Analyse d'expressions faciales en langue des signes.

Hugo Mercier[1,2] Patrice Dalle[1]

[1] Institut de Recherche en Informatique de Toulouse
118, route de Narbonne
31062 Toulouse
{mercier,dalle}@irit.fr

[2] WebSourd
99, route d'Espagne - bâtiment A
31100 Toulouse

Résumé

Dans le contexte actuel des recherches informatiques sur la langue des signes, l'analyse des expressions faciales est une composante importante qui doit être développée. Nous proposons des outils informatiques d'analyse faciale dans un contexte de communication en langue des signes utilisant des modèles statistiques déformables et mettons en relief les problèmes particuliers à traiter dans ce contexte. Nous présentons de plus, une application basée sur le résultat des méthodes d'adaptation permettant de rendre anonyme une communication signée.

Mots Clef

Langue des signes, visage, expressions, AAM, modèles déformables, anonymisation.

Abstract

In the field of computer based sign language research, facial expression analysis is an important component that has to be developed. We propose here computer tools for facial analysis in a context of sign language communication that use statistical morphable models and we emphasize specific problems that have to be addressed in this context. Moreover, we propose an application based on the fitting method results aiming at anonymizing a signed communication.

Keywords

Sign language, face, facial expressions, AAM, morphable models, anonymization.

1 Introduction

La langue des signes est une langue visuo-gestuelle, naturellement utilisée par les sourds. Elle met en jeu tout le corps du signeur et en particulier le visage par ses expressions. Les expressions faciales sont indispensables à la bonne construction et compréhension d'un discours en langue des signes.
Le développement d'outils informatiques d'analyse des expressions faciales permet d'envisager des applications intéressantes pour la population sourde ainsi que la communauté des chercheurs travaillant sur leur langue (linguistes, informaticiens, etc.)

1.1 Expressions faciales en langue des signes

En langue des signes française (LSF), les expressions faciales, encore appelées mimiques, entrent en jeu à différents niveaux de la langue : à un niveau lexical puisque la plupart des signes dits « standards » sont définis par une certaine mimique faciale en plus des autres composants (configuration des mains par exemple) ; à un niveau syntaxique puisque les modes du discours (interrogatif, conditionnel, etc.) sont introduits par des mouvements faciaux particuliers ou encore à un niveau sémantique lors par exemple de transferts personnels, où le signeur prend le rôle d'un acteur du discours et les expressions de son visage traduisent les émotions du personnage joué.
En plus des mouvements musculaires observables sur la face du signeur, la direction du regard et la pose du crâne jouent un rôle important, notamment par leur capacité à pertiniser par la désignation une région de l'espace virtuel faisant face au signeur (appelé espace de signation) et permettant la mise en place spatiale des éléments du discours.

1.2 Langue des signes et traitement d'images

Du point de vue du traiteur d'images, la langue des signes présente des problèmes particuliers. Concernant l'analyse du visage d'un signeur sur vidéo un outil d'analyse se doit de prendre en compte les faits suivants :
– le visage du signeur présente fréquemment des rotations (et en particulier des rotations hors-plan),
– les signes manuels ont une forte probabilité d'occulter une partie du visage du signeur.

2 Travaux existants

L'analyse automatique des expressions faciales est un problème émergent dans le domaine de la vision par ordinateur et du traitement d'images. La plupart des travaux existants tentent de reconnaître les émotions universelles listées par Ekman [1]. Ce sont sept émotions communes à tout humain qui se traduisent sur le visage par sept prototypes d'expressions. Les travaux traitant ce problème utilisent des mé-

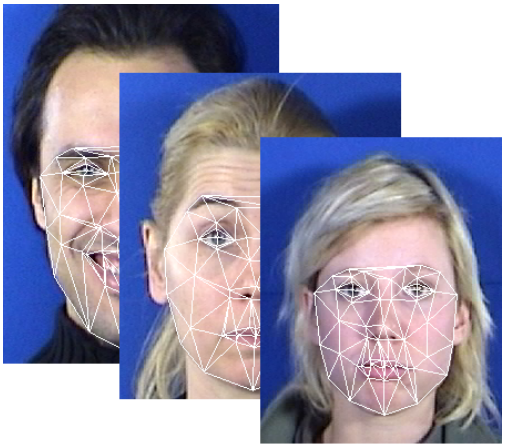

FIG. 1 – Exemple de base d'apprentissage où chaque image est associée à un modèle de forme (ici représenté via un maillage).

