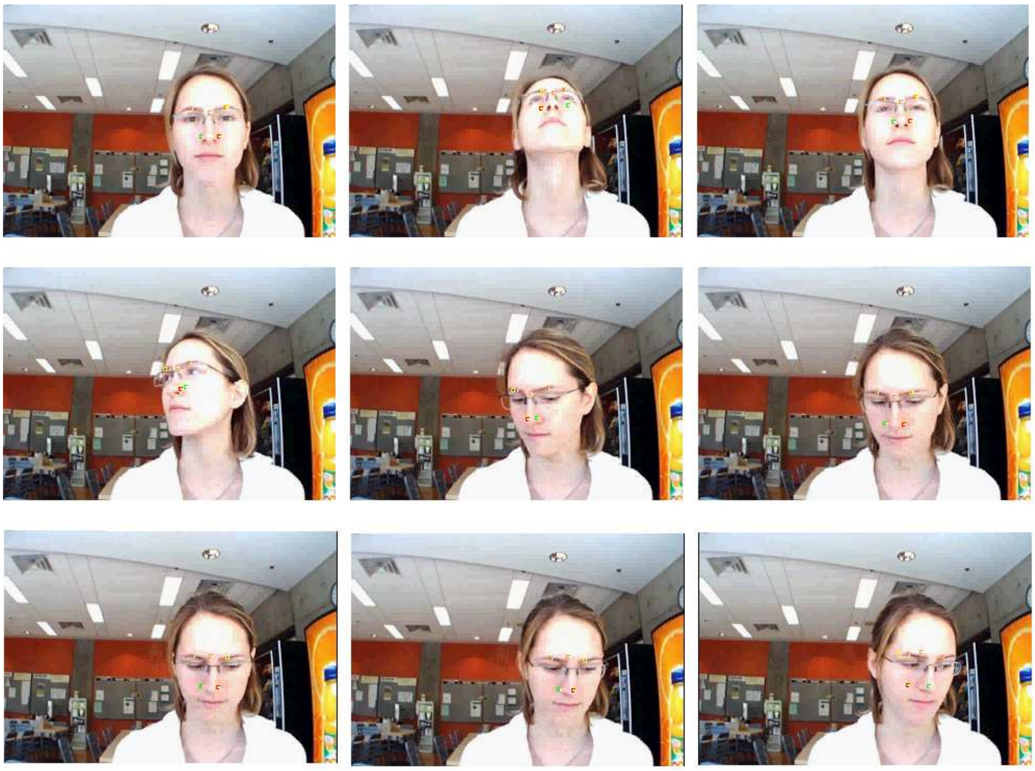


Figure A.6. One–feature correspondence results from a brightly illuminated scene.

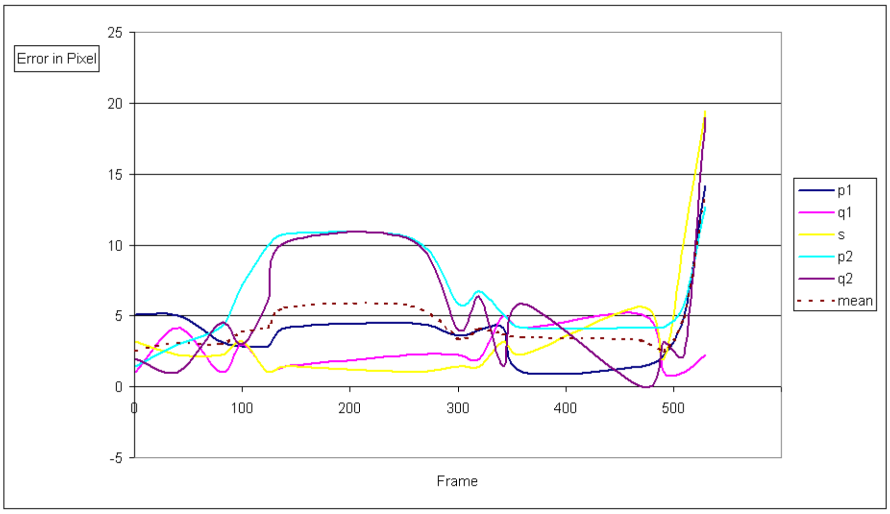


Figure A.7. Localisation errors with respect to manually selected features from a brightly illuminated scene.

can be implemented as a graph matching function.

Conclusion

This article has described a feature-based pose estimation approach using steerable filters, which have demonstrated a reasonable compromise between sensibility and invariance. A practical scheme has been proposed to express such descriptors which represent local structure information of four facial features; the fifth feature (possibly occluded) is roughly estimated using only 2D assumptions on the geometry of the face. This process allows the pose estimation procedure to achieves maximum effectiveness in the case of rotated face. When dealing with a fast movement of the face we need to increase the search area, which can turn out to be less functional in practice. In this case, a hierarchical scheme could offer a significant contribution. Also, a solution which imposes constraints on the possible geometrical configurations could be more convenient. In the other hand, instead of using a one-feature correspondence similarity function, a graph matching function which uses all of the five facial fiducial points would be more efficient in the case of a lack of local structure details particularly in highly illuminated scene. In this case the other textured facial features will positively contribute in the overall similarity function and compensate.

Annexe B

(ARTICLE) AN EFFICIENT 3D HEAD POSE INFERENCE FROM VIDEOS

Cet article [39] a été publié comme l'indique la référence bibliographique

> M. Dahmane et J. Meunier. An efficient 3d head pose inference from videos. Dans *Image and Signal Processing, 4th International Conference, ICISP 2010, Trois-Rivières, QC, Canada, June 30-July 2, 2010.*, volume 6134 de *Lecture Notes in Computer Science*, pages 368–375. Springer, 2010.

Abstract

In this article, we propose an approach to infer the 3d head pose from a monocular video sequence. First, we employ a Gabor–Phase based displacement estimation technique to track face features (two inner eye corners, two wings, tip and root of the nose). The proposed method is based on the iterative Lowe's pose estimation technique using the six tracked image facial points and their corresponding absolute location in a 3d face model. As any iterative technique, the estimation process needs a good initial approximate solution that is found from orthography and scaling. With this method, the pose parameters are accurately obtained as continuous angular measurements rather than expressed in a few discrete orientations. Experimental results showed that under the assumption of a reasonable accuracy of facial features location, the method yields very satisfactory results.

Introduction

The orientation of the human head allows inferring important non–verbal forms of communication in a face-to-face talk (e.g. spatial information about the focus of attention, agreement–disagreement, confusion. . .). The computer vision community is thus interested in automating interpersonal information captured from a scene. Thereby specific meaning can be automatically extracted from video by head pose inference, a process that presents an important challenge arising from individual appearance and personal facial dynamics.

The head pose can also be used to wrap and normalize the face for more general facial analysis, or in some augmented reality systems to construct images that wrap around the sides of the head.

For videos in which the face occupies less than 30 by 30 pixels (very low resolution), 3d model based approaches are inefficient [93]. The pose estimation problem is rather converted into a classical classification problem.

For somewhat higher resolution videos, since facial features remain hard to detect and track, adapted approaches were developed, such as appearance–based techniques (e.g. [207]).

In the case of high resolution face images, geometric–based approaches becomes relevant since facial features are visible and can be accurately recovered [87].

In recent years, a variety of methods have been introduced, the reader is therefore invited to look at the thorough review presented in [129].

In this paper we deal with high resolution videos and use a geometric approach for head pose assessment since it can directly exploit properties that are known to influence human pose [129]. A further motivation behind using feature–based approaches is their near invariance under pose and orientation changes [203]. However, their performance depends on the precise configuration of the local facial features. Therefore, an effective scheme for tracking the features on the face would be essential to compensate for this shortcoming.

Feature–based facial tracking scheme

Generally, facial feature tracking methods, mainly based on Gabor wavelets, try to perform refinement stages [121, 206, 213] by imposing geometric constraints with subspace projection techniques or by using gray–level profiles to refine and adjust the positions of the features of interest. In this paper, we adopt a facial feature–based tracking using Gabor phase–based technique. In what follows, we propose to use a personalized gallery of facial bunch graphs (sets of Gabor jets attached to each node), that we deform to fit a set of tracked points using the Procrustes transform. Six particular facial feature points (left/right eye–inner corner, left/right nose wings, root and tip of the nose) are used because they are known to be less sensitive to facial deformations (Fig. B.1).

Procrustes transform

Procrustes shape analysis is a method in directional statistics [118], used to compare two shape configurations. A two–dimensional shape can be described by a centered configurations $\mathbf{u}$ in $\mathcal{C}^k$ ($\mathbf{u}\,\mathbf{1}_k = 0$), where $\mathbf{u}$ is a vector containing `2d` shape landmark points, each represented by a complex number of the form $x + \imath y$.

The procrustes transform is the similarity transform (eq. B.1) that minimizes (eq. B.2), where $\alpha\,\mathbf{1}_k$, $|\beta|$ and $\angle\beta$, respectively, translates, scales and rotates $\mathbf{u}_2$, to match $\mathbf{u}_1$.

$$\begin{cases} \mathbf{u}_1 = \alpha\,\mathbf{1}_k + \beta\,\mathbf{u}_2 & \alpha, \beta \in \mathcal{C} \\ \beta = |\beta|\; e^{\imath\angle\beta} \end{cases} \tag{B.1}$$

$$\left\| \frac{\mathbf{u}_1}{\|\mathbf{u}_1\|} - \alpha\mathbf{1}_k - \beta\,\frac{\mathbf{u}_2}{\|\mathbf{u}_2\|} \right\|^2 \tag{B.2}$$

Tracking of facial reference–points

From frame to frame, the tracking of the 4 reference points (*left and right eye inner corner and the two nose wings*) is based on an iterative disparity estimation procedure (Gabor phase–based technique described in [38]). To ensure a high tracking efficiency, we create a personnalized bunch for each one of these features. The size of each bunch depends on the degree of deformation of the corresponding feature.

Geometric fitting

The Procrustes transform is used to adequately deform each reference graph G_i stored in the personalized gallery. The transformation, that best wraps the 4 anchor points of G_i to fit the tracked reference points (circle–dots in figure B.1), is used to adjust the positions of the two remaining feature points of G_i (diamond–dots in figure B.1). The new generated positions form the probe graph G_p.

Refinement stage

Then, Gabor jets are calculated at each point position of the generated graph G_p, and a Jet–based similarity [38] is computed between each reference graph G_i and the corresponding G_p. The probe graph with the highest similarity gives the positions of the six facial feature points. The final positions are obtained by estimating the optimal displacement of each point by using the Gabor phase–based displacement estimation technique [38].

Extracting the pose of the face

For recovering the pose of the face, we use the positions of the six facial feature points (*L–R eye-inner corner, L–R nose wings, the root and the tip of the nose*), that are less sensitive to facial deformations. The pose can be estimated from the topography

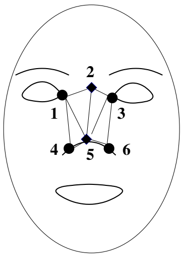


Figure B.1. The four tracked (circle) and the two adjusted (diamond) facial points.

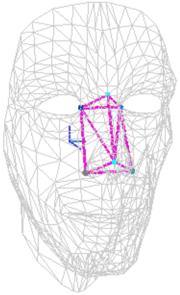


Figure B.2. The volume approaching the topography of the nose on a 3d GeoFace model.

of the nose, using the volume generated by these points as it can be seen from the face model (Fig. B.2).

Given an approximate 3d absolute position of each point of this volume and the corresponding 2d points on the face image, we estimate the pose of the head using Lowe's pose estimation algorithm [109].

Lowe's pose estimation method

Given $P_i (1 \leq i \leq n)$ the set of three–dimensional model points, expressed in the model reference frame, if P'_i represents the corresponding set, expressed in the camera reference frame, and if we denote by $p_i(x_i, y_i)$ the corresponding set of 2d image points, the pose estimation problem is to solve for the parameters[1] $\mathbf{s} = (\mathbf{R}, \mathbf{T})$ so that

$$[X'_i, Y'_i, Z'_i] \; = \; \mathbf{R}\, P_i \; + \; \mathbf{T} \tag{B.3}$$

where p_i is the perspective projection of P_i

[1] $\mathbf{s}$ is the pose vector concatenating the three rotation angles (*roll*, *pitch* and *yaw*) and the *x*,*y*,*z* translations.

$$[x_i, y_i] \;=\; f\left[\frac{X'_i}{Z'_i}, \frac{Y'_i}{Z'_i}\right] \;=\; \mathit{Proj}\left(\tilde{\mathbf{s}}, P_i\right) \qquad \text{(B.4)}$$

For $\tilde{\mathbf{s}}$, an estimate of $\mathbf{s}$, an error measurement between the observations and the locations of p_i is computed (eq. B.5).

$$\mathbf{e}_i \;=\; p_i - \mathit{Proj}\left(\tilde{\mathbf{s}}, P\right) \qquad \text{(B.5)}$$

The pose parameters correction amount $\delta\mathbf{s}$, that eliminates the residual $\mathbf{e}$, can be found via the Newton Method's [172]. Lowe's method proceeds by producing new estimates for the parameters $\mathbf{s}$ (eq. B.6) and iterating the procedure until the residual $\mathbf{e}$ (eq. B.5) drops below a given tolerance.

$$\mathbf{s}^{(i+1)} = \mathbf{s}^{(i)} + \delta\mathbf{s} \qquad \text{(B.6)}$$

Initialization

The iterative Newton's method starts off with an initial guess $\mathbf{s}^{(0)}$ which should be sufficiently accurate to ensure convergence to the true solution. As any iterative methods, choosing $\mathbf{s}^{(0)}$ from an appropriate well–behaved region is essential. For this purpose, we use the POS[2] algorithm [48, 171, 175], which gives a reasonable rough estimate for $\mathbf{s}^{(0)}$. The algorithm approximates the perspective projection with a scaled orthographic projection, and finds the rotation matrix and the translation vector of the 3d object by solving a linear system.

If p and P denote, respectively, the image points and the model points, the initial solution $\mathbf{s}^{(0)}$ can be determined, by recovering the rotation matrix $\mathbf{R}^{(0)}$ and the translation vector $\mathbf{T}^{(0)}$, from P_iP_1 and p_ip_1, two matrices constructed as follows,

[2] **P**ose from **O**rthography and **S**caling.

$$P_iP_1 = \begin{pmatrix} X_2 - X_1 & Y_2 - Y_1 & Z_2 - Z_1 \\ \vdots & \vdots & \vdots \\ X_n - X_1 & Y_n - Y_1 & Z_n - Z_1 \end{pmatrix} \quad ; \quad p_ip_1 = \begin{pmatrix} x_2 - x_1 & y_2 - y_1 \\ \vdots & \vdots \\ x_n - x_1 & y_n - y_1 \end{pmatrix}$$

The initial rotation matrix $\mathbf{R}^{(0)}$ is formed as

$$\begin{pmatrix} \mathbf{a}_N^T \\ \mathbf{b}_N^T \\ (\mathbf{a}_N^T \times \mathbf{b}_N^T) \end{pmatrix} \tag{B.7}$$

where $(\mathbf{a}\,,\mathbf{b})$ is the matrix obtained by the left division[3] $P_iP_1 \setminus p_ip_1$, $\mathbf{a}$ and $\mathbf{b}$ are three–dimensional vectors. Subscript $_N$ refers to the normalized vector and the symbol "×" to the cross product of two vectors.

The initial translation vector $\mathbf{T}^{(0)}$ is defined as

$$\frac{(\bar{p}_x, \bar{p}_y, f)^T}{sc} \tag{B.8}$$

where, sc refers to the scale of the projection that corresponds to $(\|\mathbf{a}\| + \|\mathbf{b}\|)/2$, $\bar{p} = \frac{1}{n}\sum p_i$, and f is the camera focal length in pixels.

Experimental results

To test the approach on real world data, we used representative video sequences displaying different head movements and orientations. In order to enhance the tracking performance, a personalized gallery was built with graphs[4] (Fig. B.1) from different faces under different "key orientations".

[3] The left division $A \backslash B$ is the solution to the equation $AX = B$.

[4] From our experiment, we found that about twenty graphs are sufficient to cover the different head appearances under various orientations.

First, the subgraph containing the four reference facial features (Fig. B.1) is roughly localized in the first frame of the video via an exhaustive search of the subgraph as a rigid object through the coarsest face image level. We used a three–level hierarchical image representation to decrease the inherent average latency of the graph search operation which is based on the Gabor jet similarity, by reducing the image search area and the size of the jet. For images at the finest level (640×480 pixel resolution), jets are defined as sets of 40 complex coefficients constructed from different Gabor filters spanning 8 orientations under 5 scales. Whereas those for images at the coarsest level (320×240) are formed by 16 coefficients obtained from filters spanning 8 orientations under 2 scales. The intermediate level images use jets of (8×3) coefficients.
Then, the iterative displacement estimation procedure is used as a refinement stage, performed individually on each position of all of the six feature–points, and over the three levels of the pyramid face–image. From frame to frame, only the four reference points are tracked using the iterative disparity estimation procedure. To avoid drifting during tracking, for each feature point, a local fitting is performed by searching through the gallery the subgraph that maximizes Gabor magnitude–based similarity. The rough positions of the four feature points are given by the positions of the nodes of the optimal subgraph. These are then adjusted using the displacement estimation procedure.
The entire `3d` head pose inference method is summarized in the flow diagram of figure B.3.

Figure B.4 shows the tracking and pose inference results for 3 different persons and environment conditions. Clearly, the achieved pose recovering performance are visually consistent with orientation of the head, as it is shown by the direction of the normal of the face shown in yellow color in figure B.4.

Figure B.5 gives an example where a tracking failure occurred (see the tip of the nose). In this case a correcting mechanism (e.g. reinitialization based on poor Gabor similarity) can be adopted to prevent the tracker from drifting.

