Segmentation et reconnaissance du geste : une approche par MMC et modèle probabiliste d'apparence

Raouf Hamdan, Fabrice Heitz, Laurent Thoraval

Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection
LSIIT - UPRES-A CNRS 7005 / Université Strasbourg I
Pôle API, Bd. Sebastien Brant, 67400 Illkirch, France
Raouf.Hamdan, Fabrice.Heitz, Laurent.Thoraval@ensps.u-strasbg.fr

Résumé – Cet article présente une approche statistique pour la segmentation, le suivi et la reconnaissance du geste dans des séquences longues. Une technique d'apprentissage statistique, développée récemment par Moghaddam et Pentland [6], est utilisée pour créer une représentation probabiliste compacte des apparences 2D du geste sur un sous-espace de dimension réduite. La segmentation et le suivi du geste sont basés sur un critère du maximum de vraisemblance associé à ce modèle d'apparence. La reconnaissance du geste s'appuie sur un Modèle de Markov Caché utilisant le modèle probabiliste d'apparence comme modèle d'observation. L'approche est générale et peut s'appliquer à d'autres problèmes de reconnaissance du mouvement.

Abstract – A generic approach for the extraction and recognition of gesture using raw grey-level images is presented. The probabilistic visual learning approach proposed by Moghaddam and Pentland [6], is used to create a set of compact statistical representations of gesture appearance on low dimensional eigenspaces. The same probabilistic modeling framework is used to extract and track gesture and to perform gesture recognition over long image sequences. Gesture extraction and tracking are based on Maximum Likelihood gesture detection in the input image. Recognition is performed by using the set of learned probabilistic appearance models as estimates of the emission probabilities of a Continuous Density Hidden Markov Model (CDHMM). Although the segmentation and CDHMM-based recognition use raw grey-level images, the method is fast, thanks to the data compression obtained by probabilistic visual learning. The approach is comprehensive and may be applied to other visual motion recognition tasks.

1 Introduction

La segmentation, le suivi et l'interprétation du mouvement dans une séquence d'images restent des problèmes fondamentaux en analyse de scènes dynamiques. En particulier, la reconnaissance des gestes s'est récemment imposée comme une voie attractive et intuitive pour l'interaction homme-machine [8]. Dans ce domaine, la plupart des méthodes de la littérature nécessitent le développement de systèmes d'acquisition spécifiques (stéréovision, utilisation de gants ou de marqueurs [10]) pour segmenter et appréhender le mouvement de la main. Certaines approches s'appuient sur un modèle 3D de l'objet à segmenter. Le recalage (qui reste délicat) du modèle 3D sur l'image est alors utilisé à la fois pour la segmentation et la reconnaissance. Les systèmes spécifiques de ce type ne peuvent par ailleurs être appliqués simplement à d'autres problèmes de reconnaissance du mouvement (comme la reconnaissance du mouvement des lèvres, ou celle de l'expression faciale, etc.).

Dans cet article, nous proposons une approche par Modèle de Markov Caché (MMC) et modèle probabiliste d'apparence pour la segmentation, le suivi et la reconnaissance du geste sur une séquence longue. Les MMCs, utilisés avec succès en reconnaissance automatique de la parole [9], ont été considérés plus récemment dans le cadre de la reconnaissance de l'écriture et du geste. On peut citer ici les travaux de Yamato *et al.* [12] sur la classification par MMC de mouvements de tennis, en s'appuyant sur des techniques de quantification vectorielle sur des séquences d'images binaires. Le recalage temporel par programmation dynamique, technique apparentée aux MMCs, a été proposé par Darrell et Pentland [3], pour l'appariement de séquences en niveaux de gris avec des «templates», par simple corrélation. D'autres études ont été conduites pour la reconnaissance du langage des sourds-muets. Starner et Pentland [10] ont ainsi décrit un système par MMC, pour l'interprétation d'un sous-ensemble de l'«American Sign Language». La position de la main, qui sert d'observation, est segmentée dans ce cas, en utilisant des gants colorés, ou plus récemment, en s'appuyant sur la couleur spécifique de la peau.

L'approche que nous proposons ici pour la segmentation, le suivi et la reconnaissance, s'appuie sur une exploitation directe des images en niveaux de gris, sans extraction de primitives spécifiques. La complexité calculatoire a jusqu'à récemment exclu l'utilisation directe de l'image comme vecteur d'observations. Pour répondre de manière efficace à ce problème, nous considérons une approche par modélisation probabiliste de l'apparence [6], basée sur un apprentissage statistique permettant une réduction importante de la dimension du problème de reconnaissance. L'approche est générale et peut s'adapter (en effectuant simplement un nouvel apprentissage statistique sur des séquences représentatives) à d'autres problèmes de reconnaissance.

