Combinaison de Descripteurs Hétérogènes pour la Reconnaissance de Micro-Mouvements Faciaux

Vincent Rapp, Thibaud Sénéchal, Lionel Prevost, Kevin Bailly, Hanan Salam, Renaud Seguier

Combinaison de Descripteurs Hétérogènes pour la Reconnaissance de Micro-Mouvements Faciaux.

Vincent Rapp[1], Thibaud Senechal[1], Hanan Salam[2], Lionel Prevost[3], Renaud Seguier[2], Kevin Bailly[1]

[1]ISIR - CNRS UMR 7222
Université Pierre et Marie Curie, Paris
{rapp, senechal, bailly}@isir.upmc.fr

[2]Supelec - ETR (UMR 6164)
Avenue de la Boulaie, 35511,
Cesson-Sevigne
{salam, seguier}@supelec.fr

[3]LAMIA - EA 4540
Université des Antilles et de la Guyanne
lionel.prevost@univ-ag.fr

Résumé

Dans cet article, nous présentons notre réponse au premier challenge international sur la reconnaissance et l'analyse d'émotions faciales (Facial Emotion Recognition and Analysis Challenge). Nous proposons une combinaison de différents types de descripteurs dans le but de détecter de manière automatique, les micro-mouvements faciaux d'un visage. Ce système utilise une Machine à Vecteurs Supports Multi-Noyaux pour chacune des Action Units (AU) que nous désirons détecter. Le premier noyau est calculé en utilisant des histogrammes de motifs binaires locaux de Gabor (ou Local Gabor Binary Pattern, LGBP) via un noyau d'intersection d'histogramme. Le second noyau quant à lui, est crée avec des coefficients de Modèles Actifs d'Apparence via un noyau gaussien. Les sorties de chacune des SVM sont ensuite filtrées dans le but d'inclure l'information temporelle de la séquence. Afin d'évaluer notre système, nous avons procédé à de nombreuses expérimentations sur plusieurs points clefs de notre méthode. Enfin, nous comparons nos résultats à ceux obtenus par les autres participants au challenge, tout en analysant nos performances.

Abstract

This paper presents our response to the first international challenge on Facial Emotion Recognition and Analysis. We propose to combine different types of features to automatically detect Action Units in facial images. We use one multi-kernel SVM for each Action Unit we want to detect. The first kernel matrix is computed using Local Gabor Binary Pattern histograms and a histogram intersection kernel. The second kernel matrix is computed from AAM coefficients and an RBF kernel. During the training step, we combine these two types of features using the recent SimpleMKL algorithm. SVM outputs are then filtered to exploit temporal information in the sequence. To evaluate our system, we perform deep experimentations on several key issues : influence of features and kernel function in histogram-based SVM approaches, influence of spatially-independent information versus geometric local appearance information and benefits of combining both, sensitivity to training data and interest of temporal context adaptation. We also compare our results to those of the other challengers and try to explain why our method had the best performance during the FERA challenge.

1 Introduction

Un des principaux objectifs de la création de nouveaux environnements informatisés est de placer l'utilisateur au cœur de ce système. Pour ce faire, le système doit être capable d'interagir avec l'utilisateur de façon naturelle. Les interfaces homme-machine actuelles ignorent généralement ce besoin, et l'état affectif de l'utilisateur est le plus souvent inconnu, impliquant une perte d'information importante. Pour reconnaitre l'état affectif d'une personne, le système doit être capable d'analyser les signaux non-verbaux comme le ton de la voix, les gestes ou encore les mouvements faciaux. Parmi tout ces signaux, l'expression faciale est la manière la plus naturelle de communiquer ou de transmettre une émotion à une tierce personne. Ceci explique pourquoi la reconnaissance des expressions faciales est un sujet de recherche très actif. La taxinomie des émotions à été définie par les psychologues comme un ensemble de six émotions basiques universelles (colère, dégout, peur, joie, tristesse et surprise [1]. Afin de standardiser la détection et la reconnaissance des émotions, un codage (FACS pour Facial Action Coding System) basé sur les micro-mouvements des muscles faciaux (Action Units ou AU) à été créé [2]. En combinant différentes AU, il est alors possible de reproduire chacune des six émotions basiques.

Dans la littérature, plusieurs systèmes ont déjà été pro-

posés, mais l'absence d'évaluation commune ne permet pas une réelle comparaison de ces méthodes. Le challenge "Facial Expression and Analysis" (FERA, http ://sspnet.eu/fera2011/), organisé en marge de la conférence "IEEE Face and Gesture Recognition 2011" , se propose de combler ce manque via une procédure d'évaluation commune sur la base de données GEMEP [3]. Ce challenge est divisé en deux sous-challenge représentant les principales orientations de l'analyse de l'expression faciale : détection de l'émotion et détection des AU.

