Une approche neuronale pour la classification d'actions de sport par la prise en compte du contenu visuel et du mouvement dominant

Moez Baccouche[1,2] Franck Mamalet[1] Christian Wolf[2] Christophe Garcia[1] Atilla Baskurt[2]

[1]Orange Labs - France Télécom R&D
35510 Cesson-Sévigné, France.
{prénom.nom}@orange-ftgroup.com

[2]Laboratoire d'InfoRmatique en Image et Systèmes d'information
Université de Lyon, CNRS, INSA-Lyon, F-69621, France.
{prénom.nom}@insa-lyon.fr

Résumé

Dans cet article, nous proposons une approche de classification automatique de séquences vidéo d'actions de sport. Pour cela, nous extrayons de chaque action des caractéristiques du contenu visuel, en utilisant deux approches, l'une par sac de mots, et l'autre par le mouvement dominant de la scène à chaque instant. La classification de l'évolution temporelle de ces caractéristiques extraites est gérée dynamiquement par un modèle neuronal, basé sur les réseaux de neurones récurrents à large « mémoire court-terme » (LSTM). Les expérimentations faites sur la base « MICC-Soccer-Actions-4 » montrent que l'approche neuronale de classification permet d'obtenir des résultats supérieurs à l'état de l'art ($76\ \%$ *de bonne classification), et que la combinaison des caractéristiques (information visuelle et mouvement dominant) permet un taux de bonne classification de* $92\ \%$.

Mots clefs

Classification d'actions de sport, réseaux de neurones récurrents, *Long Short-Term Memory*, sacs de mots visuels, mouvement dominant.

1 Introduction

Les volumes des contenus audio-visuels mis à disposition ne cessent de croître. Naviguer simplement et rechercher précisément ces contenus au sein de grandes collections devient un problème de première importance. L'un des enjeux majeurs des systèmes d'information s'impose donc comme étant l'indexation et la recherche des vidéos par analyse automatique de leur contenu.

Dans ce contexte, il est important de pouvoir extraire de manière automatique des informations de haut niveau pouvant décrire le contenu sémantique d'une vidéo. Ainsi, de plus en plus de travaux introduisent la notion d'« évènement » ou d'« action » dans des applications diverses (vidéo-surveillance, structuration des contenus télévisuels...). En particulier, les vidéos de sport représentent un type de contenu spécialement intéressant à traiter de part les enjeux commerciaux qui y sont liés.

Plusieurs travaux se sont ainsi intéressés à la classification automatique de séquences vidéo de sport, avec pour objectif la reconnaissance d'évènements de niveau sémantique plus ou moins élevé. On peut en effet distinguer deux catégories de travaux. La première s'intéresse à des évènements de niveau sémantique assez faible, loin de la notion d'action. On peut par exemple citer les travaux de Ekin et al. [1] dans lesquels deux types d'évènements (« phase de jeu » et « pause ») sont identifiés à partir de leurs durées et de leurs angles de prise de vue. Assfalg et al. [2] se basent sur l'extraction de primitives bas-niveau pour classer les plans de dix sports différents en trois catégories : « vue globale du terrain », « zoom sur les joueurs » et « public ». Bien que les résultats relatifs à ces approches soient assez satisfaisants, les évènements identifiés restent néanmoins sémantiquement faibles. Une deuxième catégorie de travaux [3, 4] s'intéresse à des actions complexes mais restent spécifiques vu qu'elles font intervenir des informations relatives au sport étudié (modèle du terrain, règles du jeu...).

Récemment, Ballan et al. [5] ont proposé une approche qui permet de classer des actions de haut niveau sémantique sans faire intervenir d'informations a priori. Le principe est de se baser uniquement sur des primitives décrivant l'aspect visuel de la séquence pour modéliser les actions et d'entrainer un classifieur à reconnaitre ces primitives. La méthode a été testée sur la base « MICC-Soccer-Actions-4 » [5] contenant quatre classes d'action de football différentes (*Shot-on-goal*, *placed-kick* (cf. figure 4), *throw-in* et *goal-kick*). Les taux de classification obtenus sont de $52,75\ \%$ avec un classifieur K-NN et de $73,25\ \%$ avec un classifieur SVM en se basant sur des primitives visuelles. Néanmoins, cette méthode se base sur une représentation simpliste des actions vu qu'elle ne fait intervenir aucune notion de mouvement.

