Caractérisation spatio-temporelle des co-occurrences par ACP à noyau pour la classification des actions humaines

Aznul Qalid MD SABRI[1,2], Jacques BOONAERT[1], Stéphane LECOEUCHE[1], El Mustapha MOUADDIB[2]

[1]Unité de Recherche Informatique et Automatique, Ecole des Mines de Douai
764, Boulevard Lahure, Douai, France

[2]Laboratoire Modélisation, Information et Systémes (MIS), UPJV
33, rue Saint Leu , 80039 Amiens Cedex 1, France

aznul.sabri@mines-douai.fr, jacques.boonaert@mines-douai.fr
stephane.lecoeuche@mines-douai.fr, mouaddib@u-picardie.fr

Résumé – Ce travail concerne la classification des actions humaines sur la base des co-occurrences spatio-temporelles (ST) entre labels de mots vidéos, codées au sein de corrélogrammes ST. Nous proposons un processus de coalescence s'appuyant sur la notion d'Information Mutuelle (IM) afin de réduire la taille du dictionnaire créé à partir des descripteurs initiaux. La contribution principale de ces travaux concerne l'exploitation de l'ACP à noyau (ACPN) pour réduire la taille des corrélogrammes ST qui sont de grande dimension et "creux" par nature. Ce procédé génère des vecteurs de co-occurrence destinés à caractériser les actions humaines. Nous avons testé notre approche sur la base KTH, standard du domaine, et obtenu des performances conformes à l'état de l'art.

Abstract – This work deals with human action classification by utilizing spatio-temporal (ST) co-occurrences between labels of video-words that are stored within ST correlograms. We suggest the usage of mutual information (MI) based clustering to reduce the size of vocabulary that is created from local descriptors. However, the main contribution of this work is that we propose the usage of KPCA to reduce the size of ST correlograms that are dimensionally large and sparse in nature. This produces a set of ST co-occurrence vectors that can be utilized to characterize human actions. We tested our approach on the KTH dataset which is a standard benchmark in this domain, and obtain state of the art classification performance.

1 Introduction

L'objectif initial de notre travail est d'améliorer les résultats obtenus par Savarese et al. [1] et par Sabri et al. [2] qui exploitent des techniques basées sur des co-occurrences spatio-temporelles (ST) locales. Récemment, le concept « d'Information Mutuelle » (IM) a été utilisé avec succés pour compresser le dictionnaire des mots vidéos dans le contexte de la classification d'actions humaines, [3], [4]. Ceci nous a encouragé à intégrer un processus de sélection des mots vidéos basé sur le concept d'IM dans notre approche. Un dictionnaire de mots vidéos se réfère dans ce contexte à un ensemble représentatif de caractéristiques obtenu après une étape de quantification vectorielle (grâce à un algorithme de coalescence de type « k-means ») s'appliquant aux caractéristiques brutes issues d'une collection de vidéos contenant des actions humaines. Savarese et al. [1] ont introduit l'utilisation d'éléments vectoriels quantifiés associés à des corrélogrammes spatio-temporels qui décrivent les co-occurrences de mots vidéos dans un voisinage spatial et temporel. Sabri et al. [2] ont mis en évidence que l'utilisation d'un type de descripteur discriminant affecte la « texture » globale qui est représentée par le corrélogramme ST.

Les travaux de Savarese et al. font face à une difficulté concernant l'information associée aux labels des mots vidéos, qui est perdue lors de la phase de quantification vectorielle. Pour circonvenir à ce problème, nous proposons d'extraire directement l'information utile des corrélogrammes spatio-temporels sans passer par cette phase de quantification. Nous proposons ainsi l'usage de l'Analyse en Composantes Principales à Noyau (ACPN ou KPCA) [5] afin de réduire la dimension des corrélogrammes. l'ACPN permet d'opérer une ACP sur les différents noyaux plutôt que sur les véritables corrélogrammes ST, qui sont de grande dimension et "creux" par nature. La combinaison des corrélogrammes ST et d'une approche de type « Bag Of Word » (BOW) génère une nouvelle caractérisation des actions, qui est exploitée ici pour entraîner un classifieur SVM. Des performances du niveau de l'état de l'art sont alors obtenues sur la base données KTH [6] en utilisant l'approche proposée.