Le sous-challenge dédié à la détection d'AU propose aux chercheurs de réaliser des systèmes atteignant la plus haute F1-mesure possible (voir section 6.1) , et ce pour les 12 AU les plus fréquentes. Cet article se concentre sur cette tâche. Pour ce faire, 158 séquences (87 pour l'apprentissage et 71 pour le test) de la base GEMEP furent sélectionnées. Ces séquences montrent dix personnes affichant une émotion tout en prononçant une phrase qui n'a aucun sens, impliquant une détection durant une intervention orale. Il y a 7 sujets dans la partition d'apprentissage et 6 sujets dans la partition de test, dont 3 étant absents de la partition d'apprentissage. Les données d'apprentissage étiquetées furent distribuées via le site internet des organisateurs. Concernant les données de test, elles ne furent quant à elles distribuées que sept jours avant la date limite de soumission. Afin d'obtenir les résultats sur la base de test, dont nous ne connaissions pas les étiquettes, les participants étaient invités à envoyer par email leurs prédictions obtenues grâce à leurs systèmes aux organisateurs, ces derniers se chargeant du calcul de la F1-mesure pour la détection d'AU.

L'analyse de l'existant (section 2) a montré que le fait d'utiliser plusieurs descripteurs de types différents améliorait les performances en détection. Nous proposons donc ici une combinaison originale de deux descripteurs hétérogènes. Nous utilisons les motifs binaires de gabor (LGBP) introduit par Zhang et al. [4]. Ce codage permet d'exploiter des informations multi-résolutions et multi-orientations entre les pixels, tout en étant robuste aux changements d'illuminations ou d'alignements.

Ce codage est ensuite combiné à des descripteurs de type géométrique : les modèles actifs d'apparence (AAM pour Active Appearance Models). Introduits par Cootes et al. [5], les AAM sont des modèles statistiques de formes et de niveaux de gris de visages. Les AAM apportent donc une information très importante sur la position de points clefs du visage.

Pour procéder à la détection d'AU, nous avons décidé d'utiliser les machines à vecteurs supports (SVM pour Support Vector Machine). Les LGBP et les AAM étant hétérogènes, une concaténation de ces descripteurs en un seul vecteur n'est pas le meilleur moyen de les combiner. C'est pourquoi nous utilisons ici deux noyaux par descripteur. Ces deux noyaux sont ensuite combinés via un système basé sur les SVM multi-noyaux [6]. Enfin, afin d'inclure les aspects temporels et dynamiques dans notre détecteur, les sorties du classifieur sont post-traitées en utilisant des techniques de filtrage.

Cet article est organisé de la manière suivante. La section 2 présentera un rapide état de l'art de la reconnaissance d'AU. La section 3.1 détaillera les différents descripteurs utilisés. La section 4 expliquera la procédure de classification pour la détection d'AU. La section 5 détaillera les expérimentations validant le choix des LGBP comme descripteurs. La section 6 présentera les résultats de notre détecteur sur la base de données GEME-FERA. Enfin, la section 7 conclura cet article.

2 Etat de l'Art

Les méthodes existantes pour la détection d'AU peuvent se diviser en trois catégories : celles utilisant des descripteurs géométriques, celles faisant appel à des descripteurs d'apparence, et celles combinant ces types d'information.

Les méthodes basées sur des descripteurs de type géométrique essayent d'extraire la forme de modèles locaux du visage (oeil, bouchez, nez...) et d'extraire des points de saillance (pouvant être un ensemble de points caractéristiques indépendants du visage, ou bien un modèle global). Typiquement, dans [7], l'auteur effectue un suivi de 20 points clefs du visage tout au long de la séquence. Ces points d'intérêt sont automatiquement détectés dans la première image de la séquence et sont ensuite suivis en utilisant un système basé sur des filtres à particules utilisant la combinaison de points rigides et d'un modèle morphologique. Les AU sont finalement reconnues à l'aide de SVM apprises les déplacements relatifs entre ces points.

Cohen et al. [8], quant à eux, utilisent un trackeur pour suivre le mouvement de 12 points caractéristiques du visage exprimés en terme d'"Unité de Mouvement". Ils utilisent un classifieur Bayesien et des modèles de Markov cachés (HMM pour Hidden Markov Models) pour reconnaitre les émotions dans une séquence vidéo. Les auteurs proposent également un HMM multi-niveau combinant les informations temporelles et segmentant automatiquement une séquence de longue durée.

Les méthodes basées sur des descripteurs d'apparence extraient des informations représentant la texture faciale. Un grand nombre de ces méthodes sont basées sur les ondelettes de Gabor. Bartlett et al. [9] utilisent par exemple ces descripteurs avec un classifieur GentleBoostSVM. De leur côté, Bazzo et Lamar [10] les utilisent avec des réseaux de neurones. D'autres méthodes, basées sur l'apparence, essayent d'extraire des motifs temporels dans des séquences, en utilisant par exemple un flot optique [11] ou encore des historiques de mouvements d'images (MHI pour Motion History Images) [12].

Enfin, certaines études combinent des informations géométriques et d'apparence. Par exemple, Chew et al. [13] utilisent des CLM (Contrained Local Models) pour suivre les visages et codent l'apparence à l'aide d'opérateur LBP (Local Binary Pattern), utilisant ensuite des SVM pour classifier les AU. Dans [14], le déplacement des points est utilisé pour détecter un sous-ensemble d'AU, les autres

