Action Classification in Soccer Videos with Long Short-Term Memory Recurrent Neural Networks

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

[1] Orange Labs, 4 rue du Clos Courtel, 35510 Cesson-Sévigné, France.
{firstname.surname}@orange-ftgroup.com
[2] LIRIS, UMR 5205 CNRS, INSA-Lyon, F-69621, France.
{firstname.surname}@insa-lyon.fr

Abstract. In this paper, we propose a novel approach for action classification in soccer videos using a recurrent neural network scheme. Thereby, we extract from each video action at each timestep a set of features which describe both the visual content (by the mean of a BoW approach) and the dominant motion (with a key point based approach). A *Long Short-Term Memory*-based *Recurrent Neural Network* is then trained to classify each video sequence considering the temporal evolution of the features for each timestep. Experimental results on the *MICC-Soccer-Actions-4* database show that the proposed approach outperforms classification methods of related works (with a classification rate of 77 %), and that the combination of the two features (BoW and dominant motion) leads to a classification rate of 92 %.

1 Introduction

Automatic video indexing becomes one of the major challenges in the field of information systems. Thus, more and more works focus on automatic extraction of high-level informations from videos to describe their semantic content. "*Event-based*" and "*Action-based*" classification methods are therefore progressively replacing low-level-based ones, in many applications (closed-circuit television, TV programs structuration...). Especially, sport videos are particularly interesting contents due to their high commercial potential. Several works have dealt with this problem, and can be separated into two main categories. The first one [1] tends to classify sports actions with semantically low-level labels, without using a priori information about the studied sport. On the opposite, the second one [2] extracts high-level semantic information from the sport actions and are domain knowledge-based. Recently, Ballan et al. [3] have proposed a generic approach which is able to semantically classify soccer actions without using a priori information, by relying only on visual content analysis. This approach was experimented on the *MICC-Soccer-Actions-4* database [3], which contains four action classes : *Shot-on-goal*, *Placed-kick*, *Throw-in* and *Goal-kick*. Ballan et al. obtained classification rates of $52,75$ % with a k-NN classifier and $73,25$ % with a SVM-based one.

However, most existing methods make little use of the temporal information of the video sequence. In particular, the evolution of shape over time is not treated. In this paper, we advocate the use of learning machines adapted for sequential data. In this context, *Long Short-Term Memory Recurrent Neural Networks* [4] are a particular type of recurrent neural networks that are well-suited for sequence processing due to their ability to consider the context.

In this paper, we propose an LSTM-RNN scheme to classify soccer actions of the *MICC-Soccer-Actions-4* database [3] using both visual and motion contents. The next section describes the outline of the proposed approach. Then, we present in Sect. 3 the visual and dominant motion features that will be used to fed the classifier. LSTM-RNN fundamentals and used architecture will be outlined in Sect. 4, focusing on their abilities to classify sequences. Finally, experimental results, carried out on the *MICC-Soccer-Actions-4* database, will be presented in Sect. 5.

2 Proposed Approach

The outline of the proposed approach is shown in Fig. 1. The aim is to classify soccer video sequences that are represented by a sequence of descriptors (one descriptor per image) corresponding to a set of features. The choice of those features is crucial for the successful classification (see Sect. 3). A Recurrent Neural Network (RNN) containing Long Short-Term Memory [4] (LSTM) neurons is trained to categorize each action type based on the temporal evolution of the descriptors. To that aim, descriptors are presented to the neural network (one descriptor per timestep) which makes a final decision based on the accumulation of several individual decisions (see Sect. 4).

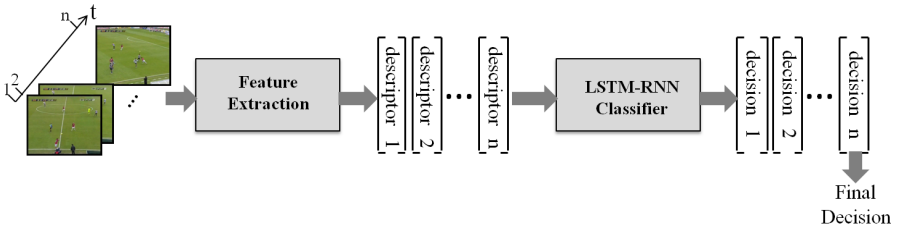

Fig. 1. Proposed classification scheme.

3 Feature Extraction for Action Representation

We have chosen to describe the content of the video sequences by considering both their visual aspect, characterizing the objects appearance, and the motion present in the scene.

