Introduction de paramètres dynamiques en reconnaissance faciale

Federico MATTA Jean-Luc DUGELAY

Institut Eurécom
2229 route des Crêtes
06904 Sophia Antipolis, France

{Federico.Matta,Jean-Luc.Dugelay}@eurecom.fr

Résumé

Dans cet article, nous présentons un système multimodal de reconnaissance de la personne à partir de vidéo, en intégrant deux approches complémentaires. Le premier module exploite l'information comportementale : il est basé sur des signaux de déplacement de la tête, automatiquement extraits à partir d'une séquence vidéo ; des mesures statistiques sont alors calculées et utilisées pour la reconnaissance. Le second module exploite l'information physique : c'est une extension probabiliste de l'approche classique d'Eigenface, dans lequel la reconnaissance est faite dans un espace réduit des visages, calculé en employant une transformation d'analyse en composantes principales (ACP). Pour une fusion cohérente, les deux systèmes partagent le même cadre probabiliste de classification : une approximation avec un modèle de mélange de gaussiennes (MMG) et un classificateur bayésien. Les résultats expérimentaux prouvent que les déplacements de la tête sont discriminants et que l'intégration multimodale apporte beaucoup en reconnaissance.

Mots clefs

Identification des personnes, reconnaissance de visage, reconnaissance d'objet, vidéo.

1 Introduction

La biométrie a connu ces dernières années un nouvel essor. Parmi les différentes biométries étudiées, le visage présente l'intérêt d'être bien accepté par les utilisateurs et sans contact ; par contre, les performances associées sont assez faibles par rapport aux empreintes digitales ou l'iris. De nombreux algorithmes ont été proposés en visage et travaillent sur l'apparence ; ces approches existantes sont sensibles à différentes sources de variabilité, par exemple : les changements d'expression, d'illumination et de pose. Une analyse détaillée de la reconnaissance de la personne en utilisant des images fixes du visage, ses résultats et ses limitations peuvent être trouvés dans [1] et [2].

La reconnaissance d'une personne par le biais de la vidéo présente quelques avantages par rapport à celle basée sur l'image. D'abord, l'information temporelle des visages peut être exploitée afin de faciliter la tâche de reconnaissance ; par exemple, les caractéristiques dynamiques, qui sont spécifiques à chaque personne, comme le mouvement de la tête, l'évolution de la pose ou la mimique du visage. En second lieu, des représentations plus efficaces, telles que des modèles 3D de visages ou des images de résolutions augmentées, peuvent être obtenus à partir des séquences vidéo et être employés pour améliorer les performances des systèmes. Enfin, la reconnaissance basée sur la vidéo permet d'apprendre ou de mettre à jour les modèles dans le temps.

En ce qui concerne les travaux sur ce sujet, une grande partie des techniques de reconnaissance de visage basée sur la vidéo sont des généralisations directes des algorithmes de reconnaissance sur images fixes. Dans ces systèmes, la stratégie de reconnaissance est appliquée indépendamment sur chaque trame, sans prendre en compte l'information temporelle incluse dans la vidéo. Parmi les tentatives visant à adresser le problème de la reconnaissance de la personne d'une façon plus systématique et plus cohérente, les méthodes par Li et Chellappa [3], Zhou et al. [4] et Lee et al. [5] sont les plus remarquables : tous développent des méthodes du suivi et reconnaissance en utilisant un cadre probabiliste unifié.

C'est une tendance commune en littérature d'exploiter seulement une partie de l'information visuelle. En fait dans nos travaux de recherche, [2] et [6], et dans la majorité des articles publiés, les systèmes de reconnaissance ont été basés soit sur l'information physique (aspect facial) soit sur l'information temporelle (mouvement facial). Vu le potentiel de ces deux modalités, l'évolution normale de la reconnaissance d'individus basée sur la vidéo est de s'orienter sur l'étude d'un système multimodal, qui exploite toute l'information visuelle. Nous présentons donc une approche originale, qui combine à la reconnaissance faciale classique basée sur l'apparence des paramètres dynamiques liés aux mouvements de la tête, afin de développer un système plus discriminant et plus robuste.

