Vérification audiovisuelle de l'identité

Rémi Landais, Hervé Bredin, Leila Zouari, et Gérard Chollet

École Nationale Supérieure des Télécommunications,
Département Traitement du Signal et des Images, Laboratoire CNRS LTCI
46 rue Barrault, 75634 PARIS Cedex 13 - FRANCE.
Tél : int+ 33 1 45 81 72 63, Fax : int+ 33 1 45 81 37 94
remi.landais@enst.fr

Résumé Un nouveau système de vérification d'identité, tirant parti de la fusion des modalités *visage*, *voix* et *synchronie* est présenté. Chacun des 3 systèmes est décrit : vérification des visages basée sur l'utilisation conjointe d'une représentation globale (par *eigenfaces*) et locale (par des descripteurs SIFT) ; vérification du locuteur par modèles de mélange de gaussiennes (GMM) à l'aide de la boîte à outil BECARS et authentification par analyse de la synchronie entre le mouvement des lèvres et la voix. L'ensemble est intégré par fusion *a posteriori* des scores de chaque système. Le système est évalué sur la base de données biométrique multimédia BANCA.

Mots clés Biométrie, multimodalité, vérification d'identité, fusion, vidéo, BANCA.

1 Introduction

La méthode de vérification d'identité multimodale proposée dans cet article s'appuie sur trois modalités différentes : *visage*, *voix* et *synchronie*. Cette dernière modalité utilise la corrélation entre le signal de parole acoustique et le signal visuel (mouvement des lèvres). L'utilisation de la synchronie permet de faire face à de nouvelles menaces, notamment lorsqu'un imposteur dispose d'un enregistrement vocal de la personne dont il cherche à usurper l'identité.
Nous reviendrons successivement sur chacune de ces modalités. La vérification du *visage* utilise conjointement une représentation globale et une représentation locale des visages. La vérification du locuteur est basée sur les modèles de mélange de gaussiennes (GMM) avec la boîte à outil BECARS [2]. La méthode d'analyse de la synchronie est, quant à elle, basée sur l'estimation de la corrélation *lèvres/voix*. La dernière partie portera sur l'exposé des résultats obtenus sur la base vidéo BANCA [1].

2 Modalité *visage*

La base de données BANCA est constituée de séquences vidéo de personnes énonçant un texte devant une caméra équipée d'un microphone. De cette façon, toutes les trames de la vidéo sont disponibles pour mener à bien la vérification d'identité.

2.1 Méthode de détection de visages

La première étape obligatoire consiste à détecter le visage dans chacune des trames de la vidéo. Elle est ici réalisée à l'aide de la boîte à outil *Machine Perception Toolbox* dont l'algorithme de détection repose sur les modèles génératifs [6]. Sachant que chaque vidéo fait apparaître un unique visage, la majorité des fausses alarmes est supprimée en conservant la plus grande zone détectée. Un filtre temporel médian évite finalement d'obtenir des positions aberrantes et permet de produire une position du visage dans les images où celui-ci n'a pas été initialement détecté.
Le système détecte ensuite la position des yeux en s'appuyant sur une méthode similaire en prenant l'*a priori* que deux yeux exactement doivent être détectés. La position des yeux ainsi obtenue permet de normaliser les visages par alignement spatial et égalisation d'histogramme.
Une phase de filtrage des visages détectés permet de réduire le nombre de résultats erronés pouvant nuire au processus de vérification. L'inverse de la distance entre chaque zone détectée et sa projection dans l'espace de visage défini par les *eigenfaces* (voir la section suivante) est utilisé comme indice (noté *Rel*) de la qualité de la détection. En effet, une zone détectée représente effectivement un visage si celle-ci est proche (au sens euclidien) de sa projection dans l'espace des visages. L'ensemble des résultats est alors classé selon l'indice *Rel* et seules les zones dont l'indice est compris entre l'indice maximal Rel_{max} (estimé sur la séquence vidéo) et $\alpha \times Rel_{max}$ (α fixé expérimentalement à 2/3) sont conservées. Par la suite, seuls les 100 meilleurs visages pour l'analyse par *eigenfaces* et les 5 meilleurs pour l'analyse par descripteurs SIFT (un plus grand nombre dégradant les résultats) sont conservés.

2.2 Représentation globale des visages

La méthode des *eigenfaces* [14] constitue un ajustement de l'analyse en composantes principales (PCA) à la reconnaissance de visages. Etant donné un ensemble d'apprentissage de visages de face, la PCA permet d'obtenir un espace dans lequel la dispersion des données est maximisée. Les directions définissant cet espace constituent les *eigenfaces*. Une fois la dimension de l'espace de visage (le nombre d'*eigenfaces*) choisie, chaque nouvelle image peut y être projetée, les coefficients de projection constituant alors sa représentation globale. L'ensemble d'apprentissage utilisé contient 300 visages de la base BANCA (30 personnes) ainsi que 500 environ de la base BIOMET [7] (130 personnes).