3.1 Visual Content Representation : A Bag of Words Approach

Bag of words (BoW) are widely used models in image processing, and particularly in object recognition. The main idea is to represent an image by means of an histogram of visual words, corresponding each to a set of local features extracted from the image. In most cases, these features are SIFT descriptors [5].

In the proposed work, the appearance part of our descriptor is inspired by the work of Ballan et al. [3] where a video is represented by means of a sequence of visual BoW (one BoW per frame). To that aim, we generate a codebook of 30 words (empirical choice) resulting of a K-means classification applied to a large number of images extracted from the database. Then, for each video we associate a sequence of descriptors (one per image) having the same size as the codebook and containing values that encode the occurrence frequency of words present in the sequence. Such a representation allows us to take into account the visual content relative to the scene and also to modelize transitions between images by means of the appearance and the disappearance of words.

3.2 A SIFT-Based Approach for Dominant Motion Estimation

In addition to the appearance descriptor described above, we propose to introduce another feature, that we called *dominant motion*, to describe the movement represented by the largest number of elements of the scene. Obviously, for a sport video with a global view of the playing field (which is the case of all actions of the database *MICC-Soccer-Actions-4*), the *dominant motion* is assumed to be the one related to the camera. We made the assumption that the camera's movement is affine, which is generally true. The idea is then to estimate the affine transformation T between an image I_t at time t and an image I_{t+1} at time $t+1$.

To that aim, we tend to match SIFT points extracted from each two successive frames of the video. A *Kd-tree* algorithm is used to accelerate the nearest neighbor search process. We reject the interest points corresponding to the TV logos which tend to impose a null motion. Thus, we perform a pre-processing step, inspired by the work in [6], which consists in detecting and blurring these logos. Once SIFT matches are computed, we robustly estimate the affine transformation while ignoring outliers (e.g. moving players) using the RANSAC algorithm [7], aiming at only preserving matches corresponding to the dominant motion.

4 Action Classification using LSTM-RNN

Once the descriptors presented in the previous section are calculated, image by image, for each feature (bag of visual words and dominant motion), the next step consists in using them to classify the actions of the video sequences. We propose to use a particular recurrent neural network classifier, namely Long Short-Term Memory, in order to take benefits of its ability to use the temporal evolution of the descriptors for classification.

4.1 Long Short-Term Memory Recurrent Neural Networks

Recurrent Neural Networks (RNN) are a particular category of Artificial Neural Networks which can *remember* previous inputs and use them to influence the network output. This can be done by the use of recurrent connections in the hidden layers. Nevertheless, even if they are able to learn tasks which involve short time lags between inputs and corresponding teacher signals, this *short-term memory* becomes insufficient when dealing with long sequence processing.

The Long Short-Term Memory (LSTM) recurrent architecture was introduced by Schmidhuber et al. [4] in order to provide remedies for the RNN's problem of *exponential error decay*. This is achieved by adding a special node, namely *constant error carousel* (CEC), that allows for constant error signal propagation through time. The second key idea is the use of multiplicative gates to control the access to the CEC.

LSTM have been tested in many applications (CSL learning, music improvisation, phoneme classification...) and generally outperformed existant methods. LSTM have also been used in [8] to structure tennis videos by modelizing transitions between shots, but without analysing their content. In this paper we propose to give as input to the LSTM the extracted features presented in section 3 at each timestep, and train the LSTM network to classify the sport's video sequences.

4.2 Network Architecture and Training

In our experiments, we used a recurrent neural network architecture with one hidden layer of LSTM-cells. The input layer has a variable size depending on which features are set as input (see Sect. 5). For the output layer, we used the *softmax* activation function, which is standard for 1 out of K classification tasks [9]. The *softmax* function ensures that the network outputs are all between 0 and 1, and that their sum is equal to 1 at every timestep. These outputs can then be interpreted as the posterior probabilities of the actions at a given timestep, given all the inputs up to the current one. Finally, the hidden layer contains several one-cell unidirectional LSTM neurons fully inter-connected and fully connected to the rest of the network. We have tested several configuration of networks, varying the number of hidden LSTM, and verified that a large number of memory blocks leads to overfitting, and the opposite leads to divergence. Thus, a configuration of 150 LSTM was found to be a good compromise for this classification task. This architecture corresponds to about 10^5 trainable weights depending on the input size. The network was trained with Online-BPTT with $learning\ rate = 10^{-4}$ and $momentum = 0.9$.