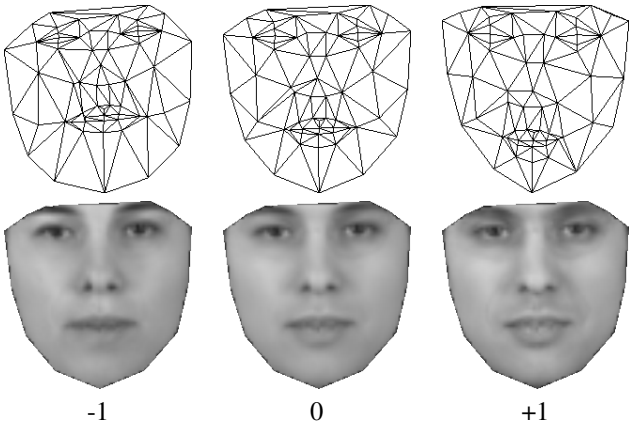

FIG. 2 – Examples de modes de variation d'un AAM. La première composante de variation de forme est sur la première ligne. La seconde ligne représente un vecteur de variation d'apparence. Sur la colonne de gauche sont visibles des visages générés avec un poids négatif pour le vecteur de déformation. Le poids est positif sur la colonne de droite et la colonne centrale représente les forme et apparence moyennes.

thodes de classification : laquelle des sept émotions universelles correspond le plus à l'expression observée sur un visage donné. Cette vision des choses est trop limitée pour un contexte de communication où l'ensemble des expressions affichées ne sont sans doute pas universelles.
De manière plus générique, une expression faciale peut être vue comme une combinaison de mouvements faciaux unitaires (nommés *Action Units* dans la terminologie FACS [1]). Le but d'un outil d'analyse est alors de mesurer la présence et l'intensité des *action units* du visage au cours d'une vidéo.

2.1 Suivi des mouvements faciaux

Nous utilisons ici un formalisme nommé *Active Appearance Models* (AAM - [2, 3]) dans lequel un visage et ses déformations de forme et d'apparence sont apprises à partir de données manuellement annotées. Les détails sur la construction d'un AAM et la manière de l'utiliser sont présentés dans cette section.

Modélisation. Un AAM décrit un objet d'une classe prédéfinie comme une instance de forme et une instance d'apparence. Chaque objet, pour une classe donnée, peut être représentée par sa forme, décrite par un ensemble de coordonnées 2D et une apparence, décrite par un ensemble de pixels. Il est donc défini par :

1. une forme $\boldsymbol{s} = \boldsymbol{s}_0 + \sum_{i=1}^{n} v_i \boldsymbol{s}_i$, où $\boldsymbol{s}_0$ est la forme moyenne, $\boldsymbol{s}_i$ sont les vecteurs de déformations et v_i les coefficients pondérateurs de ces déformations ;
2. une apparence $A(x) = A_0(x) + \sum_{i=1}^{m} \lambda_i A_i(x)$, où $A_0(x)$ est l'image d'apparence moyenne, $A_i(x)$ sont les vecteurs de variation d'apparence et λ_i sont leurs coefficients pondérateurs.

Le modèle est construit à partir d'une base d'apprentissage de visages. Une annotation *i.e.*, les coordonnées de tous les points d'intérêt du visage, est associée à chaque image de la base (Fig. 1). Il est alors possible de définir, via une analyse en composantes principales, une forme moyenne $\boldsymbol{s}_0$, une apparence moyenne $A_0(x)$ et leurs vecteurs de variations correspondants $\boldsymbol{s}_i$ et $A_i(x)$ (Fig. 2). Un visage (de la base ou proche) peut donc être décrit par un vecteur de forme v et un vecteur d'apparence λ
Sur une base de n visages, on retiendra $m \leq n$ vecteurs de variations permettant de décrire au mieux un visage.

