Prédiction Temporelle de Descripteurs Visuels pour la Mesure de Similarité entre Vidéos

Eric Bruno, Stéphane Marchand-Maillet
Laboratoire de Vision par Ordinateur et Mutimédia
Département d'Informatique, Université de Genève
25 rue du Général Dufour, 1211 Genève 4, Suisse
Eric.Bruno@unige.ch, marchand@cui.unige.ch

Résumé – Nous abordons dans cet article le problème de la mesure de similarité entre les contenus visuels spatio-temporels de vidéos. L'information visuelle peut-être caractérisée par un ensemble de descripteurs extraits des images tout au long de la séquence. La dimension temporelle des descripteurs nous permet alors de les considérer comme une série temporelle dont on cherche à modéliser le comportement par une fonction de prédiction. Cette fonction est nonlinéaire et est estimée par la technique des *Machines à Vecteurs Supports*. La comparaison des modèles temporels ainsi estimés sur différentes vidéos permet, par l'intermédiaire de l'erreur de prédiction, de définir une mesure de similarité qui tienne à la fois compte des descripteurs visuels et de leur évolution temporelle tout au long de la séquence. Des résultats expérimentaux sur des séquences réelles montrent la validité de cette mesure de similarité.

Abstract – This paper deals with the problem of the similarity measure between spatio-temporal visual contents of videos. Visual information could be characterized by a set of descriptors extrated from images along the sequence. The temporal dimension of these descriptors allows to consider them as temporal series that we want to model the behavior by a prediction function. This function is nonlinear and is estimated by a *kernel Support Vector Machine*. The comparison between temporal models estimated from different sequences enables us, via the predicton error, to define a similarity measure where visual descritpors and their temporal behavior is taking into account. Experimental results on real image sequences show the efficiency of this similarity measure.

1 Introduction

La définition de mesures de similarité appliquées aux séquences d'images est fondamentale pour la mise en place de systèmes de recherche et d'exploration de vidéos par le contenu visuel. La nature spatio-temporelle de ces documents nécessite de spécifier une mesure qui tienne compte des propriétés à la fois spatiales et temporelles de la séquence d'images.

D'une manière générale, une séquence d'images est caractérisée par un ensemble de descripteurs visuels estimés sur une ou plusieurs images de la séquence. La relation temporelle existant entre les descripteurs nous amène à considérer cet ensemble comme une série temporelle multidimensionelle sur laquelle la mesure de similarité doit s'appliquer.

Nous présentons ici une solution consistant à modéliser le comportement temporel par une fonction de prédiction nonlinéaire. Celle-ci est estimée sur les séquences de descripteurs par l'algorithme des Machines à Vecteurs Supports. L'erreur obtenue en appliquant la fonction de prédiction sur une autre série de descripteurs, fournit directement une mesure de similarité spatio-temporelle entre deux vidéos. Des expériences sur des séquences réelles sont présentées pour évaluer la pertinence de la mesure de similarité.

2 Prédiction de séries temporelles de descripteurs

Considérons une séquence d'images S décrite par une série de descripteurs quelconques $\{\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_N\}$, $\mathbf{X}_t \in \mathbb{R}^D$, N étant la longueur de la séquence de descripteurs. Soit $\mathbf{F} : \mathbb{R}^{D\times H} \rightarrow \mathbb{R}^D$ la fonction de prédiction d'ordre H de la série temporelle $\{\mathbf{X}_t\}_{t=0}^N$

$$\mathbf{X}_t = \mathbf{F}(\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \ldots, \mathbf{X}_{t-H}) \; \forall t \in [H, N]. \quad (1)$$

La fonction de prédiction multidimensionelle $\mathbf{F}$ fournit un *modèle* de l'évolution temporelle des descripteurs et *caractérise* le contenu dynamique de la séquence. L'ordre H détermine la mémoire du modèle : un événement à l'instant T, décrit par $\mathbf{X}_T$, dépend des H évenements précédents, décrits par $\{\mathbf{X}_t\}_{t=T-H}^{T-1}$. Plus H est important, plus le modèle est spécifique à la séquence traitée. A l'inverse, lorsque la mémoire du modèle diminue, l'information caractérisée par la fonction de prédiction tend à être nulle.