2 Approche proposée

La figure 1 présente la chaîne de traitement globale permettant d'associer des descriptions des actions humaines basées sur les co-occurrences spatio-temporelles à un ensemble de séquences vidéo. Ces derniéres sont représentatives de différentes actions

comme marcher, trotter et applaudir :

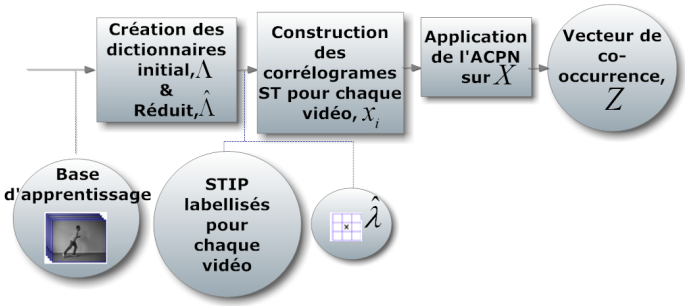

FIGURE 1 – Caractérisation basée sur les co-occurrences spatio-temporelles

2.1 Dictionnaire de mots vidéo et sélection des mots vidéo

Ce paragraphe détaille la construction des mots vidéo dans le but de représenter une séquence par l'intermédiaire d'une approche de type BOW. Celle-ci débute par la détection de points d'intérêt de type STIP (pour « Spatio Temporal Interest Point ») et par la construction des descripteurs qui leur sont associés. Ces STIP correspondent à des zones de saillance maximale dans une vidéo et constituent des indices de mouvement, tandis que le descripteur est un « patch » associé à cette zone et contenant les informations permettant de décrire le mouvement. Dans le cadre de nos travaux et dans l'optique d'une validation sur la base KTH, nous avons choisi d'utiliser des STIP de type Harris3D combinés avec un descripteur construit en concaténant un HOG (« Histogram of Oriented Gradient ») et un HOF (« Histogram Of Flow ») [7] . Les détails d'implémentation de chacun de ces descripteurs sont disponibles sur les sites web de leurs auteurs respectifs [1]. Dans le cadre de nos expérimentations, nous avons utilisé les paramètres par défaut.

En bref, un ensemble de mots vidéo, Λ est construit en appliquant une approche « k-means » sur un ensemble de descripteurs , D, représentant un ensemble, P, de STIP extraits de séquences contenant des actions humaines. Chacun des points d'intérêt est alors labellisé avec le label l de son mot vidéo le plus proche. Cette proximité est évaluée par le biais de la distance Euclienne entre le descripteur D (associé à chacun des points d'intérêt), et les mots vidéo mémorisés Λ. Les fréquences des différents mots au sein d'une séquence sont comptabilisées dans un histogramme qui, utilisé conjointement avec les co-occurrences spatio-temporelles, permet de caractériser les actions humaines.

2.2 Sélection des mots vidéo

Dans le cadre de l'approche BOW, il est fréquent qu'un grand nombre de mots vidéo soi extrait afin de définir les ensembles représentant les différentes actions. La taille de ces collections pouvant être un handicap pour la détermination des corrélogrammes ST, il peut sembler plus judicieux de chercher à travailler avec un sous-ensemble du dictionnaire initial qui préserve cependant le taux de classification. Afin de diminuer la taille de ce dictionnaire initial, nous avons appliqué l'approche développée par Liu et al. dans [3] qui opère la sélection des mots vidéo Λ sur la base de l'Information Mutuelle entre ces derniers et les différentes classes d'action Y.

Ainsi, à partir du dictionnaire initial Λ obtenu après application d'un algorithme de coalescence « k-means », la perte d'information mutuelle est calculée à chaque itération pour toute paire de mots vidéo λ_1 et λ_2. Les mots de la paire sont fusionnés dès lors qu'ils génèrent une perte d'IM minimale. Ce processus d'association est poursuivi jusqu'à ce que la taille désirée K^* pour le dictionnaire soit atteinte. Dans le cadre de nos essais, celle-ci a été fixée entre 20 et 70 pourcents de la taille originale. Le lecteur intéressé par le détail des aspects théoriques concernant l'approche IM pourra se référer à Liu et al. [3] . Suite à cette étape de sélection des mots vidéo sur Λ nous disposons donc du « dictionnaire réduit », noté $\hat{\Lambda}$ dans ce qui suit.