2.3 Représentation locale des visages

Les descripteurs SIFT comptent parmi les descripteurs locaux les plus efficaces [11]. Dans un premier temps, les points robustes aux changements d'échelle sont extraits des images en s'appuyant sur leur représentation *scale-space* [9]. L'étape suivante consiste à préciser la position de ces points et à déterminer l'échelle leur étant associée. L'ensemble de points obtenu est filtré selon des contraintes liées au contraste et à la géométrie locale.

Finalement, un vecteur de description (de dimension 128 dans cette étude) est associé à chaque point retenu en analysant l'orientation et la magnitude du gradient dans son voisinage. Par ailleurs, un vecteur à 4 composantes, incluant la position, l'échelle ainsi que l'orientation, est associé à chaque point clé extrait.

2.4 Comparaison des représentations par une méthode de *matching* basée sur la décomposition SVD

La méthode de *matching* basée sur la décomposition en valeurs singulières (SVD) [13] permet d'associer les points d'intérêt extraits de deux images différentes. Elle s'appuie sur l'analyse d'une matrice de proximité : $G_{ij} = g(R_{ij}) = \exp^{-R_{ij}^2/2\sigma^2}$ où R_{ij} définit la distance euclidienne entre les points i et j. Les associations recherchées sont *exclusives* : un point d'une image est associé à un unique point de l'autre image et inversement. Pour faciliter la recherche de ces paires de points, l'idée est de rechercher une projection permettant de *rapprocher* la matrice G de la matrice identité. Cette recherche est facilitée par une procédure dérivée de la solution au *problème orthogonal de Procrustes* [8] :

1. Calculer la SVD de la matrice R : $G = UDV'$
2. Calculer la matrice Q définie par $Q = UV'$
3. Rechercher les paires (i, j) telles que Q_{ij} soit le maximum de la ligne i et de la colonne j.

Une première amélioration est proposée dans [12] : la matrice G prend la forme $G_{ij} = f(C_{ij}) * g(R_{ij})$, où C_{ij} définit la corrélation entre les niveaux de gris des voisinages des points d'intérêt i et j ; et où f peut prendre une forme exponentielle ($f(C_{ij}) = \exp^{-(C_{ij}-1)^2/2\gamma 2}$) ou linéaire $f(C_{ij}) = (C_{ij} + 1)/2$. Un seuil sur la valeur de la corrélation C_{ij} permet de conserver uniquement les meilleurs *matchings*.
Une seconde amélioration consiste à prendre en compte la corrélation entre les descripteurs SIFT associés aux points d'intérêt [4]. Le *matching* de notre méthode basée sur les descripteurs SIFT relève de cette dernière amélioration, tout en considérant pour le calcul des distances R_{ij}, les vecteurs à 4 composantes précédemment cités.
Concernant les représentations *eigenfaces*, la même méthode de *matching* est mise en oeuvre à ceci prêt que la composante *euclidienne* de la matrice G n'est plus prise en compte. À la différence des *matchings* SIFT établis entre différents points d'intérêts, les associations produites par l'application de la méthode de *matching* SVD dans le cas de représentations *eigenfaces* sont établies entre les images des vidéos considérées lors de la vérification (la vidéo *modèle* et la vidéo de *test*).
Cette méthode met en jeu de nombreux paramètres. Les paramètres σ et γ sont fixés selon les recommandations de Pilu [12], validées expérimentalement. Les autres paramètres sont estimés par validation croisée entre les deux groupes composant la base BANCA [10]. Le seuil sur la corrélation C_{ij} est ainsi fixé à 0.4 ; la forme de la fonction f optimale est exponentielle pour la méthode SIFT et linéaire pour la méthode *eigenfaces*. Enfin, la dimension de l'espace de projection pour la méthode *eigenfaces* est fixée à 97.
Quelle que soit la méthode *visage* appliquée, le nombre de *matchings* calculés constitue le

score de vérification. Dans le cas des descripteurs SIFT, les représentations des 5 images retenues dans chaque vidéo sont comparées deux à deux. 25 scores de vérification sont ainsi obtenus pour chaque test. Chacun de ces 25 scores de matching est normalisé relativement au nombre de descripteurs contenus dans chacune des deux images concernées. Le score moyen constitue par la suite le score de vérification final. En ce qui concerne le *matching* des représentations *eigenfaces*, un unique score de *matching* (entre images des deux vidéos) est mesuré et utilisé comme score final de vérification.