Les résultats préliminaires sur des séquences d'images réelles montrent une bonne qualité de segmentation et de reconnaissance (de l'ordre de 90 % avec un corpus de 5 gestes visuellement proches, 80 séquences d'apprentissage avec 8 personnes différentes et 120 séquences de test).

2 Modèle probabiliste d'apparence pour la segmentation et le suivi du geste

Les modèles d'apparence [7, 11] permettent de coder la forme, la pose et l'illumination d'objets 2D ou 3D dans une représentation compacte. Murase et Nayar [7] ont ainsi développé avec succès un système de reconnaissance d'objets 3D à partir de leurs apparences 2D. Turk et Pentland [11] ont décrit une technique d'identification des visages dans l'espace propre à partir des images en niveaux de gris. Des extensions robustes de ces méthodes ont été proposées par Black [1] et Dahyot [2] pour gérer les données manquantes, occlusions ou les fonds très structurés.

Récemment Moghaddam et Pentland [6] ont proposé une approche probabiliste pour la représentation de l'apparence, par des modèles gaussiens ou multi-gaussiens, identifiés dans une phase préalable d'apprentissage. Dans le phase d'apprentissage, on réunit un ensemble représentatif d'images en niveaux de gris $\mathbf{x}$, de dimensions N, présentant les différentes apparences 2D de la structure à modéliser. Les statistiques du premier et second ordre (moyenne μ et matrice de covariance $\mathbf{Q}$) sont estimées à partir de cet ensemble d'apprentissage, pour élaborer un modèle d'apparence gaussien (à N variables) $\mathcal{N}(\mathbf{x}|\mu, \mathbf{Q})$. La vraisemblance d'une image $\mathbf{x}$ (de dimension N) s'exprime par :

$$P(\mathbf{x}|\mu, Q) = \frac{\exp[-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{Q}^{-1}(\mathbf{x}-\mu)]}{(2\pi)^{N/2}|\mathbf{Q}|^{1/2}} \qquad (1)$$

L'évaluation directe de la vraisemblance (1) est coûteuse, en raison de la dimension du vecteur image $\mathbf{x}$ (N est de l'ordre de 100×100 dans notre cas). En pratique, les images du corpus d'apprentissage sont très corrélées. Une décorrélation de ces images, par transformée de Karhunen-Loeve (TKL) du vecteur aléatoire $\mathbf{x}$, permet classiquement de réduire la dimension du problème [4]. Le calcul de la TLK implique la diagonalisation de la matrice de covariance :

$$\mathbf{Q} = \Phi\ \Lambda\ \Phi^T$$

où Φ désigne la matrice des vecteurs propres, et Λ est la matrice diagonale des valeurs propres. La TKL est alors définie par la projection :

$$\mathbf{y} = \Phi^T(\mathbf{x}-\mu) \qquad (2)$$

La réduction de la dimension du problème est obtenue en approchant $P(\mathbf{x}|\Omega)$ à partir des $M (M \ll N)$ composantes principales associées à la TKL. Moghaddam et Pentland [6] proposent, sur ce principe, une «bonne» approximation de $P(\mathbf{x}|\mu, Q)$, sur l'espace engendré par les M premiers vecteurs propres et son orthogonal :

$$\hat{P}(\mathbf{x}|\mu, Q) = \left[\frac{\exp(-\sum_{i=1}^{M}\frac{y_i^2}{2\lambda_i})}{(2\pi)^{M/2}\prod_{i=1}^{M}\lambda_i^{1/2}}\right]\left[\frac{\exp(-\sum_{i=M+1}^{N}\frac{y_i^2}{2\rho})}{(2\pi\rho)^{(N-M)/2}}\right] \qquad (3)$$

où les y_i désignent les composantes du vecteur $\mathbf{y}$ et les λ_i sont les valeurs propres. ρ est un paramètre, dont la valeur optimale (qui minimise la distance de Kullback-Leibler entre $P(\mathbf{x}|\mu, Q)$ et $\hat{P}(\mathbf{x}|\mu, Q)$) est donnée dans [6] : $\rho = \frac{1}{N-M}\sum_{i=M+1}^{N}\lambda_i$. Le résiduel de la reconstruction dans l'espace orthogonal, peut être calculé efficacement à partir des M premières composantes principales :

