Robustification des signatures de couleurs par modélisation de leurs variabilités intra-plan-vidéo

Riad HAMMOUD, Roger MOHR, Christophe BIERNACKI

Laboratoire Gravir, projet Movi de l'Inria Rhône-Alpes
655, avenue de l'Europe
38330 Montbonnot Saint Martin, France
riad.hammoud@inrialpes.fr, roger.mohr@imag.fr

Résumé – Dans cet article, nous proposons une nouvelle approche probabiliste de structuration de la vidéo. Notre approche consiste à capturer la variabilité temporelle intra-plan de l'objet par une gaussienne et ensuite apparier les objets vidéo par un modèle de loi de mélange gaussien. Cette approche permet de robustifier notablement les descripteurs utilisés dans la reconnaissance d'objets vidéo.

Abstract – In this paper, we propose a new probabilistic approach for video structuring. Our appoach consists in capturing the temporal intra-shot variability of object by a gaussian and then to match based-objects video by a mixture of gaussians. The approach allows to robustify existing descriptors used for video object recognition using temporal information.

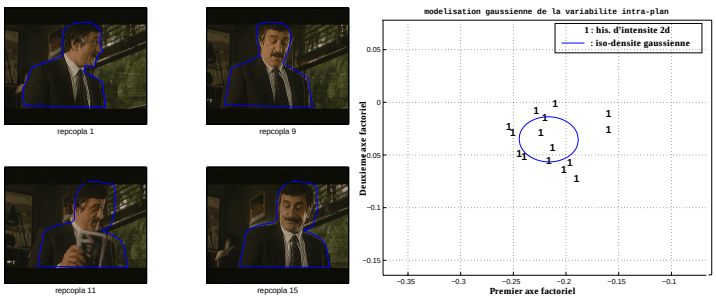
1 Introduction

L'évolution rapide de la vidéo numérique comme un moyen informatif riche, largement intégré dans des nombreuses applications sur le web, mit en lumière la nécessité de méthodes automatiques de structuration et d'indexation, offrant, à la fois, une meilleure accessibilité en navigation et en recherche [7]. La fabrication de résumés vidéo et plus particulièrement de résumés des objets d'intérêt (acteur, ...) désignés par l'utilisateur, représente une des applications récentes et très démandées sur la vidéo [1]. Par exemple, l'utilisateur veut regarder dans un film de deux heures toutes les scènes d'un acteur particulier.

Pour réaliser ce genre d'application, **hypervidéo**, des techniques de reconnaissance d'objets et de classification seront mises en place. Cependant, l'indexation des objets vidéo, semble une tâche très délicate suite à deux raisons :

- **Apparence intra et inter-plan de l'objet :** un objet particulier peut apparaître dans le même plan [1] ainsi que dans plusieurs plans sous différentes prises de vues, différents effets d'éclairages et sous une variété de mouvements de la caméra (zoom, ...). La figure 1.a illustre l'apparence d'un personnage segmenté dans un plan vidéo. Notons les effets dûs à la rotation et aux occultations partielles, la mise en correspondance par différentes techniques d'appariement [9] [11] montre ces limites.
- **Taille de la base d'index :** indexer les objets d'un film de 90 minutes revient en réalité à indexer au moins environ 129600 objets différents [2]. La taille énorme de la base d'index (nb d'objets + dimension de l'espace de représentation de l'index) semble être difficile à manipuler même avec des techniques d'indexation avancées [3].

1. un plan est une suite d'images (durée moyenne 1 seconde)
2. un objet d'intérêt par plan et 24 images par seconde



(a) (b)

FIG. 1 – *(a) Apparence sous rotations 3D et occultations partielles des parties du personnage détouré aux instants 1, 9, 11 et 15. (b) iso-densité de la distribution gaussienne des histogrammes d'intensité aux instants ([1..15]) d'apparence du personnage dans le plan.*