Adaptation. Le but d'un algorithme d' adaptation est, en supposant une première estimation (v_0, λ_0) des paramètres de forme et d'apparence connue, de trouver (v, λ) qui représente au mieux le visage observé sur l'image. C'est un processus d'optimisation itératif.
La pose originelle du problème [2] ne permet pas de trouver une solution de manière efficace. Les gradients sont estimés numériquement et considérés comme étant constant d'une itération à l'autre. Une implémentation plus récente [3] pose le problème de manière à ce que l' approximation sur la constance des gradients soit réaliste. De plus, il est possible de les dériver analytiquement. Les méthodes posées dans ce cadre (appelé *inverse compositional*) sont efficaces et précises.
Il existe plusieurs variantes de cet algorithme de base en particulier dans le cas où l'apparence n'est pas considérée connue ou fixe sur l'image. C'est le cas par exemple lors de l'adaptation d'un AAM à un visage dont on ne connait pas l'apparence : l'algorithme doit être initialisé avec une première estimation de l'apparence « éloignée » de l'apparence optimale. L'apparence doit alors évoluer à chaque itération.
Les deux principales variantes de l'algorithme permettant

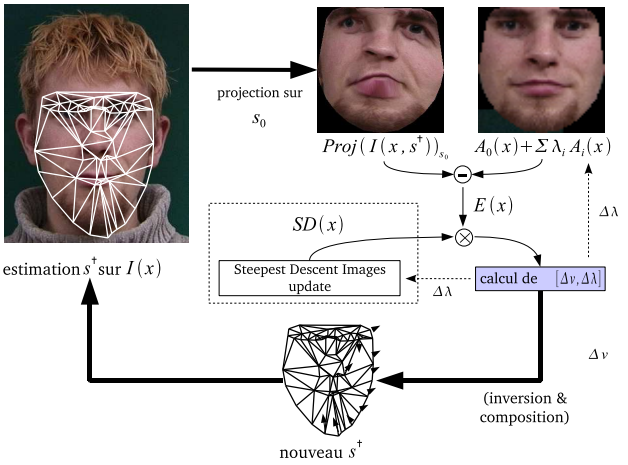

FIG. 3 – Vue schématique du *simultaneous algorithm*. A partir de l'estimation courante de la forme $\boldsymbol{s}^\dagger$, la forme moyenne $\boldsymbol{s}_0$ doit être modifiée par $\Delta\boldsymbol{s}_0$ (calculé à partir de $\Delta\boldsymbol{v}$) dans le but de rendre l'apparence courante $A(x)$ plus proche de $Proj(I(x,\boldsymbol{s}^\dagger))_{\boldsymbol{s}_0}$.

de traiter le cas où l'apparence varie sont le *project-out* et le *simultaneous*. Le premier est très efficace en temps de calcul mais ne permet de traiter que des variations faibles d'apparence. Le second permet quant à lui de traiter des variations importantes d'apparence mais souffre d'une plus grande complexité algorithmique.
Voir la Fig. 3 pour une illustration du *simultaneous*. A chaque pas de l'algorithme, l'apparence et la forme sont optimisées.

Suivi. Les algorithmes d'adaptation permettent le suivi des mouvements faciaux sur une vidéo. A chaque image, l'algorithme est relancé en partant de la position attente à l'image précédente.