La suite de notre article est organisée comme suit : dans la section 2, nous détaillons notre système de reconnaissance, puis dans la section 3, nous présentons et commentons des expériences. Finalement, dans la section

4, nous concluons cet article par des remarques et la présentation de nos travaux futurs.

2 Description du système de reconnaissance

Notre système de reconnaissance de l'individu est composé de trois modules : un module de reconnaissance statique, qui reconnait les personnes à partir de leur apparence, un autre module de reconnaissance temporelle, qui les reconnaît en utilisant le mouvement de la tête, et enfin, un module de fusion, qui réalise l'identification et la vérification du sujet en intégrant les deux modalités précédentes.

2.1 Système de reconnaissance statique

Notre algorithme de reconnaissance, basé sur l'apparence, est une extension probabiliste de l'approche classique d'Eigenfaces, présentée par Turk et Pentland dans [7].
En reprenant la technique originale, nous calculons l'analyse en composantes principales (APC) sur un assortiment d'images de visage, afin d'obtenir un ensemble de vecteurs orthogonaux (les vecteurs propres) qui représentent de façon optimale la distribution des données au sens des moindres carrées ; ces vecteurs définissent le sous-espace des images de visage, que nous appelons l'espace propre des visages. Une nouvelle image du visage est transformée en ses composantes dans l'espace propre des visages par une simple projection ; nous noterons $\mathbf{y}_k$ le vecteur projeté pour l'image k.
Puis, nous améliorons le système de reconnaissance original en employant un cadre bayésien pour l'étape de classification. Premièrement, pour chaque individu, nous voulons modéliser la distribution de ses images dans l'espace propre des visages. Nous approximons la fonction de densité de probabilité conditionnelle à la classe de chaque individu en employant un modèle de mélange de gaussiennes (GMM). Nous pouvons exprimer cela en utilisant la formule suivante :

$$P(\mathbf{y}_k \mid \varpi_q) = \sum_{c=1}^{C} \alpha_c \aleph(\mathbf{y}_k, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

où ϖ_q exprime la classe (l'individu) q et α_c est le poids de la c-ième composante gaussienne, $\aleph$. Ensuite, pour tester une image donnée, nous calculons les probabilités logarithmiques β a posteriori pour chaque classe q :

$$\beta_{q,k} = \log\left(P(\varpi_q \mid \mathbf{y}_k)\right) = \log\left(\frac{P(\mathbf{y}_k \mid \varpi_q)P(\varpi_q)}{P(\mathbf{y}_k)}\right)$$

Notons que les probabilités a priori et les facteurs d'échelle, présentés dans la formule précédente, sont directement estimés à partir de la base de données d'entraînement.

2.2 Système de reconnaissance temporel

Pour l'utilisation de l'information temporelle contenue dans la vidéo, nous proposons un nouveau système d'identification d'individus, basé sur les signaux de déplacement de quelques éléments de la tête, automatiquement extraits de la séquence vidéo. Notre système de reconnaissance temporelle est composé de trois sous-modules : un analyseur visuel afin d'obtenir des signaux de déplacement, un extracteur de vecteurs caractéristiques pour calculer des paramètres discriminants et un classificateur de personne pour discerner les identités.

2.2.1 Le module d'analyse visuel

Le module d'analyse visuel prend en entrée une séquence vidéo de quelques secondes d'un présentateur de télévision. La détection de la tête est semi-automatique : l'utilisateur doit manuellement cliquer sur des points d'intérêt (du visage) dans la première image ; ensuite, un algorithme de suivi opère jusqu'à la fin de la séquence. En fait, les signaux de déplacement sont automatiquement extraits en utilisant une technique d'appariement de blocs dans l'espace couleur RVB. La mesure de similarité est obtenue en additionnant les distances euclidiennes calculées pour chaque composante de couleur (poids des composantes égaux). Si $\mathbf{T}_t$ est le bloc courant, $\mathbf{T}_{t-1}$ le précédent, $\mathbf{M}_{t-1}$ le dernier appariement et α une constante de pondération, alors le contenu du bloc est mis à jour avec la formule suivante :

$$\mathbf{T}_t = \alpha \mathbf{M}_{t-1} + (1-\alpha)\mathbf{T}_{t-1}$$

On peut facilement vérifier que le bloc courant est une somme pondérée de tous les précédents et que la formule de calcul inclut les cas limites d'aucune mise à jour ($\alpha = 0$) et de mise à jour complète ($\alpha = 1$).