$$\sum_{i=M+1}^{N} y_i^2 = ||\tilde{\mathbf{x}}||^2 - \sum_{i=1}^{M} y_i^2$$

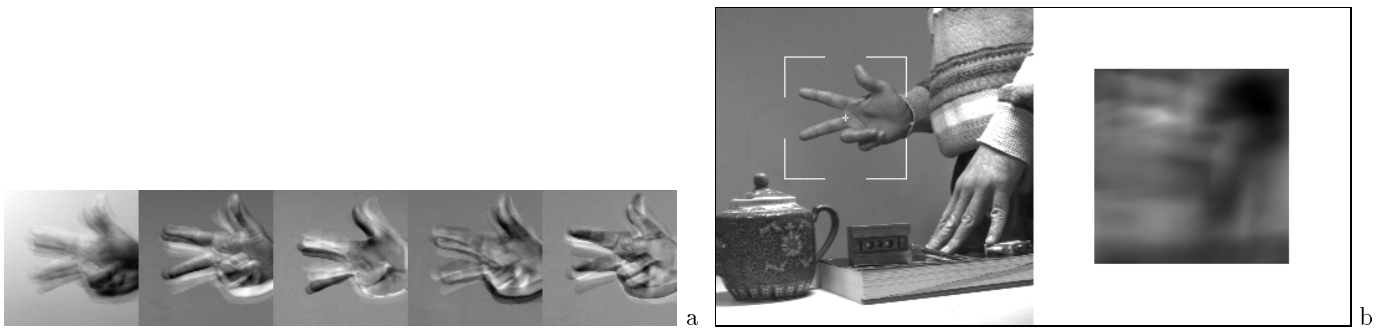
La moyenne et les premiers vecteurs propres sont présentés à titre d'illustration sur la figure 1(a), pour un échantillon d'apprentissage correspondant à une position de la main dans le geste «trois».

Une fois que l'on a construit un modèle d'apparence caractérisé par $(\mu, \mathbf{Q})$, la segmentation du geste sur la *première* image de la séquence est obtenue par une estimation au sens du maximum de vraisemblance. Une fenêtre glissante de N pixels centrée en (i, j) balaye l'ensemble de l'image. L'observation est constituée par le vecteur $\mathbf{x}^{(i,j)}$ des pixels de la fenêtre, ordonnés selon l'ordre lexicographique. La segmentation est obtenue en détectant la position de la fenêtre (i, j) conduisant à la vraisemblance maximale, d'après le modèle (3). Pour cette phase de segmentation, le modèle d'apparence considéré est construit à partir de l'ensemble des images de la base d'apprentissage (espace propre «universel»). La procédure est par ailleurs complétée par une recherche multiéchelle pour la gestion du facteur d'échelle, ainsi que par une recherche locale pour gérer les légères rotations de la main. La figure 1(b) montre un exemple de segmentation et la carte de vraisemblance associée.

Pour les images suivantes de la séquence, un suivi du geste est réalisé par un filtre de Kalman intégrant le modèle d'apparence probabiliste dans la phase de mesure. Cette procédure de suivi permet de gérer de grands changements d'apparence, des angles de rotation ou des facteurs d'échelles importants. La figure 2 présente un exemple de segmentation et de suivi par un modèle à vitesse constante.

3 Reconnaissance du geste par modèle d'apparence et MMC

L'approche de reconnaissance automatique de gestes mise en œuvre est similaire à celle empruntée par la communauté parole en reconnaissance automatique de mots isolés [9]. Un ensemble de gestes $\{G_1, G_2, \ldots, G_P\}$, observés au travers de séquences d'images $\mathbf{X}$ en niveaux de gris, sont analysés concurremment par P MMC $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_P\}$ à lois d'observation continues. Comme en parole, la reconnaissance automatique de gestes procède en 3 étapes : modélisation markovienne cachée, apprentissage des modèles, reconnaissance proprement dite. Dans notre implantation actuelle, les gestes à reconnaître sont au nombre de



a b

FIG. 1: (a) Moyenne et premiers vecteurs propres associés au modèle d'apparence de la main (geste «trois»). (b) Segmentation du geste par MV (à droite, la carte de «log-vraisemblance»).

FIG. 2: Suivi de la main sur une séquence longue présentant des changements d'apparence, des grandes rotations et des variations d'échelle (trames: 10, 20, 30, 40, 50).