Traitement des occultations. Il arrive qu'une partie du visage observé ne soit pas visible par la caméra. Soit parce qu'elle est cachée par une main dans le contexte d'une narration en langue des signes, soit parce que le visage a subi des rotations hors plan trop importantes (on parle alors d'auto-occultations).
Une variante de l'algorithme utilisé (détaillée dans [4]) permet de remplacer la fonction d'erreur (somme des pixels au carré) par une fonction d'erreur robuste (évoluant moins vite que le carré à partir d'un certain seuil). Les modifications apportées à l'algorithme originale sont relativement faibles et permettent de traiter les deux cas d'occultations.
Il est cependant nécessaire de fournir à l'algorithme un paramètre de seuillage, décidant si une erreur est à considérer comme une erreur « utile » à l'évolution du modèle ou une erreur due à une occultation. Des travaux en cours tentent de déterminer automatiquement si une erreur est due à une occultation ou à un mauvais placement du modèle.

Limites. L'algorithme *simultaneous* permet de suivre précisement un ensemble de déformations faciales au cours d'une vidéo. La base d'apprentissage doit être la plus fidèle possible de ce qui est suivi.
En particulier, pour que l'algorithme soit capable de suivre un ensemble d'expressions d'une personne, il est nécessaire d'avoir précédemment annoté manuellement un ensemble d'images de cette même personne correspondant le plus possible aux expressions que l'on désire suivre et ce dans des conditions d'acquisition très proches des conditions de test.

3 Contributions

3.1 Algorithme d'adaptation / suivi

L'algorithme de adaptation / suivi de mouvements faciaux, *simultaneous* souffre d'une complexité algorithmique élevée. Nous en avons proposé une modification afin de tenter de pallier ce problème tout en conservant sa précision. L'idée est de remplacer la partie la plus coûteuse de l'algorithme par une heuristique.
La partie la plus coûteuse correspond au calcul des paramètres de mise à jour $(\Delta v, \Delta\lambda)$. Dans la version originale, elle utilise une matrice approximant (au sens de Gauss-Newton) une hessienne :

$$[\Delta\boldsymbol{v}, \Delta\boldsymbol{\lambda}]^T = -\boldsymbol{H}^{-1}\sum_x SD^T(x)E(x)$$

où

$$\boldsymbol{H} = \sum_x SD(x)^T SD(x)$$

et est calculée en $O((n+m)^2N)$ pour n vecteurs de forme, m vecteurs d'apparence et une image s_0 de résolution N pixels.

Régulation. Nous proposons de remplacer cette formulation par la suivante :

$$[\Delta\boldsymbol{v}(t), \Delta\boldsymbol{\lambda(t)}]^T = -\boldsymbol{C}(t-1)\odot\sum_x SD^T(x)E(x)$$

où $\boldsymbol{C}$ est un vecteur de coefficients régulateurs et $\odot$ le produit membre à membre. L'idée est alors de calculer indépendamment l'évolution de chaque paramètre (chacun des v_i ou λ_i). En considérant l'espace global d'optimisation, si le sens de recherche d'un paramètre n'a pas changé d'une itération à l'autre, cela signifie qu'il n'a pas encore atteint son minimum. Dans ce cas, sa progression est accélérée en augmentant le poids de son coefficient correspondant c_i. En revanche, si le sens de recherche a changé, c'est qu'un optimum a été passé. La paramètre est envoyé dans le sens inverse et dans ce cas il est freiné en diminuant son coefficient c_i. L'algorithme est résumé (pour les paramètres de forme, il en va de même pour les paramètres d'apparence) de la manière suivante :

for $i = 1$ to n **do**
  if $\Delta v_i(t-1)\Delta v_i(t) > 0$ **then**
    $c_i(t) \leftarrow c_i(t-1)\eta_{inc}$
  else
    $c_i(t) \leftarrow c_i(t-1)/\eta_{dec}$
  end if
end for