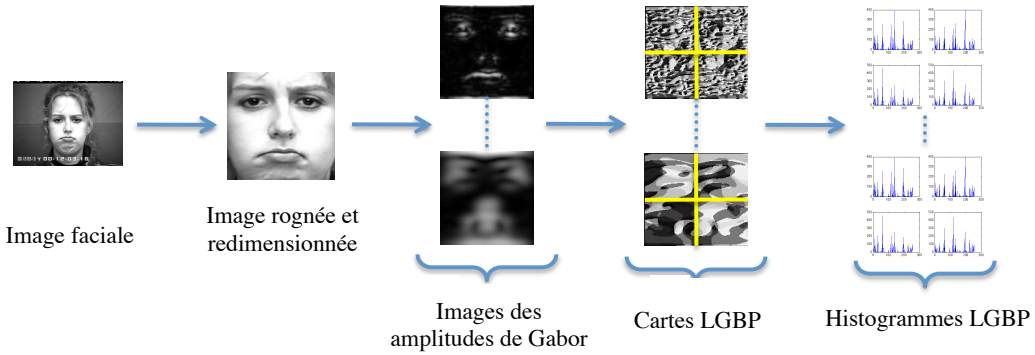


FIGURE 1 – Calcul des histogrammes LGBP.

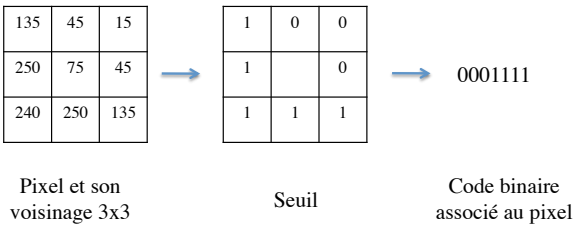


FIGURE 2 – Opérateur LBP.

étant détectées à l'aide de filtre de Gabor et de SVM.

3 Descripteurs

3.1 histogrammes LGPB

Pour pré-traiter les données nous avons automatiquement détecté le centre des yeux à l'aide de notre détecteur de points caractéristiques [15]. La position des yeux est ensuite utilisée pour normaliser les variations d'échelles ainsi que les rotations dans le plan. Nous obtenons ici des images de visages avec la même taille de $128 * 128$ pixels, avec la position du centre des yeux aux mêmes coordonnées. La figure 1 décrit l'ensemble du processus d'extraction des caractéristiques.

Images de Gabor. Les images de Gabor sont obtenues en convoluant les images de visages avec les filtres de Gabor. Nous utilisons trois fréquences spatiales et six orientations, pour un total de 18 filtres de Gabor. La phase étant très sensible, seul le module est généralement gardé. Ainsi, nous obtenons un total de 18 images de Gabor.

Motif Binaire Local (LBP). L'opérateur LBP a été popularisé par Ojala et al. [16] pour analyser les textures. Chaque pixel de l'image est codé en seuillant son voisinage 3×3 par sa valeur et en considérant que le résultat est un nombre binaire comme montré sur la figure 2.

Motif Binaire de Gabor (LGBP). L'opérateur LBP est appliqué sur les images de Gabor pour obtenir autant de cartes LGBP qu'il y a d'images de Gabor. La combinaison de l'opérateur LBP et des ondelettes de Gabor exploite les liens entre les pixels pour plusieurs résolutions et orientations. Cela s'est révélé être très robuste à l'illumination et aux erreurs d'alignement [4].

Histogrammes LGBP. Chaque zone du visage contient différentes informations pertinentes. Ainsi, nous avons choisi de diviser le visage en plusieurs zones et de calculer un histogramme pour chacune d'elles. Ce découpage a été choisi du fait du changement local que peut impliquer l'apparition d'une AU. Par exemple, l'AU1 ne modifie que le coin intérieur du sourcil. Nous avons choisi de diviser chaque image de visage en $4 \times 4 = 16$ régions.
Pour un visage i, nous obtenons ainsi un vecteur H_i par concaténation des $16 \times 6 \times 3 = 288$ histogrammes de dimension 256 calculés pour chacune des régions, orientations et fréquences spatiales. Nous obtenons finalement un vecteur de $288 \times 256 = 73728$ caractéristiques par image de visage.

Méthode de réduction d'histogramme. Afin de réduire significativement les temps de traitement, nous procédons à la réduction de la dimension de nos descripteurs. Pour ce faire, Ojala et al.[16] proposent de ne garder que les motifs uniformes.
Pour réaliser cette réduction, nous avons opté pour une approche légèrement différente. En effet, désirant garder la contribution de tous les motifs, cette réduction propose de grouper les occurrences des différents motifs dans les mêmes classes d'histogrammes. La méthode est détaillée dans [17].
Nous réduisons ainsi la dimension d'un histogramme de 256 classes à 26. L'histogrammes LGBP d'un visage n'est donc plus que de dimension $288 \times 26 = 7488$.
Cette réduction de classe d'histogramme permet de réduire les temps de calcul des matrices noyaux de façon significative, et ce sans dégrader les résultats.

3.2 $2.5D$ Active Appearance Model

Afin d'extraire les coefficients AAM, nous avons appris deux AAM $2.5D$ locaux [18] : un pour la bouche et un autre pour les yeux.
Ces AAM sont appris sur un total de 466 images expres-

sives et neutres provenant de la base de données Bosphorus [19].
Une fois cette étape réalisée, nous alignons les modèles AAM sur tous les visages de la partition d'apprentissage de GEMEP-FERA en gardant les paramètres C d'apparence obtenus (extraits à partir des paramètres de formes et de texture). Quelques résultats obtenus sur des images de la base GEMEP-FERA sont présentés sur la figure 3. Nous pouvons voir dans la dernière image l'un des désavantages d'utiliser des modèles locaux par rapport à un modèle global : pour un modèle des yeux, la texture du front est relativement importante et une perturbation causée par des cheveux cachant le front peut causer des localisations imprécises. Par contre, l'avantage est que la bouche est décorrélée des yeux et n'est donc pas affectée par cette perturbation engendrant ainsi un bon alignement du modèle local de la bouche.