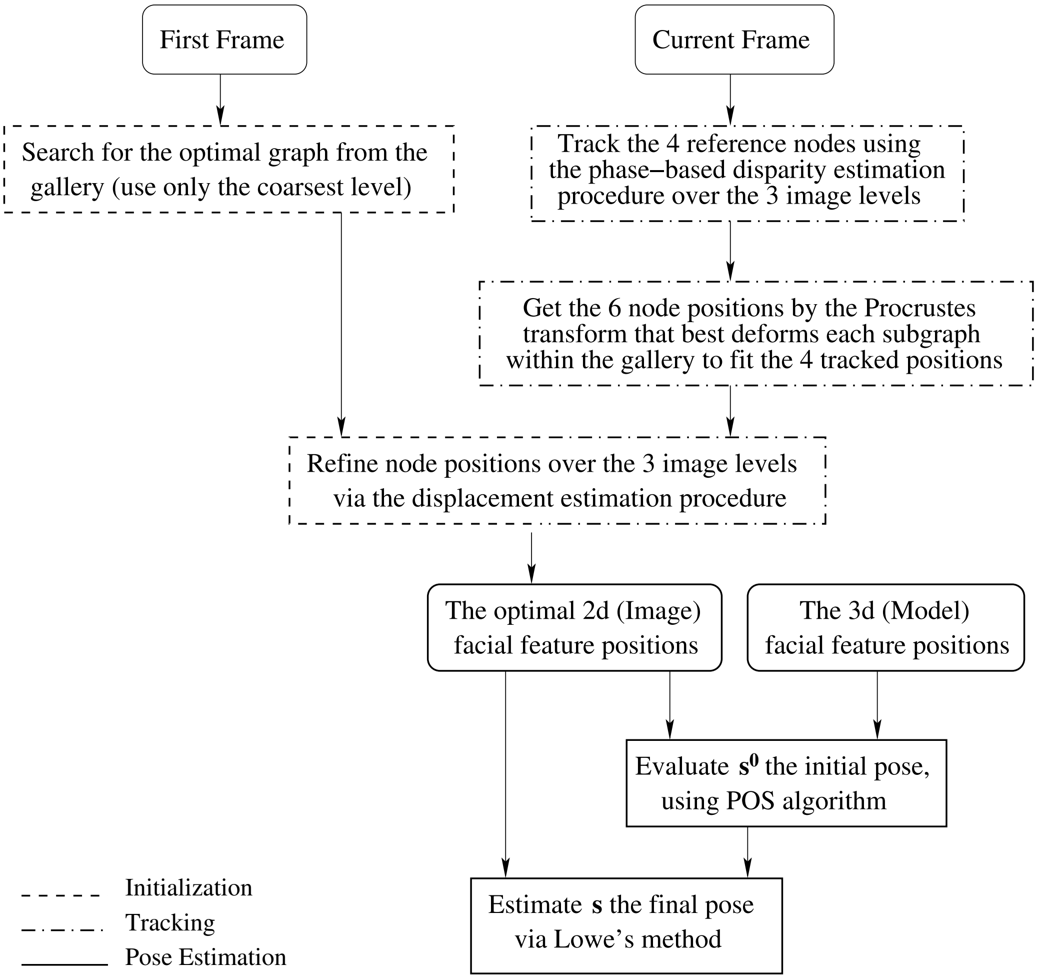

First Frame
Current Frame
Search for the optimal graph from the gallery (use only the coarsest level)
Track the 4 reference nodes using the phase–based disparity estimation procedure over the 3 image levels
Get the 6 node positions by the Procrustes transform that best deforms each subgraph within the gallery to fit the 4 tracked positions
Refine node positions over the 3 image levels via the displacement estimation procedure
The optimal 2d (Image) facial feature positions
The 3d (Model) facial feature positions
Evaluate $\mathbf{s^0}$ the initial pose, using POS algorithm
Estimate $\mathbf{s}$ the final pose via Lowe's method
Initialization
Tracking
Pose Estimation

Figure B.3. Flow diagram of the entire 3d pose inference approach.

Conclusions

This article has described a feature–based pose estimation approach using the iterative Lowe's pose estimation technique, with six tracked image facial points and their corresponding locations in a 3d face model. The estimated orientation parameters are given as continuous angular measurements rather than expressed in a few discrete orientations such as (left, up, front . . .), furthermore the solution provides information about the 3d position of the object. As any iterative solution, the estimation process needs an accurate initialization seed, which can be done using the pose from

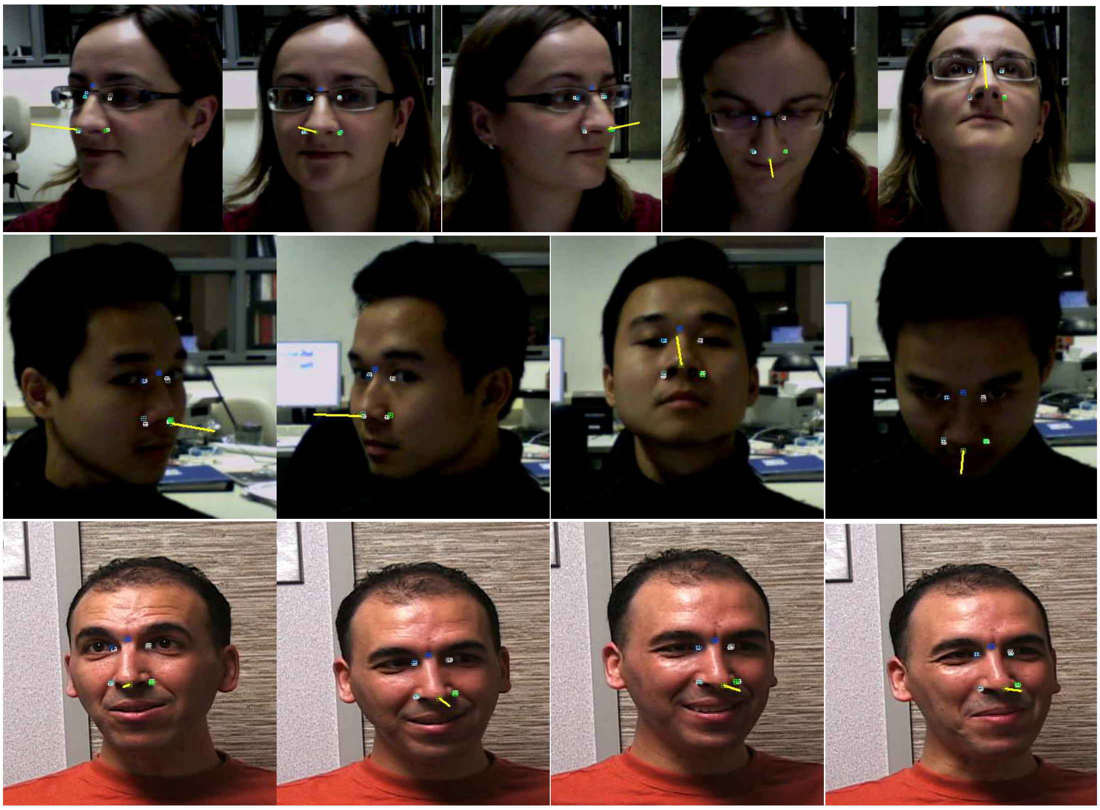


Figure B.4. Tracking and pose recovery performances

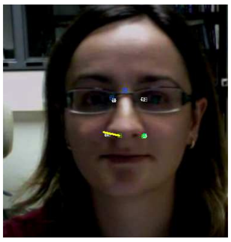


Figure B.5. A tracking failure of the nose tip affects the pose recovery.

orthography and scaling algorithm. The facial features' tracking is accomplished by using the phase–based displacement estimation technique and a personalized gallery of facial bunch graphs to further enhance the tracking efficiency. In the future, we plan to assess non–verbal communications between a patient and his health care professional in clinical setting.

Annexe C

(ARTICLE) OBJECT REPRESENTATION BASED ON GABOR WAVE VECTOR BINNING: AN APPLICATION TO HUMAN HEAD POSE DETECTION

Cet article [43] a été publié comme l'indique la référence bibliographique

> M. Dahmane et J. Meunier. Object representation based on gabor wave vector binning: An application to human head pose detection. Dans *IEEE International Conference on Computer Vision Workshops, ICCV 2011 Workshops, Barcelona, Spain*, pages 2198–2204, 2011.

Abstract

Visual object recognition is a hard computer vision problem. In this paper, we investigate the issue of the representative features for object detection and propose a novel discriminative feature sets that are extracted by accumulating magnitudes for a set of specific Gabor wave vectors in 1-D histogram defined over a uniformly-spaced grid.

A case study is presented using radial-basis-function kernel SVM as base learners of human head poses. In which, we point out the effectiveness of the proposed descriptors, relative to related approaches. The average performance reached 65% for yaw and 73.3% for pitch, which are better than the (40.7% and 59.0%) accuracy achieved by calibrated people. A substantial performance gain as higher as (1.18% for yaw and 1.27% for pitch) is achievable with the proposed feature sets.

Introduction

The human head pose allows inferring important non verbal information between individuals such as spatial information about the focus of attention, agreement-disagreement, confusion, people nod etc.... Computerized extraction of such specific meanings from videos presents an important challenge arising mainly from personal appearance (identity) and individual facial dynamics. Among adopted approaches, geometric methods are particularly interesting, since they can directly exploit properties that are known to influence human pose [129]. However, their performance is very sensitive to location of the local facial points since these are hard to detect and track. The alternative is to view the head pose detection problem as a classification problem; however, this requires a robust discriminative feature sets for object representation.

In this article, we study the issue of feature sets for object detection (head pose in our case), showing that the Gabor wave vector binning based descriptors, which use Gabor magnitude processed over dense grid of uniformly spaced cells, provide superior performance relative to state of the art methods. This feature set can be considered as orientation/scale tunable line and edge detectors, since they only respond to some specifically oriented patterns in some specific scale, whereas some related descriptors such as histogram of oriented gradient (HoG) [44] uses edge information by considering the local intensity gradient distribution over the edge directions.

The test case consists of detecting head pose from images to evaluate the performances of the proposed descriptors. Experiments are performed to compare our features sets relative to the histogram of oriented gradient descriptors. The effectiveness of the proposed approach is evaluated by comparing it to other algorithms using the same dataset.

As base learner classifier, we use a multiclass radial-basis-function kernel Support Vector Machines which are known as stronger classifiers for high dimensionality problems, and powerful algorithms for solving difficult classification problems [34, 180]. We have used the Intel OpenCV Library implementation.

Related work

In a computer vision context, head pose estimation is the process of inferring the orientation of a human head from digital imagery. In this context, we are interesting to investigate inter-persons interaction patterns from video recordings, such as the visual focus of attention, people nod, quick head movement that may be sign of surprise or alarm, other head movements that indicate dissent, confusion, consideration, agreement [129].

Recently, a wide variety of computerized methods for extracting the orientation of the head were developed. The existing approaches can be roughly classified into three main approaches:

- *Appearance-based* methods [131, 200, 207] establish image view comparison between training examples and a probe head image.
- *Tracking-based* methods [146, 208] use the image intensity information and/or inter-frame correspondence between features of interest as input of different tracking algorithms to get the head orientation at each frame.
- *geometric/feature-based* methods [189, 206, 213] use general prior knowledge of geometrical face structure, (e.g. eye and/or mouth lines) to infer the orientation of the human face.

See [129] for a more elaborated categorization.Feature-based methods require precise and stable localization of facial landmarks which is often not ensured, the drawbacks that can be naturally avoided by the appearance based-methods, which have attracted more and more attention, but in contrast, are in the need to a suitable feature set closely relevant to pose variations and reliably insensitive to facial variations irrelevant to pose [117]. This is precisely the issue that we are trying to address by designing the Gabor wave binning based feature sets.

Authors in [1] propose a feature-based face recognition method that directly uses jets generated from Gabor filtering, and use Euclidean distance to match Gabor jets. What we propose is an appearance based approach, our descriptors are based on wave vector based histograms that use the amplitude of the Gabor filter response. Voit et. al. [185] use a neural network-based approach to estimate the head pose, whereas Tu et. al. [173] propose a method based on graphs and PCA. Gourier in [74] use associative neural networks to train their winner-takes-all classifier.

Histogram of Oriented Gradients

The (HoG) feature descriptors were proposed by Dalal and Triggs [44]. This method that was originally developed for person detection is used in more general object detection algorithms. The main idea behind the HoG descriptors is based on the edge information. That is each window region can be described by the local distribution of the edge orientations and the corresponding gradient magnitude. The local distribution is described by the local histogram of oriented gradients which is formed by dividing the detection window into a set of small region called cells, and within each cell integrate the magnitude of the edge gradient for each orientation bin. This is done for all of the pixels within the cell. To provide better invariance, the local HoG is normalized over the block of neighboring cells.

Dalal and Triggs [44] have shown that HoG outperforms wavelet, PCA–SIFT [111] and Shape Context [10] approaches. Different implementations of HoG were considered to study the influence of different descriptor parameters. For good performance, the authors highlight the importance of fine scale gradients, fine orientation binning, relatively coarse spatial binning, and high-quality local contrast normalization in overlapping descriptor blocks.

The default typical parameters include $[-1, 0, 1]$ gradient filter with no smoothing, linear gradient voting into 9 orientation bins in $0° - 180°$, 16×16 blocks of four overlapping 8×8 pixel cells, $L_2 - norm$ block normalization, block spacing stride of 8 pixels, and 64×128 detection window. In this implementation, we used 8 by 4 blocks of 12×10 pixels cell, with a cell spacing stride of 6 pixels.

The Gabor wave vector binning based descriptors

Gabor wave vector binning based descriptors are Gabor wavelet transform based feature sets generated from a set of given filters corresponding to a selected wave vectors at specific orientations and scales.

Gabor wave vectors

The Gabor wavelet transform family is defined as:

$$\psi_{\mu,\nu}\left(z\right) \;=\; \frac{\left\|k_{\mu,\nu}\right\|^2}{\sigma^2}\, e^{-\frac{\|k_{\mu,\nu}\|^2\,\|z\|^2}{2\sigma^2}} \left[e^{\imath\, k_{\mu,\nu}\, z} - e^{-\frac{\sigma^2}{2}}\right] \tag{C.1}$$

where $z = (x, y)$, and

$$k_{\nu,\mu} = k_\nu\, e^{\imath \phi_\mu} \tag{C.2}$$

Equation C.2 defines the wave vector $k_{\nu,\mu}$ relative to the orientation μ and the scale ν, with $k_\nu = k_{\max}/f^\nu$ and $\phi_\mu = \pi\,\mu/8$. $k_{\max}$ refers to the maximum frequency and f to the proportionality factor between the wavelet frequencies.

Commonly, we have $\mu \in \{0, \ldots, 7\}$ and $\nu \in \{0, \ldots, 4\}$ defining 40 wave vectors. Figure C.1 shows an example of the transformed images including the gradient magnitude image and the magnitude of the filter responses of the GWT at $k_{1,4}$ and $k_{4,5}$. σ was set to 2π.

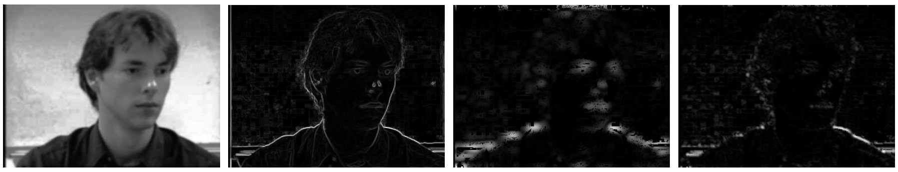


Figure C.1. Examples of the transformed images: the leftmost image is the original image; the second image represents the gradient magnitude; the two last images are the magnitude of the GWT responses for $k_{1,4}$ and $k_{4,5}$ respectively

Gabor wave vector binning

The method uses an image window to evaluate local histograms of GWT magnitude responses at selected frequencies and orientations, considering a first–order image gradients. For HoG, the gradients capture contours, and some edge information, whereas for our descriptors, the gradient map provides salient image locations at which the GWT magnitude will be computed. Thereby, reinforcing the identification task by selecting the pixels that are potentially more informative.

The Basic key idea is that local object appearance and shape can often be learned from local window on the image using the spatial distribution of magnitude over different frequencies and orientations. The method constructs local histograms of Gabor magnitude responses, as image features. We consider a map of potentially "textured" pixels by computing the edge-gradient magnitude window-image using Sobel operator. Features are extracted by dividing the image window into small spatial regions called "cells", and by accumulating, over the pixels of each cells, a local 1D magnitude histogram for a specific GWT wave vector referring to specific scale and orientation. Within each cell, only pixels with a dominant local gradient magnitude are considered. The number of bins is specified by the number of the considered wave vectors. Thereby, the resulting combined cumulative histograms over the cells of the detection window provide the image representation.

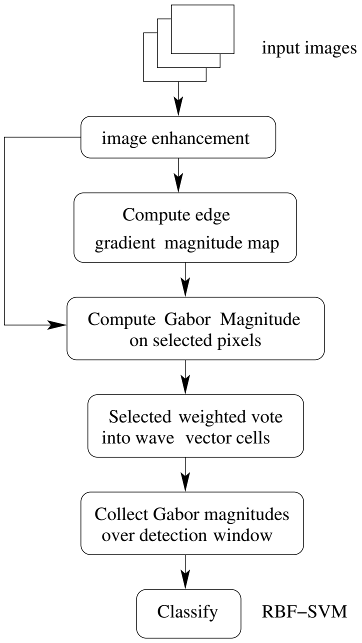

input images
image enhancement
Compute edge gradient magnitude map
Compute Gabor Magnitude on selected pixels
Selected weighted vote into wave vector cells
Collect Gabor magnitudes over detection window
Classify
RBF–SVM

Figure C.2. The outline of our proposed system.