Les paramètres η_{inc} et η_{dec} sont fixés empiriquement.
Avec une telle reformulation, le coût de calcul d'une itération est négligeable devant la complexité originale. Cependant, plus d'itérations sont nécessaires pour atteindre un résultat similaire à l'algorithme original.
Les deux variantes de l'algorithme ont été testées sur 40 images de visage frontaux inexpressif (cas particulier simple à mettre en place). Afin de tester la qualité d'adaptation atteinte par les deux algorithmes, une nouvelle définition de la vérité terrain a été définie. En effet, une erreur commune consite à considérer une seule annotation manuelle de visage comme étant la vérité terrain. Or le processus d'annotation manuelle introduit du bruit. Certains points sont en effet difficiles à identifier. Pour pallier ce problème, chaque visage a été annoté 11 fois et les coordonnées de référence sont constitués de la moyenne des coordonnées de toutes les annotations. De cette manière, le bruit est réduit.
De plus, il est possible de considérer les matrices de covariance associées à chacune des coordonnées dans le but de définir une mesure de tolérance. Si le résultat d'une adaptation donne un point mal placé alors qu'il a été précédemment bien placé 11 fois, ce point est penalisé. A l'inverse une erreur de placement sur un point mal défini est moins pénalisée.
Ainsi, le score de qualité d'adaptation d'un modèle à un visage est basé sur la somme des distances de Mahalanobis.
Il s'avère que les deux algorithmes sur le cas particulier de visages inconnus de la statistique ont des performances similaires en temps de calcul et en précision (voir [5] pour détails). Cependant, la variante basée sur une heuristique reste prometteuse, en particulier dans le cas où beaucoup de vecteurs d'apparence sont retenus, ce qui est un scénario réaliste quand il s'agit de traiter les variations importantes d'apparence.

3.2 Application : anonymisation

Bien qu'il reste des problèmes à traiter pour être capable d'étendre les algorithmes d'adaptation / suivi au cas de la langue des signes (en particulier concernant l'adaptation automatique en présence d'occultations), il est possible d'envisager dès à présent des applications intéressantes.
En attendant une forme écrite de la langue des signes, la vidéo est le médium de prédiléction des sourds signants pour la communication. Un besoin se fait sentir cependant : la protection de l'anonymat. En effet, alors qu'il est trivial de masquer l'identité d'une personne sur une vidéo, via le recours à des traitements simples (flou, mosaïque), la transposition au contexte de la langue des signes n'est pas possible, puisque dégrader le visage d'un locuteur dégraderait par la même occasion ses expressions et gênerait donc la compréhension.

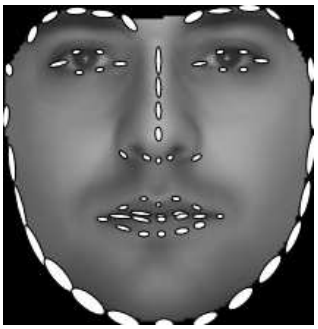

FIG. 4 – Représentation de la verité terrain d'un visage (ici le visage moyen). Les ellipses sont centrées sur les coordonnées des points de référence et leur taille représente la covariance.

Des résultats préliminaires permettent d'envisager un traitement ne modifiant que l'identité d'un visage sans en modifier l'expression.
Grâce aux AAM, un visage peut être representé par un vecteur de forme v et un vecteur d'apparence λ. Par la suite, nous désignons par p le vecteur de forme ou le vecteur d'apparence.
Un visage quelconque peut donc être représenté comme étant ce même visage neutre en expression auquel est ajouté une expression :

$$\boldsymbol{p} = \bar{\boldsymbol{p}}_i + \sum_{i=1}^{m} e_i \boldsymbol{p}_{e_i}$$

Ou, en notation matricielle :

$$\boldsymbol{p} = \bar{\boldsymbol{p}}_i + \boldsymbol{P}\boldsymbol{e}$$

où $\boldsymbol{P}$ est une matrice qui contient tous les $\boldsymbol{p}_{e_i}$ et $\boldsymbol{e}$ est un vecteur contenant tous les e_i.
Trouver les inconnues de ce problème nécessite d'ajouter des contraintes. Nous contraignons donc le problème de telle manière que les vecteurs p_{e_i} soient orthonormaux. Par conséquent, la solution est donnée par analyse en composantes principales (voir [6] et [7] pour les détails).
Pour un nouveau visage $\boldsymbol{q}$, en supposant que son visage neutre correspondant $\bar{\boldsymbol{q}}$ est connu, il est possible d'extraire le vecteur d'expression e et alors de changer la partie identitaire p_i par une autre identité p_j.