3 Modalité *voix*

La vérification du locuteur repose sur l'utilisation des modèles de mélange de gaussiennes (GMM) et est réalisée à l'aide de la boîte à outils *open-source* BECARS [2]. Dans un premier temps, un modèle du monde Ω est construit à partir de paroles prononcées par un grand éventail de locuteurs. Ensuite, un modèle de locuteur λ est estimé par adaptation MAP (Maximum A Posteriori) du modèle du monde à l'aide de données propres à ce locuteur. Cette technique permet de surmonter le manque de données d'apprentissage disponibles pour chaque locuteur. De façon classique, 13 MFCC (*Mel Frequency Cepstral Coefficients*) sont calculés toutes les 10ms sur une fenêtre glissante de 20ms, auxquels sont concaténés les dérivées premières et secondes. Au moment du test, étant donnée l'observation x des MFCC sur un segment de test, le rapport de vraisemblance $P(x|\lambda)/P(x|\Omega)$ fournit le score de la modalité *voix*.

4 Modalité *synchronie*

L'objectif est ici de reconnaître une personne par sa façon de synchroniser sa voix et ses lèvres. La méthode est décrite en détails dans [3]. Étant donnée la séquence d'enrôlement de la personne λ, les signaux acoustique X_λ et visuel Y_λ sont extraits. Il s'agit des coefficients MFCC extraits toutes les 10ms et des coefficients DCT (*Discrete Cosine Transform*) de la zone de la bouche (localisée à l'aide de la méthode détaillée dans [3]). L'analyse de coinertie [5] (CoIA pour *CoInertia Analysis*) permet de calculer les matrices de projection A_λ et B_λ maximisant la covariance des projections des vecteurs X_λ et Y_λ, ces matrices constituant alors le modèle de la personne λ. Au moment du test, une personne ϵ prétend être la personne λ. Les caractéristiques X_ϵ et Y_ϵ associées sont calculées et transformées à l'aide des matrices de projection A_λ et B_λ de la personne λ. Le score de la modalité *synchronie* est alors obtenu par la formule suivante :

$$S_{X_\epsilon,Y_\epsilon} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{cov}(a_{\lambda,k}^T X_\epsilon, b_{\lambda,k}^T Y_\epsilon) \tag{1}$$

où K désigne la dimension de l'espace de projection retenue (c'est à dire le nombre de colonnes $a_{\lambda,k}$ et $b_{\lambda,k}$ conservés dans les matrices A_λ et B_λ).

5 Expérimentations

Les expérimentations ont été menées sur la base BANCA [1]. Elle regroupe 52 personnes divisées en 2 groupes disjoints (G1 et G2). Chacune des personnes a été enregistrée à 8 reprises (4 accès client et 4 accès imposteur) dans 3 conditions d'enregistrement différentes (*controlled*, *degraded* et *adverse*). Le protocole d'évaluation utilisé (le protocole P pour *Pooled*) met en oeuvre des enregistrements produits selon ces trois conditions et constitue ainsi le protocole le plus strict. Pour chaque groupe, 234 accès client et 312 accès imposteurs sont réalisés à l'aide de chacun des 4 systèmes (visage PCA, visage SIFT, voix et synchronie). Les 4 scores sont fusionnés à l'aide d'un SVM (pour *Support Vector Machine*) à noyau RBF : la séparation entre les classes *client* et *imposteur* est apprise sur G1 et testée sur les scores obtenus sur G2 et inversement. L'ensemble des résultats obtenus est résumé dans la figure 1.