À ce stade, nous proposons dans cet article une approche probabiliste pour apparier les objets inter-plans. On suppose que la vidéo est segmentée au préalable en plans et en objets mobiles [4] [6]. On séléctionne d'abord des objets de sémantiques différentes dans la vidéo (premier acteur, deuxième acteur, voiture telle couleur, ...). Ces objets sont considérés comme des modèles ou des classes d'objets. Ensuite, on modélise la variabilité temporelle du descripteur (ou signature: histogramme de couleurs, contour, ...) de chaque objet modèle par une gaussienne multidimensionnelle. Dans l'espace de représentation de descripteurs ($\Re^2$ dans fig.1.b), la matrice de variance capture la dispersion du descripteur extrait des différentes prises de vues de l'objet dans le plan. La figure 1.b illustre la dispersion de l'histogramme d'intensité extrait de toutes les occurences du personnage de la figure 1.a dans un plan vidéo. L'histogramme d'intensité étant

projeté dans le premier plan factoriel. Ensuite, on collecte les K classes d'objets modélisés par un modèle de mélange de gaussien. Ce modèle de mélange nous permet de partitionner l'espace de représentation de descripteurs en zones d'appartenance aux K classes d'objets. En effet, on sera capable, en appliquant la règle de probabilité de maximun à posteriori [8] à classer un objet quelconque de la vidéo dans la classe d'objet la plus probable. La section suivante détaille cette approche.

La modélisation de la variabilité temporelle du descripteur dans le plan nous permet à la fois de bien gérer nos documents vidéo (les paramètres du modèle de mélange sont seulement indexés) et d'améliorer le processus d'appariement d'objets inter-plans. Ceci est confirmé par des expérimentations sur une base d'objets vidéo de 1016 images segmentées en 1745 objets. Les résultats de l'approche proposée sont comparés à ceux de la méthode classique, largement adaptée dans la littérature, qui consiste à représenter le plan vidéo par une image représentative souvent l'image médiane [10].

2 Modèle de mélange gaussien multivarié

2.1 Définition

Ayant un ensemble de n individus $\mathbf{x_1}, ..., \mathbf{x_n}$ décrits par d variables, c'est-à-dire un échantillon de taille n dans $\Re^d$, le modèle de mélange correspond à l'hypothèse que la population de référence est formée de K sous-populations Π_1, ..., Π_K de densités $\varphi(\mathbf{x}, \mathbf{a_k})$ et avec des proportions p_1, ..., p_k ($p_k \in]0,1]$ et $\sum_{k=1}^{K} p_k = 1$): les n individus constituent donc un échantillon ($\mathbf{x_1}$, ..., $\mathbf{x_n}$) de réalisations indépendantes d'un vecteur aléatoire $\mathbf{X}$ de $\Re^d$ et ayant une partition représentée par un ensemble de n labels $Z = (z_1, ..., z_n)$ ($z_i \in [1...n]$). La densité de mélange peut donc s'écrire:

$$f(x, \theta) = \sum_{k=1}^{\mathbf{K}} p_k \varphi(\mathbf{x}, \mathbf{a_k}) \qquad (1)$$

où φ appartient à une famille paramétrée de densités sur $\Re^d$, $\theta = (\mathbf{a}, \mathbf{p})$, $\mathbf{a} = \mathbf{a_1}, ..., \mathbf{a_k}$ et $\mathbf{p} = (p_1, ...p_k)$.

Les $p_k = P(\mathbf{Z} = k)$ ($k \in 1..K$) représentent les probabilités à priori d'appartenance des $\mathbf{x_i}$ aux classes (sous-populations) Π_i ($i \in 1..n$).

2.2 Mélange gaussien

Les lois de mélanges gaussiens sont souvent utilisées grâce à leur identifiabilité où il existe une correspondance unique entre la distribution d'un mélange gaussien et les composantes de ce mélange. Cependant la densité gaussienne multidimensionnelle de chaque population k s'écrit

$$\phi(\mathbf{x} \mid \mathbf{a}_k) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} exp\left(-\frac{1}{2}(\mathbf{x} - \mu_k)' \mathbf{\Sigma}_k^{-1} (\mathbf{x} - \mu_k)\right) \qquad (2)$$

avec $\mathbf{a_k} = (\mu_k, \mathbf{\Sigma}_k)$, μ_k moyenne de la classe k ($\mu_k \in \Re^d$) et $\mathbf{\Sigma}_k$ sa matrice de variance-covariance (matrice symétrique définie positive $d \times d$).