Dans cet article, nous proposons une nouvelle méthode de classification d'actions de sport. Dans la section suivante,

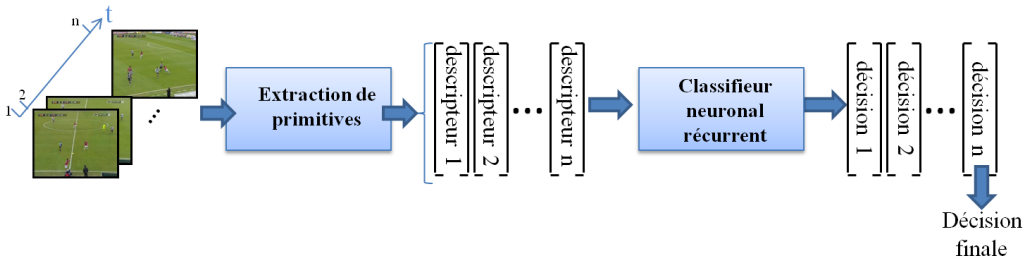

Figure 1 – *Schéma synoptique de la méthode proposée.*

nous présentons l'idée générale de la méthode proposée. Nous décrivons ensuite dans la section 3 les primitives visuelles introduites par Ballan et al. [5], et proposons un ensemble de caractéristiques décrivant le mouvement global à chaque instant de la vidéo, ainsi que la fusion des deux types de caractéristiques. Dans la section 4, nous présentons une méthode de classification, basée sur les réseaux de neurones récurrents. Enfin, nous présentons les résultats expérimentaux sur la base « MICC-Soccer-Actions-4 » dans la section 5.

2 Principe général de la méthode proposée

Le schéma général de la méthode proposée est présenté dans la figure 1. L'idée est de classer des séquences d'actions de sport issues d'une segmentation en plans. Nous calculons, pour chaque image de chaque séquence, un descripteur relatif à une primitive particulière (cf. section 3). Chaque séquence est alors représentée par une succession de descripteurs (un vecteur par image). Le nombre de ces descripteurs peut-être variable selon la longueur de la séquence. La classification est réalisée ensuite en considérant de manière dynamique l'évolution temporelle des primitives extraites. Concrètement, le classifieur devra considérer un à un les descripteurs et prendre une décision finale calculée à partir de plusieurs décisions individuelles accumulées tout au long de la séquence.

3 Extraction de primitives pour la représentation des actions de sport

Afin de décrire le contenu des séquences vidéos, en prenant en compte aussi bien leur aspect visuel que le mouvement, nous avons opté pour une approche par sacs de mots, puis nous avons introduit la notion de mouvement dominant.

3.1 Représentation du contenu visuel : une approche par sacs de mots

Les sacs de mots sont des modèles largement utilisés en traitement d'images en général, et en classification d'objets en particulier. Plusieurs travaux ont essayé d'étendre ces modèles au cas de la vidéo. Par exemple dans [6], les points d'intérêts 2D sont remplacés par des points d'intérêts spatio-temporels, et un modèle par sac de mots spatio-temporels est utilisé pour la classification d'actions humaines. Même si ce type d'approches donne de bons résultats pour les actions simples (action d'une seule personne), la généralisation aux cas des actions de sport, où les mouvements locaux sont en général trop complexes et non représentatifs de la scène, ne donne pas de résultats. Dans cet article, nous avons repris le modèle de Ballan et al. [5] dans lequel une vidéo est représentée par une séquence de sacs de mots visuels (un histogramme de mots SIFT par image). Le dictionnaire de mots est généré en appliquant une classification par k-moyennes sur un large nombre d'images extraites de la base. Le descripteur pour chaque image a donc la taille du dictionnaire et chaque valeur représente la fréquence d'occurrence du mot du dictionnaire dans l'image. Cette représentation permet à la fois de prendre en compte le contenu visuel de la vidéo, mais aussi de modéliser les transitions entre les images à travers l'apparition ou la disparition des mots.
Cette approche va aussi par la suite nous servir de base pour évaluer, sur la base *MICC-Soccer-Actions-4* et dans les mêmes conditions que [5], les performances de la classification neuronale (cf. sous-section 5.2).