The underlying motivation for using Gabor–based descriptors

The use of GWT is motivated by the fact that the Gabor transformed face images reveals strong characteristics of spatial locality, and orientation selectivity, and therefore are consistent with intrinsic characteristics of human face images. Researchers have used 2D Gabor filters as "biological motivated" filters, based on what Daugman (1984) cited about the visual system. Concerning the multi-scale scheme, in the Gabor wavelet transformation we use a term of (k/σ) (equation C.1) which keeps the continuously changing window $(radius = \sigma/k)$ within the Gaussian envelope over the topdown Gabor filter processing. In contrast with the standard image down-sampling, the Gabor pyramid image filtering maintains not only continuity in the spatial frequency of the Gabor feature but also detection ability. Clearly, with a standard image down-sampling, HoG cannot ensure this continuity. Besides, it has been proven that HOG/SIFT representation for facial analysis (eg. gender recognition) has several problems among which, the feature matching is assumed that the faces are well registered.

Pointing'04 benchmark

For our experiments, we have used the pointing'04 benchmark [73] in order to get results comparable to many other algorithms tested on the same database. The training data was collected from the PRIMA Team in INRIA Rhone-Alpes by the FAME Platform where persons are asked to gaze successively at 93 markers that cover a half-sphere in front of the person corresponding to the set of poses.

The head pose database consists of 15 sets of images. Each set contains 2 series of 93 images of the same person at different poses. There are 15 people in the database, wearing glasses or not and having various skin color. The pose, or head orientation is determined by 2 angles varying in 15 degree pan angle intervals from -90 degrees to +90 degrees, and in 15 degree tilt angle intervals from -30 degrees to +30 degrees in addition to a tilt of +60 and -60 degrees. Also, there exists two poses with tilt angles of ±90 degrees at a pan angle of 0 degree. Figure C.3 shows an example of person 13.

As it can be seen from figure C.4, the poses from the three image pairs seem to be nearly identical, the leftmost image pair are supposed to be different by 15 degrees in tilt and pan direction. Also, the 15-degree pose difference in the middle and the right image pairs, are quite close, which makes the dataset more difficult.

Technical implementation and evaluation

Gabor wave vector based descriptors

The color images with a resolution of 384 × 288 are converted to grayscale, and enhanced by contrast amelioration and brightness normalization. We used a detection window with 100 × 40 pixels size. Since the proposed approach is window-based detection, we have to deal with the alignment problem by searching for the eyes

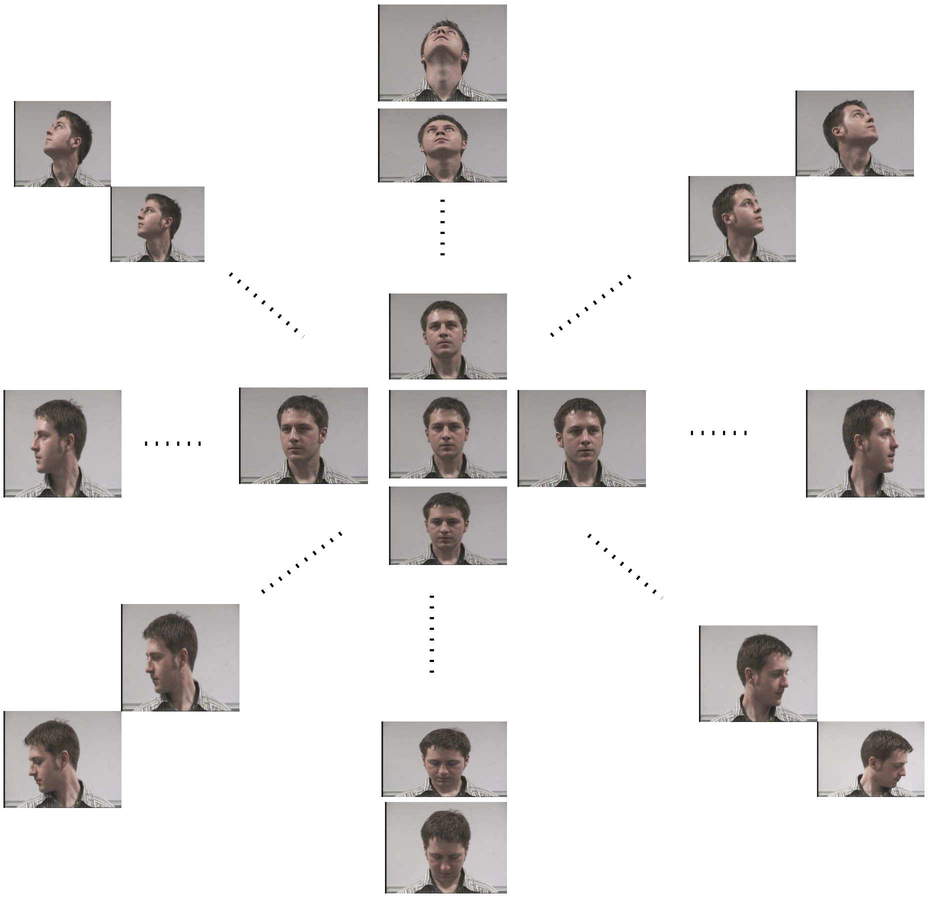


Figure C.3. The head images of the person13 in Pointing'04 dataset.

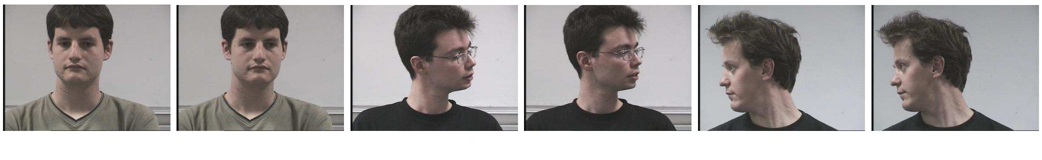


Figure C.4. Near poses in Pointing'04 dataset.

Table C.1. Comparing the performance of different pose detection techniques on POINTING'04 setup.

	Mean Absolute Error (°)		Classification Accuracy (%)	
	Yaw	Pitch	Yaw	Pitch
Human [74]	11.8	9.4	40.7	59.0
Voit [185]	12.3	12.7	–	–
Tu [173]	14.1	14.9	55.2	57.9
Gourier [74]	10.1	15.9	50.0	43.9
Our method	**5.7**	**5.3**	**65.0**	**73.3**

region over the entire image. Each eye is represented by 4 Gabor jets. The searching procedure uses a metric for Gabor jets similarity.

The detection window is partitioned into 8 by 4 cells of 12 × 10 pixels, without overlapping stride, nor grouping blocks. We compute the horizontal and vertical image gradient to generate the edge gradient map, that is used for selecting informative pixels (pixels with an edge-gradient beyond a given magnitude, 30 for instance, that will be considered in the subsequent stages). The voting strategy is based on the Gabor magnitudes that are processed at each wave vector $k_{\nu,\mu}$ and for each informative pixel. Magnitudes are collected into 40 histogram bins (each bin corresponds to one wave vector). Histograms are then integrated over the cell to form the local histogram.

After that for each block of 2×2 neighboring cells the 4 corresponding histograms are concatenated in a block- histogram. Then, the normalized 8 block-histograms were concatenated into a single 1280 dimensional feature vector.

For the multiclass SVM, we use Radial Basis Function (RBF) kernel (equation C.3).

$$K(\mathrm{x}, \mathrm{x}_i^*) = \exp\left(-\gamma \|\mathrm{x} - \mathrm{x}_i^*\|^2\right) \tag{C.3}$$

We have proposed an innovative simple epoch-based strategy to determine the optimal values for the kernel parameter (σ) and the penalty multiplier parameter (C) of the SVM: At each epoch, if we put $\sigma = (1/\sqrt{2 * \pi}(V_{max} - V_{min})$ with V_{max} (resp. V_{min}) the maximum (resp. minimum) value of the feature vector elements over the training dataset, then γ_e will be given by $1/(e * \sigma^2)$ and the multiplier C by $C_e = e * \sigma^{1.75}$. The tuple (γ_e, C_e) corresponding to the epoch e with highest training accuracy and a reasonable number of support vectors relatively to the cardinality of the training set, was selected. Generally, the stopping criterion is met between $e = 10$ and $e = 15$ epochs .

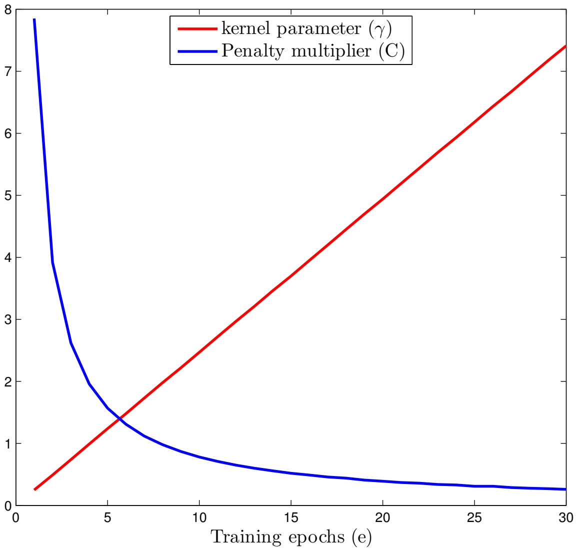


Figure C.5. The SVM parameters evolution over training epochs.

We choose to use the 1/3 of the dataset for testing and the rest of data to train the system. The two sets were split according to the training set having the highest accuracy. At each split the optimal SVM parameters were found.

Performance comparison with existing algorithms are reported in Table C.1. As shown, our feature descriptor achieves better accuracy than state-of-the art approaches [129] that are using the same setup. The {pitch,yaw} angles are accurately

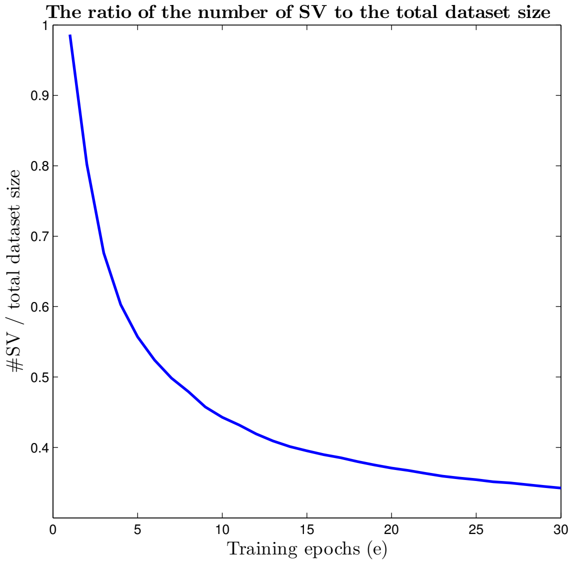


Figure C.6. Stabilization of the number of support vectors from e=25.

estimated (about {5.3,5.7} degree on absolute mean error). The classification performance reached 65% for yaw and 73.3% for pitch are better than the (40.7% and 59.0%) accuracy achieved by calibrated people as reported in [73].

Given the ambiguity cases in the training data (see figure C.4 and in figure C.7 which are showing clearly inside the detection window the difference in appearance for the same pose), we considered a relaxation scheme of the accuracy constraints on the output of the classifier . We assumed that any ±15 neighboring poses is correct classification as well. In Table C.2 we can see that the mean absolute error on the yaw angles drops to 0.9, it drops by half for the pitch angle. The overall classification accuracy was 96.4%, whereas Wu et al. [198] obtained 90% of correct classification using the same dataset.

Continuous poses

The continuity of Gabor response in the neighborhood permits to establish a mapping between the space of the discreet poses and the descriptors space. Interpolating the

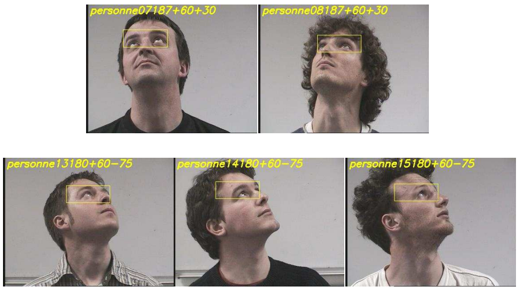


Figure C.7. Ambiguity cases in Pointing'04 dataset: Obviously, the appearances of the eyes region inside the detection window are different for the same pose class.

Table C.2. Considering ±15 neighboring poses as correct classification.

	Mean Absolute Error (°)		Classification Accuracy (%)	
	Yaw	Pitch	Yaw	Pitch
±0°	**5.7**	**5.3**	**65.0**	**73.3**
±15°	**0.9**	**2.5**	**97.5**	**91.8**

Table C.3. HoG vs. proposed descriptor performance comparison.

		M.A.E (°)		C. Acc. (%)	
		Yaw	Pitch	Yaw	Pitch
±0°	HoG	5.6	6.3	66.4	70.7
	Our feature sets	**5.7**	**5.3**	**65.0**	**73.3**
±15°	HoG	0.9	3.7	97.5	88.1
	Our feature sets	**0.9**	**2.5**	**97.5**	**91.8**

Table C.4. Interpolated pose : Overall classification performance.

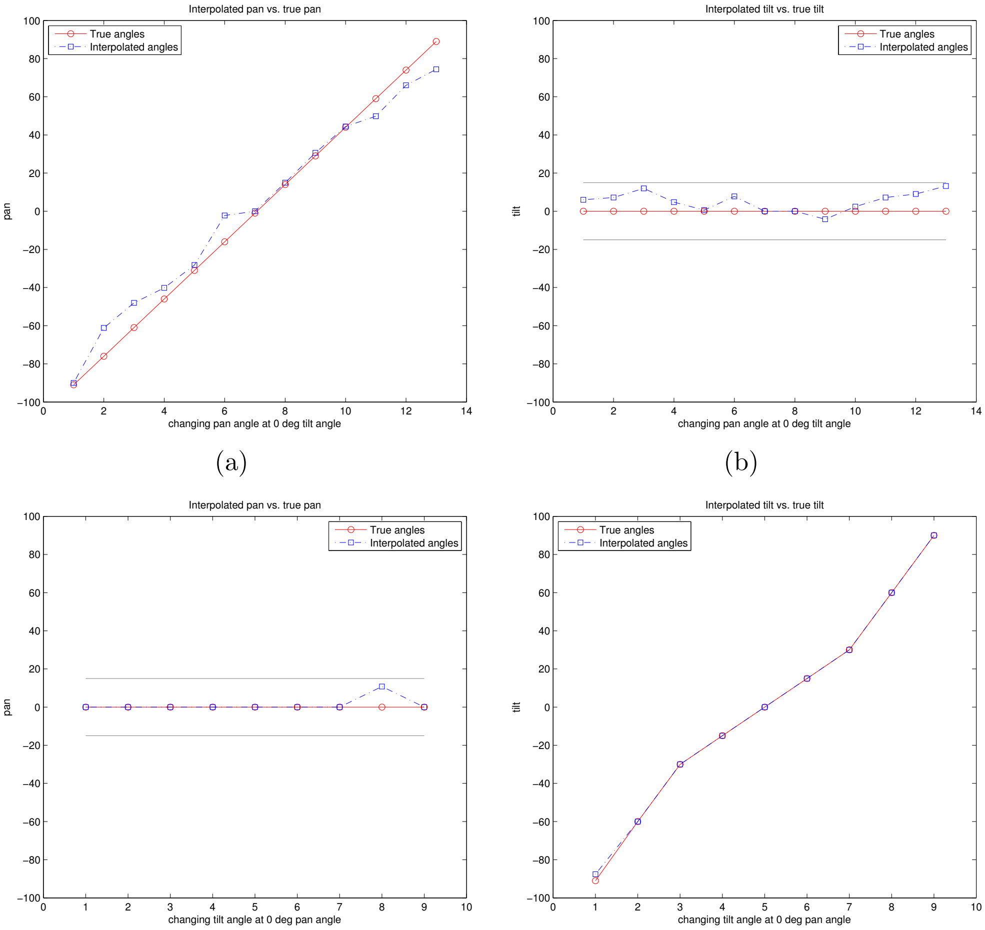
		C. Acc. (%)
±0°	HoG	74.1
	Our feature sets	**80.1**
±15°	HoG	96.8
	Our feature sets	**97.3**

3×3 neighboring poses (i.e. poses within ±15° range) of the winner pose (pose having the maximum SVM score) using the respective scores as weights, gives the continuous pose.

As described, the classification system permits to estimate discrete poses (i.e. classes) as discrete angular measurements. Since the number of fixed poses to sufficiently sample the continuous pose space is not sufficient, we introduced a commitee method to infer a smooth continuous head pose estimate by interpolating the $N \times N$ angular values. The “in-between” pose estimation shows that the feature descriptors corresponding to two neighboring poses are closer in the feature space. Figure C.8(a) and C.8(b) show the profiles of the interpolated angles versus true angles corresponding to the middle horizontal line in figure C.3 (i.e. pan=0°). The interpolated pan (resp. tilt) at 0° tilt degree which corresponds to the middle column in figure C.3, is shown in figure C.8(c) (resp. figure C.8(d)).