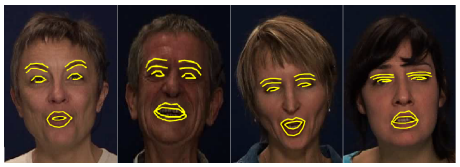


FIGURE 3 – Exemples d'alignements de visages via les AAM locaux.

4 Classification

Afin de procéder à la reconnaissance des AU, nous avons décidé d'utiliser une SVM par AU. Pour l'apprentissage d'une SVM, toutes les images contenant l'AU désirée sont utilisées comme exemples positifs, toutes les autres images comment exemples négatifs.

4.1 SVM multi-noyau

Etant donné un exemple d'apprentissage composé d'histogrammes LGBP et d'un vecteur de coefficients d'AAM, $x_i = (H_i, C_i)$, associé aux étiquettes y_i (exemple positif ou négatif), la fonction de classification de la SVM associe un score s à un nouvel exemple $x = (H, C)$:

$$s = \left(\sum_{i=1}^{m} \alpha_i k(x_i, x) + b \right) \quad (1)$$

avec α_i la représentation duale du vecteur orthogonal à l'hyperplan [20]. k étant la fonction noyau réalisant un produit scalaire dans un espace de plus grande dimension.
Dans le cas d'une SVM multi-noyau, le noyau k peut être toute combinaison convexe de fonctions semi-définies positives vérifiant la condition de Mecer.

$$k(x_i, x) = \sum_{j=1}^{K} \beta_j k_j \text{ with } \beta_j \geq 0, \sum_{j=1}^{K} \beta_j = 1 \quad (2)$$

Ici, nous avons un noyau par type de descripteur :

$$k = \beta_1 k_{LGBP}(H_i, H) + \beta_2 k_{AAM}(C_i, C) \quad (3)$$

Les poids α_i et β_j sont optimisés pour avoir un hyperplan maximisant la marge (distance minimum d'un exemple à l'hyperplan). Ce problème d'optimisation est conjointement convexe en α_i et β_j [21].
β_1 représente le poids donné aux descripteurs LGBP et β_2 celui donné pour le vecteur d'apparence des AAM. Ainsi, après apprentissage, le système est capable, de lui même, de trouver la meilleur combinaison de ces deux types de descripteurs.

4.2 Fonctions noyaux

Dans la section 5, les résultats expérimentaux montre que dans la reconnaissance d'AU basée sur des histogrammes utilisés avec un noyau d'intersection d'histogramme conduit aux meilleurs performances :

$$K_{LGBP}(H_i, H_j) = \sum_{k} min(H_i(k), H_j(k)) \quad (4)$$

Pour le vecteur d'apparence des AAM, nous utilisons un noyau gaussien (RBF pour Radial Basis Function) :

$$K_{AAM}(C_i, C_j) = e^{-\frac{\|s_i - s_j\|_2^2}{2\sigma^2}} \quad (5)$$

avec σ un paramètre à optimiser sur une base de cross-validation durant l'apprentissage.

4.3 Filtrage Temporel

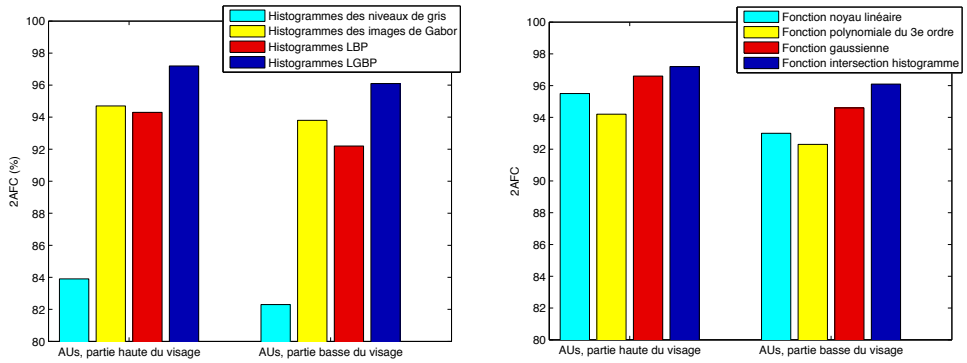
Afin de prendre en compte l'information temporelle dans notre système, nous appliquons, pour chaque AU, un filtre moyenneur sur chacune des sorties des SVM. La taille du filtre à été choisie de manière à maximiser la F1-mesure obtenue sur la base d'apprentissage.

5 Evaluation des descripteurs et des fonctions noyaux

Dans cette section, nous résumons des expérimentations précédemment effectuées [17] sur la base de donnée Cohn-Kanade. Ce sont ces résultats qui nous ont orienté vers le choix des histogrammes LGBP associé à un noyau d'intersection d'histogrammes pour le challenge FERA. Nous rapportons dans cette partie les aires sous la courbe ROC (2AFC) obtenues durant un processus de validation "un sujet exclu" (leave-one-subject-out) pour 16 AU : 7 AU de la partie haute du visage (1, 2, 4, 5, 6, 7 et 9) et 9 de la partie basse (11, 12, 15, 17, 20, 23, 24, 25 et 27). Plusieurs types d'histogrammes et différentes fonctions noyaux sont ici comparés.