2.2.1 Choix du modèle de mélange de gaussiens

Des connaissances à priori peuvent être disposées sur le paramètre du mélange. L'avantage d'introduire ces informations est de mieux estimer le paramètre θ de mélange et la règle de discrimination. Il s'agit aussi de protéger le modèle de mélange contre une surparamétrisation. Le choix du modèle influence sur la frontière de discrimination : linéaire, quadratique, etc. Dans la littérature, 28 modèles de loi de mélange ont été détaillés dans [5] et [2]. À l'aide d'une décomposition spectrale (décomposition en valeurs propres et vecteurs propres) de la matrice de variance $\Sigma_k = \lambda_k D_k A_k D_k'$ on obtient 8 modèles différents. Finalement, le modèle de mélange est paramétré par :

- les centres des classes μ_1, ..., μ_K
- les proportions p_1, ..., p_K
- les volumes λ_1, ..., λ_K
- les formes A_1, ..., A_k
- les orientations D_1, ..., D_k

Les proportions représentent les taux des individus présents dans les classes classes et les volumes représentent les places occupées par les classes dans l'espace.

En utilisant cette paramètrisation, il est possible de proposer des situations intermédiaires entre des hypothèses restrictives (matrice de variance proportionnelle à la matrice identité ou matrice de variance identique pour toutes les classes) et les hypothèses très générales (aucune contrainte).

2.3 Classement par MAP

Connaissant le paramètre θ, l'affectation de tout élément $\mathbf{x}$ de $\Re^d$ à une classe k ($k \in 1, ..., K$) peut être réalisée en utilisant la méthode de *maximum a posteriori*, dite aussi méthode du MAP. Cette méthode consiste simplement à affecter l'élément $\mathbf{x}$ à la classe la plus probable a posteriori. La probabilité a posteriori que $\mathbf{x}$ appartient à la classe k se calcule directement par le théorème de Bayes :

$$t_k(\mathbf{x}, \theta) = \frac{p_k \varphi_k(\mathbf{x}, \mathbf{a}_k)}{f(\mathbf{x}, \theta)} \qquad (3)$$

3 Expérimentation

Séquence vidéo Nous avons expérimenté l'approche proposée dans cet article sur des séquences vidéo de différentes tailles. Nous présentons ici les résultats obtenus sur "Avengers" [3] : une vidéo de 1016 images ségmentée au préalable en plans et objets mobiles par les outils fournis par [6]. Les autres séquences vidéo experimentées ne dépassent pas une centaine d'images. La séquence "Avengers" est découpée en 31 plans et comporte 1745 objets mobiles et statiques. Les objets statiques sont segmentés manuellement et d'autres objets mobiles ont eu des corrections manuelles. Un objet est identifié par un label et un numéro d'image. Deux objets ayant un même identificateur (label) représentent deux occurences dans le plan d'un même objet.

3. extrait du film "chapeau mellon et bottes de cuire" fourni par l'INA

Préparation et réduction des données De cette séquence vidéo on a séléctionné 15 ($K = 15$) classes d'objets de sémantique différente. Les descripteurs extraits de la base d'objets vidéo tout entière sont des mesures globales de la distribution des couleurs et de niveaux de gris représentés par des histogrammes normalisés. La méthode d'indexation par les histogrammes est remarquablement fiable par rapport aux changements d'orientation d'objets, aux variations d'échelle, aux occultations partielles et aux changements de points de vue [13]. Chaque descripteur est ensuite vu comme un point dans $\Re^d$ où d représente la taille du descripteur (63 par exemple pour un histogramme RGB tridimensionnel normalisé de 64 classes de couleurs ou bins). Une réduction de l'espace de représentation des descripteurs a été effectuée par l'élimination des axes décorrélés à l'aide d'une analyse en composantes principales (ACP). La réduction de l'espace de représentation nous a permis de mieux estimer les paramètres du mélange dans le cas où le nombre des points appartenant à une classe d'objet n'est pas assez grand.

Le modèle de mélange qu'on a fixé lors de l'étape d'éstimation était $[p_k \lambda D_k A_k D_k']$ où on dispose des volumes égaux (λ) pour les 15 classes d'objets. Les autres paramètres du mélange (proportions, orientations, formes) ont été laissés libres. Ce modèle à été choisi empiriquement : on a remarqué sur plusieurs bases d'objets vidéo que le taux de reconnaissace d'objets par les autres modèles est moins bon qu'avec ce modèle.

Le nombre total de points utilisés dans l'estimation était de 554 points. Cependant, 1745 points on été classés durant la phase de test. Chaque objet dans la vidéo est considéré comme un objet requête. Cependant, on utilise la règle de MAP pour classer un objet requête à l'une des classes apprises. Le tableau 1 illustre la performance de notre approche en utilisant le critère d'évaluation suivant.