En faisant l'hypothèse qu'il n'y a pas de corrélation entre les composantes de la série $\{\mathbf{X}_t\}_{t=0}^N$, la fonction de prédiction mulidimensionelle $\mathbf{F}$ (eq. 1) peut être estimée sur chacune de ses dimensions, indépendemment des autres. En notant x^l la $l^{\text{ième}}$ composante de $\mathbf{X}$, le

problème consiste alors à estimer f^l tel que

$$x_t^l = f^l(x_{t-1}^l, x_{t-2}^l, \ldots, x_{t-H}^l). \quad (2)$$

Alors

$$\mathbf{F} = [f^1, f^2, \ldots, f^D]^T. \quad (3)$$

Pour simplifier les écritures, nous pouvons définir le vecteur

$$\mathbf{x}_{t-1}^l = [x_{t-1}^l, \ldots, x_{t-H}^l] \; \forall t \in [H, N], \quad (4)$$

de telle sorte que l'équation (2) se réécrive $x_t^l = f^l(\mathbf{x}_{t-1}^l)$. La difficulté principale de cette approche réside dans l'estimation de la fonction de prédiction. La séquence des descripteurs est nonstationnaire, ce qui implique que $\mathbf{F}$ est une fonction nonlineaire. Ce problème de prédiction nonlineaire a été largement abordé dans différents domaines de recherche, et l'approche basée sur les Machines à Vecteurs Supports (acronyme anglais SVM) semble se distinguer par la robustesse des résultats obtenus [3].

3 Machines à Vecteurs Supports pour la prédiction nonlinéaire

Cette section décrit rapidement le principe des SVM pour la régression.On peut trouver une description plus détaillée de cette théorie dans [4]. Le problème consiste à approximer une fonction inconnue $g : \mathbb{R}^D \rightarrow \mathbb{R}$ à partir des observations $\{\mathbf{x}_i, y_i\}_{i=1}^N$ telles que $y_i = g(\mathbf{x}_i) + \eta$, η étant du bruit. Afin d'approximer g, la technique des SVM utilise une fonction paramétrique de la forme

$$f(\mathbf{x}) = \sum_{i=1}^{L} c_i \phi_i(\mathbf{x}) + b, \quad (5)$$

où $\{\phi_i\}_{i=1}^L$ est un ensemble de fonctions de bases. Les paramètres b et $\{c_i\}_{i=1}^L$ sont les inconnues à estimer en minimisant la fonctionnelle

$$R(f) = \frac{1}{N} \sum_{i=1}^{N} |y_i - f(\mathbf{x}_i)|_\epsilon + \lambda ||\mathbf{c}||^2, \quad (6)$$

avec $\mathbf{c} = [c_1, \ldots, c_N]$ et λ une constante imposant une contrainte de lissage sur la solution. La fonction d'erreur est définie par

$$|x|_\epsilon = \begin{cases} 0 & \text{si } x < \epsilon \\ x & \text{sinon.} \end{cases} \quad (7)$$

Dans [5], Vapnik a prouvé que la fonction minimisant la fonctionnelle (6) peut s'écrire sous la forme

$$f(\mathbf{x}, \alpha, \alpha^*) = \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) K(\mathbf{x}, \mathbf{x}_i) + b \quad (8)$$

avec $\alpha_i^* \alpha_i = 0$, $\alpha_i, \alpha_i^* \geq 0 \; i = 1, \ldots, N$ et $K(\mathbf{x}, \mathbf{y})$ une fonction appelée *noyau* définissant le produit scalaire dans l'espace multidimensionel défini par les fonctions de bases ϕ_i

$$K(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{L} \phi_i(\mathbf{x}) \phi_i(\mathbf{y}). \quad (9)$$

L'interêt de la technique SVM vient du fait que seul le noyau K doit être défini alors que l'espace defini par les fonctions ϕ_i n'est jamais explicitement calculé. Ceci offre un grand choix de fonctions de bases nonlinéaires, incluant les bases de dimension infinie.