5.1 Descripteurs

Dans la figure 4-a, nous comparons les performances de différents types d'histogramme en utilisant le noyau d'intersection d'histogramme. Les histogrammes calculés à partir des images LBP et des images de Gabor donnent de bien meilleurs résultats que les histogrammes calculés directement sur les niveaux de gris. C'est avec la combinaison des deux approches, à travers les LGBP, que nous obtenons les meilleurs résultats.



a) b)

FIGURE 4 – Scores 2AFC pour la détection des AU de la partie haute et basse du visage pour a) une fonction noyau intersection d'histogrammes et différents histogrammes et b) des histogrammes LGBP et différentes fonctions noyaux.

5.2 Fonctions noyaux

La figure 4-b montre les résultats obtenus avec les histogrammes LGBP associés à différentes fonction noyaux. Pour le noyau polynomial et le noyau gaussien, différentes valeurs d'ordre du polynôme ou d'écart-type ont été testés. Les meilleurs performances sont obtenues avec la fonction d'intersection d'histogramme. Comparée à un noyau linéaire, l'aire moyenne sous la courbe ROC passe de 95.5% à 97.2% pour les AU du haut du visage et de 93.0% à 96.1% pour les AU du bas. Le noyau gaussien donne quant à lui de bon résultats, mais les paramètres ont été directement optimisés sur la base d'apprentissage.

6 Résultats expérimentaux sur la base GEMEP-FERA

Nous présentons ici les scores obtenus sur la partition de test de la base GEMEP-FERA. Ces scores ont été calculés par les organisateurs de la compétition FERA à partir des sorties de nos classifieurs.

6.1 Protocole expérimental

Mesure de performance. L'évaluation de nos résultats à été calculée en terme de F1-mesure. Il s'agit d'une mesure réalisant une moyenne harmonique de la précision p et du rappel r :

$$F1 = 2 \cdot \frac{p \cdot r}{p + r} \tag{6}$$

Pour calculer cette mesure, il nous faut seuiller les sorties des SVM. Mais ce seuillage pose problème : plus ce seuil est bas et plus on aura de détection donc plus de chance d'avoir un taux de rappel haut, et plus ce seuil est haut, moins on aura de détections mais une plus grande chance d'avoir des détections correctes, donc une meilleure précision. Dans ce cas précis, la mesure F1 varie beaucoup avec le choix de ce seuil, la rendant sensible aux optima locaux C'est pourquoi nous avons choisi une seconde mesure ne dépendant pas du seuil : l'aire sous la courbe (2AFC). Cette courbe est tracée en reportant le taux de bonne détection ou rappel en fonction du taux de fausse alarme pour différents points de fonctionnement du détecteur (dans l'exemple précédent, en faisant varier la valeur du seuil).

Cross-Validation. Pour les expérimentations, les bases de donnée utilisées sont les suivantes :

- Cohn-Kanade [22] : la dernière image (*apex* de l'expression) de toutes les 486 séquences. Cette base contient des séquences d'images de 97 étudiants de 18 à 30 ans. Les conditions d'éclairage sont relativement uniformes et les images présentent de faibles variations de poses.
- Bosphorus [19] : Environ 500 images choisies en fonction des combinaisons d'AU représentées. Les conditions d'éclairage sont relativement uniformes et les images présentent de faibles variations de poses.
- GEMEP-FERA [3] : une image de chaque séquence pour toutes les combinaisons d'AU présentes. Soit un total de 600 images.

Le système a été appris sur la partition d'apprentissage de la base GEMEP-FERA. La phase de validation s'est effectuée en deux étapes :

- Les paramètres externes du système, le paramètre C de la SVM et l'écart- type σ de la fonction noyau gaussienne, ont été choisis pour maximiser le score 2AFC lors de la validation croisée un-sujet-exclu sur la base d'apprentissage GEMEP-FERA. Ce score a été choisi car il ne dépend pas de la valeur du seuil en sortie de la SVM.
- Pour régler la taille du filtre moyenneur et du seuil (un par AU), nous avons utilisé les sorties signées des images de la partition d'apprentissage GEMEP- FERA obtenues lors de la validation croisée, puis nous avons choisi la taille du filtre et la valeur de seuil permettant d'obtenir le score F1 maximum. Nous avons ensuite appliqué les filtres moyenneurs et les seuils aux sorties signées des images de la partition de test GEMEP-FERA.