2.2.2 Le module d'extraction des vecteurs caractéristiques

Le module d'extraction des vecteurs caractéristiques analyse les signaux bruts du suivi des différents éléments de la tête, calculés à partir d'une séquence vidéo.
Afin d'obtenir les vecteurs caractéristiques, le système applique quelques transformations globales aux signaux de déplacement pour les normaliser et fournir une meilleure représentation pour la classification. Par défaut, ce module centre les signaux et uniformise les échelles pour enlever toute dépendance par rapport à la position absolue et à la dimension de la tête dans la vidéo. Il est également possible d'imposer une contrainte de variance uniforme pour tous les signaux, d'utiliser des coordonnées polaires ou de rajouter des paramètres obtenus en calculant des dérivés (vitesses ou accélérations). Il est important de noter que chaque signal de déplacement a deux composantes : horizontale et verticale. De ce fait, la dimension du vecteur F est le double du nombre des

points d'intérêt du visage analysé. Dans la section suivante, nous allons exprimer tous les vecteurs caractéristiques d'une personne extraits à partir de la k – ième vidéo, avec la notation suivante :

$$\mathbf{X}^{(k)} = \begin{bmatrix} \mathbf{x}_1^{(k)} \\ \vdots \\ \mathbf{x}_T^{(k)} \end{bmatrix}$$

où T est le nombre total de trames dans la vidéo et $\mathbf{x}_t$ est un vecteur ligne contenant les valeurs des paramètres caractéristiques normalisés pour la trame t.

2.2.3 Le module de classification d'individu

Le dernier module exploite les vecteurs caractéristiques calculés à partir d'une séquence vidéo pour la reconnaissance.

Les vecteurs caractéristiques, qui contiennent l'information des déplacements de la tête, sont utilisés pour entraîner un modèle de mélange de gaussiennes (MMG) pour chaque personne présente dans la base de données afin de modéliser le mouvement caractéristique pour cet utilisateur. Plus précisément, l'algorithme estime la fonction de densité de probabilité conditionnelle dans un classificateur bayésien. La probabilité a posteriori pour la classe ω_q s'exprime sous la forme :

$$P(\omega_q \mid \mathbf{x}_t) = \frac{P(\mathbf{x}_t \mid \omega_q) P(\omega_q)}{P(\mathbf{x}_t)}$$

La mesure de similarité pour chaque vidéo est calculée en faisant l'hypothèse que les déplacements sont indépendants (hypothèse généralement non vérifiée) et en prenant le produit des différentes probabilités :

$$P(\omega_q \mid \mathbf{X}) \cong \prod_{t=1}^{T} P(\omega_q \mid \mathbf{x}_t)$$

Les fonctions de densité de probabilité conditionnelle $P(\omega_q \mid \mathbf{x}_t)$ pour chaque trame sont approximées avec un modèle de mélange de gaussiennes (MMG) qui s'exprime sous la forme :

$$P(\varpi_q \mid \mathbf{x}_t) = \sum_{c=1}^{C} \alpha_c \aleph(\mathbf{x}_t, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

où α_c représente le poids de la c –ième composante gaussienne, $\aleph(\mathbf{x}_t, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$.

Pour intégrer ce module dans notre système multimodal, nous récupérons l'ensemble des probabilités logarithmiques a posteriori, calculées lors de la classification bayésienne. Le produit de ces log-probabilités sera noté γ :

$$\gamma_{q,k} = \log\left(P(\omega_q, \mid \mathbf{X}^{(k)})\right)$$

dans laquelle ϖ_q exprime la classe (l'individu) q et $\mathbf{X}^{(k)}$ contient les vecteurs caractéristiques de la vidéo k.