3.2 Estimation du mouvement dominant par appariement de points SIFT

Nous proposons d'introduire également un autre type de primitive décrivant le mouvement dominant de la scène. Celui-ci est défini comme étant le mouvement représenté par le plus grand nombre d'éléments de cette scène. Typiquement, pour une action de sport avec une vue globale du terrain (ce qui est le cas pour les actions de la base *MICC-Soccer-Actions-4*), le mouvement dominant se confond avec celui de la caméra, et celui-ci est très caractéristique du type d'actions. L'idée est donc d'estimer ce mouvement puis de l'exploiter pour la classification. Nous avons fait l'hypothèse d'un mouvement affine de la caméra, ce qui est généralement vérifié. Le principe est donc d'estimer la transformation affine T qui permet de passer d'une image I_t d'une vidéo à l'image I_{t+1}.
Pour ce faire, nous effectuons un appariement des points SIFT entre deux images successives de la vidéo. Nous avons recours à l'algorithme *kd-tree* pour une recherche rapide du voisin le plus proche de chaque point. L'algorithme

RANSAC [7] est ensuite appliqué pour séparer le mouvement dominant (celui de la caméra) des mouvements locaux (ceux des joueurs par exemple). Enfin les N paires de points qui ont été considérées comme *inliers* (points conformes) sont utilisées pour estimer les paramètres de la transformation T.
Si l'on pose : $T = [a_1\ a_2\ a_3\ a_4\ t_1\ t_2]^T$ où les a_i sont les coefficients relatifs à la rotation et au facteur d'échelle et les t_i ceux relatifs à la translation, et si l'on note par $\left(x_i^{(t)}, y_i^{(t)}\right)$ pour $i \in \{1, \ldots, N\}$ les N *inliers* relatifs à l'image I_t, la relation entre $\left(x_i^{(t)}, y_i^{(t)}\right)$ et $\left(x_i^{(t+1)}, y_i^{(t+1)}\right)$ sera de la forme :

$$\begin{bmatrix} x_i^{(t+1)} \\ y_i^{(t+1)} \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \begin{bmatrix} x_i^{(t)} \\ y_i^{(t)} \end{bmatrix} + \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$$

En ré-écrivant l'équation précédente pour les N *inliers*, on peut se ramener à un système linéaire d'inconnue T et de la forme :

$$A\,T = B$$

avec :

$$A = \begin{bmatrix} & & \cdots & & & \\ x_i^{(t)} & y_i^{(t)} & 0 & 0 & 1 & 0 \\ 0 & 0 & x_i^{(t)} & y_i^{(t)} & 0 & 1 \\ & & \cdots & & & \end{bmatrix}$$

et :

$$B = \begin{bmatrix} \cdots \\ x_i^{(t+1)} \\ y_i^{(t+1)} \\ \cdots \end{bmatrix}$$

La résolution de ce système se fait par moindres carrés en décomposant la matrice A en valeurs singulières, puis en calculant sa matrice pseudo-inverse. La figure 2 montre un exemple d'appariement de points entre deux images issues de la même vidéo, ainsi que la compensation du mouvement affine estimé.
Pour éviter que les logos soient pris en considération dans l'estimation, nous effectuons un pré-traitement sur toutes les vidéos pour les détecter et les flouter. Pour ce faire, nous avons recours à une combinaison de deux approches. La première permet de détecter, à partir de l'analyse des statistiques de plusieurs images sélectionnées aléatoirement, les pixels immobiles. La deuxième se base sur l'approche introduite dans [8] et qui permet de détecter les textes horizontaux.
Une fois les transformations estimées, les six coefficients sont normalisés, colonne par colonne, entre -1 et 1. Pour cela, pour chaque colonne, nous calculons la moyenne m et l'écart type σ sur toute la base et nous normalisons et tronquons les valeurs en fixant les extremums à $m \pm 2\sigma$ afin de prendre en compte 98 % de la masse, en faisant l'hypothèse d'une distribution Gaussienne. Les descripteurs pour chaque image auront donc une taille de 6 valeurs. Enfin nous proposons de fusionner les deux primitives précédemment décrite pour alimenter le clasifieur avec des vecteurs de taille 36 (qui correspond à la concaténation entre les 6 coefficients précédents et le nombre de mots du dictionnaire).