		LBP (baseline)	LGBP	y. AAM	b. AAM	AAMs	LGBP + y. AAM	LGBP + b. AAM	LGBP + AAMs
Haut du visage	AU1	**79.0**	78.8	60.5	54.4	**62.3**	77.6	**81.8**	80.3
	AU2	76.7	**77.1**	**57.4**	51.3	**57.4**	57.0	**83.4**	82.7
	AU4	52.6	**62.9**	**62.0**	56.4	59.3	**62.4**	61.3	58.5
	AU6	65.7	**77.0**	56.4	75.8	**80.9**	79.3	80.9	**81.0**
	AU7	55.6	**68.5**	**67.8**	54.3	54.4	72.5	71.0	**71.2**
	Moy. haut	62.7	**71.4**	60.9	59.5	**63.0**	67.8	**74.2**	73.4
Bas du visage	AU10	**59.7**	51.9	43.2	**56.3**	51.9	49.9	**52.7**	52.1
	AU12	72.4	**79.9**	63.9	69.9	**79.5**	81.7	**82.3**	82.2
	AU15	56.3	**63.0**	58.3	68.7	**71.2**	**67.1**	59.6	61.4
	AU17	64.6	**65.8**	51.9	**67.1**	66.3	67.2	**72.5**	70.7
	AU18	61.0	**70.4**	51.9	75.3	**75.8**	54.8	**79.0**	78.5
	AU25	59.3	**59.8**	55.6	**69.6**	63.6	56.0	63.5	**65.6**
	AU26	50.0	**64.3**	57.0	**67.5**	58.6	58.7	**64.8**	62.9
	Moy. bas	60.6	**67.2**	56.4	**69.7**	69.2	64.3	**70.3**	70.2
Moy. pers. spec.		63.1	**67.8**	57.4	63.6	**65.5**	65.5	**71.6**	71.5
Moy. pers. ind.		61.1	**68.0**	57.4	65.3	**66.1**	65.9	**69.5**	69.0
Moyenne		62.8	**68.3**	56.8	63.9	**65.1**	65.3	**71.1**	70.6

TABLE 1 – Scores 2AFC obtenus sur la base de test GEMEP-FERA en utilisant les LGBP, les coefficients du modèle AAM des yeux (y. AAM), du modèle AAM de la bouche (m. AAM), la concaténation des coefficients des modèles des yeux et de la bouche (AAMs) et la fusion des LGBP avec les AAM.

6.2 Descripteurs et évaluation de la stratégie de fusion.

Nous évaluons différents descripteurs sur la partition de test de la base GEMEP-FERA. Lorsque nous avons envoyé, après compétitions, nos résultats signés, les organisateurs nous ont renvoyé le score 2AFC pour chaque AU dans les deux cas suivants : indépendant (du sujet : les sujets de la base de test ne sont pas dans la base d'apprentissage), spécifique (à un groupe de sujets : les sujets de la base de test sont dans la base d'apprentissage) et la moyenne des deux scores. Nous reportons dans le tableau 1 les scores 2AFC sur l'ensemble des sujets pour chaque AU et la moyenne des scores 2AFC dans les cas "indépendant" et "spécifique".

En utilisant seulement les LGBP, nous pouvons voir que les résultats sont déjà bien supérieurs à ceux proposés par les organisateurs qui utilisent des LBP (68.3% contre 62.8%). Les deux méthodes sont similaires : même protocole, même base d'apprentissage, même classifieur, la seule différence vient du descripteur et du choix de la fonction noyau. Ces résultats confirment donc les conclusions de la section 5 sur les meilleurs performances des LGBP par rapport aux LBP et l'importance d'une fonction noyau adaptée.

Avec les coefficients AAM de la bouche, nous obtenons de bon résultats pour détecter les AU localisées dans la partie basse du visage. Les résultats sont même meilleurs qu'avec les LGBP pour les AU 15, 17, 18, 25 et 26 (68.7% 67.1% 75.3% 69.6% et 67.5% contre respectivement 63% 65.8% 70.4% 59.8% et 64.3%). Les résultats obtenus pour les AU de la partie haute du visage sont, en toute logique, assez proches de ceux obtenus avec un système naïf répondant au hasard. Il n'y a que l'AU 6 qui soit bien détectée car elle représente le lever des pommettes qui apparait souvent en même temps que l'AU 12 (le sourire).

Avec les coefficients AAM des yeux, les résultats ne sont que légèrement supérieurs à ceux d'un système naïf (56.8% là où un système naïf fait 50%). Cela peut être expliqué par la faible précision de l'alignement du modèle des yeux. Les sourcils, par exemple, sont difficiles à localiser, surtout quand ils sont cachés par des cheveux.

Concernant la fusion, nous notons que les coefficients AAM des yeux n'améliorent pas les performances lorsqu'ils sont couplés avec les histogrammes LGBP. Les meilleurs résultats sont obtenus lorsque les coefficients AAM de la bouche ou des deux modèles sont combinés avec les LGBP. Plus surprenant, la détection des AU de la partie haute du visage est améliorée avec l'AAM de la bouche (81.8%, 83.4%, 80.9% et 71.0% pour respectivement les AU 1, 2, 6, et 7 contre 78.8%, 77.1%, 77.0% et 68.5% en utilisant seulement les LGBP). L'amélioration de la reconnaissance de l'AU 6 peut être expliquée comme précédemment, car elle apparaît en même temps que l'AU 12 qui elle est détectée facilement par un modèle de la bouche. Pour les autres AU, cela est plus difficile à interpréter. Le classifieur multi-noyaux peut utiliser l'information donnée par l'AAM de la bouche non pas pour détecter directement ces AU mais pour obtenir une information à propos du sujet (par exemple, une information sur son identité, son type de peau...). Cela se traduit par une augmentation de la dimension dans laquelle les données sont représentées permettant d'améliorer la détection de ces AU. Cela montre ainsi l'intérêt de combiner deux types de descripteurs.