Conclusion

In this article, we presented a Gabor wave vector binning based descriptors. To our knowledge, it is the first time that Gabor wave vector based histogramming is addressed. We show that they present, for a hard estimation problem such as pose estimation, a suitable feature set closely relevant to pose variations and reliably insensitive to facial variations irrelevant to pose, such as identity, lighting etc. We also show its robustness with respect to possible bias of the training data. In comparison with some of the state of the art algorithms, we showed that our descriptors



(a)

(b)

(c)

(d)

Figure C.8. Comparison of interpolated angles.

set performs better classification accuracy, achieving a substantial performance gain (as higher as 1.18% for yaw and 1.27% for pitch) against state of the art algorithms that used the same benchmark (table C.1). It achieves high recognition rates relative to human classification, the average performance reached 65% for yaw and 73.3% for pitch, which are better than the (40.7% and 59.0%) accuracy achieved by calibrated people. Better classification accuracy against the HoG detector is obtained (table C.4), particularly, for the pitch angle with a benefit of 2.6% at a tolerance of 0° and 3.7% at ±15°. The overall classification performance of interpolated poses reached (80.1% at a tolerance of 0° and 97.3% at ±15°) with the our feature sets, and respectively (74.1%, and 96.8%) with the HoG feature set. The average processing time is 10.92 *sec* on a 3GHz x86-CPU to generate 40 integral images of Gabor magnitude, and is 0.104 *sec* to process 9 integral images of gradient magnitude required for the HoG features. Finally, we should mention that the system does not need any non-pose negative examples, and is able to provide a smooth continuous estimate of the yaw and pitch angles from discrete poses.

Annexe D

(ARTICLE) ENHANCED PHASE–BASED DISPLACEMENT ESTIMATION: AN APPLICATION TO FACIAL FEATURE EXTRACTION AND TRACKING

Cet article [38] a été publié comme l'indique la référence bibliographique

> M. Dahmane et J. Meunier. Enhanced phase–based displacement estimation - An application to facial feature extraction and tracking. Dans *Proc. of Int. Conference on Computer Vision Theory and Applications*, pages 427–433, 2008.

Abstract

In this work, we develop a multi–scale approach for automatic facial feature detection and tracking. The method is based on a coarse to fine paradigm to characterize a set of facial fiducial points using a bank of Gabor filters that have interesting properties such as directionality, scalability and hierarchy. When the first face image is captured, a trained grid is used on the coarsest level to estimate a rough position for each facial feature. Afterward, a refinement stage is performed from the coarsest to the finest (original) image level to get accurate positions. These are then tracked over the subsequent frames using a modification of a fast phase–based technique. This includes a redefinition of the confidence measure and introduces a conditional disparity estimation procedure. Experimental results show that facial features can be localized with high accuracy and that their tracking can be kept during long periods of free head motion.

INTRODUCTION

The computer vision community is interested in the development of techniques to figure out the main element of facial human communication in particular for HCI applications or, with additional complexity, meeting video analysis. In both cases, automatic facial analysis is highly sensitive to face tracking performance, a task which is rendered difficult due principally to environment changes and particularly to its great appearance variability under different head orientations, its non–rigidity adds yet another degree of difficulty. To overcome these problems, a great number of techniques have been developed which can be divided into four categories: knowledge–, feature–, template– and appearance–based [203].
Among these techniques, it is known that face analysis by feature point tracking demonstrates high concurrent validity with manual FACS (Facial Action Coding System) coding [27], which is promising for facial analysis [32]. Moreover, when facial attributes are correctly extracted, geometric feature–based methods typically share some common advantages, such as explicit face structure, practical implementation, collaborative feature-wide error elimination [87]. In this context, several concepts were developed.

The classical matching technique extracts features from two frames and tries to establish a correspondence, whereas correlation-based techniques compare windowed areas in two frames, and the maximum cross correlation value provides the new relative position. However, recent techniques have been developed to determine the correct relative position (disparity[1]) without any searching process as it is required by the conventional ones. In this category, phase–based approaches have attracted attention because of their biological motivation and robustness [66, 166].

In the literature, one can find several attempts at designing non–holistic methods based on Gabor wavelets [158]. Due to their interesting and desirable properties

[1] we use interchangeably the words “disparity” and “displacement”

including spatial locality, self similar hierarchical representation, optimal joint uncertainty in space and frequency as well as biological plausibility [64]. However, most of them are based on the magnitude part of the filter response [102, 107, 168, 178]. In fact, under special consideration, particularly because of shift–variant property, the Gabor phase can be a very discriminative information source [208].

In this paper, we use this property of Gabor phase for facial feature tracking. In section 2, we describe the Gabor-kernel family we are using. In section 3, we introduce the adopted strategy for facial features extraction. The tracking algorithm is given in section 4, including technical details and a discussion on its derivation. Finally, we apply the approach to a facial expression database, in section 5.

LOCAL FEATURE MODEL BASED ON GABOR WAVELETS

Gabor Wavelets

A Gabor jet $J(\mathbf{x})$ describes via a set of filtering operation (eq. D.1), the spatial frequency structure around the pixel $\mathbf{x}$, as a set of complex coefficients.

$$J_j(\mathbf{x}) = \int_{N^2} I(\mathbf{x}')\Psi_j\left(\mathbf{x} - \mathbf{x}'\right) d\mathbf{x}' \tag{D.1}$$

A Gabor wavelet is a complex plane wave modulated by a Gaussian envelope:

$$\Psi_j\left(\mathbf{x}\right) = \eta_j\, e^{-\frac{\|\mathbf{k}_j\|^2\,\|\mathbf{x}\|^2}{2\sigma^2}} \left[e^{\imath\,\mathbf{k}_j\cdot\mathbf{x}} - e^{-\frac{\sigma^2}{2}}\right] \tag{D.2}$$

where $\sigma = 2\pi$, and $\mathbf{k}_j = (k_{jx}, k_{jy}) = (k_\nu \cos(\phi_\mu), k_\nu \sin(\phi_\mu))$ defines the wave vector, with

$$k_\nu = 2^{-\frac{\nu+2}{2}}\,\pi \qquad \text{and} \qquad \phi_\mu = \mu\frac{\pi}{8}$$

Notice that the last term of equation D.2 compensates for the non-null average value of the cosine component. We choose the term η_j so that the energy of the wavelet Ψ_j

is unity (eq. D.3).

$$\int_{N^2} |\Psi_j\,(\mathbf{x})\,d\mathbf{x}|^{\,2} \;=\; 1 \tag{D.3}$$

A *jet* $J(\mathbf{x}) = \{a_j\ e^{\imath\phi_j}\ /\ j = \mu + 8\nu\}$, is commonly defined as a set of 40 complex coefficients constructed from different Gabor filters spanning different orientations ($\mu \in [0, 7]$) under different scales ($\nu \in [0, 4]$).

AUTOMATIC VISUAL ATTRIBUTE DETECTION

Rough face localization

When the first face image is captured, a pyramidal image representation is created, where the coarsest level is used to find near optimal starting points for the subsequent individual facial feature localization stage. Each trained grid (Fig. D.1) from a set of pre-stored face grids is displaced as a rigid object over the image. The grid position that maximizes the weighted magnitude–based similarity function (eq. D.4 and D.5) provides the best fitting node positions.

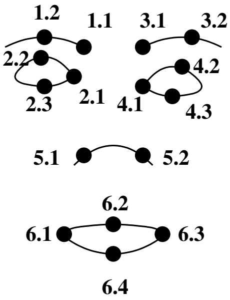


Figure D.1. Facial nodes with their respective code.

$$Sim(\mathrm{I}, \mathrm{G}) \;=\; \prod_{l}^{L} S(J_l, J_l') \tag{D.4}$$

$S(J, J')$ refers to the similarity between the jets of the corresponding nodes (eq. D.5), L stands for the total number of nodes.

$$S(J, J') = \sum_j c_j \frac{a_j\, a'_j}{\sqrt{\sum {a_j}^2 \sum {a'_j}^2}} \text{ with } c_j = \left(1 - \frac{\left|a_j - a'_j\right|}{a_j + a'_j}\right)^2 \tag{D.5}$$

The role of the weighting factor c_j is to model the magnitude–distortion δ as illustrated in figure D.2.

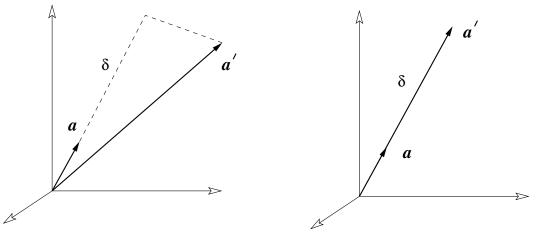


Figure D.2. Two different 3–dimensional jets. In the right subfigure, a not–weighted magnitude–based similarity $S(J, J')$ would have given an incorrect perfect match value 1.

Local Facial feature position refinement

The rough facial grid-node positions are then independently refined by estimating the displacement using a hierarchical selective search. The calculated displacements are propagated to subsequent hierarchy level, and a refinement operation is again performed. The optimal displacements are, finally, given at the finest image level.

The selective local search can be described as a local 3 × 3 neighborhood search, which allows distorting the grid until the maximum similarity value is reached. The search is then refined by propagating, to the next finer level, the three positions giving the highest similarity values. For each propagated potential position $P(x, y)$ the three adjacent neighboring positions $P(x + 1, y), P(x, y + 1)$ and $P(x + 1, y + 1)$ are also explored. The selective search continues downward until the finest level of the pyramid image is reached, where the optimal position is maximum (eq. D.5).

This procedure permits to decrease the inherent complexity required to calculate the convolution under an exhaustive search, first by reducing the search area (e.g. a 12×12 neighborhood on the finest level will correspond only to a 3×3 on the coarsest one) (Fig. D.3), and second by using smaller–size jets in coarser levels.

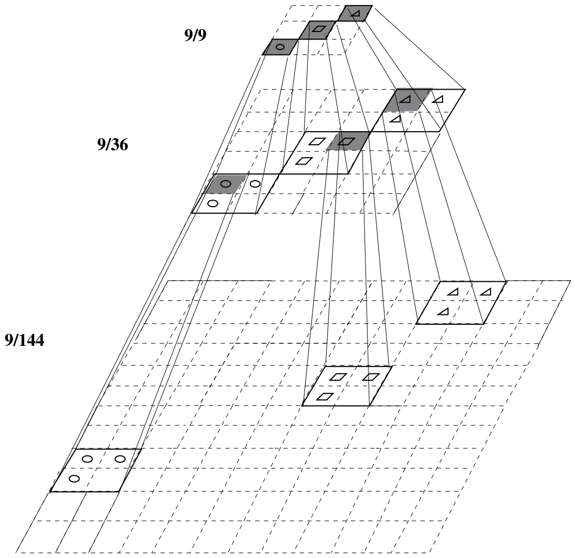


Figure D.3. Hierarchical–selective search. The values in left side denote the number of explored positions vs. the total number that would be explored in the case of an exhaustive search.

FACIAL ATTRIBUTES TRACKING

Facial features tracking is performed by estimating a displacement $\mathbf{d}$ via a disparity estimation technique [166], that exploits the strong variation of the phases of the complex filter response [119].

Later adopted by [215], this framework investigated in [119, 197] is based on the maximization of a phase–based similarity function which is nothing else than a modified way to minimize the squared error, within each frequency scale ν given two jets

J and J' (eq. D.6), as it has been proposed in [166].

$$e_\nu^2 = \sum_\mu c_{\nu,\mu} (\Delta\phi_{\nu,\mu} - \mathbf{k}_{\nu,\mu} \cdot \mathbf{d}_\nu)^2 \tag{D.6}$$

However, we assume that the merit of that framework is the use of a *saliency* term (eq. D.7) as weighting factor $c_{\nu,\mu}$, privileging displacement estimation from filters with higher magnitude response. Also, for such response it seems that phase is more stable [121].

$$c_j \ = \ a_j \, a'_j \tag{D.7}$$

In [166], the weighting factor c_j represents a confidence value (eq. D.8), that assesses the relevance of a single disparity estimate, and tends to reduce the influence of erroneous filter responses.

$$c_j \ = \ 1 - \frac{\left| a_j - a'_j \right|}{a_j + a'_j} \tag{D.8}$$

Both saliency term and normalized confidence ignore the phase of the filter response. In the present work, we try to penalize the response of the erroneous filters by using a new confidence measure that combines both magnitude and phase (eq. D.9).

$$c_j \ = \ {a_j}^2 \left(1 - \frac{\left| a_j - a'_j \right|}{a_j + a'_j} \right)^2 \ \frac{\pi - \left| \lfloor \Delta\phi_j \rfloor_{2\pi} \right|}{\pi} \tag{D.9}$$

The first term in this formulation represents the saliency term that is incorporated as a squared value of only the magnitude of the *reference jet* J which –contrary to the *probe jet* J'– necessarily ensures high confidence. We mean here by the reference jet the jet calculated from the previous frame or even a pre-stored one. The second bracket squared-term holds the normalized magnitude confidence. While, the last term, where $\lfloor \Delta\phi_j \rfloor_{2\pi}$ denotes the principal part of the phase difference within the interval $[-\pi, \pi)$, allows giving more weight to filters where the phase difference has a favorable convergence while, at the same time, limiting the influence of outlier filters.

The displacements can then be estimated with sufficient accuracy by minimizing (eq. D.6) which leads to a set of linear equations for $\mathbf{d}$, that can be directly resolved from (eq. D.10).

$$\mathbf{d}(J, J') = \begin{pmatrix} \sum_j \; c_j \, {k_{jx}}^2 & -\sum_j \; c_j \, k_{jx} \, k_{jy} \\ -\sum_j \; c_j \, k_{jx} \, k_{jy} & \sum_j \; c_j \, {k_{jy}}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_j \; c_j \, k_{jx} \lfloor \Delta\phi_j \rfloor_{2\pi} \\ \sum_j \; c_j \, k_{jy} \lfloor \Delta\phi_j \rfloor_{2\pi} \end{pmatrix} \tag{D.10}$$

Iterative Disparity computation

In [166], to obtain the disparity within one scale, the feature displacement estimates for each orientation were combined into one displacement per scale ($\mathbf{d}_\nu$) using the least squared error criterion (eq. D.6). The optimal disparity is then calculated by a combination of these estimates as an average value over all scales with appropriate weights (eq. D.8). Whereas in various approaches, a least squared solution is obtained in one pass, over the overall considered frequencies [197], some of them propose at first to use the lower frequencies subset (e.g. $\nu \in [2, 4]$), and then to resolve for higher frequencies subset (e.g. $\nu \in [0, 2]$).

These resolutions may carry an additive risk of unfavorable results; that is knowing that at each scale, there exists a displacement value above which its estimation would not be reliable, due to the lack of a large overlap of the Gabor kernels. Obviously, this value depends on the radius (σ/k_ν) of the Gaussian envelope.

As the power spectrum of the Gabor signal (eq. D.2) is concentrated in the interval $[-\sigma/\left(2k_\nu\right), \sigma/\left(2k_\nu\right)]$, we can compute the maximum disparity $\mathbf{d}_\nu^{max}$ that can be estimated within one scale (eq. D.11).

$$d_\nu^{max} = \frac{\sigma}{2\, k_\nu} = \frac{\pi}{k_\nu} \tag{D.11}$$

If for example the true displacement is $d = 7\,pixels$, then according to the Gabor–kernel family we used (section D), only the lowest frequency band filter gives a reliable estimation of the disparity.

So, the trick consists in estimating the disparity iteratively, from the lowest frequency to a highest *critical* frequency, depending on a stopping criterion involving the maximum allowed disparity value that can be effectively estimated. Some values are shown in table D.1 as a function of scale.

Table D.1. Critical displacement for each frequency.

ν	0	1	2	3	4
$d_\nu^{\max}(pixel)$	2	≈ 3	4	≈ 6	8

Given $J(\mathbf{x}) = \{a_j\, e^{\imath\phi_j}\}$ the *reference jet* and $J'(\mathbf{x}+\mathbf{d}) = \{a'_j\, e^{\imath\phi'_j}\}$ the *probe jet* i.e. the jet calculated at the probe position $(\mathbf{x}+\mathbf{d})$, using the j^{th} wavelet, an iterative disparity estimation algorithm (Fig. D.4) gives the optimal displacement $\mathbf{d}_{opt}$, that makes the two jets the most similar possible.

Iteratively, the conditional iterative disparity estimation (Fig. D.4) will unroll on the novel position $\mathbf{x}^{new} \leftarrow \mathbf{x} + \mathbf{d}_{opt}$ until a convergence criterion is achieved i.e. $\mathbf{d}_{opt}$ tends to $\mathbf{0}$ or the maximum number of iterations $l_{max_{iter}}$ is reached. Herein, ν_{critic} could keep its previous value, instead of starting, for each new position, with the coarsest scale (i.e. $\nu_{critic} = N_f - 1$).

Table D.2. Percentage of used frames to handle local facial deformations.

facial feature															
1.1	1.2	2.1	2.2	2.3	3.1	3.2	4.1	4.2	4.3	5.1	5.2	6.1	6.2	6.3	6.4
(%) of used frames															
2.5	1.8	3.9	4	3	2.3	3.4	4.2	3.7	2.7	1.5	2.4	3.8	8	2	9

Algorithm 1. IterativeDisparityEstimation $(\mathbf{x})$

```
1 Initially set $\nu$ with the lowest frequency index;

2 Calculate $J'_\nu(\mathbf{x})$ for the components that refer to $\nu$ at different
  orientations;

3 Estimate the disparity $\delta\mathbf{d}$ using equation (D.10) by considering
  all the processed frequencies at different orientations;

4 Compensate for the phase $\phi'_j = \lfloor \phi'_j - \mathbf{k}_j \cdot \delta\mathbf{d} \rfloor_{2\pi}$;

5 Cumulate the disparity $\mathbf{d} = \mathbf{d} + \delta\mathbf{d}$;

6 Perform the convergence test, if $\delta$d is greater than a threshold
  goto (3);

7 If the stopping criterion is not met, i.e. the overall
  displacement d is less than the critical displacement value $d_\nu^{\max}$,
  see Table (D.1), then put $\nu = \nu + 1$ (the next higher frequency)
  and goto (2).
```

Figure D.4. Conditional iterative disparity estimation algorithm.

EXPERIMENTAL RESULTS

The Hammal–Caplier face database [81] is used to test the proposed approach. In this database, each video contains about 120 frames for each of the 15 distinct subjects that are acting different facial expressions (neutral, surprise, disgust and joy) with some tolerance on rotation and tilting. We used 30 videos with spatial resolution of (320×240).