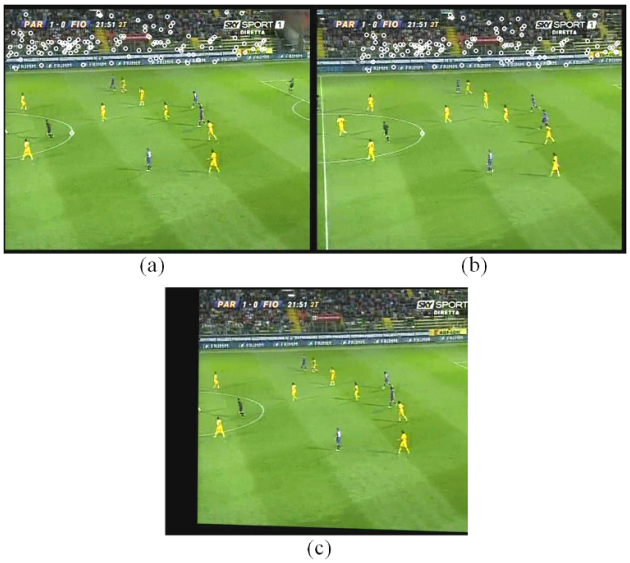

Figure 2 – *Exemple d'estimation du mouvement affine entre deux images : (a,b) - Inliers appariés entre les deux images (c) - Compensation du mouvement sur la première image.*

4 Réseaux de neurones récurrents à base de LSTM pour la classification de séquences

Nous allons nous baser sur les descripteurs extraits pour chaque image représentant les primitives décrites dans la section précédente pour classer les actions de la base *MICC-Soccer-Actions-4*. Le choix de la méthode de classification est alors primordial. Dans [5], deux schémas de classification ont été testés, l'un basé sur les K-NN et l'autre sur les SVM. Dans cet article, nous avons opté pour une approche neuronale récurrente qui permet, comme présenté dans la section 2, d'analyser l'évolution temporelle de ces primitives.
Les réseaux de neurones récurrents (RNN : *Recurrent Neural Networks*) sont des réseaux possédant des connexions récurrentes qui permettent de prendre en compte à un instant t un certain nombre d'états passés. On parle alors de « mémoire à court-terme ». De ce fait, les RNN sont particulièrement adaptés aux applications faisant intervenir le contexte, et plus particulièrement au traitement des séquences.
Néanmoins, pour les applications faisant intervenir de longs écarts temporels (typiquement la classification de séquences vidéos), cette « mémoire à court-terme » n'est pas suffisante. En effet, les RNN « classiques » ne sont capables

de mémoriser que le passé dit « proche », et commencent à « oublier » au bout d'une cinquantaine d'itérations environ. Ce phénomène a été mis en évidence par Hochreiter et al. dans [9]. Les auteurs ont étudié plusieurs algorithmes d'apprentissage (BPTT, RTRL...) pour les RNN et montrent que l'erreur rétro-propagée liée à une entrée du réseau à l'instant t décroit de manière exponentielle après un certain nombre d'itérations. Ce phénomène est particulièrement problématique pour la classification de vidéos (contrairement à la classification de chaque image), puisque le réseau doit attendre la fin de la séquence avant de lui attribuer un label, et par conséquent, si la « mémoire à court-terme » est négligeable devant la taille de la séquence, la mise-à-jour des poids pendant l'apprentissage par rétro-propagation ne prendra en compte que les premiers instants.

Pour remédier à ce problème, Hochreiter et al. [9] ont mis au point des neurones particuliers, à large mémoire court-terme : les LSTM (*Long Short-Term Memory*). Intuitivement, un LSTM peut être vu comme un neurone ayant, en plus des connexions externes, une connexion auto-récurrente de coefficient constant égal à 1. Ceci permet de sauvegarder d'une itération à l'autre les états successifs du neurone. Des portes multiplicatives au niveau de l'entrée (*input gate*) et de la sortie (*output gate*) permettent de protéger respectivement l'état actuel de la mémoire et celui du reste du réseau. Dans cet article, nous allons nous appuyer sur l'architecture proposée par Gers et al. dans [10] (cf. figure 3), dans laquelle une nouvelle porte, dite « *forget gate* », permet de réinitialiser l'état de la mémoire au cours de la séquence.