2.3 Construction des corrélogrammes spatio-temporels et extraction des vecteurs de co-occurrence

Dans cette section, nous reprenons les notations introduites dans l'article de Savarese et al [1] pour ce que touche à la génération des corrélogrammes ST. Sauf précision contraire, l'ensemble des mots constituant le vocabulaire se réfère en définitive au dictionnaire $\hat{\Lambda}$ présenté précédemment.

2.3.1 Corrélogrammes spatio-temporels

A l'issue de la construction du dictionnaire réduit $\hat{\Lambda}$, nous procédons de manière similaire à l'approche BOW. Ainsi, à l'aide de leurs descripteurs respectifs contenus dans D, chacun des points d'intérêt (STIP) de l'ensemble P se voit tout d'abord affecter un label correspondant à celui du mot vidéo le plus proche (au sens d'une distance) de $\hat{\Lambda}$. Pour chaque vidéo issue de la collection contenant les différentes classes d'actions, un histogramme local $H(\Pi, p)$ est défini sous la forme d'une fonction vectorielle qui comptabilise le nombre de STIP dotés du même label l au sein d'un même volume spatio-temporel (« noyau ») Π centré sur le point p.

A l'instar de [1], pour chaque position associée à un point d'intérêt p, nous exploitons un ensemble J de noyaux de tailles différentes centrés sur celui-ci. Les volumes associés à ces noyaux sont rectangulaires et s'étendent de 20 à 40 pixels dans les directions spatiales et couvrent de 2 à 60 trames le long de l'axe temporel. Le r^{eme} noyau de cet ensemble est noté Π_r . L'histogramme local moyen est alors défini par :

$$\hat{H}(\Pi_r, l) = \sum_{p \in P_l} \frac{H(\Pi_r, p)}{|P_l|}; 1 \leq r \leq J; 1 \leq l \leq K^* \qquad (1)$$

1. http ://www.di.ens.fr/ laptev/download.html

où P_l désigne l'ensemble des STIP s'étant vus attribuer le label l tandis que $|P_l|$ représente son cardinal. Un corrélogramme ST, x, associé à une vidéo particulière est construit en concaténant dans un tableau les histogrammes locaux correspondant à toutes les combinaisons possibles de labels et de noyaux. Ce que nous traduisons ici par :

$$x = \begin{bmatrix} \hat{H}(\Pi_1,1) & \cdots & \hat{H}(\Pi_1,K^*) \\ \vdots & \ddots & \vdots \\ \hat{H}(\Pi_J,1) & \cdots & \hat{H}(\Pi_J,K^*) \end{bmatrix} \quad (2)$$

2.3.2 Construction du vecteur de co-occurrence spatio temporelle

Dans ce paragraphe, nous expliquons comment nous convertissons le corrélogramme ST en vecteur de co-occurrence spatio-temporelle. Le but de cette approche est d'éviter la quantification vectorielle qui favorise la perte de l'information portée par la co-occurrence ente les mots vidéos. A ce stade, nous disposons d'un ensemble de corrélogrammes ST extraits d'une collection de séquences vidéo contenant différentes classes d'actions, $X = \{x_1, x_2, ...x_M\}$, où M est le nombre de vidéos dans un ensemble particulier. Un « simple » corrélogramme ST est constitué de $K^* \times K^* \times J$ éléments. Chaque corrélogramme, x_i, est tout d'abord réécrit sous la forme d'un vecteur $\hat{x_i}$ en concaténant toutes ses colonnes en une seule. A partir d'une collection $\hat{X} = \{\hat{x_1}, \hat{x_2}, ..., \hat{x_M}\}$ de tels vecteurs, nous appliquons une ACP à noyau [5] de sorte à réduire la dimension de chaque $\hat{x_i}$. Dans le cadre de nos travaux, ceci est réalisé en fixant par avance le nombre de composantes principales, S, puis en projetant chaque élément de $\hat{X}$ sur une nouvelle base orthogonale par le biais d'une fonction noyau donnée. Nous créons ainsi un nouvel ensemble Z formé par ces vecteurs réduits :

$$\begin{aligned} KPCA : \hat{X} &\longmapsto Z \\ \hat{x_i} &\mapsto z_i \\ |\hat{x_i}| = (1 \times K^* \times K^* \times J); &|z_i| = (1 \times S); \\ i \in \{1,...,M\}; &S \ll (K^* \times K^* \times J) \end{aligned} \quad (3)$$