5 Experimental Results

All the experiments presented in this paper were carried out on the *MICC-Soccer-Actions-4* dataset [3] with a *3-fold cross* validation scheme. In order to

Table 1. Summary of obtained results.

	Classification rate
BoW + k-NN [3]	52,75 %
BoW + SVM [3]	73,25 %
BoW + LSTM-RNN	**76 %**
Dominant motion + LSTM-RNN	**77 %**
BoW + dominant motion + LSTM-RNN	**92 %**

study the neural classifier's efficiency and to compare to those used in [3], we have learnt such a network taking as input only the BoW descriptors. The codebook described in subsection 3.1 was used to calculate visual word frequency histograms, retaining 30 entries that we use as input of the network. Classification results are reported in table 1, and compared to those presented in [3]. We also present the confusion matrix in Fig. 2-(a).

	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)

	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**

(c)

Fig. 2. Confusion matrices : (a) - BoW-based approach (b) - Dominant motion-based approach (c) - Combination of the BoW and the dominant motion.

Table 1 shows that the neural classification scheme largely outperforms the k-NN-based approach and gives better results than the SVM-based one. We have then tested the contribution of the dominant motion descriptors using a network with only 6 inputs (see subsection 3.2). The confusion matrix relative to the classification results is shown in Fig. 2-(b). Results are comparable to those obtained by the BoW-based approach - this is a surprisingly good result given that only camera motion information has been used without any appearance information or local (player) motion.

Furthermore, Fig. 2-(a,b) shows that informations provided by the visual appearance and the dominant motion are complementary. Indeed, the dominant motion-based approach is particularly suited for the classes *throw-in* and *shot-on-goal* because of the representative camera motion existing in these actions (non-moving camera for the first and zoom on the goal-keeper at the end of

the action for the last). On the other hand, the classes *goal-kick* and *placed-kick* present highly similar camera movements but distinct characteristic visual words apparition's order.

Therefore, we propose to combine both informations and train a network with an input layer's size of 36 (which corresponds to the concatenation of the dominant motion and the BoW). This network enables us to reach a classification rate of 92 % (see table 1 and Fig. 2-(c)), which outperforms the results corresponding to the use of only one type of features, and is, to our knowledge, the best published result on the *MICC-Soccer-Actions-4* dataset.

6 Conclusion and Future Work

In this paper, we have presented a recurrent neural scheme for soccer actions classification by considering both visual and dominant motion aspects. Experimental results (see table 1) on the *MICC-Soccer-Actions-4* database show that the LSTM-RNN proposed approach is superior, for this application, to SVM-based and k-NN-based ones. Furthermore, we have demonstrated that camera motion descriptors contain as many discriminant information as visual ones (reaching a classification rate of 77 %). We have also shown that the combination of the two information leads to a classification rate of 92 %, which is the best published result on this dataset. More generally, we have demonstrated that LSTM-RNN are able to learn to classify variable length video sequences taking as input features of different nature automatically extracted from the video.

As future work, we plan to verify the genericity of the approach by testing it on other, more-complex video databases. We also plan to jointly learn feature extractors and classification network using a Convolutional Neural Network-LSTM approach.

References

1. Ekin, A., Tekalp, A., Mehrotra, R.: Automatic Soccer Video Analysis and Summarization. IEEE Transactions on Image Processing **12**(7) (2003)
2. Gong, Y., Lim, T., Chua, H.: Automatic Parsing of TV Soccer Programs. In: IEEE International Conference on Multimedia Computing and Systems. (1995) 167–174
3. Ballan, L., Bertini, M., Del Bimbo, A., Serra, G.: Action categorization in soccer videos using string kernels. In: Proc. of IEEE CBMI. Chania, Crete. (2009)
4. Gers, F., Schraudolph, N., Schmidhuber, J.: Learning precise timing with LSTM recurrent networks. The Journal of Machine Learning Research **3** (2003) 115–143
5. Lowe, D.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2) (2004) 91–110
6. Wolf, C., Jolion, J., Chassaing, F.: Text Localization, Enhancement and Binarization in Multimedia Documents. In: Proc. of ICPR. (2002)
7. Fischler, M.: RANSAC: A Paradigm for Model Fitting With Applications to Image Analysis and Automated Cartography. Communications of the ACM (1981)
8. Delakis, E.: Multimodal Tennis Video Structure Analysis with Segment Models. PhD thesis, Université de Rennes 1 (2006)
9. Bishop, C.: Neural networks for pattern recognition. Oxford Univ Press, Inc (2005)