Le noyau que nous avons choisi est la fonction gaussienne radiale $K(\mathbf{x}, \mathbf{y}) = exp(-\gamma ||\mathbf{x} - \mathbf{y}||^2)$. Cette fonction permet de modéliser les variations temporelles complexes des descripteurs avec un minimum de contraintes *a priori* sur la solution. La détermination de la valeur du paramètre d'échelle γ dépend de l'espace des observations $\{\mathbf{x}_t^l\}_{t=H}^N$ et de la métrique que l'on souhaite imposer dans cet espace. La valeur de ce paramètre a un impact sur la regularité de la fonction estimée. Nous le définissons par

$$\gamma = \frac{1}{2} \left(\frac{1}{N-1} \sum_t ||\Delta_t \mathbf{x}_t||^2 \right)^{-1}, \quad (10)$$

où $\Delta_t \mathbf{x}_t = \mathbf{x}_{t-1} - \mathbf{x}_t$. Cette définition assure que, en moyenne, la distance entre les observations temporellement voisines ($\mathbf{x}_t$ et $\mathbf{x}_{t+1}$) soit suffisamment faible pour obtenir une solution régulière.

4 Mesure de similarité par l'erreur de prédiction

Soient $\mathbf{F}$ et $\mathbf{G}$ les fonctions de prédiction estimées respectivement sur les séries de descripteurs $\{\mathbf{X}_t\}_{t=0}^N$ et $\{\mathbf{Y}_t\}_{t=0}^M$ associées aux séquences d'images S_1 et S_2. Nous pouvons alors construire les deux séries temporelles en croisant modèles et descripteurs

$$\tilde{\mathbf{X}}_t = \mathbf{G}(\mathbf{X}_{t-1}, \ldots \mathbf{X}_{t-H}), \forall t \in [H, N]$$
$$\tilde{\mathbf{Y}}_t = \mathbf{F}(\mathbf{Y}_{t-1}, \ldots \mathbf{Y}_{t-H}), \forall t \in [H, M]. \quad (11)$$

La mesure de distance entre les séquences S_1 et S_2 est définie par

$$D(X, Y) = \frac{1}{2} \left[d\left(\{\tilde{\mathbf{X}}_t\}_t, \{\mathbf{X}_t\}_t \right) + d\left(\{\tilde{\mathbf{Y}}_t\}_t, \{\mathbf{Y}_t\}_t \right) \right], \quad (12)$$

avec $d(\cdot, \cdot)$ la distance L_2 définie dans l'espace des descripteurs normalisée par la durée de la séquence

$$d\left(\{\tilde{\mathbf{X}}_t\}_t, \{\mathbf{X}_t\}_t \right) = \frac{1}{N} \sqrt{\sum_{l=1}^{D} \sum_{t=1}^{N} \left(\tilde{x}_t^l - x_t^l \right)^2}. \quad (13)$$

Dans le cas où la séquence de descripteurs $\{\mathbf{Y}_t\}_t$ est semblable à $\{\mathbf{X}_t\}_t$, la séquence $\{\mathbf{X}_t\}_t$ peut être approximativement prédite par la fonction $\mathbf{G}$ estimée sur $\{\mathbf{Y}_t\}_t$. L'erreur de prédiction $d\left(\{\tilde{\mathbf{X}}_t\}_t, \{\mathbf{X}_t\}_t \right)$ est d'autant plus faible que les deux séries $\{\mathbf{X}_t\}_{t=0}^N$ et $\{\mathbf{Y}_t\}_{t=0}^M$ sont similaires. Le raisonnement est identique pour $\{\mathbf{Y}_t\}_t$ et $\{\tilde{\mathbf{Y}}_t\}_t$.

5 Résultats Expérimentaux

Nous présentons dans cette section différentes expériences réalisées dans le but d'évaluer la pertinence de notre

mesure de similarité.