Comme précisé dans [5], il existe une variété de fonctions noyau possibles, parmi lesquelles les noyaux polynomiaux qui ont été adoptés dans nos travaux. Les vecteurs projetés ainsi obtenus, bien que de dimension moindre, intègrent cependant l'essentiel des informations de co-occurrence des labels des mots vidéo qui peut être utilisée pour caractériser les actions humaines. Cette représentation sera exploitée de manière conjointe à une approche BOW pour « décrire » ces dernières.

3 Résultats expérimentaux

Dans ce paragraphe, nous détaillons les résultats expérimentaux obtenus sur la base KTH [6] . Celle-ci contient 6 types d'actions avec des scénarii variables correspondant à des changements d'éclairage, des modifications des personnages (sujets), d'échelle et de localisation (en particulier « extérieure » ou « intérieure »). Dans le cadre de nos expérimentations, nous avons suivi le même protocole que les auteurs [6]. La base est ainsi divisée en un ensemble de test (9 sujets), tandis que le reste (16 sujets) constitue l'ensemble d'apprentissage. La taille initiale (non compressée) du dictionnaire Λ exploité pour produire l'histogramme BOW est de 1000 éléments. Le taux de classification obtenu est alors de 90.74% . En appliquant l'approche IM décrite ici, la taille du dictionnaire réduit $\hat{\Lambda}$ vaut alors 700 éléments tandis que le taux de classification atteint 91.67%. Dans le cadre de nos expérimentations, le critére d'arrêt du processus de sélection des mots vidéos est la taille du dictionnaire réduit, qui doit atteindre une valeur fixée à l'avance (ici 70%). Nous évaluons ensuite les effets de la réduction du dictionnaire par l'intermédiaire des performances obtenues en classification. Sur la base de nos expérimentations, nous avons constaté qu'un dictionnaire ne conservant que 20% des mots initiaux procurait encore des résultats honnorables ($> 75\%$) tandis qu'un ratio de 70% conduisait à des taux tels que 91% et plus. La tendance constatée est similaire à celle décrite dans [3], où une légère amélioration est en effet obtenue en exploitant l'approche IM. Ce dictionnaire réduit $\hat{\Lambda}$ est ensuite utilisé pour produire les vecteurs de co-occurrence spatio-temporelle représentatifs de l'information de co-occurrence entre les labels des mots vidéo de chaque séquence.

Pour fusionner le descripteur basé sur l'approche BOW avec le vecteur de co-occurrence, nous avons exploité la méthode décrite dans [8], qui fait usage d'un SVM à noyau multicanaux. Ici chaque "canal" correspond à une des approches utilisées pour la caractérisation (à savoir, "BOW" et vecteur de co-occurrence spatio-temporelle). Nous avons alors été capables de porter le taux de reconnaissance jusqu'à 93.06 % , (avec, rappelons-le, un dictionnaire réduit) ce qui est comparable à l'état de l'art en la matière, dont [3] qui atteint 94.16% est un exemple. L'examen du tableau 1 nous permet de noter que l'amélioration du taux de reconnaissance est effectif pour la quasi-totalité des classes d'actions notoirement problèmatiques. Ceci est particulièrement notable pour les classes « onduler des bras » et « trotter » qui connaissent un accroissement de ce taux situé entre 5 et 12 %. Un léger progrés est aussi à remarquer pour la classe « courir ». Par ailleurs, l'exploitation des co-occurrences ST a permis de réduire à néant la confusion entre les classes « courir » et « marcher ».

Les résultats obtenus montrent que l'approche que nous proposons est capable de faire face à certaines des difficultés identifiées dans la littérature. Dans ce cas précis, l'amélioration du taux de classification est particulièrement nette, puisqu'il passe de 86.63% à 93.04%. Ceci corrobore indirectement notre postulat qui est qu'il est important de minimiser l'usage de la quantification vectorielle des corrélogrammes ST, ceci afin d'éviter la perte de l'information qui est « intégrée » au sein des co-occurrences entre les labels des mots vidéo.