A generic face grid (Fig. D.1) is created using one frame from each subject (frontal view). In order to handle the facial deformation and prevent drifting, facial feature bunches are generated. Table D.2 shows each landmark and the percentage of the total number of frames required to create its own representative facial bunch. As we can see the number increases with the degree of variability of the local deformation that can be observed for each facial feature. These percentages were set empirically.

To locate the face grid, a search is performed over the coarsest level of the 3 image-levels that we used. Then a hierarchical selective refinement is performed using a weighted magnitude–based similarity to get the optimal node positions. Figure D.5 shows the results corresponding to the position refinement after rough node positioning.

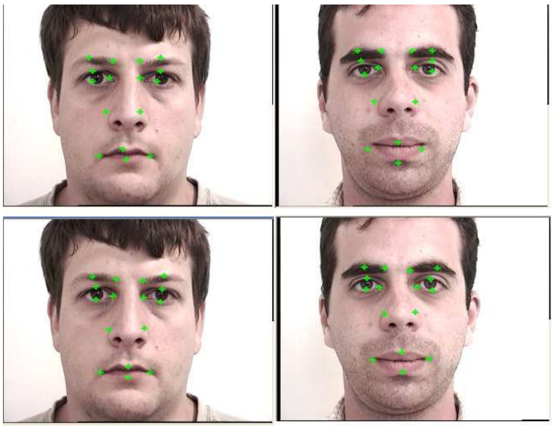


Figure D.5. Nodes position refinement (bottom) after rough positioning (top).

Figure D.6 shows the magnitude profile corresponding to $(\mu, \nu) = (0, 0)$ for node 2.1 (right inner–eye) from a video where the subject is performing a disgust expression. Figure D.7 illustrates the phase profile of the same subject with and without phase compensation ($\phi'_j \leftarrow \lfloor \phi'_j - \mathbf{k}_j \cdot \mathbf{d}_l \rfloor_{2\pi}$) in Algorithm 1. One can observe some large and sharp phase variations when non compensation is used, corresponding to tracking failure.

Figure D.8 shows three shots of a video showing a subject performing a disgust expression, the top subfigure presents the last frame. In this figure, we can see that the tracking has failed with a single jet (instead of a bunch). It's easy to see that the drifting can not be measured from the magnitude profile only (middle row), because

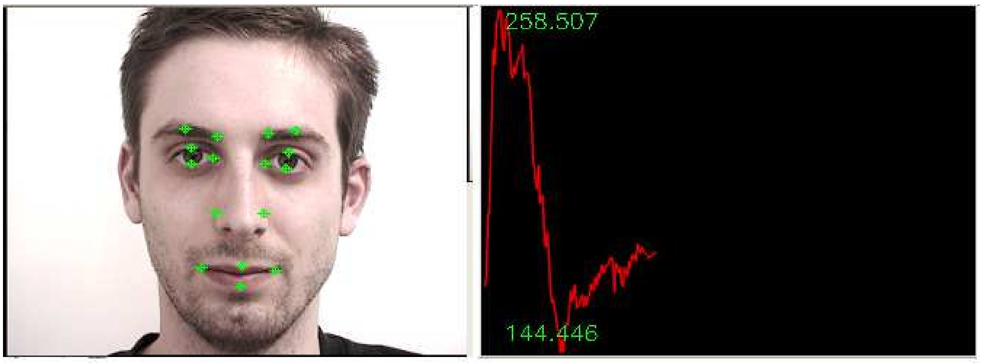


Figure D.6. Magnitude profile over time of Node 2.1 (right inner–eye).

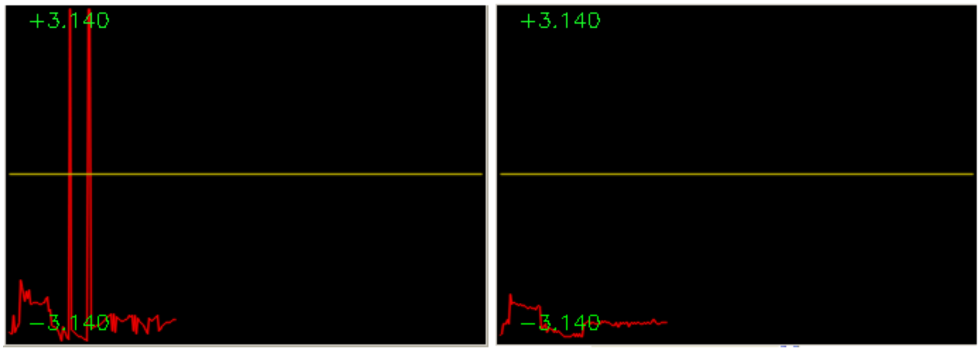


Figure D.7. Phase profile: not–corrected (left) vs. corrected (right) phase.

the magnitude changes smoothly with the position. This is not the case for the phase (bottom row) which is shift–variant, however by using a shift–compensation and facial bunches as described in Algorithm 1, we can correctly track the facial landmarks (Fig. D.9). In comparison with figure D.8, the bottom graph shows a horizontal and correct phase profile (without node drifting). The reader can appreciate the impact of such correction by looking in particular at node 2.1 (right inner–eye) and 2.3 (right lower eyelid) in figures D.8 and D.9.

In table 3, we summarize the tracking results of 16 facial features of 10 different persons with different expressions. The mean error of node positions using the proposed approach is presented in pixels. From the last column, we can see how the use of facial bunches appreciably increases nodes positioning and consequently the

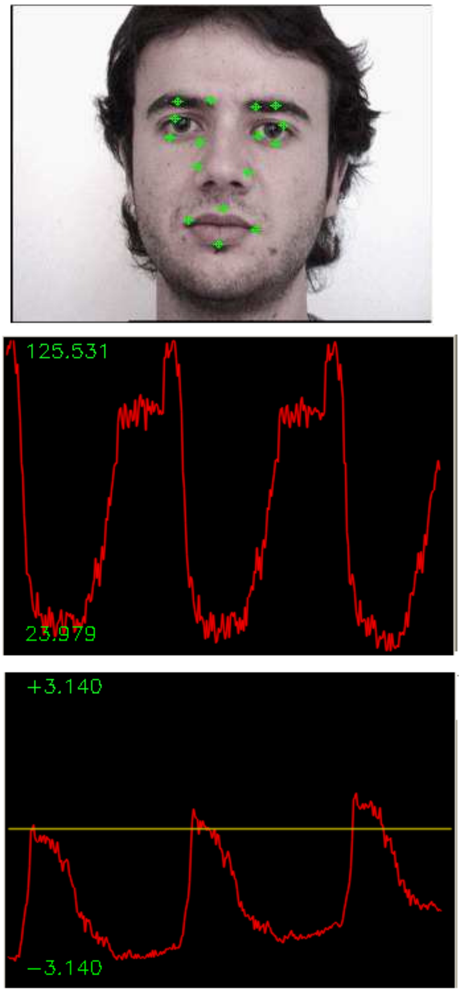


Figure D.8. A drifting case: magnitude vs. phase profile.

tracking accuracy.

CONCLUSION

In this work, we present a modification of a phase–based displacement estimation technique using a new confidence measure and a conditional disparity estimation. The proposed tracking algorithm permits to eliminate accumulation of tracking errors to avoid drifting, so offering a good facial landmark localization, which is a crucial

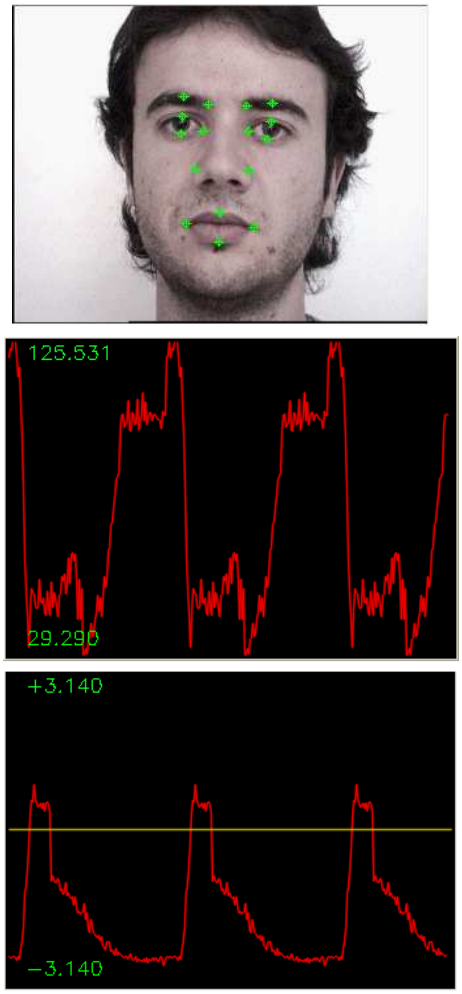


Figure D.9. Drift avoidance.

task in a feature–based facial expression recognition system. We notice that in these experiments, excepts for the first frame, no geometry constraints were used to enforce the facial shape configuration, especially for features that are difficult to track.

More training sessions could be needed to obtain pre-stored grids and features bunches that are representative of the variability of the human face appearance for initialisation and tracking respectively. In this context, through available face databases, advanced statistical models of data can be obtained using learning algorithms, such

Table D.3. Mean position error (pixels).

Subject	Without bunches	With bunches
#1	4.28	1.78
#2	3.98	1.37
#3	5.07	2.03
#4	4.44	1.9
#5	4.17	1.7
#6	4.05	1.63
#7	4.69	1.5
#8	4.1	1.75
#9	5.85	2.49
#10	6.93	2.47

as EM [92].

To reinforce the refinement step we are working on improving the local structure by providing an alternative appearance model which focuses more on high frequency domain without necessarily altering the relevant low frequency texture information, instead of modeling the grey level appearance [209] or exploiting the global shape constraint [121] which tends to smooth out important details. As future work, we plan to use facial feature bunches to generate for each facial expression and for each facial attribute what could constitute "Expression Bunches" for facial expression analysis.

Annexe E

(ARTICLE) INDIVIDUAL FEATURE–APPEARANCE FOR FACIAL ACTION RECOGNITION

Cet article [42] a été publié comme l'indique la référence bibliographique

> M. Dahmane et J. Meunier. Individual feature-appearance for facial action recognition. Dans *Proc. of Int. Conference on Image Analysis and Recognition*, pages 233–242, 2011.

Abstract

Automatic facial expression analysis is the most commonly studied aspect of behavior understanding and human-computer interface. Most facial expression recognition systems are implemented with general expression models. However, the same facial expression may vary differently across humans, this can be true even for the same person when the expression is displayed in different contexts. These factors present a significant challenge for recognition. To cope with this problem, we present in this paper a personalized facial action recognition framework that we wish to use in a clinical setting with familiar faces; in this case a high accuracy level is required. The graph fitting method that we are using offers a constrained tracking approach on both shape (using procrustes transformation) and appearance (using weighted Gabor wavelet similarity measure). The tracking process is based on a modified Gabor-phase based disparity estimation technique. Experimental results show that the facial feature points can be tracked with sufficient precision leading to a high facial expression recognition performance.

Introduction

The computer vision community is interested in the development of techniques, such as automated facial expression analysis (AFEA), to figure out the main element of facial human communication, in particular for HCI applications or, with additional complexity, in meeting video analysis, and more recently in clinical research.

AFEA is highly sensitive to face tracking performance, a task which is rendered difficult owing principally to environment changes, and appearance variability under different head orientations, and non–rigidity of the face. To meet these challenges, various techniques have been developed and applied, that we can divide into two main categories: model–based and image–based approaches [57, 139]. However, these methods are still providing inaccurate results due to the variation of facial expression across different people, and even for the same individual since facial expression is context–dependent. Thus, the overall performance of the AFEA systems can severely be affected by their variability across humans and even inter–individual. To establish an accurate *subject-independent* facial expression recognition system, the training data must include a significant number subjects covering all possible individual expression appearances across different people. This is why, in most general systems, accurate recognition is made so difficult.

It is advantageous to identify a subject face before attempting facial expression recognition, since facial physiognomy that characterizes each person leads to a proper facial action display [56]. In [113], for each individual in the datasets a subject-dependent AAM was created for Action Units (AUs) recognition by using similarity normalized shape and similarity normalized appearance. By integrating user identification, the subject–dependent method, proposed in [22], performs better than conventional expression recognition systems, with high recognition rate reaching 98.9%. In [56], the best facial expression recognition results were obtained by fusing facial expression and face identity recognition. In [80], the authors conclude that the recognition rates for familiar faces reached 85%. For unfamiliar faces, the rate does not exceed 65%. This low performance may be justified by the fact that the training collection was less representative.

For both subject–independent and subject–dependant approaches, expression recognition rate is highly dependent on facial tracking performance. Among the existing tracking techniques, the feature–based techniques demonstrate high concurrent validity with manual FACS (Facial Action Coding System) coding [27, 32]. Furthermore, they have some common advantages such as explicit face structure, practical implementation, and collaborative feature–wide error elimination [87]. However, the tracking performance depends on the precise configuration of the local facial features, which poses a significant challenge to automated facial expression analysis, since subtle changes in the facial expression could be missed due to errors in facial point localization [137]. Though an effective scheme for tracking the facial attributes can compensate for this shortcoming, geometric approaches including only shape information may be rather irrelevant [113]. Fig. E.1 shows an example of two different facial expressions (*fear* vs. *happy*) where the respective appearances are significantly different, while the two expressions have a high degree of shape similarity. Therefore, including appearance matching should improve the recognition rate, which can be done by including the local appearance around each facial feature.

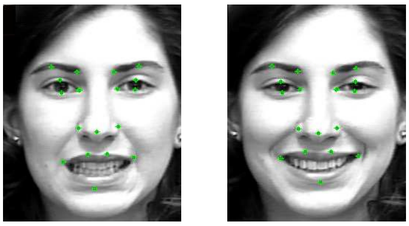


Figure E.1. Facial point position may not be sufficient to achieve reliable FE recognition (e.g. fear vs happy).

In this paper, a subject–dependent facial action recognition system is described using a facial features–based tracking approach and a personalized gallery of facial action graphs. At the first frame of the video, four basic points are automatically found, and then tracked over time. At each frame, from these reference points, the most similar graph through a gallery is chosen. The search is based on the Procrustes transformation and the set of Gabor jets that are stored at each node of the graph. The rest of the paper is organized as follows. First, we present the approach to identify a familiar face, and we outline the facial features we are using. In section (E), we illustrate the facial localization and tracking approach, then we present the facial action recognition process. Experimental results are presented in section (E), while section (F) concludes this work.

Face localization

An individualized facial expression recognition needs a face identify stage, for this purpose a facial graph gallery is constructed as a set of pre–stored facial graphs Fig. E.2. Each graph represents a node configuration that characterizes the appearance of a facial expression to recognize. A set of jets J, describing the appearance around each point of interest, is generated and attached to the corresponding node. In our implementation, a set of 28 nodes defines the facial graph of a given expression Fig. E.2.

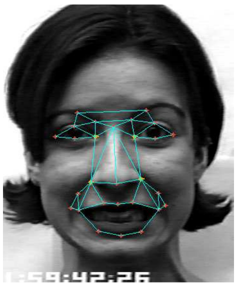


Figure E.2. The facial graph corresponding to the $fear$ **expression.**

Rough face localization

The face localization stage is done only once, in the first frame as an initialization step. For this purpose, we consider a set of subgraphs that includes four facial reference points (*left and right eye inner corners and the two nose wings*) Fig. E.3, and two other points, specifically, the tip and the root of the nose. The two additional features are used only to further enhance the localization stage. If the first frame contains an expressive face, a 6–node facial subgraph that refers to a graph of that specific expression can be used to identify the person.

When the first face image is captured, a pyramidal image representation is created, where the coarsest level is used to find near optimal starting points for the subsequent facial feature localization and refinement stage. Each graph from the gallery is displaced over the coarsest image. The graph position that maximizes the weighted magnitude–based similarity function (E.1 and E.2) provides the best fitting node positions.

$$Sim(\mathrm{I}, \mathrm{G}) = \frac{1}{L} \sum_{l}^{L} S(J_l, J_l') \tag{E.1}$$

$S(J, J')$ refers to the similarity between the jets of the corresponding nodes (E.2), $L = 6$ stands for the total number of nodes.

$$S(J, J') = \sum_j c_j \, \frac{a_j \, a_j'}{\sqrt{\sum {a_j}^2 \sum {a_j'}^2}} \text{with} \qquad c_j = \left(1 - \frac{\left|a_j - a_j'\right|}{a_j + a_j'}\right)^2 \tag{E.2}$$

In (E.2), c_j is used as a weighting factor and a_j is the amplitude of the response of each Gabor filter [38].

Position refinement

The rough localizations of the facial nodes are refined by estimating the displacement using the iterative phase–based disparity estimation procedure. The optimal displacement $\mathbf{d}$ is estimated, an iterative manner, through a minimization of a squared error given two jets J and J' corresponding to the same node (E.3).

$$\mathbf{d}(J,J') = \begin{pmatrix} \sum_j \ c_j \, {k_{jx}}^2 & -\sum_j \ c_j \, k_{jx} \, k_{jy} \\ -\sum_j \ c_j \, k_{jx} \, k_{jy} & \sum_j \ c_j \, {k_{jy}}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_j \ c_j \, k_{jx} \lfloor \Delta\phi_j \rfloor_{2\pi} \\ \sum_j \ c_j \, k_{jy} \lfloor \Delta\phi_j \rfloor_{2\pi} \end{pmatrix} \tag{E.3}$$

where (k_{jx}, k_{jy}) defines the wave vector , and $\lfloor \Delta\phi_j \rfloor_{2\pi}$ denotes the principal part of the phase difference. For more details, the reader is referred to [38]. The procedure is also used to track the positions of the nodes over time. During the refinement stage, the two jets are calculated in the same frame, in this case the disparity represents the amount of position correction. Whereas, during tracking, the two jets are processed from the two consecutive frames, in this case the disparity represents the displacement vector.