La première expérimentation porte sur la reconnaissance d'activité. Nous utilisons pour cela une collection de dix séquences d'images contenant deux types d'activités : cinq séquences contiennent l'activité *Aller* (personnes s'éloignant de la caméra) et cinq séquences contiennent l'activité *Venir* (personnes se rapprochant de la caméra) (Fig. 1). La caméra et fixe et la longeur de ces séquences est comprise entre 30 et 40 images. Le contenu dynamique des vidéos est caractérisé par le mouvement. Les descripteurs que nous utilisons correspondent à l'énergie calculée sur chaque sous-bande d'une décomposition en ondelette du flot optique estimé entre chaque image de la séquence. Sous la forme d'un vecteur à 10 composantes, ils représentent, pour chaque image, une mesure d'activité sensible à l'amplitude, l'échelle et l'orientation des mouvements dans la scène. Les détails de l'implantation de ces descripteurs sont presentés dans [1].

La deuxième ligne du tableau 1 représente les séries de descripteurs de mouvement estimés sur les séquences *Venir1*, *Venir2* et *Aller1* (blanc pour les valeurs élevées et noir pour les valeurs nulles). Chaque vecteur de descripteurs correspond à une colonne de la matrice. L'indice de la colonne indique la position temporelle dans la séquence d'images. On peut observer que l'énergie des descripteurs croît dans le cas de l'activité *Venir* et décroit dans le cas de l'activité *Aller*. A partir du modèle temporel $\mathbf{F}_{Venir1}$ (estimé sur la séquence *Venir1*), d'ordre $H = 15$, nous avons prédit les séries de descripteurs pour les 3 séquences (3ème ligne) ainsi que l'erreur quadratique de prédiction correspondante (4ème ligne). A titre de comparaison, la dernière ligne du tableau donne la distance calculée entre la moyenne temporelle des descripteurs *Venir1* et la moyenne des deux autres séries de descripteurs.

Du fait de la description grossière du mouvement dans les séquences, la distance basée sur la moyenne temporelle ne permet pas de différencier les deux types d'activités présents dans ces trois séquences. En revanche, en conservant l'information temporelle des descripteurs, l'erreur de prediction est capable de mesurer une similarité plus importante entre les séquences *Venir1* et *Venir2* que entre *Venir1* et *Aller1*. Nous avons généralisé ce resultat en classant les dix séquences de la collection (5 *Venir* et 5 *Aller*) par un algorithme de classification par agglomération [6], en utilisant comme mesure de similarité soit la distance définie en (12), soit la distance entre moyennes temporelles des descripteurs. Dans le premier cas, le taux de classification est de 100%, alors qu'il n'est que de 60% dans le deuxième.

La deuxième expérience porte sur la représentation de vidéo par le contenu. La base de vidéos considérée est composée de 40 documents contenant principalement des vidéos de sport (football, basketball, windsurf) et des séquences présentant des personnes en gros plan (extraits de journaux télévisés) (Fig.2). Les vidéos sont des plan-séquences et sont caractérisées par les descripteurs de mouvements décrits précédemment.

Nous avons calculé la matrice de similarité de la base en utilisant la distance (12), avec l'ordre $H = 20$. Cette matrice représente les distances entre les vidéos de la base dans un espace de description de grande dimension. Afin de visualiser la distribution des éléments dans cet espace, nous le projetons dans un espace 2D à l'aide de l'algorithme d'Analyse en Composante Curviligne (ACC) [2]. Le résultat de cette projection, présenté figure 2.a, montre que la distribution des vidéos dans l'espace de description est fortement liée au contenu dynamique. La figure 2.b montre la projection obtenue à partir de la matrice de similarité calculée sur les moyennes temporelles des descripteurs. La comparaison de ces deux résultats fait clairement apparaître que la distance basée sur l'erreur de prédiction est plus efficace pour représenter les similarités existantes entre les différentes vidéos.

6 Conclusion

Basée sur une modélisation temporelle du contenu visuel, la distance presentée ci-dessus permet une analyse fine des similarités entre vidéos. Cette distance peut-être employée dans un grand nombre d'applications, allant de la reconnaissance non-supervisée d'activités à la recherche de vidéos par le contenu. Nous travaillons actuellement sur le développement d'outils pour l'exploration de vidéos par le contenu basés sur cette mesure de similarité.