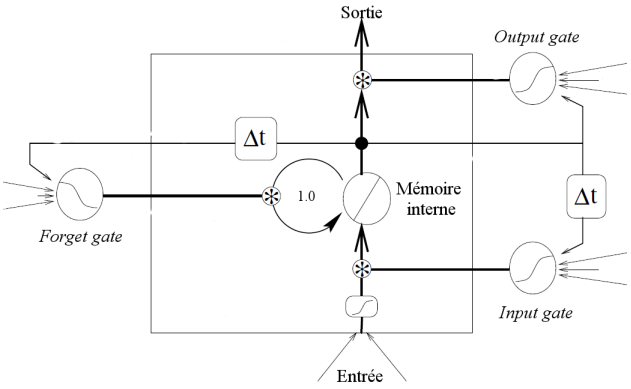

Figure 3 – *Exemple d'un neurone LSTM : architecture proposée dans [10].*

Les LSTM et ont été testés sur différentes applications (apprentissage de CSL, improvisation automatique de musique, classification de phonèmes, reconnaissance d'écriture manuscrite...) avec à chaque fois des résultats au moins aussi bons que ceux de l'état de l'art. Ils ont aussi été utilisés avec succès dans des applications de structuration de vidéos de tennis [11], pour modéliser l'évolution temporelle des transitions entre les plans de la vidéo sans en analyser le contenu. Dans cet article, nous proposons d'utiliser les LSTM pour analyser directement le contenu des vidéos.

5 Résultats expérimentaux

5.1 Données utilisées

Afin d'évaluer notre méthode de classification, nous avons effectué plusieurs tests sur la base publique [1] *MICC-Soccer-Actions-4* [5]. Cette base comprend 100 séquences vidéo au format MPEG-2 pleine résolution PAL (720×576 pixels, 25 images/seconde). La base contient quatre actions : *Shot-on-goal*, *placed-kick*, *throw-in* et *goal-kick* (cf. figure 4).

Figure 4 – *Les quatre actions constituant la base MICC-Soccer-Actions-4 [5] : Shot-on-goal, placed-kick, throw-in et goal-kick.*

Les séquences ont été générées par une détection de plans à partir de 5 vidéos de matchs, faisant intervenir 7 équipes et 5 stades différents, ainsi que des conditions d'éclairage assez variables (notamment 4 matchs se jouant sous un éclairage naturel à différentes heures de la journée, et le dernier sous un éclairage artificiel). Chacune des quatre classes est représentée par 25 séquences de longueurs variables, allant de 100 images à 2500 images. Même si toutes les séquences représentent des vues globales du terrain, la variabilité intra-classe reste très importante puisque les actions se déroulent suivant plusieurs scénarios différents.

Nous présentons dans un premier temps l'évaluation de la classification neuronale en comparant, dans les mêmes conditions, les résultats avec ceux présentés dans [5]. Puis nous étudierons l'apport du mouvement dominant. Enfin nous évaluerons notre proposition de fusionner les deux types de primitives. Toutes les expérimentations ont été effectuées par une validation croisée avec un partitionnement de la base en 3 parties (*3-fold cross validation*), selon la répartition décrite dans le tableau 1, de manière à ce que toutes les séquences soient testées une seule fois. Dans

1. Disponible sur : *www.micc.unifi.it/vim*

ce qui suit, les résultats présentés sont calculés en prenant compte les trois configurations.

	Apprentissage	Test
Config. 1	68 (17/classe)	32 (8/classe)
Config. 2	68 (17/classe)	32 (8/classe)
Config. 3	64 (16/classe)	36 (9/classe)

Tableau 1 – *Répartition du nombre de séquences entre « apprentissage » et « test » pour la validation croisée.*