Facial feature tracking

Tracking of facial reference points

From frame to frame, the four reference points Fig. E.3 which are known to have relatively stable local appearance and to be less sensitive to facial deformations, are tracked using the phase–based displacement estimation procedure. The new positions are used to deform each graph from the collection, using the Procrustes transformation. In this way, a first search-fit strategy using shape information provides a localized appearance–based feature representation in order to enhance the recognition performance.

Procrustes transform

Procrustes shape analysis is a method in directional statistics [118], used to compare two shape configurations. A two–dimensional shape can be described by a centered

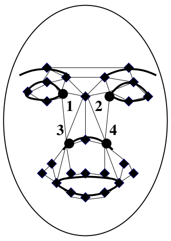


Figure E.3. The four tracked (circle) and the twenty four adjusted (diamond) facial points.

configuration $\mathbf{u}$ in $\mathcal{C}^k$ $(\mathbf{u}\,\mathbf{1}_k = 0)$, where $\mathbf{u}$ is a vector containing `2d` shape landmark points, each represented by a complex number of the form $x + \imath y$.

The procrustes transform is the similarity transform (E.4) that minimizes (E.5), where $\alpha\,\mathbf{1}_k$, $|\beta|$ and $\angle\beta$, respectively, translates, scales and rotates $\mathbf{u}_2$, to match $\mathbf{u}_1$.

$$\begin{cases} \mathbf{u}_1 = \alpha\,\mathbf{1}_k + \beta\,\mathbf{u}_2 & \alpha, \beta \in \mathcal{C} \\ \beta = |\beta|\; e^{\imath\angle\beta} & \end{cases} \tag{E.4}$$

$$\left\| \frac{\mathbf{u}_1}{\|\mathbf{u}_1\|} - \alpha\mathbf{1}_k - \beta\,\frac{\mathbf{u}_2}{\|\mathbf{u}_2\|} \right\|^2 \tag{E.5}$$

First fit-search

The Procrustes transform is used to adequately deform each reference graph G_i stored in the gallery. Given the position of the four tracked reference–point (circle–dots in Fig. E.3), the transformation that "best" wraps the corresponding points in G_i to fit these points, is used to adjust the positions of the twenty–four remaining points of G_i (diamond–dots in Fig. E.3). The new adjusted positions form the distorted graph G_d. "best" refers to a minimal cost that allows to transform the four–node subgraph of G_i to match the corresponding subgraph of G_d.

Facial action recognition

The facial action recognition process is based on evaluating the weighted similarity (E.1) between the reference graph and its distorted version.

The reference graph is the best representative given by the distorted graph with the highest similarity. The node positions give the initial coarse positions of the twenty–eight facial feature points. Refinement stage is performed to obtain the final positions by estimating the optimal displacement of each point by using the Gabor phase–based displacement estimation technique.

Then, at each refined position the attached set of jets is recalculated and updated. The similarity measure between these jets and those of G_i defines the facial expression that closely corresponds to the displayed expression. The scoring value may indicate the intensity of a given expression for a possible graph G_i referring to the peak of a given facial action.

The entire facial action recognition process is summarized in the flow diagram of Fig. E.4.

Results

The videos we used in this work for testing are from the Cohn–Kanade Database [95]. The sequences consist of a set of prototypic facial expressions (happiness, anger, fear, disgust, sadness, and surprise) that were collected from a group of psychology students of different races with ages ranging from 18 to 30 years. Each sequence starts from a neutral expression, and terminates at the peak of a given expression.

First, in order to select the appropriate graphs from the gallery (only one graph per person displaying a given expression at its peak is required), and to properly initialize the positions of the four reference points (the points to track), the subgraph containing the four reference facial features Fig. E.3 is roughly localized in the first frame of the video via an exhaustive search of the subgraph through the coarsest

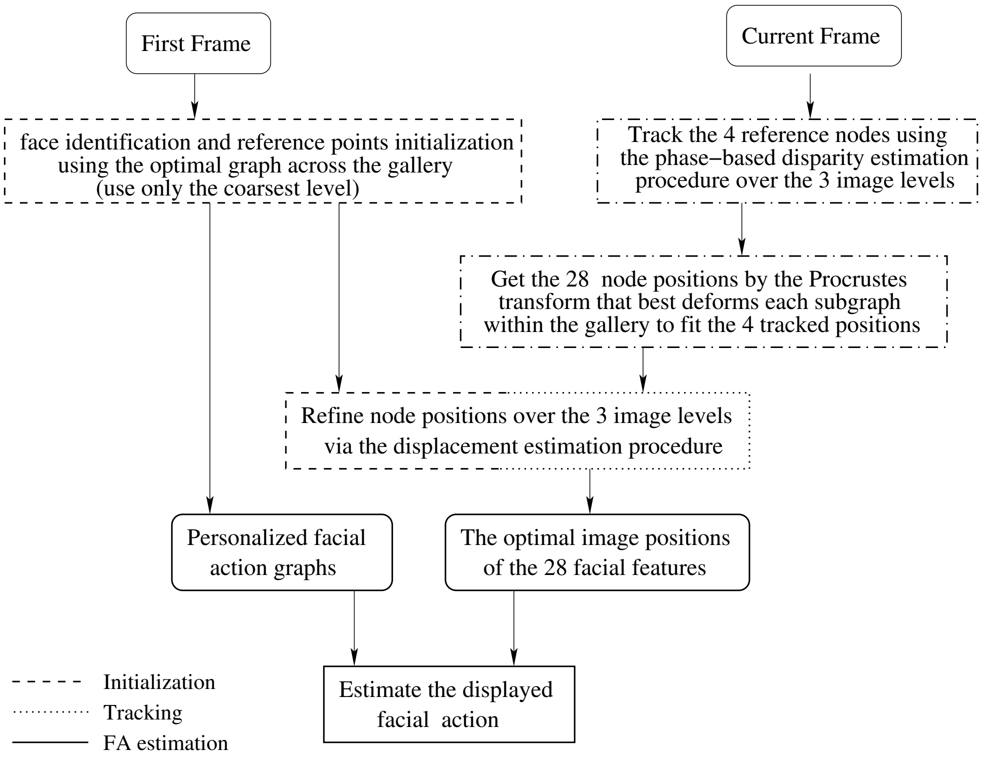

First Frame
Current Frame
face identification and reference points initialization using the optimal graph across the gallery (use only the coarsest level)
Track the 4 reference nodes using the phase−based disparity estimation procedure over the 3 image levels
Get the 28 node positions by the Procrustes transform that best deforms each subgraph within the gallery to fit the 4 tracked positions
Refine node positions over the 3 image levels via the displacement estimation procedure
Personalized facial action graphs
The optimal image positions of the 28 facial features
Estimate the displayed facial action
Initialization
Tracking
FA estimation

Figure E.4. Flow diagram of the entire facial action estimation process.

face image level. We used a three–level hierarchical image representation to decrease the inherent average latency of the graph search operation, by reducing the image search area and the Gabor–jet sizes. For images at the finest level (640 × 480 pixel resolution), jets are defined as sets of 40 complex coefficients constructed from a set of Gabor filters spanning 8 orientations and 5 scales. Whereas those for images at the coarsest level (320 × 240) are formed by 16 coefficients obtained from filters spanning 8 orientations under 2 scales. The intermediate level images use jets of (8 × 3) coefficients.

From frame to frame, only the four reference points are tracked using the iterative disparity estimation procedure. To avoid drifting during tracking, for each feature point, a local fitting is performed by searching through the gallery for the subgraph that maximizes the jet–based similarity. The rough positions of the four feature points are given by the positions of the nodes of the optimal subgraph. These are then adjusted using again the displacement estimation procedure. Fig. E.5 shows snapshots of the tracking of the four reference facial feature points. The bottom subfigure shows, which facial graph (neutral expression or peak happiness) one should be used to refine the positions of the four facial reference points. The position error of the tracked feature points (calculated by erroneous tracked frames / the total frames) was calculated to be 2.4 pixels (measured on level 0).

The twenty–eight facial point positions are obtained by the transformed positions of the reference graph using the Procrustes transformation that wraps the four–point positions to fit the four tracked positions. Then, the iterative displacement estimation procedure is used as a refinement stage, which is performed individually on each position of all of the twenty–eight feature–points, and over the three levels of the pyramid face–image. At each refined position a Gabor–jet is calculated. The reference graph that maximizes the similarity (E.1) over the gallery defines the final facial action that corresponds to the displayed expression.

The two curve profiles in the bottom subgraph of Figure E.5 illustrate that in the first 7 frames the displayed expression is *neutral* with a decreasing intensity, whereas the last 8 ones display a *happiness* expression with a gradually increasing intensity as it is expected. Figure E.6 shows how *fear* and *hapiness* expressions evolve in time with the intensity profile from a *neutral* display to its peak. The facial action recognition process, clearly, differentiates between the two different facial displays, as it is shown by the two respective curve profiles in Figure E.7. The overall performance on the six basic facial expression (*Angry, Fear, Hapiness, Sadness, Surprise, and disgust*) reached 98.7%. The most difficult expression to recognize was the *disgust* expression with a rate of 90.0%. Table E.1 shows the mean score of the correctly recognized expressions over the 81 videos we have used for testing.

Table E.1. The mean score of the correctly recognized expressions.

Expressions	Disgust	Angry	Fear	Hapiness	Sadness	Surprise
scores	0.95	0.97	0.97	0.97	0.96	0.96

Conclusions

Most facial expression recognition systems are based on general model, leading to a poor performance when using familiar human face databases. Within this context, a personalized feature–based facial action recognition approach was introduced and evaluated in this paper. A facial localization step permits to select the most similar graphs from a set of familiar faces. The node positions are used to initialize the positions of the reference points which are tracked using the iterative phase–based disparity estimation procedure. A Procrustes transformation is used to distort each facial action graph according to the positions of the tracked reference points. The facial action recognition process is based on local appearances, provided by the Gabor–jets given by the twenty–eight fiducial points corresponding to the refined node positions of the distorted graph. The adopted facial tracking scheme provides a noticeable

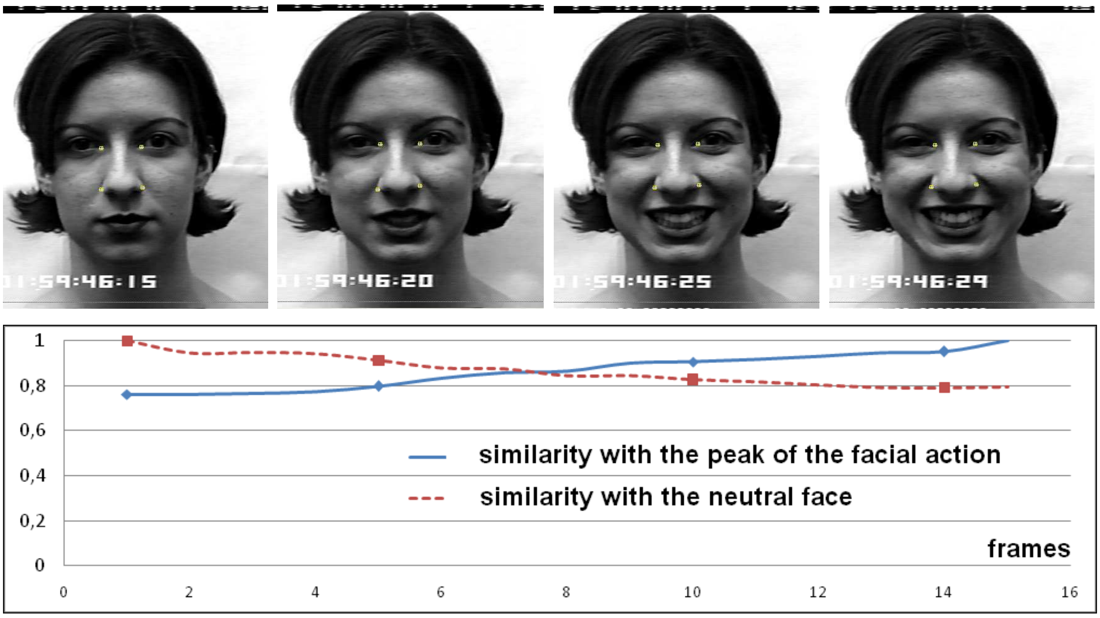


Figure E.5. The tracking performance of the facial reference points. Notice that similarity is computed for each reference graph in the gallery, only two facial graphs (neutral and happiness) are shown for clarity.

performance for the facial action recognition due to the localized appearance–based feature representations. Further work will involve assessing non–verbal communications between a patient and his health care professional in clinical setting, in which the proposed framework can be easily integrated with familiar patient faces to detect some events (eg. subject's smile).

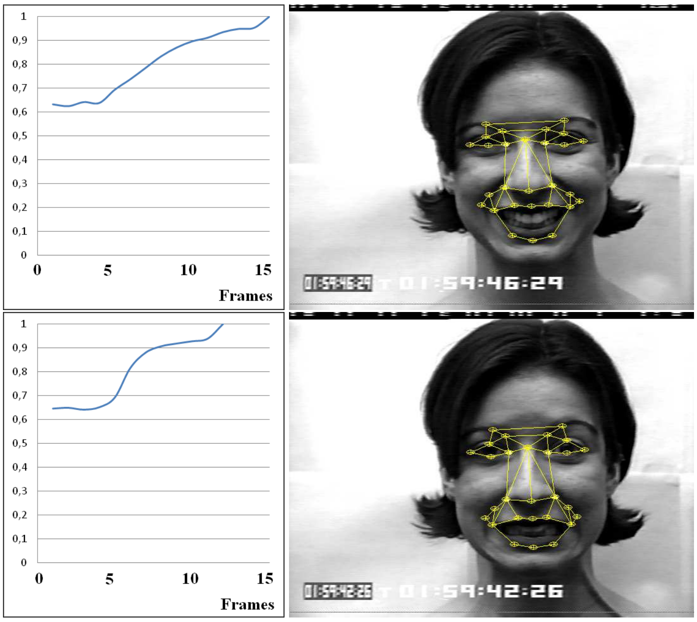


Figure E.6. Facial action (FA) recognition performance. The similarity curve reflects the intensity of the FA (peak $hapiness$**: top – peak** $fear$**: bottom).**

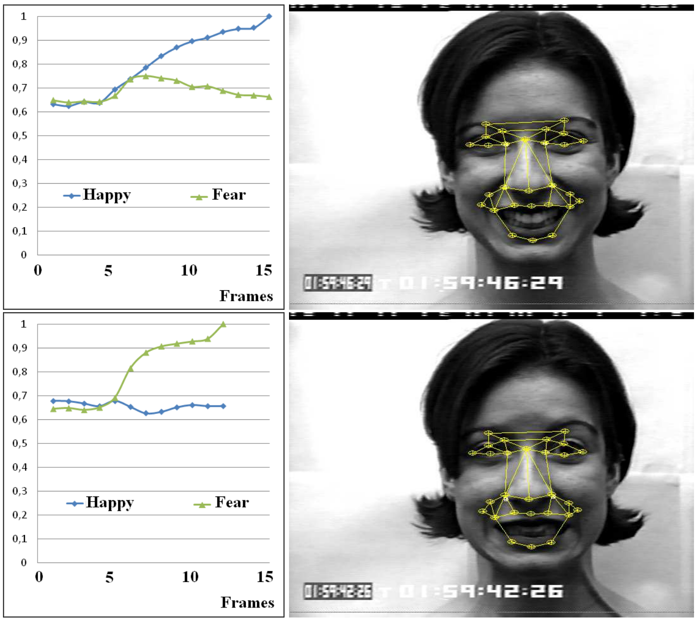


Figure E.7. The mutual exclusion $happiness$ **(top) vs.** $fear$ **(bottom).**

Annexe F

(ARTICLE) EMOTION RECOGNITION USING DYNAMIC GRID–BASED HOG FEATURES

Cet article [41] a été publié comme l'indique la référence bibliographique

> M. Dahmane et J. Meunier. Emotion recognition using dynamic grid-based HoG features. Dans *IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, pages 884–888, 2011.

Abstract

Automatic facial expression analysis is the most commonly studied aspect of behavior understanding and human-computer interface. The main difficulty with facial emotion recognition system is to implement general expression models. The same facial expression may vary differently across humans; this can be true even for the same person when the expression is displayed in different contexts. These factors present a significant challenge for the recognition task. The method we applied, which is reminiscent of the "baseline method", utilizes dynamic dense appearance descriptors and statistical machine learning techniques. Histograms of oriented gradients (HoG) are used to extract the appearance features by accumulating the gradient magnitudes for a set of orientations in 1-D histograms defined over a size-adaptive dense grid, and Support Vector Machines with Radial Basis Function kernels are the base learners of emotions. The overall classification performance of the emotion detection reached 70% which is better than the 56% accuracy achieved by the "baseline method" presented by the challenge organizers.

Introduction

The computer vision community is interested in the development of techniques, such as automated facial expression analysis (AFEA), to figure out the main element of facial human communication, in particular for HCI applications or, with additional complexity, in meeting video analysis, and more recently in clinical research.

AFEA is highly sensitive to face tracking performance, a task which is rendered difficult owing principally to environment changes, and appearance variability under different head orientations, and non–rigidity of the face. To meet these challenges, various techniques have been developed and applied, that we can divide into two main categories: model–based and image–based approaches [57, 139]. These methods are still providing inaccurate results due to the variation of facial expression across different people, and even for the same individual since facial expression is context–dependent.