5 et consistent, par analogie avec la parole, en la prononciation gestuelle, avec la main, des mots «*un*», «*deux*», «*trois*», «*deux puis trois*» et «*trois puis cinq*», chaque geste débutant le poing fermé.

L'étape de modélisation consiste à adopter des chaînes de Markov cachées à topologie gauche-droite fondée sur l'entrelacement d'états «*statiques*» et d'états «*transitoires*», les premiers modélisant les positions quasi stables de la main au cours du temps, les seconds les phases transitoires entre ces positions. Afin d'interdire le saut d'états stables, constitutifs de chaque modèle, tout en autorisant le saut d'états transitoires, de durée d'occupation souvent non significative, la contrainte supplémentaire $a_{ij} = 0, j > i + 2$, est appliquée à la topologie des chaînes. Les lois d'observation associées aux états sont quant à elles continues et correspondent au modèle d'apparence développé précédemment: la probabilité $b_j(\mathbf{x}_t)$ d'un état q_j d'avoir produit à l'instant t l'image en niveaux de gris $\mathbf{x}_t$ est calculée suivant l'équation (3), q_j modélisant indifféremment un état stable ou transitoire.

De par l'introduction du modèle d'apparence dans le formalisme markovien traditionnel, les lois d'observation $\{b_j(.)\}$ et les probabilités de transitions a_{ij} associées sont apprises séparément, pour chaque modèle $\mathcal{M}_i$, sur la base d'un même corpus G_i constitué de 16 séquences d'apprentissage (8 personnes répétant 2 fois le même geste). Chaque loi d'observation $b_j(.)$ est apprise conformément au modèle d'apparence, à partir de l'ensemble des images de G_i représentatives des positions de la main modélisées par q_j. L'apprentissage des probabilités de transitions $A = [a_{ij}]$ s'effectue par application itérative des formules de réestimation de Baum-Welsh [9] au corpus de gestes G_i supposé produit par $\mathcal{M}_i$. Avant apprentissage, les termes a_{ij} associés à chaque état q_i sont initialisés de façon équiprobable. Les probabilités initiales $\{\pi_j\}$ ne requièrent quant à elles aucun apprentissage de par la topologie gauche-droite retenue pour les chaînes. La figure 3 représente, de gauche à droite, le modèle de geste «*deux:trois*» avant puis après apprentissage ainsi que le vecteur moyenne et les premiers vecteurs propres associés à chaque état q_j, utilisés dans le calcul des lois d'observation $b_j(.)$.

Finalement, la reconnaissance des gestes proprement dite est effectuée suivant le critère du maximum de vraisemblance. Le geste G_i observé au travers d'une séquence d'image d'entrée $\mathbf{X}$ est déclaré reconnu si le modèle $\mathcal{M}_i$ conduit à la vraisemblance $P(\mathbf{X}|\mathcal{M}_i)$ maximum. Les vraisemblances sont calculées de façon classique par l'algorithme de calcul des probabilités avant-arrière [9].

4 Résultats expérimentaux

A titre de validation, une base de données de 200 séquences vidéo (25 images/s) a été réalisée à partir de huit personnes différentes. Les séquence varient en longueur entre 25 et 45 images et ont une résolution de 256×256. Nous disposons de cinq réalisations pour chaque geste et pour chaque personne, parmi lesquelles deux ont été réservées à l'apprentissage. Le temps de calcul par trame a été évalué à environ 200 ms (sur une station du travail standard) pour une dimension de la fenêtre d'analyse de 100×100 pixels. Toutes les segmentations obtenues par le modèle ont été satisfaisantes et un taux de reconnaissance de 100% a été constaté pour les séquences qui appartiennent à l'ensemble d'apprentissage ; un taux de 90% est obtenu pour les 120 séquences de test. Une implantation «temps réel» du système est envisagée sur un PC.