5.2 Résultats

Evaluation de la classification neuronale basée sur les LSTM. Afin d'évaluer la classification neuronale, nous nous sommes placés dans les mêmes conditions que dans [5] pour pouvoir comparer les résultats. Ainsi, nous avons généré un dictionnaire de 30 mots visuels à partir d'une partie de la base (5 images extraites aléatoirement de chacune des 100 vidéos) avec une classification k-moyennes. Nous avons vérifié que l'augmentation de la taille du dictionnaire n'améliorait pas les résultats, mais augmentait considérablement la complexité, ce qui est en conformité avec les observations présentées dans [5]. Le dictionnaire est ensuite utilisé pour générer des histogrammes de mots visuels comme décrit dans la sous-section 3.1.
Pour le réseau, nous avons utilisé un RNN avec une couche en entrée de taille 30 (une entrée par mot visuel à chaque intervalle de temps de la stimulation du neurone), une couche de sortie de taille 4 (une sortie par classe) et une couche cachée comportant des neurones LSTM unidirectionnels totalement inter-connectés et connectés au reste du réseau. Nous avons noté qu'une augmentation importante du nombre de neurones LSTM conduisait à un sur-apprentissage du réseau (et augmentait considérablement la complexité). De même, un réseau de taille réduite conduit à une divergence de l'apprentissage. Nos expérimentations on montré que 150 LSTM pour la couche cachée est un bon compromis. Ainsi le nombre de poids à optimiser est de 109 654. Enfin, pour la couche de sortie, elle correspond à des fonctions d'activations de type *softmax*.
Le résultat de la classification est reporté sur le tableau 2. Pour comparaison, nous présentons aussi les résultats de Ballan et al. [5], qui correspondent respectivement à des classifications par *K-plus proches voisins* (K-NN) et *machine à vecteur de support* (SVM), les deux combinés avec une distance d'édition pour comparer des vecteurs de tailles différentes.
Le tableau 2 montre que l'approche neuronale est largement plus performante que les méthodes de type K-NN, et est comparable aux méthodes de type SVM. Même si le résultat est supérieur à celui obtenu par les méthodes basées sur les SVM, la différence n'est pas assez importante pour pouvoir généraliser. Néanmoins, les résultats permettent de valider cette approche.

Etude de l'apport du mouvement dominant. Nous allons évaluer l'apport du mouvement dominant pour la classification, dans un premier temps seul, puis en étudiant la possibilité d'une combinaison entre les deux informations. Pour ce faire, nous allons utiliser le même réseau que pour les tests précédents, en modifiant la taille de la couche d'entrée. Nous avons utilisé la répartition apprentissage / test décrite dans le tableau 1, en rajoutant des versions symétriques par rapport à la verticale pour les vidéos de la base d'apprentissage. Ceci permet au réseau d'apprendre pour chaque séquence, deux directions du mouvement. La figure 5-(b) montre la matrice de confusion relative à la classification basée sur le mouvement dominant.

	Taux de classification
k-NN [5]	52,75 %
SVM [5]	73,25 %
Méthode proposée	**76 %**

Tableau 2 – *Evaluation de la classification neuronale RNN-LSTM par rapport aux autres méthodes de classification utilisées dans [5].*

	Goal-kick	Placed-kick	Shot-on-goal	Throw-in
Goal-kick	**0,92**	0,08	0	0
Placed-kick	0,08	**0,8**	0	0,12
Shot-on-goal	0	0,2	**0,72**	0,08
Throw-in	0,12	0,12	0,16	**0.6**

(a)

	Goal-kick	Placed-kick	Shot-on-goal	Throw-in
Goal-kick	**0.64**	0.28	0.08	0
Placed-kick	0.08	**0.68**	0.08	0.16
Shot-on-goal	0.08	0	**0.88**	0.04
Throw-in	0.08	0	0.04	**0.88**

(b)

Figure 5 – *Matrices de confusion : (a) - Classification RNN-LSTM basée sur les sacs de mots visuels (b) - Classification RNN-LSTM basée sur le mouvement dominant.*

Les résultats de la classification sont comparables à ceux obtenus par les approches basées sur les sacs de mots (taux de classification de 77 %) et montrent que le mouvement de la caméra contient beaucoup d'informations discriminantes.
De plus, la comparaison avec la figure 5-(a) montre une complémentarité entre l'information visuelle et le mouvement dominant. En effet, les classes *throw-in* et *shot-on-goal* sont très adaptées à l'approche basée sur le mouvement de la caméra, vu que ce dernier est très caractéristique de l'une et de l'autre (mouvement quasi inexistant pour la première, et très caractéristique pour la deuxième, notamment à cause des zooms sur la cage de but). En revanche, les classes *goal-kick* et *placed-kick* sont caractérisées par des scénarios très variables (en terme de mouvement de la caméra) mais sont particulièrement adaptées à l'approche par le contenu visuel vu que dans les deux cas, l'ordre dans lequel apparaissent / disparaissent les mots est très caractéristique.

Nous proposons donc de concaténer les deux types de caractéristiques en entrée d'un réseau de neurones récurrent. Nous avons repris les expériences précédentes, dans les mêmes conditions, mais en concaténant les vecteurs d'entrée qui sont maintenant de taille 36 (Sacs de mots visuels + mouvement dominant). les résultats présentés dans la figure 6 montrent que ce système est capable de classer correctement 92 % des séquences.