Among the existing facial expression recognition techniques, the feature–based techniques demonstrate high concurrent validity with manual FACS (Facial Action Coding System) coding [27, 32]. Furthermore, they have some common advantages such as explicit face structure, practical implementation, and collaborative feature–wide error elimination [87]. However, the tracking performance depends on the precise localization of the local facial features, which poses a significant challenge to automated facial expression analysis, since subtle changes in the facial expression could be missed due to the errors in facial point localization [137]. Though an effective scheme for tracking the facial attributes can compensate for this shortcoming, geometric approaches including only shape information may be rather irrelevant [113] by requiring accurate and stable localization of facial landmarks. These drawbacks that can be naturally avoided by the appearance based-methods, which have attracted more and more attention, but in contrast, are in need of a suitable feature set closely relevant to facial expression variations and reliably insensitive to facial variations irrelevant to facial expression. This is precisely the issue that we are trying to address by designing histogram of oriented gradient processed on dynamic dense grids. Experiments are performed to compare our feature set relative to the static Local Binary Pattern (LBP) method [91] descriptors. Indeed, although designed for upper face Action Units detection, the authors have suggested that LBP can be readily used for all other AUs.

The challenge context

Most Facial Expression Recognition and Analysis systems proposed in the literature resemble each other not only by means of processing and ever-present data sparseness, but also by the lack of standardised evaluation procedures. They therefore suffer of low comparability. This is in stark contrast with more established problems in human behaviour analysis from video such as face detection and face recognition. Yet at the same time, this is a rapidly growing field of research, due to the constantly increasing interest in applications for human behaviour analysis, and technologies for human-machine communication and multimedia retrieval.

In these respects, the FG 2011 Facial Expression Recognition and Analysis challenge (FERA2011) shall help bridging the gap between excellent research on facial expression recognition and low comparability of results, for two selected tasks: the detection of FACS Action Units and the detection of a set of emotional categories. This paper focuses on the second task (sub-challenge).

Overview of the data

The GEMEP-FERA dataset consists of recordings of 10 actors displaying a range of expressions, while uttering a meaningless phrase, or the word “Aaah”. There are 7 subjects in the training data, and 6 subjects in the test set, 3 of which are not present in the training set.

Because of the nature of the emotion categories in this challenge, it is not possible to use other training data for the emotion recognition sub-challenge.

The baseline method

The "baseline method" applied by the challenge organizers utilizes static dense appearance descriptors and statistical machine learning techniques. They used Uniform Local Binary Patterns (Uniform LBP) with 8 neighbours and radius 1 to extract the appearance features, principal component analysis (PCA) to reduce the dimensionality of the descriptor, and Support Vector Machines with Radial Basis Function kernels to classify the data. The method used is generally that described for the static LBP method [91].

Data was pre-processed as follows. They used the OpenCV face detector to extract the face location in each frame. The detected face was scaled to be 200 by 200 pixels. Then, they applied the OpenCV implementation of eye-detection, and used the detected eye locations to remove any in-plane rotation of the face. They also translated the face so that the subject's right eye centre is always at the coordinates $x = 60$, $y = 60$. They did not use the detected eye locations to normalise for scale, as the OpenCV eye detection is too inaccurate for this. For training, the pre-processed images were manually verified. Incorrectly pre-processed images were removed from the training set but were not replaced. For the test set, no manual verification was done.

Emotion recognition sub-challenge

To extract features for the emotion recognition sub-challenge, the authors used the entire face. The face area was divided into 100 squares: 10 rows and 10 columns with a side of 20 pixels. Uniform LBP was applied, and the histograms of all 100 blocks were concatenated into a single 5900 dimensional feature vector.

As the videos do not have a clear neutral element, all frames in a video of a certain emotion are assumed to depict that emotion, and thus all frames were used to train the classifiers for emotions. They trained five one-versus-all binary classifiers. During testing, every frame from the test video was passed to the five classifiers, and the emotion belonging to the classifier with the highest decision function value output was assigned to that frame. To decide which emotion label should be assigned to the entire test video; they applied majority voting over all frames in the video. In case of a tie, the emotion that occured first in an alphabetically sorted list was chosen (i.e. if the emotions "anger" and "fear" would tie for having the highest amount of detected frames, then "anger" would be chosen as it occurs before "fear" in alphabetical order). PCA was applied to the training set, retaining 95% of the variance in the reduced set. This was used to train one support vector machine (SVM) for every emotion, using radial basis function (RBF) kernels. The optimal values for the kernel parameter and the slack variable were found using 5-fold subject-independent cross-validation on the training set. After the optimal parameter values were found, the classifiers were trained on the entire data.

If during pre-processing the face detector failed, they decided that that frame had no emotion in it. If the eye-detection was off, they simply use the detection that the classifiers provided given the mis-aligned face.

The classification rate for every emotion is given in Table F.1. The confusion matrix was not provided to the challenge participants by the organizers, as it would reveal too many details about the test set.

Our method

The method we applied, which is reminiscent of the "baseline method" that uses a static Uniform LBP, utilizes a dynamic dense area-based appearance descriptors and statistical machine learning techniques. We applied histograms of oriented gradients

Table F.1. Independent/specific partition classification rate for every emotion ("baseline method").

	Person Independent	Person specific	Overall
anger	0.86	0.92	0.89
fear	0.07	0.40	0.20
joy	0.70	0.73	0.71
relief	0.31	0.70	0.46
sadness	0.27	0.90	0.52
avg.	0.44	0.73	**0.56**

(HoG) to extract the appearance features by accumulating the gradient magnitudes for a set of orientations in 1-D histograms defined over a size-adaptive dense grid, and Support Vector Machines with Radial Basis Function kernels as base learners of emotions.

Histogram of Oriented Gradients

Histogram of oriented gradient (HoG) feature descriptors were proposed by Dalal and Triggs [44]. This method that was originally developed for person detection is used in more general object detection algorithms. The main idea behind the HoG descriptors is based on edge information. That is each window region can be described by the local distribution of the edge orientations and the corresponding gradient magnitude. The local distribution is described by the local histogram of oriented gradients which is formed by dividing the detection window into a set of small regions called cells, where the magnitude of the edge gradient for each orientation bin is computed for all pixels. To provide better invariance, the local HoG is normalized over a block of neighboring cells.

Dalal and Triggs [44] have shown that HoG outperforms wavelet, PCA-SIFT and Shape Context approaches.

Different implementations of HoG were considered to study the influence of different descriptor parameters. For good performance, the authors highlighted the importance of fine scale gradients, fine orientation binning, relatively coarse spatial binning, and high-quality local contrast normalization in overlapping descriptor blocks.

The default typical parameters include $[-1, 0, 1]$ gradient filter with no smoothing, linear gradient voting into 9 orientation bins in $0° - 180°$, 16×16 blocks of four overlapping 8×8 pixel cells, $L_2 - norm$ block normalization, block spacing stride of 8 pixels, and 64×128 detection window. In our application, we used 8 rows by 6 columns cells of $a \times a$ pixels, with a cell spacing stride of $a/2$ pixels (see sec. F).

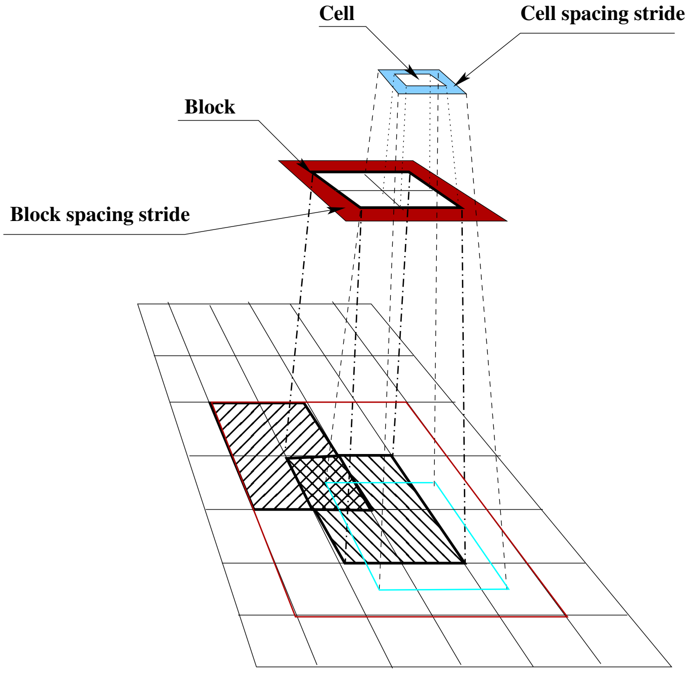


Figure F.1. HoG feature levels.

Data Preprocessing

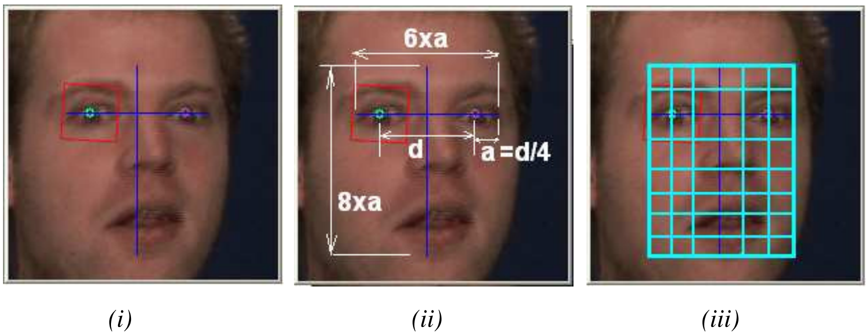
Data were pre-processed as follows. We used the OpenCV face detector to extract the face location in each frame on which a contrast amelioration and brightness normalization have been performed. The detected face was scaled to be 200 by 200 pixels. We then applied the OpenCV implementation of eye-detection, and use the detected eye locations to, firstly, remove any in-plane rotation of the face, and, to crop the region of interest. In contrast to the “baseline method”, we did not need to translate the face so that the subject’s right eye centre be always at given coordinates. We did not use the detected eye locations to normalise for scale, as the OpenCV eye detection is too inaccurate for this. For training, the pre-processed images were manually verified. Incorrectly pre-processed images were removed from the training set but were not replaced. In the absence of labeled video (not furnished by the organizers) of the test set, no manual verification was done.

Specific details

To extract features for the emotion recognition sub-challenge, we used a cropped region from the aligned face, the region of interest is obtained from physiognomic facial measurements as shown in Fig. F.2. The aligned face area is divided into (48) squares: 8 rows and 6 columns with a variable side of a pixels. The adaptive grid-size depend on the distance d between the two eyes as $a = d/4$.

For each cell a local histogram is processed (Fig. F.1). Then for each block of 2×2 cells, the 4 local histograms are concatenated and the resulting histogram is normalized using the $L_2 - norm$. The twelve normalized histograms are then concatenated into a global histogram giving a single 432 dimensional feature vector. We recal that the dimension of the feature vector generated by the Uniform LBP used in the “baseline method” (Sec. F) has a much higher dimension of 5900 (i.e. 13.66 times higher). The “baseline method” uses PCA to reduce the training set, but no information was given by the organizers about the final feature set dimensionality.

The experimental protocol of the “baseline method” was followed to decide which emotion label should be assigned to the entire test video, we applied majority voting over all frames in the video. In case of a tie, the emotion that occured first in an



(i) *(ii)* *(iii)*

Figure F.2. Eye distance adaptive window–based HoG : (i). **In-plane rotated face,** (ii). **The baseline distance** a **extracted from physiognomic measurements of the rotated face,** (iii). **The cropped ROI based on** a**.**

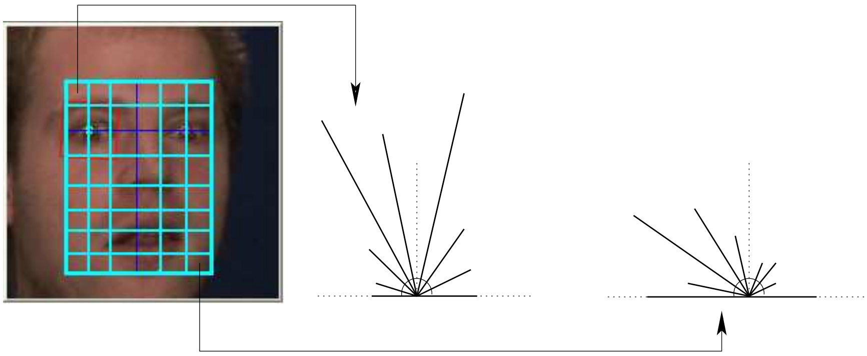


Figure F.3. A Global HoG concatenating the local HoGs extracted from each cell.

alphabetically sorted list was chosen. As base learner classifier, we use a multiclass radial-basis-function kernel Support Vector Machines which are known as stronger classifiers for high dimensionality problems, and powerful algorithms for solving difficult classification problems [34, 180]. We have used the Intel OpenCV Library implementation.

All frames are used to train the multiclass RBF-SVM. If during pre-processing the face detector failed, we decided that the frame had no emotion in it. If the eye-detection was off, we simply discarded the frame in the case of training, in the testing period the emotion was classified as “unknown”. This mis-classification situation occured twice during the testing phase.

The optimal values for the kernel (F.1) parameter γ and the penalty multiplier parameter C for outliers were found using a certain number of training epochs.

$$K(\mathrm{x}, \mathrm{x}_i^*) = \exp\left(-\gamma \|\mathrm{x} - \mathrm{x}_i^*\|^2\right) \tag{F.1}$$

If we put $\sigma = (1/\sqrt{2 * \pi})(V_{max} - V_{min})$ with V_{max} (resp. V_{min}) the maximum (resp. minimum) value of the feature vector elements over the training dataset, then γ_e will be given by

$$\gamma_e = \frac{1}{(e * \sigma^2)} \tag{F.2}$$

and the multiplier C by

$$C_e = e * \sigma^{1.75} \tag{F.3}$$

The tuple (γ_e, C_e) corresponding to the epoch e with highest training accuracy, and a reasonable number of Support Vectors (SV) relatively to the training set, was selected. Generally, the stopping criterion is met between $e = 10$ and $e = 15$ epochs (Fig. F.4 and Fig. F.5). Equations F.2 and F.3 were derived emperically.

The confusion matrices are presented in Tables F.2, F.3, and F.4, respectively, for person independent, person specific, and both. Table F.5 shows the indepen-

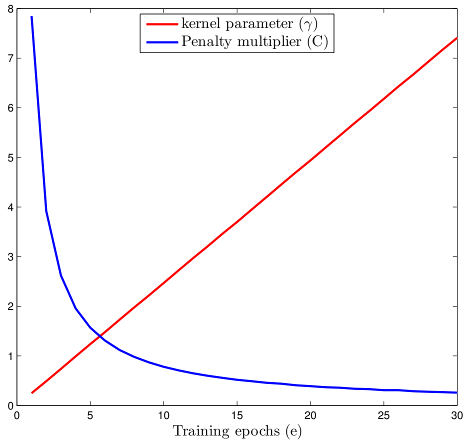


Figure F.4. The SVM parameters evolution over training epochs.

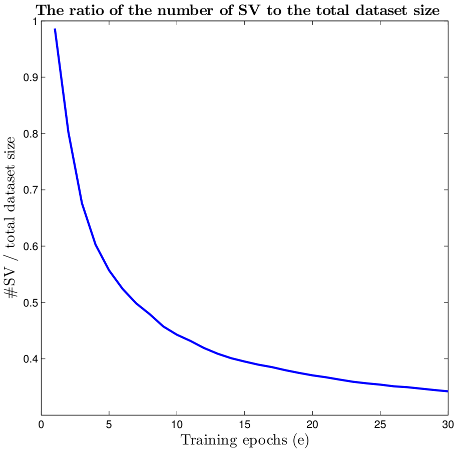


Figure F.5. Stabilization of the number of support vectors from e=25.

dent/specific partition classification rate for every emotion. These scores were computed by the organizers from the results that we emailed to them.

Table F.2. Confusion matrix, person independent (our method).

pred/truth	Anger	Fear	Joy	Relief	Sadness
Anger	13	10	1	3	7
Fear	0	1	0	0	0
Joy	1	4	19	1	1
Relief	0	0	0	12	4
Sadness	0	0	0	0	3

Table F.3. Confusion matrix, person specific (our method).

pred/truth	Anger	Fear	Joy	Relief	Sadness
Anger	12	1	0	1	0
Fear	1	9	0	1	0
Joy	0	0	8	0	0
Relief	0	0	2	8	0
Sadness	0	0	1	0	10

Table F.4. Confusion matrix, overall (our method).

pred/truth	Anger	Fear	Joy	Relief	Sadness
Anger	25	11	1	4	7
Fear	1	10	0	1	0
Joy	1	4	27	1	1
Relief	0	0	2	20	4
Sadness	0	0	1	0	13

Clearly, from the score tables we can see that our results are in agreement with the “baseline method”, but are always more accurate. For instance, “fear” was the most difficult emotion to classify for both methods, and “anger” the easiest, but in both cases we performed better. The overall classification performance of the proposed emotion detection system reached 70% (Table F.5) which is better than the 56% (Table F.1) accuracy achieved by the “baseline method”.

Table F.5. Independent/specific partition classification rate for every emotion (our method).

	Person Independent	Person specific	Overall
anger	0.93	0.92	0.93
fear	0.07	0.90	0.40
joy	0.95	0.73	0.87
relief	0.75	0.80	0.77
sadness	0.20	1.00	0.52
avg.	0.58	0.87	**0.70**

Conclusions

In this article, we elaborated an emotion recognition framework using dynamic dense grid-based HoG features. We showed that the proposed method performs better than static Uniform LBP implementation, used in the "baseline method" offered by the challenge organizers. The overall classification performance of emotion detection reached 70% with the HoG feature set, and 56% with the LBP feature set. This performance is especially due to the HoG global characteristics that are based on orientation binning of the local edges and the corresponding gradient magnitude. This is also due to the relative alignment and dynamic scaling of the grid, whereas, even with a resized face image, the "baseline method" uses an absolute unscaled alignment that translates the position of the center of the right eye to a fixed coordinates.