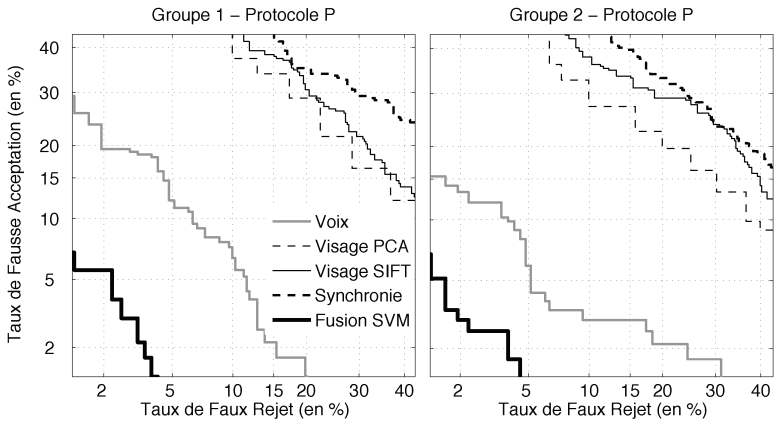

Fig. 1. Résultats des systèmes de vérification sur les données BANCA

Les courbes de la figure 1 montrent l'impact de la fusion sur les performances globales du système : si la modalité *voix* s'avère produire les meilleurs résultats, il n'en demeure pas moins que les performances sont accrues par l'ajout des modalités *visage* et *synchronie*. Conjointement aux courbes DET, nous avons calculé les Taux d'Erreur Pondérée (WER) : $WER(R) = (P_{FR} + RP_{FA})/(1 + R)$ pour plusieurs valeurs de R (0.1, 1 et 10), où P_{FR} désigne le taux de faux rejet et P_{FA} le taux de fausse acceptation. Ces résultats sont résumés dans le tableau 1 dans lequel les intervalles de confiance des mesures obtenues sont précisés. Le non-recouvrement de ces derniers permet de conclure sur le caractère significatif de l'amélioration apportée par la fusion multimodale.

6 Conclusion

Le système de vérification présenté dans cet article s'appuie sur trois modalités : *voix*, *visage* et *synchronie*. Une originalité de ce travail relève de l'utilisation conjointe

Tab. 1. WER obtenus selon les différentes modalités et selon leur fusion.

	R=0.1		R=1		R=10		
Groupe	G1	G2	G1	G2	G1	G2	Moyenne
Voix	2.79	3.65	8.29	5.14	3.79	2.37	4.34 [3.69-5.09]
Eigenfaces	8.25	10.19	23.44	19.80	7.10	6.33	12.52 [11.43-13.70]
SIFT	8.98	8.51	25.76	24.66	7.73	7.69	13.89 [12.75-15.12]
Synchronie	9.73	9.03	27.24	26.10	7.17	7.41	14.45 [12.49-16.66]
Fusion	**0.90**	**2.59**	**3.74**	**2.51**	**2.43**	**0.78**	**2.16 [1.45-3.21]**

de représentations globale et locale pour la vérification des visages. L'utilisation de la modalité *synchronie* constitue la seconde contribution. Les résultats obtenus sur la base BANCA sont satisfaisants puisque ils montrent clairement l'apport de la fusion multimodale relativement à l'utilisation de systèmes mono-modaux. Le travail futur consistera à améliorer chacune des briques constituant le système actuel. Il sera notamment envisagé de prendre en compte un ensemble d'apprentissage plus grand pour la détermination des *eigenfaces* et de définir une méthode de *matching* qui puisse prendre en compte des associations multiples.

Références

1. E. Bailly-Baillière and S. Bengio *et al.* The BANCA Database and Evaluation Protocol. In *Lecture Notes in Computer Science*, volume 2688, pages 625 – 638, January 2003.
2. R. Blouet, C. Mokbel, H. Mokbel, E. Sanchez, and G. Chollet. BECARS : a Free Software for Speaker Verification. In *ODYSSEY 2004*, pages 145 – 148, 2004.
3. H. Bredin and G. Chollet. Audio-Visual Speech Synchrony Measure for Talking-Face Identity Verification. In *Proc. of the IEEE Int. Conf. on Acoustic, Speech and Signal Processing*, 2007.
4. E. Delponte, F. Isgrò, F. Odone, and A. Verri. SVD-Matching using SIFT Features. In *Proc. of the Int. Conf. on Vision, Video and Graphics*, pages 125–132, 2005.
5. S. Dolédec and D. Chessel. Co-Inertia Analysis : an Alternative Method for Studying Species-Environment Relationships. *Freshwater Biology*, 31 :277–294, 1994.
6. I. Fasel, B. Fortenberry, and J.R. Movellan. A Generative Framework for Real-Time Object Detection and Classification. *Computer Vision and Image Understanding*, pages 182–210, 2004.
7. S. Garcia-Salicetti and C. Beumier *et al.* BIOMET : a Multimodal Person Authentication Database including Face, Voice, Fingerprint, Hand and Signature Modalities. *Audio- and Video-Based Biometric Person Authentication*, pages 845 – 853, June 2003.
8. G.H. Golub and C.F. Van Loan. *Matrix Computations 3rd Edition*.
9. J. Koenderink. The Structure of Images. *Biological Cybernetics*, 50 :363–370, 1984.
10. R. Landais, H. Bredin, and G. Chollet. Multilevel Face Verification involving SIFT Descriptors and Eigenfaces, 2007. (soumis à IAPR/IEEE International Conference on Biometrics).
11. D.G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Int. Journal of Computer Vision*, 60(2) :91–110, 2004.
12. M. Pilu. A Direct Method for Stereo Correspondence based on Singular Value Decomposition. In *Proceedings of CVPR*, pages 261–266, 1997.
13. G.L. Scott and H.C. Longuet-Higgins. An Algorithm for Associating the Features of Two Images. *Proc. of the Royal Society of London. Series B. Biological Sciences*.
14. M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1) :71 – 86, 1991.