Globalement, la fusion des AAM avec les LGBP a amélioré significativement les résultats pour 9 AU sur 12, les AU 1 2 6 7 12 17 18 et 25. Les performances moyennes sont, elles aussi, améliorées.

Finalement, en comparant les résultats dans le cas "spécifique" à ceux dans le cas "indépendant", nous pouvons no-

	Sans filtrage	Filtre moyen
Cas "spécifique"	75.4	77.0
Cas "indépendante"	74.7	76.6
Moyenne	75.0	76.7

TABLE 2 – Influence du filtrage sur les scores 2AFC sur la partition de test GEMEP-FERA

ter que la fusion est meilleure particulièrement dans le cas "spécifique".

6.3 Filtrage temporel

Pour tester l'impact du filtre moyenneur, nous l'appliquons au détecteur appris sur les bases GEMEP-FERA et Cohn Kanade combinant les histogrammes LGBP aux AAM. Nous reportons dans le tableau 2, les scores 2AFC avec et sans filtrage.

6.4 Résultats sur la base de test et comparaisons

Pour la compétition FERA'11, nous ne connaissions pas les résultats présentés dans les sections précédentes, car obtenus sur la partition de test GEMEP-FERA. Nous ne pouvions donc pas les utiliser pour optimiser notre système en choisissant, par exemple, d'utiliser uniquement la base Cohn-Kanade en plus des images de la base GEMEP-FERA. Nous avions choisi d'utiliser une combinaison de LGBP et les coefficients des deux AAM et avons entrainé les classifieurs sur les bases GEMEP-FERA, Cohn-Kanade et Bosphorus pour les AU 1, 2, 3, 12, 15, 17 et les bases GEMEP-FERA et Cohn-Kanade pour les autres. Nous avons choisi de ne pas utiliser la base Bosphorus pour les AU non-représentées dans cette base de données. Le système présenté dans cette section surpasse tous les autres systèmes dans les deux cas, cas 'indépendant" et cas "spécifique".

Le tableau 3 présente une synthèse des résultats et des méthodes utilisées par les compétiteurs.

Premièrement, si nous comparons simplement les approches basées sur des descripteurs d'apparence, la méthode 5a utilisant des histogrammes LGBP et une fonction noyau d'intersection d'histogramme, donne de meilleurs résultats que la méthode 1 utilisant des histogrammes LBP et une fonction noyau gaussienne (respectivement 68.3% comparé à 62.8%). Ceci peut s'expliquer par la supériorité des filtres de Gabor sur les LBP (aussi remarquée dans [24]) et l'utilisation d'une fonction noyau bien adaptée aux descripteurs de type histogramme.

La méthode 3 peut être comparée avec la méthode 5b. Le première utilise des modèles locaux contraints (CLM) et la seconde des AAM. Ces descripteurs utilisent l'information géométrique et d'apparence. Malheureusement, la mesure de performance utilisée n'est pas la même, on ne peut donc pas les comparer entre elles. Par contre, ces expériences donnent des résultats moins bons que celles utilisant des descripteurs d'apparence seulement.

Ces résultats sont surpassés avec la méthode 5c, combinant des descripteurs d'apparence (LGBP) avec une information géométrique et d'apparence (AAM) grâce à des classifieurs SVMs multi-noyaux.

Finalement, nous pouvons noter, que de prendre en compte le contexte temporel de la séquence, améliore la précision globale du système comme montré par les expériences de la méthode 4 (amélioration de 72.3% à 75.8%) et les expériences 5c et 5d (de 75.0% à 76.7%).

7 Conclusion

Nous avons proposé ici une méthode originale pour la détection d'AU. Le système repose sur la combinaison de deux types d'information : des descripteurs d'apparence, ne prenant pas d'information spatiale en compte (histogrammes LGBP) et des descripteurs géométrique contenant une information statistique sur les formes et les textures (coefficients AAM). Afin de gérer ces deux descripteurs hétérogènes et de les combiner de façon optimisée, nous proposons d'utiliser un système d'apprentissage avancé. Les SVM multi-noyau permettent de sélectionner les informations les plus importantes grâce à une pondération des noyaux en fonction de leur pertinence.

Les résultats expérimentaux dans les cas "spécifique" et "indépendant" ont montré que les histogrammes LGBP convergent vers de meilleurs résultats que les AAM. Ce phénomène s'explique principalement par le manque de précision des AAM lorsque qu'ils s'alignent sur des visages expressifs, cette précision étant un élément crucial pour la détection d'AU dans une approche géométrique. Néanmoins, nous avons tout de même montré que, malgré ce manque de précision, la combinaison des AAM avec les histogrammes LGBP augmente l'aire sous la courbe ROC pour 9 AU sur 12.

Finalement, les résultats en terme de F1-mesure obtenus par ce système ont permis de remporter le challenge FERA. Ce succès s'explique par :

- La capacité des histogrammes LGBP, couplés avec un noyau d'intersection d'histogrammes, a généraliser sur de nouveaux exemples.
- L'utilisation du multi-noyau pour fusionner différents types d'information.
- L'optimisation des scores 2AFC en premier lieu et de la F1-mesure durant un processus de validation croisé.

Mais les résultats globaux restent tout de même insuffisants. Certaine AU sont détectées avec une précision tout juste supérieure à celle d'un système naïf. Plusieurs améliorations sont à envisager. L'utilisation d'un système dynamique (plutôt que statique), basé par exemple sur des modèles graphiques, permettrait d'utiliser, et ce de façon optimisée, l'aspect temporel des émotions.