2.3 Module de fusion statique et temporelle

Le module de fusion intègre les deux mesures de similarité (probabilités logarithmiques a posteriori) des sous-systèmes précédents et calcule les taux d'identification et vérification du système multimodal. Dans cet article, les valeurs de similarité multimodales sont obtenues par fusion en utilisant deux versions de la sommation [8] qui, dans notre cas, a la forme générale suivante :

$$\theta_{q,k} = b_{q,k}\beta_{q,k} + g_{q,k}\gamma_{q,k}$$

ou $b_{q,k}$ et $g_{q,k}$ sont deux poids.

Dans la première version, nous calculons la moyenne entre les valeurs de similarité des deux sous-systèmes (poids égaux) :

$$b_{q,k} = g_{q,k} = 0.5 \quad \forall q,k$$

Cette technique simple a une interprétation probabiliste intéressante. Si nous supposons que $\mathbf{X}^{(k)}$ et $\mathbf{y}_k$ sont indépendants entre eux et également distribués, et que toutes les classes sont équiprobables (un scénario commun dans des applications réelles) alors la mesure de similarité $\theta_{q,k}$ est la probabilité logarithmique a posteriori conjointe de $\mathbf{X}^{(k)}$ et $\mathbf{y}_k$:

$$\theta_{q,k} = 0.5\log\left(P(\omega_q \mid \mathbf{y}_k, \mathbf{X}^{(k)})\right) + Q$$

où Q est une constante de translation.

La deuxième version de la mesure de similarité pour le système multimodal est une pondération adaptative, proposé par Chang et al. [9] et calculée comme suit :

$$b_{q,k} = \frac{\beta_k^{1st} - \beta_k^{2nd}}{\beta_k^{1st} - \beta_k^{3rd}} \quad \forall q$$

$$g_{q,k} = \frac{\gamma_k^{1st} - \gamma_k^{2nd}}{\gamma_k^{1st} - \gamma_k^{3rd}} \quad \forall q$$

où β_k^i et γ_k^i sont le i –ièmes meilleures valeurs pour le test k. L'idée générale de ce choix des poids est que, si la différence entre les premières et deuxièmes valeurs de similarité est grande par rapport à la similarité moyenne, alors la modalité peut être considérée comme fiable et son poids est grand. Dans notre calcul, nous normalisons les poids pour faire en sorte que la somme soit égale à 1 :

$$b_{q,k} + g_{q,k} = 1 \quad \forall q,k$$

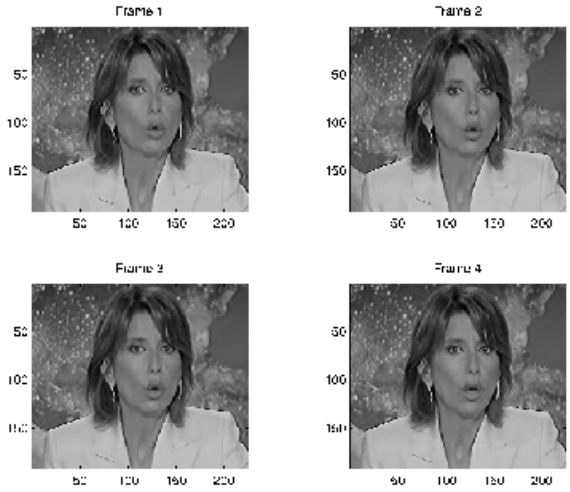

Figure 1 – *Les 4 premières images d'une séquence vidéo.*

3 Expériences et résultats

3.1 Base de données

3.1.1 Vidéo

Malheureusement, les bases de données vidéo existantes ne sont pas adaptées pour tester efficacement nos algorithmes. En effet, le système de reconnaissance temporelle (i.e. biométrie comportementale) a besoin de quelques minutes pour chaque individu afin d'extraire l'information temporelle et entraîner les modèles de MMG. Pour cette raison, nous avons rassemblé un ensemble de 192 séquences vidéo de 12 personnes différentes pour pouvoir entraîner et tester notre système. Les différents extraits vidéo montrent des présentateurs du télévision annonçant les nouvelles du jour. Ces vidéos ont été extraites des différents journaux télévisés sur une période de 14 mois. Une séquence typique a une résolution spatiale de 352x288 pixels, une résolution temporelle de 23.97 trames/seconde et une durée d'environ 14 secondes (se référer à la Figure 1 pour un exemple). Bien que les vidéos soient de mauvaise qualité, comprimées à 300 kbits/seconde (y compris l'audio), l'approche comportementale de notre système est moins affectée par les défauts visuels, introduits pendant le processus de compression, que les méthodes basées sur l'apparence. D'ailleurs, les vidéos sont réalistes, dans la mesure où le comportement des présentateurs est normal, sans aucune contrainte imposée au niveau des mouvements, de la pose ou de l'action.