Critère d'évaluation Pour une méthode d'indexation, la formule 4 représente le taux moyen des bons classements des objets requêtes, jusqu'au rang R. Cette mesure globale nous permet d'évaluer la performance de notre approche vis-à-vis des autres méthodes. Dans notre cas, on sera intéressé par les résultats du classement du premier rang seulement ($R = 1$).

$$\text{r-measure} = \frac{\sum_{r=1}^{R} \text{nb. des objets bien classés dans } \mathbf{r}}{\text{nb. totale des objets requêtes}} \tag{4}$$

Descripteurs	*dim*	**r-measure**
hist. d'intensité	7	42.10
hist. RGB-3D	10	58.90

TAB. 1 – *Taux de bon classement des objets requetês par modèle de mélange gaussien*

3.1 Analyse Comparative

Rappelons que notre approche de *robustification des signatures par intégration de l'aspect temporel* repose sur le principe suivant : modéliser la classe d'un objet qui se meut dans un plan, et lui *affecter l'objet requête à la classe la plus probable selon les modèles construits*. L'évaluation de notre approche est effectuée en comparant les résultats de classement probabiliste à ceux de deux autres méthodes : *la méthode de représentation classique des plans et la méthode de la moyenne que nous avons dévellopée.*

- *La méthode classique* consiste à représenter le plan vidéo par une image représentative [10]. Souvent l'image médiane est considérée comme image représentative. Selon cette stratégie la base d'index contient les descripteurs extraits sur les K classes d'objets médians.
- *La méthode de la moyenne* consiste à calculer la moyenne du descripteur extrait de toutes les occurrences dans les plans des classes d'objets. Soit n le nombre d'occurrences de l'objet k dans le plan p et soit $\{H_{ki}^d\}$ la suite des d-histogrammes[4] extraits de l'objet k aux instants i ($i \in [t+1, ..., t+n]$). L'histogramme moyen sera donné par la formule suivante (Eq. 5):

$$\begin{cases} \overline{H_k^d} = \{\overline{h_i^d}\} \\ \textbf{et} \\ \overline{h_i^d} = \frac{\sum_{j=1}^{n} h_{ij}^d}{n} \end{cases} \tag{5}$$

Finalement, les deux méthodes ci-dessus représentent un objet-plan par un seul point dans l'espace $\Re^d$ (où d est la taille du descripteur extrait). En d'autres termes, la variabilité de ce descripteur de l'objet-plan n'est pris en compte que dans moyennage. Pour apparier un objet requête, avec les bases d'indexes de ces deux méthodes, on utilise comme métrique la distance χ^2 donnée par la formule 6. En effet, selon [12] (chapitre 5), les expérimentations montrent, dans la plupart des cas, une meilleure reconnaissance par l'usage de la distance du χ^2 que par les autres fonctions des comparaisons.

$$\chi^2(Q, V) = \sum_i \frac{(q_i - v_i)^2}{(q_i + v_i)} \tag{6}$$

où q_i représente la fréquence de la couleur i de l'histogramme Q.

L'idée du calcul d'une distance de similarité est d'*affecter l'objet requête à la classe la plus proche*. Dans ce cas une classe d'objet (objet indexé dans la base) est représentée par un seul point.

La classe la plus proche en terme de distance est souvent aussi la plus proche au sens de MAP. Mais ce qui fait l'écart entre les résultats de classement n'est pas tant la règle de calcul de la distance mais plutôt la modélisation de la classe, c'est le point le plus fondamental de notre approche.

Le tableau 2 représente les taux de bon classement des objets requêtes obtenus par la méthode classique de représentation du plan vidéo. Pour chaque objet requête les 15 classes d'objets sont ordonnées dans un ordre croissant de valeurs de χ^2.

Le tableau 3 illustre la performance de la deuxième méthode concurente : celle du descripteur moyen.

4. d représente dans ce cas le type de l'histogramme mono ou tri-dimensionnels

Descripteurs	*dim*	*distance*	r-measure
hist. d'intensité	8	χ^2	31.12
hist. RGB-3D	64	χ^2	47.16

TAB. 2 – *Taux de bon classement des objets requêtes par la méthode classique de représentation du plan*

Une évaluation de la performance du modèle de mélange de gaussien peut être vérifiée en comparant les taux de bon classement des différentes méthodes présentées dans les tableaux 1, 2 et 3.