	Goal-kick	Placed-kick	Shot-on-goal	Throw-in
Goal-kick	**1**	0	0	0
Placed-kick	0.04	**0.84**	0.08	0.04
Shot-on-goal	0	0.12	**0.88**	0
Throw-in	0.04	0	0	**0.96**

Figure 6 – *Matrice de confusion relative à la classification RNN-LSTM basée sur la combinaison entre les sacs de mots visuels et le mouvement dominant.*

6 Conclusion et discussion

	Taux de classification
Sacs de mots visuels + K-NN + Distance d'édition [5]	52,75 %
Sacs de mots visuels + SVM + Distance d'édition [5]	73,25 %
Sacs de mots visuels + RNN-LSTM	**76 %**
Mouvement dominant + RNN-LSTM	**77 %**
Sacs de mots visuels + Mouvement dominant + RNN-LSTM	**92 %**

Tableau 3 – *Récapitulatif des résultats obtenus.*

Dans cet article, nous nous sommes intéressés à la problématique de la classification des actions de sport. Pour ce faire, nous avons mis au point une méthode de classification neuronale prenant en compte aussi bien l'aspect visuel que le mouvement dominant. Les résultats de nos expérimentations sur la base « MICC-Soccer-Actions-4 », résumés sur le tableau 3, permettent de conclure que la classification neuronale à base de LSTM est aussi performante que celle basée sur les SVM, et dépasse clairement pour cette problématique les approches de type K-NN. De plus, nous avons démontré que le mouvement de la caméra est une information discriminante entre les classes, et permet à elle seule d'avoir des taux de classification équivalents à ceux utilisant les sacs de mots visuels ($\approx$ 77 %). Enfin, la complémentarité entre l'aspect visuel et le mouvement de la caméra a été prouvée, consuisant à un système de classification « hybride » des séquences vidéo capable d'obtenir un taux de classification de 92 %.
Plusieurs pistes peuvent être envisagées à l'issue de ce travail. Nous prévoyons d'abord de tester cette méthode sur d'autres sports que le football, afin de vérifier la généricité de l'approche. Une autre piste serait d'appliquer la méthode sur des données présentant des scénarios plus complexes, et en considérant plus de classes. Dans ce cas, d'autres informations de mouvement locaux pourraient être intégrées pour différentier entre les actions sémantiquement très proches (par exemple entre un « tir au but » et un « but »).

Références

[1] A. Ekin et A.M. Tekalp. Automatic soccer video analysis and summarization, Août 1 2003. US Patent App. 10/632,110.

[2] J. Assfalg, M. Bertini, C. Colombo, et A. Del Bimbo. Semantic annotation of sports videos. *IEEE MULTIMEDIA*, pages 52–60, 2002.

[3] Y. Gong, TS Lim, et HC Chua. Automatic Parsing of TV Soccer Programs. Dans *IEEE International Conference on Multimedia Computing and Systems*, pages 167–174, 1995.

[4] L.Y. Duan, M. Xu, et Q. Tian. Semantic Shot Classification in Sports Video. *Storage and retrieval for media databases 2003 : 22-23 January 2003, Santa Clara, California, USA*, 5021 :300, 2003.

[5] L. Ballan, M. Bertini, A. Del Bimbo, et G. Serra. Action categorization in soccer videos using string kernels. Dans *Proc. of IEEE international workshop on content-based multimedia indexing (CBMI). Chania, Crete*, 2009.

[6] P. Dollár, V. Rabaud, G. Cottrell, et S. Belongie. Behavior recognition via sparse spatio-temporal features. *ICCV VS-PETS*, 2005.

[7] M. Fischler. Random Sample Consensus : A Paradigm for Model Fitting With Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, 24(6) :381–395, 1981.

[8] C. Wolf, J. Jolion, et F. Chassaing. Text Localization, Enhancement and Binarization in Multimedia Documents. Dans *Proceedings of the International Conference on Pattern Recognition*, pages 1037–1040, 2002.

[9] S. Hochreiter et J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8) :1735–1780, 1997.

[10] F.A. Gers, N.N. Schraudolph, et J. Schmidhuber. Learning precise timing with LSTM recurrent networks. *The Journal of Machine Learning Research*, 3 :115–143, 2003.

[11] E. Delakis. *Structuration multimodale des vidéos de tennis en utilisant des modèles segmentaux*. Thèse de doctorat, Université de Rennes 1, 2006.