Besides, with respect to the sensitivity of the recognition algorithm to the critical parameters of the multi-class SVM, instead of using the classical n-fold cross-validation, we have proposed an innovative simple epoch-based strategy to determine the optimal values for the kernel parameter and the penalty multiplier parameter.

Finally, with integral images [184], HoG features can be completely computed in real time.

Annexe G

(ARTICLE) CONTINUOUS EMOTION RECOGNITION USING GABOR ENERGY FILTERS

Cet article [40] a été publié comme l'indique la référence bibliographique

> M. Dahmane et J. Meunier. Continuous Emotion Recognition using Gabor Energy Filters. Dans *Proceedings of the 4th international conference on Affective computing and intelligent interaction - Volume Part II*, ACII'11, pages 351–358, Berlin, Heidelberg, 2011. Springer-Verlag.

Abstract

Automatic facial expression analysis systems try to build a mapping between the continuous emotion space and a set of discrete expression categories (e.g. happiness, sadness). In this paper, we present a method to recognize emotions in terms of latent dimensions (e.g. arousal, valence, power). The method we applied uses Gabor energy texture descriptors to model the facial appearance deformations, and a multiclass SVM as base learner of emotions. To deal with more naturalistic behavior, the SEMAINE database of naturalistic dialogues was used.

Introduction

Within the affective computing research field, researchers are still facing a big challenge to establish automated systems to recognize human emotions from video sequences, since human affective behavior is subtle and multimodal. Recently, efforts have been oriented on how to modelize affective emotions in terms of continuous dimensions (e.g. activation, expectation power, and valence), rather than resolving the most known classification problems (e.g. happiness, sadness, surprise, disgust, fear, anger, and neutral). Prior works on human facial emotion recognition have focused on both images and video sequences. Different approaches were investigated including feature-based and appearance-based techniques [57, 139]. Most of these techniques use databases that were collected under non realistic conditions [194]. An advance emotion recognition system needs to deal with more natural behaviors in large volumes of un-segmented, un-prototypical, and natural data [156]. The challenge data were collected from the SEMAINE database [122] which consists of natural dialogue video sequences that are more challenging and more realistic.

Overview of the database

The dataset is based on the first part of the SEMAINE database called the solid-SAL part. This part consists of 24 recordings, which were splitted into three partitions: training (31 sessions: 501277 frames), development (32 sessions: 449074 frames), and test partition (32 sessions: 407772 frames). Each partition consists of 8 recordings. Videos were recorded at 50 f/s and at a resolution of 780 × 580 pixels.

Overview of the baseline method

The challenge organizers method for the video feature sub-challenge is based on dense appearance descriptors. They used Uniform Local Binary Pattern (Uniform LBP). For each frame a preprocessing stage permits to extract the face region using the OpenCV face detector. Then a registration stage is performed by finding the two eye positions which were used, first, to remove any in-plane rotation of the face, and second the distance between these two locations was used to normalize for scale. The face was translated so that the subject's right eye centre is always at the coordinates $x = 80$, $y = 60$. The registered image was cropped to 200 × 200 pixels face region.

Challenge baselines

For the emotion analysis problem, four classification sub-problems need to be solved: the originally continuous dimensions: *arousal*, *expectancy*, *power*, and *valence*, which were redefined as binary classification tasks by checking at every frame whether they were above or below average.

The cropped face area was divided into 100 squares with side of 20 pixels. Uniform LBP was applied and histograms of all 100 blocks were concatenated into a relatively high (5900) dimensional feature vector. Besides, as the video sequences were very large, the authors chose to sample periodically the data by selecting 1000 frames from the training partition and 1000 frames from the development partition. The extracted frame descriptors were used to train a Support Vector Machine (SVM) using Radial Basis Function (RBF) kernels.

No further details were given about possible preprocessing failure, for instance, if the face detector failed or if the eye detection was off.

The classification rate for every dimension is given in Table G.1. It shows the accuracy for training on the training partition and testing on the development partition, as well as for training on the unification of the training and the development partitions and testing on the test partition sub-set.

Table G.1. Baseline results (WA stands for weighted accuracy, and UA for unweighted).

Accuracy	Activation		Expectation		Power		Valence	
	WA	UA	WA	UA	WA	UA	WA	UA
Devel	60.2	57.9	58.3	56.7	56.0	52.8	63.6	60.9
Test	**42.2**	**52.5**	**53.6**	**49.3**	**36.4**	**37.0**	**52.5**	**51.2**

Technical approach

Our method uses dense facial appearance descriptors. We applied Gabor energy to extract the facial appearance features by calculating the responses to a set of filters. A multiclass Support Vector Machine with polynomial kernels was used as base learner of the affective dimensions.

Gabor filtering

In the literature, one can find several attempts at designing feature–based methods using Gabor wavelets [50, 102, 107, 115, 158, 168, 178, 194], due to their interesting and desirable properties including spatial locality, self similar hierarchical representation, optimal joint uncertainty in space and frequency as well as biological plausibility [64].

Gabor filtering permits to describe via image convolution, with Gabor wavelet (Eq. G.1), the spatial frequency structure around the pixel $\mathbf{x}$.

$$\Psi_{\lambda,\theta,\varphi,\sigma,\gamma}(\mathbf{x}) = \exp\left(-\frac{x'^2 + \gamma^2\, y'^2}{2\sigma^2}\right)\cos\left(2\pi\,\frac{x'}{\lambda} + \varphi\right) \tag{G.1}$$

with

$$\begin{aligned} x' &= x\,\cos\theta + y\,\sin\theta \\ y' &= -x\,\sin\theta + y\,\cos\theta \end{aligned}$$

and where $\lambda, \theta, \varphi, \sigma$, and γ stand, respectively, for wavelength, orientation, phase offset, aspect ratio, and the bandwidth.

Gabor energy filter

Relative to the simple linear Gabor filtering, the energy filter gives a smoother response to an edge or a line of appropriate width with a local maximum exactly at the edge or in the center of the line [76, 142] (see figure G.1). The energy filter response

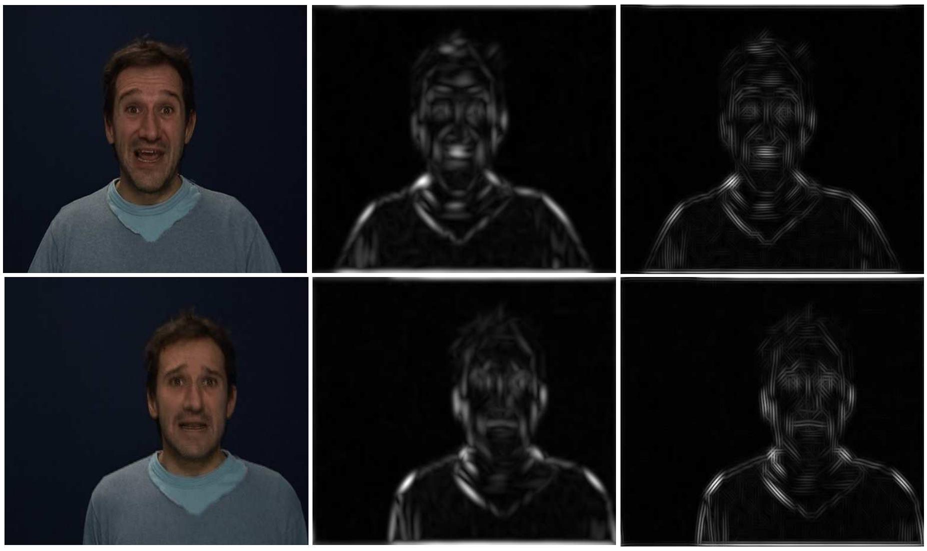


Figure G.1. Left column: facial images with a given emotion. Middle : Energy filter output. Right: Standard Gabor filter output.

is obtained by combining the convolutions obtained from two different phase offsets ($\varphi_0 = 0$) and ($\varphi_1 = \pi/2$) using the L_2-*norm*. Figure G.1 shows the outputs of the Gabor energy filter vs. the standard Gabor filter applied on facial images with certain emotions, using 5 different orientations ($\theta \in [0 : \pi/4 : \pi]$) with $\lambda = 8$, $\varphi \in \{0, \pi/2\}$, $\gamma = 0.5$, and $b = 1$.

Data processing

First, we used the OpenCV face detection to extract the face region in each frame, which is scaled to 200 by 200 pixels. We did not use eye positions to normalize for scale, as the OpenCV eye detector is too inaccurate to do this, particularly for oriented head or closed eyes. The Gabor energy filter output of the cropped face region was divided into 10×10 cells then, 25 blocks were defined by gathering 2×2 neighboring cells . For each cell, a local histogram was processed in which each bin represented a frequency at a selected orientation. The parameters were set as follows $\lambda \in \{5, 10, 15\}$ and $\theta \in [0 : \pi/8 : 7\pi/8]$. So each histogram had 24 bins (3 frequencies×8 orientations). After that, for each block, the 4 local histograms were concatenated to generate a block-histogram that was locally normalized using the L_2-*norm*. Thus the local values were not affected by the extreme values in other blocks. The 25 block-histograms were then concatenated into a region-histogram giving a single 2400 dimensional feature vector which remained relatively small compared to the baseline feature vector.

The Experimental protocols

- The protocol to decide which affective dimension label should be assigned to every frame consisted of a majority voting. In case of a tie, the class that occurred first in the alphabetical order was chosen. This does not introduce a bias in the results; since in this case by default, SVM assigns labels considering the class label order.

- Because of the large amount of data (over 1.3 million frames) and relatively high feature dimensionality (2400 features per frame) we decided to sample frames from the videos sequences at interval of 10 frames. However, if a new class label appeared in the video, we decreased the sampling rate to 4 frames, after a while the rate was again increased to 10 frames. Thus, we reduced the risk of missing transitions in the affective continuous dimensions over the stream.

- If during preprocessing, the face detector failed, we decided in the case of training to simply discard the frame. Whereas, for the testing stage, the code label of the preceding frame was used to label the current frame.

- For practical considerations, we used a 4-bit binary code for class labeling. A '1' in the leftmost bit position indicated that the affective element "*valence*" was above average.

The SVM kernel that we used in this work, is given by equation (G.2). The penalty parameter for outliers was fixed to $C = 10$.

$$K(\mathrm{x}, \mathrm{x}_i^*) = (8.\ (\mathrm{x} \cdot \mathrm{x}_i^*) + 1.)^{10.} \tag{G.2}$$

The classification accuracies of the video sub-challenge over the different affective dimensions are shown in Table G.2. The τ values indicate the different thresholds we used to translate the SVM hyperplane. Two tests were performed, on two different partitions: development and test subsets. For the test set, the scores were computed by the organizers from the results that we emailed to them.

Table G.2. The results of our proposed method (WA stands for weighted accuracy, and UA for unweighted).

Accuracy(%)		Activation		Expectation		Power		Valence	
		WA	UA	WA	UA	WA	UA	WA	UA
Devel ($\tau = 0.0$)		54.9	55.0	51.8	51.2	53.2	52.8	56.6	55.5
Test	$\tau = 0.0$	**63.4**	**63.7**	**35.9**	**36.6**	**41.4**	**41.1**	**53.4**	**53.6**
	$\tau = -.2$	**58.0**	**58.4**	**41.0**	**41.0**	**50.5**	**49.7**	**48.6**	**50.5**
	$\tau = -.3$	**55.1**	**55.7**	**46.7**	**43.0**	**50.4**	**49.5**	**48.6**	**51.0**
	$\tau = -.4$	**53.2**	**53.9**	**54.8**	**46.6**	**50.7**	**49.7**	**47.7**	**50.4**

Table (G.2), shows that, on the development partition, all the different affective dimensions were recognized at approximately the same rates (Fig. G.2). All obtained rates are better than chance but somewhat lower than the ones obtained by the baseline method probably due the RBF kernel that this method used.

Figure (G.3) shows that our method performs better than the baseline method for the *arousal*, *expectation*, and *power* classes. The three dimension accuracies are above 50%. *Power* was the most difficult affective class to recognize for the baseline method with only 36%. The lowest score we obtained was for *valence* with 48%.

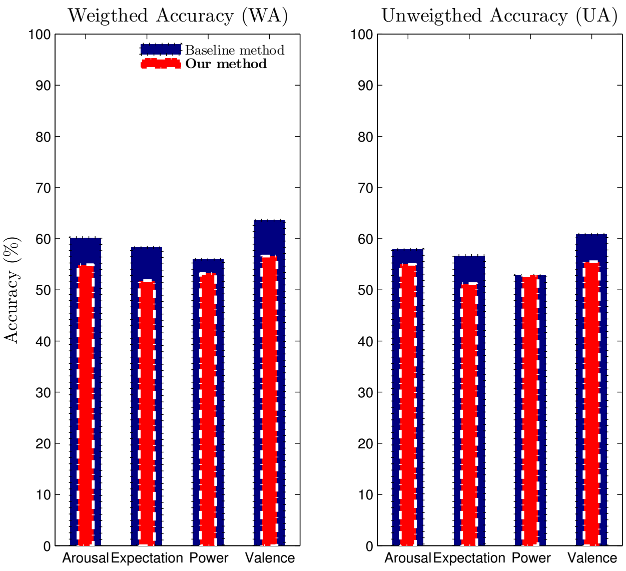


Figure G.2. Comparison of results: Test on the development partition.

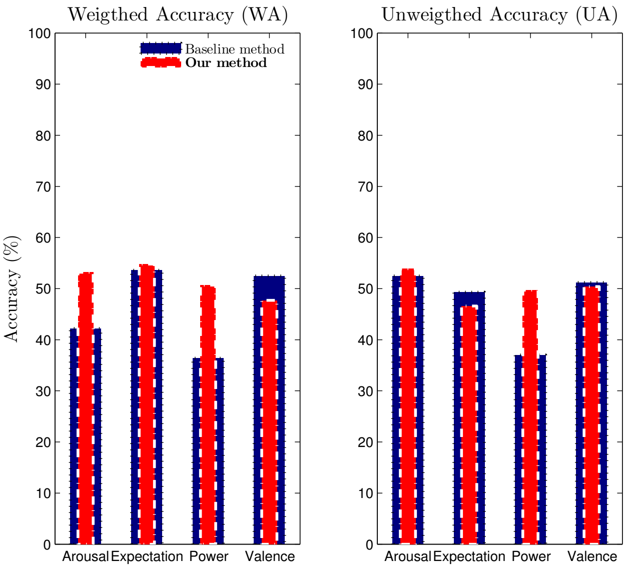


Figure G.3. Comparison of results: Test on the testing partition.

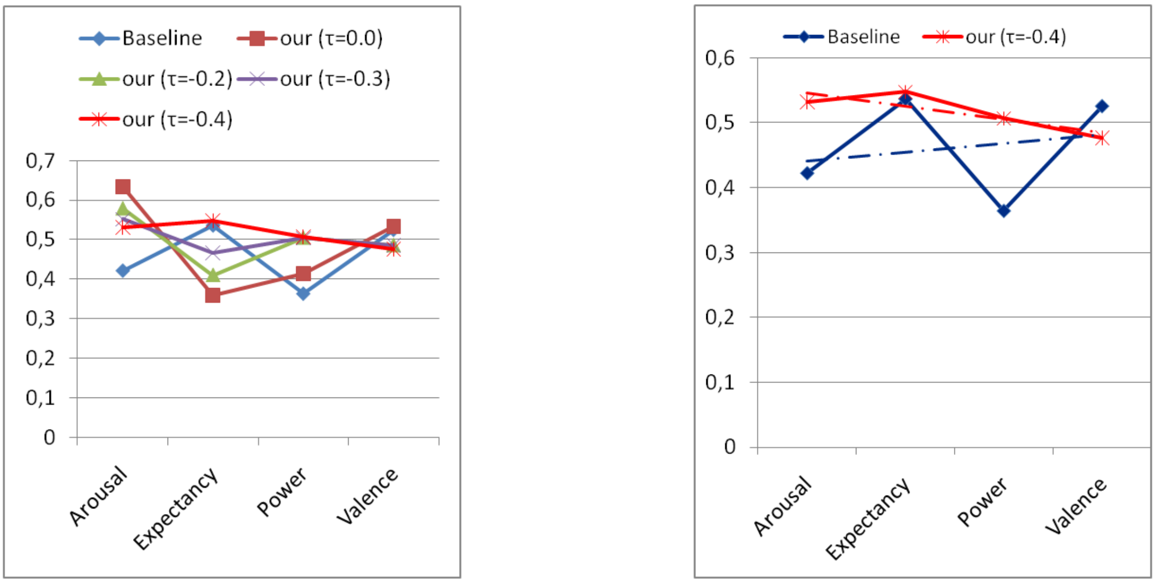


Figure G.4. Comparison of results: Test on the test partition at different SVM hyperplanes

The overall classification performance of the proposed approach reached 51.6% with almost equally distributed classification errors with a mean absolute deviation of 1 %. That is better than the 46.2% achieved by the baseline method with a higher MAD of 7 % (Fig. G.4).

Conclusions

In this paper, we presented a method to recognize different emotions explained in terms of latent dimensions. We considered Gabor energy filters for image representation. For each face region, we used a uniform grid to integrate the filter response in one histogram by cell. The concatenated histogram represented the emotion descriptor. We used one multiclass polynomial-SVM to classify the extracted emotion descriptors. The experiment evaluation on both development and test partitions, showed promising results on the challenging AVEC 2011 dataset. The overall classification performance of the affective dimensions classification reached 51.6% with the Gabor energy filters, and 46.1% with the LBP feature set. In future, we intend to study the performance of radial-basis-function kernel SVM type, where the optimal parameters will be fixed by using what we called "epoch" based strategy [41]. Also, we will investigate the use of a threshold adjusting procedure that relaxes the SVM threshold from zero.