Nous remarquons que notre approche probabiliste d'appariement des objets vidéo apporte plus de 10% de bon classement par rapport aux deux autres méthodes concurrentes.

Descripteurs	*dim*	*distance*	r-measure
hist. d'intensité	8	χ^2	30.09
hist. RGB-3D	64	χ^2	43.09

TAB. 3 – *Taux de bon classement des objects requêtes par la méthode du descripteur moyen*

4 Conclusion

Nous venons de présenter une approche statistique pour modéliser la variabilité intra-plan d'un descripteur d'un objet. Une classe d'objet-plan n'est plus résumée par un seul point multidimensionnel mais par une loi de mélange qui permet de capturer l'évolution temporelle intra-plan de la signature à indexer. L'avantage de cette modélisation est double: d'une part elle modélise explicitement des représentations différentes, ce qui correspond bien au fait que les apparences sont différentes; d'autre part elle compactifie les données en les rassemblant en quelques lois pour des échantillons homogènes observés.

Les résultats expérimentaux présentés dans cet article sont limités mais illustrent bien cet aspect: quand les aspects varient peu, les résultats de classement sont équivalents à une représentation moyenne des données (mais en plus compact), et quand les aspects varient beaucoup, notre représentation est plus performante.

Nos expérimentations ont montré que le modèle de mélange gaussien avec une hypothèse de volumes égaux donne le meilleur taux de bon classement des objets requêtes parmi d'autres modèles. Cependant, il existe des critères qui permettent de choisir d'une manière automatique le modèle de mélange gaussien le plus approprié aux données. Parmi ces critères, on cite la validation croisée [2]. Toutefois, d'après nos expérimentations, ce critère ne choisit pas toujours le bon modèle de mélange gaussien.

Remerciements

`Riad Hammoud` est financé par `Alcaltel CRS`. Les auteurs tiennent également à remercier l'`Institut National` de l'`Audiovisuel` pour la permission d'utiliser leur vidéo MPEG.

Références

[1] S. Benayoun, H. Bernard, P. Bertolino, M. Gelgon, C. Schmid, and F. Spindler. Structuration de vidéos pour des interfaces de consultation avancées. In *CORESA 98 – Journées d'études et d'échanges COmpression et REprésentation des Signaux Audiovisuels.*, June 1998.

[2] C. Biernacki and G. Govaert. Choosing Gaussian Models in Discriminant Analysis. In *IV International Meeting of Multidimentional Data Analysis*, Bilbao, Spain, September 10-12 1997.

[3] S. Blott and R. Weber. A simple vector-approximation file for similarity in high-dimensional vector spaces. Technical Report 19, ESPRIT project HERMES (no. 9141), March 1997. Postscript version available by `ftp` [5].

[4] P. Bouthemy and F. Ganansia. Video partitionning and camera motion characterisation for content-based video indexing. *Proc. 3rd IEEE Int. Conf. Image Processing.*, september 1996.

[5] G. Celeux and G. Govaert. Gaussian Parsimonious Models. *Pattern Recognition*, 28(5):781–783, 1995.

[6] M. Gelgon and P. Bouthemy. A region-level graph labeling approch to motion-based segmentation. In *CVPR*, pages 514–519, Puerto Rico, June 17-19 1997.

[7] S.W. Smoliar H. J. Zhang, C. Y. Low and J.H. Wu. Video parsing, retrieval and browsing: An integrated and content-based solution. *ACM Multimedia*, pages 15–24, 1995.

[8] G.J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York, 1992.

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

[10] B.C. O'Connor. Selecting key frames of moving image documents: A digital environment for analysis and navigation. *Microcomputers for Information Management*, 8(2):119–133, 1991.

[11] L.A. Rowe, J.S. Boreczky, and C.A. Eads. Indexes for user access to large video databases. *Proc. IS.T SPIE Conf. on Storage and Retrieval for Image and Video Databases II*, pages 150–161, 1994.

[12] B. Schiele. *Reconnaissance d'objets utilisant des histogrammes multidimensionnels de champs réceptifs.* Thèse de doctorat, GRAVIR – IMAG – INRIA Rhône–Alpes, July 1997.

[13] M.J. Swain and D.H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

5. http://www-dbs.ethz.ch/~weber/paper/VAFILE.ps.gz