Références

[1] E. Bruno and D. Pellerin. Video structuring, indexing and retrieval based on global motion wavelet coefficients. In *Proceedings of International Conference of Pattern Recognition (ICPR)*, Quebec City, Canada, August 2002.

[2] P. Demartines and J. Herault. Curvilinear component analysis : A self-organising neural network for non linear mapping of data sets. *IEEE Transactions on Neural Networks*, 8(1) :148–154, 1997.

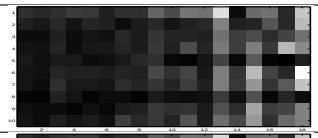
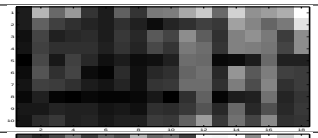
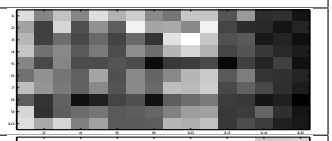
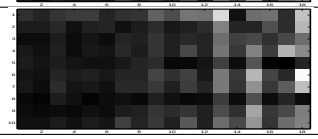
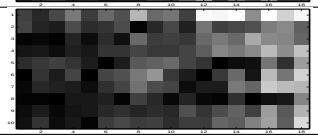
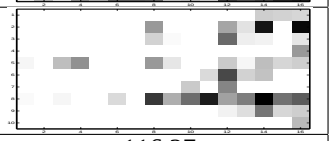
[3] S. Mukherjee, E. Osuna, and F. Girosi. Nonlinear prediction of chaotic time series using support vector machines. In *Proceeding of IEEE Neural Networks for Signal Processing, NNSP'97*, pages 24–26, September 1997.

[4] A. Smola and B. Schölkopf. A tutorial on support vector regression. Neurocolt2 technical report nc2-tr-1998-030, 1998.

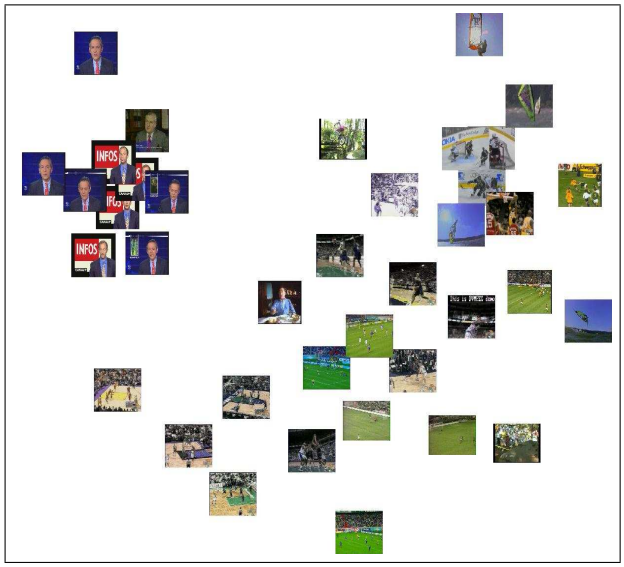
[5] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

[6] J.H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58 :234–244, 1963.

FIG. 1 – Extrait d'une collection de séquences d'images contenant 2 classes d'activité : 5 séquences "*Aller*" et 5 séquences "*Venir*".

Sequences	*Venir1*	*Venir2*	*Aller1*
Descripteurs réels			
Descripteurs prédits par $\mathbf{F}_{Venir1}$			
Erreurs de prédiction (13)	5.77	64.57	116.27
Erreurs sur la moyenne $\|\|\bar{\mathbf{X}} - \bar{\mathbf{Y}}\|\|$	0	114.65	36.33

TAB. 1 – Prédiction des descripteurs des séquences *Venir1*, *Venir2* et *Aller1* par le modèle temporel estimé sur la séquence *Venir1*.



(a)

(b)

FIG. 2 – Projection 2D de l'espace de description avec comme métrique : a) l'erreur de prédiction et b) la distance sur les moyennes temporelles des descripteurs. Chaque vidéo est représentée par son image médiane. La position de l'image dans le plan 2D correspont aux coordonnées des éléments après projection.