Références

[1] P. Ekman and E. Rosenberg, *What the Face Reveals*. Oxford University Press, 2005.

CompÈtiteurs	Type	Descripteurs	Classifieurs	Bases d'app.	F1	2AFC
1-Valstar et al.[23]	A, I	LBP + PCA	SVM	G	0.45	0.63
2-Baltrusaitis et al.[14]	G+A, S	points + Gabor	Règles + SVM	G	0.46	
3-Chew et al.[13]	G, S	CLM	SVM	G+C	0.51	
4a-Wu et al.[24]	A, I	Gabor	AdaBoost + SVM	G + C + M + pr	0.55	0.72
4b	A, C	Gabor	AdaBoost + SVM	G + C + M + pr	0.58	0.76
5a-**Notre méthode**	A, I	LGBP	SVM	G		0.68
5b	G+A, I	AAM	SVM	G		0.65
5c	G+A, I	LGBP + AAM	MKL SVM	G + C+ Bos		0.75
5d	G+A, C	LGBP + AAM	MKL SVM	G + C + Bos	0.62	0.77

TABLE 3 – Tableau résumant les différentes approches des participants à la campagne FERA'11. G/A : Géométrique/Apparence. I/S/C : Image-par-Image/Suivi (tracking)/Prise en compte du contexte. G/C/M/Bos/pr : bases GEMEP-FERA, Cohn-Kanade, MMI, Bosphorus et privée.

[2] P. Ekman and W. Friesen, "Facial action coding system (facs) : A technique for the measurement of facial actions," *Consulting Psychologists Press*, 1978.

[3] T. Banziger and K. R. Scherer, "Using actor portrayals to systematically study multimodal emotion expression : The gemep corpus," 2007.

[4] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Lgbphs : A novel non-statistical model for face representation and recognition," in *Proc. IEEE Int'l Conf. on Computer Vision (ICCV '05)*, 2005.

[5] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," in *Proc. IEEE European Conf. on Computer Vision (ECCV '98)*, p. 484, Springer, 1998.

[6] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[7] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR '06)*, 2006.

[8] I. Cohen, N. Sebe, A. Garg, L. Chen, and T. Huang, "Facial expression recognition from video sequences : Temporal and static modeling," *Computer Vision and Image Understanding*, vol. 91, no. 1-2, pp. 160–187, 2003.

[9] M. Bartlett, G. Littlewort, M. Frank, I. Fasel, and J. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, 2006.

[10] J. Bazzo and M. Lamar, "Recognizing facial actions using gabor wavelets with neutral face average difference," in *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition (AFGR '04)*, pp. 505–510, 2004.

[11] I. Essa and A. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp. 757–763, 1997.

[12] M. Valstar, M. Pantic, and I. Patras, "Motion history for facial action detection in video," in *Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics (SMC '04)*, vol. 1, 2004.

[13] S. Chew, P. Lucey, S. Lucey, J. Saragih, J. Cohn, and S. Sridharan, "Person-independent facial expression detection using constrained local models," in *Proc. IEEE FG'11, Facial Expression Recognition and Analysis Challenge (FERA'11)*, 2011.

[14] T. Baltrusaitis, D. McDuff, N. Banda, M. Mahmoud, R. Kaliouby, P. Robinson, and R. Picard, "Real-time inference of mental states from facial expressions and upper body gestures," in *Proc. IEEE FG'11, Facial Expression Recognition and Analysis Challenge (FERA'11)*, 2011.

[15] V. Rapp, T. Senechal, K. Bailly, and L. Prevost, "Multiple kernel learning svm and statistical validation for facial landmark detection," in *Proc. IEEE Int'l. Conf. Face and Gesture Recognition (FG'11)*, 2011.

[16] T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[17] T. Senechal, K. Bailly, and L. Prevost, "Automatic facial action detection using histogram variation between emotional states," in *Proc. Int'l Conf. Pattern Recognition (ICPR'10)*, 2010.

[18] A. Sattar, Y. Aidarous, S. Le Gallou, and R. Seguier, "Face alignment by 2.5 d active appearance model optimized by simplex," in *Proc. Int'l Conf. on Computer Vision Systems (ICVS'07)*, 2007.

[19] A. Savran, N. Alyuz, H. Dibeklioglu, O. Celiktutan, B. Gokberk, B. Sankur, and L. Akarun, "Bosphorus database for 3d face analysis," *Biometrics and Identity Management*, pp. 47–56, 2008.

[20] B. Scholkopf and A. Smola, *Learning with Kernels*. MIT Presss, 2002.

[21] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, and M. Jordan, "Learning the kernel matrix with semidefinite programming," *The Journal of Machine Learning Research*, vol. 5, p. 27, 2004.

[22] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition (AFGR '00)*, pp. 46–53, 2000.

[23] M. Valstar, B. Jiang, M. Méhu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," in *IEEE Int'l. Conf. Face and Gesture Recognition (FG'11)*, 2011.

[24] T. Wu, N. Butko, P. Ruvolo, J. Whitehill, M. Bartlett, and J. Movellan, "Action unit recognition transfer across datasets," in *Proc. IEEE FG'11, Facial Expression Recognition and Analysis Challenge (FERA'11)*, 2011.