3.1.2 Image

Pour ce qui concerne le système de reconnaissance statique, nous avons créé une base de données d'image dérivée de la base de données vidéo présentée auparavant. Pour chaque vidéo d'entraînement, nous avons extrait 28 images (2 trames/seconde) tandis que, pour l'ensemble de test, nous utilisons seulement la première trame. En raison de la sensibilité élevée à l'alignement facial, la variation de la pose et de l'échelle, bien connue pour les algorithmes de reconnaissance basés sur l'ACP, nous avons manuellement normalisé la base de données d'image en isolant la région du visage, en alignant les yeux puis en positionnant les têtes bien horizontalement.

3.2 Cadre expérimental

3.2.1 Système temporel

Dans nos expériences, nous avons choisi 96 séquences vidéo pour l'entraînement de notre système de reconnaissance temporelle (8 pour chacun des 12 individus), et les 96 restantes (sur un total de 192) ont été laissées pour le test. Il est important de préciser qu'il n'y a aucune contrainte théorique en ce qui concerne le nombre de vidéos nécessaire par utilisateur et leurs durées. Par contre, il est bien nécessaire d'avoir quelques minutes par individu pour apprendre les mouvements caractéristiques et entraîner le MMG. Dans nos expériences, nous avons choisi d'extraire les déplacements horizontaux et verticaux de 4 éléments du visage - les yeux, le nez et la bouche – et obtenir ainsi 8 signaux au total.
Pour ce qui concerne la normalisation des signaux, les résultats les plus intéressants ont été obtenus en centrant les signaux et en uniformisant les échelles. En fait, des contraintes plus fortes, comme une variance fixée, ont réduit l'information distinctive et ont été abandonnées. Nous avons également essayé de calculer nos vecteurs caractéristiques en incluant des paramètres dérivés des signaux, comme vitesse et accélération, mais il n'y avait pas d'amélioration des résultats de reconnaissance par rapport à l'utilisation seule des déplacements.
Pour l'entraînement des MMG individuels, nous avons obtenu de meilleurs résultats en utilisant un algorithme classique d'Espérance-Maximisation (EM) avec 4 composantes gaussiennes par modèle. Dans nos expériences, nous ne pouvions pas utiliser plus de 9 composantes du fait de la dimension réduite de notre base de données pour un entraînement fiable de MMG.

3.2.2 Système statique

Pour ce qui concerne le système statique de reconnaissance, nous avons choisi un total de 2688 images pour l'entraînement (224 par individu), et de 96 pour le test. En réalité, l'algorithme se sert de 5376 images (448 par individu) en appliquant un effet miroir vertical sur les images originales. Dans nos expériences, nous avons été obligés de choisir un espace propre de visages de dimension 13 seulement, en raison de la difficulté d'estimation des distributions dimensionnellement élevées avec une quantité limitée de données d'entraînement. Pour la même raison, nous avons également considéré des MMG avec $1 \div 3$ composantes par modèle.

3.3 Résultats

Les résultats de reconnaissance des deux sous-systèmes statique et temporel seuls et ceux du système multimodal sont présentés dans le Tableau 1. La deuxième et troisième colonne de ce tableau représentent respectivement les valeurs d'identifications correctes de la personne : en considérant les meilleures mesures de similarité et ensuite plus largement les trois meilleures. La quatrième colonne contient le taux égal d'erreur en mode de vérification.