1. $e = \boldsymbol{P}^T(\boldsymbol{q} - \bar{\boldsymbol{q}})$
2. $\hat{\boldsymbol{q}} = \bar{\boldsymbol{p}}_j + \boldsymbol{P}\boldsymbol{e}$

Des résultats préliminaires ont été obtenus sur la base de visages MMI ([8]) sur 15 *actions units* (choisies pour leur fréquence d'occurence en langue des signes) et 6 individus. Les résultats sont encourageants puisque les expressions sont conservées alors que l'identité est complètement modifiée.

FIG. 5 – Résultat d'anonymisation. A partir du résultat de l'adaptation d'un AAM à un visage (à gauche), il est possible de modifier la partie identitaire du visage sans en modifiant l'expression.

4 Conclusion et perspectives

Nous avons présenté le problème d'analyse d'expressions faciales dans le contexte d'une communication vidéo en langue des signes. En utilisant un formalisme basé modèles (AAM), il est possible de suivre les déformations faciales d'un locuteur au cours d'une vidéo. Bien que des problèmes restent à résoudre (généralisation d'un modèle à un visage inconnu, voire à des expressions inconnues, prise en compte automatique des occultations des mains), les algorithmes existants permettent déjà d'envisager des traitements intéressants.

Nous avons présenté un traitement possible exploitant le résultat de ces algorithmes d'adaptation / suivi spécifique à la langue des signes et concernant le rendu anonyme d'un visage sur une vidéo sans dégradation de la partie expressive langagière.

Concernant le traitement des occultations par les méthodes d'adaptation / suivi, des travaux sont en cours aussi bien concernant la modification des algorithmes originaux (par la prise en compte d'une fonction d'erreur robuste et d'une connaissance a priori sur la provenance des *outliers*) que par l'ajout de connaissances spécifiques au contexte : les occultations sont générées par les mains et uniquement par les mains dans notre cas.

L'application d'anonymisation ne peut être envisagée qu'avec une chaîne de traitements complète permettant de modifier chaque image d'une vidéo. La qualité langagière du rendu doit être validée ainsi que le taux de dégradation de l'identité (jusqu'à quel point le locuteur est-il difficile à identifier ?).

De plus, d'autres applications utilisant le résultat des algorithmes d'adaptation / suivi restent à développer, en particulier l'annotation automatique des expressions en termes d'*action units* à des fins d'études pédagogiques ou linguistiques ou encore l'animation faciale de personnages virtuels signants.

Références

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

[2] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 681–685, 2001.

[3] I. Matthews and S. Baker, "Active appearance models revisited," Tech. Rep. CMU-RI-TR-03-02, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, April 2003.

[4] S. Baker, R. Gross, and I. Matthews, "Lucas-kanade 20 years on : A unifying framework : Part 3," Tech. Rep. CMU-RI-TR-03-35, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, November 2003.

[5] H. Mercier, J. Peyras, and P. Dalle, "Toward an efficient and accurate AAM fitting on appearance varying faces," in *7th International Conference on Automatic Face and Gesture Recognition - FG2006*, (Southampton, United Kingdom), April 2006.

[6] H. Mercier and P. Dalle, "Face analysis : identity vs. expressions," in *2e Congrès de l'International Society for Gesture Studies (ISGS) : Interacting Bodies / Corps en interaction*, (Lyon, Ecole normale supérieure Lettres et Sciences humaines), Juin 2005.

[7] N. Costen, T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Automatic extraction of the face identity-subspace.," *Image Vision Comput.*, vol. 20, pp. 319–329, 2002.

[8] L. Maat, R. Sondak, P. Gaia, and M. Pantic, "Mmi face database." BS Thesis, 2004. Man-Machine Interaction Group, Delft University of Technology, Delft.