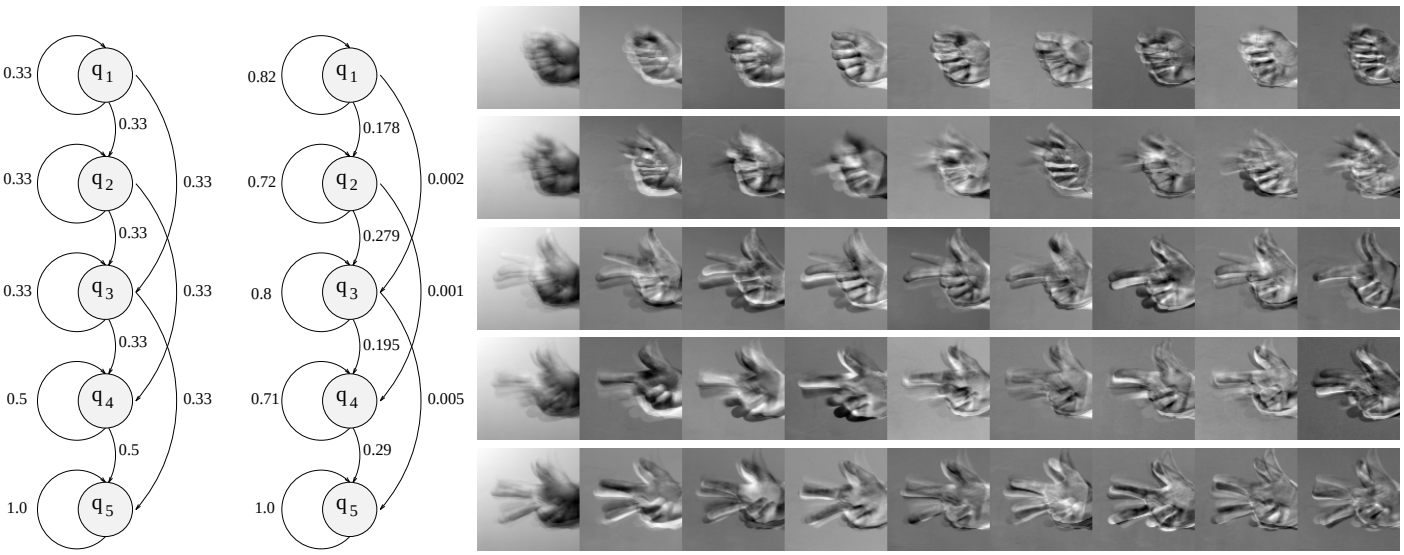


FIG. 3: Le modèle MMC de cinq états utilisé pour le geste «deux:trois». Chaque état est représenté par un modèle d'apparence gaussien dont la moyenne et les premiers vecteurs propres sont affichés à droite.

geste	un	deux	trois	deux:trois	trois:cinq
bonne classif.	87.5%	87.5%	100%	91.7%	95.8%

TAB. 1: *Taux de reconnaissance pour les séquences de test.*

Le tableau 1 présente les taux de reconnaissance pour les séquences de test.

5 Conclusion

Dans cet article, nous décrivons une approche générale pour la segmentation et la reconnaissance du geste sur séquences lonques, utilisant directement les images en niveaux de gris. Grâce à une représentation probabiliste compacte de l'apparence sur un sous-espace de dimension réduite, une reconnaissance rapide peut être obtenue sur une station de travail standard. Des résultats expérimentaux préliminaires ont éte présentés à titre illustratif sur des séquences d'images réelles. Ils montrent des performances encourageantes en reconnaissance et une bonne robustesse du système dans la segmentation et le suivi du geste. Pour valider l'approche dans un cadre plus général, nous envisageons de l'appliquer à d'autres problèmes d'analyse du mouvement, avec un nombre plus significatif de séquences de test (par exemple en reconnaissance du mouvement des lèvres).

Références

[1] M. J. Black and A. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26(1):63–84, January 1998.

[2] R. Dahyot, P. Charbonnier et F. Heitz. Reconnaissance robuste non supervisée d'images couleurs utilisant la théorie semi-quadratique. In *17ème Colloque GRETSI*, Vannes, Sept. 1999.

[3] T. Darrell and A. Pentland. Space-time gestures. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 335–340, June 1993.

[4] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Acadamic Press, 1990.

[5] M. Isard and A. Blake. Visual tracking by stochastic propagation of conditional density. In *Proc. 4th European Conf. on Computer Vision*, pages 343–356, 1996.

[6] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):696–710, July 1997.

[7] H. Murase and S. K. Nayar. Visual learning and recognition of 3-D objects from appearance. *International Journal of Computer Vision*, 14(1):5–24, 1995.

[8] V. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for human-computer interaction: a review. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):677–695, July 1997.

[9] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, February 1989.

[10] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, December 1998.

[11] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[12] I. Yamato, I. Ohya, and K. Ishii. Recognizing human action in time-sequential images using Hidden Markov Model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 379–385, June 1992.