METHODE	IDENTIFICATION		VERIFICATION
	Meilleur	3 meilleurs	Taux d'erreur égal
Système statique	93,75%	97,92%	2,37%
Système temporel	92,71%		6,91%
Fusion avec moyenne	94,97%		2,18%
Fusion avec pondération adaptive	96,88%		2,75%

Tableau 1 – *Résultats de reconnaissance.*

Les résultats obtenus pour le sous-système temporel, basé sur les déplacements de la tête, sont intéressants. En fait, même si ces signaux pourraient être considérés comme des modalités faibles, ils obtiennent des taux de reconnaissance similaire au système statique ; ce qui montre que le comportement des personnes peut être un identificateur biométrique. D'ailleurs, notre système est appliqué dans des cas réels, en utilisant des vidéos compressées et sans aucune contrainte sur les actions de l'individu (pas de scénario ou de gestuelles prédéfinis). Notre approche comportementale a également montré une grande tolérance par rapport aux changements d'aspect, comme, par exemple, la présence de lunettes, de barbe ou pour des variations sur les coupes de cheveux et d'illumination.
En regardant les résultats, il est clair que l'intégration multimodale des systèmes spatial et temporel augmente les taux d'identification et de vérification d'individus. D'autre part, les deux méthodes de fusion ont montré des résultats similaires.

4 Conclusion et travaux futurs

Dans cet article, nous avons analysé les effets de l'utilisation combinée de l'information physique et comportementale, présentes dans les signaux vidéo, pour la reconnaissance des personnes. Tout d'abord, ce travail est original du fait de l'utilisation des déplacements de la tête et il prouve que le comportement et le mouvement humain peuvent être utiles pour distinguer les personnes. Ensuite, nos résultats expérimentaux montrent que l'intégration temporelle et statique permet une amélioration significative des taux de reconnaissance. Enfin, il faut tenir compte du fait que notre travail a encore besoin d'une plus grande validation expérimentale et cela devrait être fait en utilisant des bases de données vidéo plus complètes mais malencontreusement encore rares aujourd'hui.
Notre système peut être amélioré à plusieurs niveaux. Une possibilité serait de modifier le sous-module statique, en remplaçant l'approche basée sur l'ACP avec un algorithme de reconnaissance plus performant. Puis, le module temporel du système de reconnaissance pourrait être amélioré en ajoutant des paramètres de la mimique faciale : il pourrait intégrer le clignotement de l'œil ou le mouvement des lèvres avec les déplacements de la tête. En conclusion, il y a une variété de techniques de fusion qui peuvent être étudiées et probablement appliquées à notre approche multimodale.
A plus long terme, ces travaux en visage, en introduisant un aspect comportemental, devraient faciliter la reconnaissance automatique des personnes en vidéo surveillance.

Références

[1] Zhao W., Chellappa R., Phillips P.J. and Rosenfeld A. Face recognition: a literature survey. *ACM Computing Surveys*, vol. 35, iss. 4, pag. 399-458, December 2003.

[2] Dugelay J.-L., Junqua J.-C., Kotropoulos C., Kuhn R., Perronnin F. and Pitas I. Recent advances in biometric person authentication. *IEEE Proceedings on Acoustics, Speech, and Signal Processing (ICASSP2002)*, pag. 4060-4063, 2002.

[3] Li B. and Chellappa R. A generic approach to simultaneous tracking and verification in video. *IEEE Transactions on Image Processing*, vol. 11, iss. 5, pag. 530-544, May 2002.

[4] Zhou S., Krueger V. and Chellappa R. Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding*, vol. 91, iss. 1-2, pag. 214-245, July-August 2003.

[5] Lee K.-C., Ho J., Yang M.-H. and Kriegman D. Visual tracking and recognition using probabilistic appearance manifolds. *Computer Vision and Image Understanding*, vol. 99, iss. 3, pag. 303-331, September 2005.

[6] Matta F. and Dugelay J.-L. Person recognition using human head motion information. *International Conference on Articulated Motion and Deformable Objects (AMDO2006), LNCS*, vol. 4069, July 2006.

[7] Turk M. and Pentland A. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, vol. 3, iss. 1, pag. 71-86, 1991.

[8] Sanderson C. and Paliwal K.K. Identity verification using speech and face information. *Digital Signal Processing*, vol. 14, iss. 5, pag. 449-480, September 2004.

[9] Chang K.I, Bowyer K.W. and Flynn P.J. An evaluation of multimodal 2D+3D face biometrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, iss. 4, pag. 619-624, April 2005.