Il est par ailleurs important de noter que, bien que les performances globales de notre approche en termes de classification soient très légèrement inférieures à celle de [3], cette différence

TABLE 1 – Matrice de confusion combinée. Pour chaque case, la première valeur est obtenue en exploitant les mots vidéo seuls, $K = 1000$, moyenne diagonale : 90.74 %. La seconde valeur est obtenue en combinant les mots vidéo du dictionnaire réduit $K^* = 700$, et le vecteur de co-occurrence spatiale, moyenne diagonale : 93.06 %

Boxer	100.00/ 100.00	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0
Applaudir	2.78 / 2.78	97.22 / 97.22	0 / 0	0 / 0	0 / 0	0 / 0
Onduler	0 / 0	8.33 / 2.78	91.67 / 97.22	0 / 0	0 / 0	0 / 0
Trotter	0 / 0	0 / 0	0 / 0	83.33 / 94.44	13.89 / 2.78	2.78 / 2.78
Courir	0 / 0	0 / 0	0 / 0	25.00 / 30.56	72.22 / 69.44	2.78 / 0
Marcher	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	100.00/ 100.00
	Boxer	Applaudir	Onduler	Trotter	Courir	Marcher

est plutôt minime (approximativement 1%). Qui plus est, les travaux développés dans [3] exploitent une technique permettant de tirer parti de la structure spatio-temporelle de la distribution de caractéristiques locales, motivés en cela par les apports de [9] (mise en correspondance de pyramides spatiales), conceptuellement différents de ce que nous avons mis en oeuvre. En définitive, la seule manière permettant d'évaluer pleinement les performances de notre approche serait de la confronter à une base de données vidéo plus réaliste, telle que « UCF-Sports dataset » [10]. Ceci fait partie de nos futurs travaux à cours termes.

4 Conclusion & travaux futurs

Dans cet article, nous avons proposé l'utilisation de l'ACP à noyau pour produire un vecteur de co-occurrence spatio-temporel qui peut être utilisé pour caractériser les actions humaines. La combinaison de ce type de descripteur avec une approche « Bag Of Words » est capable de produire des résultats du niveau de l'état de l'art avec un dictionnaire réduit tout en diminuant les risques de confusion (par exemple entre "courir" et "marcher"). Les résultats obtenus sur la base KTH nous laissent entrevoir un accroissement de performances significatif sur des bases plus « réalistes », ce qui est l'objet de nos travaux en cours. Qui plus est, comme cette approche est basée sur des techniques de co-occurrence spatio-temporelles locales, il y a ici un potentiel d'amélioration en matière de classification, ne serait-ce qu'en autorisant la détection d'un nombre plus important de STIP afin d'enrichir encore l'information de co-occurrence.

Références

[1] S. Savarese, A. DelPozo, J. Niebles, and L. Fei-Fei, "Spatial-temporal correlatons for unsupervised action classification," in *IEEE Workshop on Motion and video Computing, 2008. WMVC 2008*, 2008, pp. 1–8.

[2] A. Q. Md Sabri, J. Boonaert, S. Lecoeuche, and E. Mouaddib, "Human action classification using surf based spatio-temporal correlated descriptors," in *IEEE International Conference on Image Processing, 2012. ICIP 2012*, 2012.

[3] J. Liu and M. Shah, "Learning human actions via information maximization," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, Jun. 2008, pp. 1–8.

[4] Q. Qiu, Z. Jiang, and R. Chellappa, "Sparse dictionary-based representation and recognition of action attributes." in *ICCV*, D. N. Metaxas, L. Quan, A. Sanfeliu, and L. J. V. Gool, Eds. IEEE, 2011, pp. 707–714.

[5] B. Schlkopf, A. J. Smola, and K. R. Müller, "Kernel principal component analysis," *Advances in kernel methods : support vector learning*, pp. 327–352, 1999.

[6] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions : A local svm approach," in *In Proc. ICPR*, 2004, pp. 32–36.

[7] I. Laptev and T. Lindeberg, "Space-time interest points," *Computer Vision, IEEE International Conference on*, vol. 1, p. 432, 2003.

[8] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories : a comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, Jun. 2007.

[9] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features : Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, ser. CVPR '06. Washington, DC, USA : IEEE Computer Society, 2006, pp. 2169–2178.

[10] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach : a spatio-temporal maximum average correlation height filter for action recognition," in *In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.