SOCCER VIDEO RETRIVAL USING ADAPTIVE TIME-FREQUENCY METHODS

Jonathan Marchal, Cornel Ioana*, Emanuel Radoi*, André Quinquis*, Sridhar Krishnan***

* : ENSIETA, E3I2 Laboratory, 2 rue François Verny, Brest - FRANCE
E-mails : marchajo@ensieta.fr, ioanaco@ensieta.fr, radoiem@ensieta.fr, quinquis@ensieta.fr
** : Dept. Electrical and Computer Engineering, Ryerson University, Toronto – CANADA
E-mail : krishnan@ee.ryerson.ca

ABSTRACT

The retrieval of soccer highlights is a suitable technique for video indexing, required by the multimedia database management or for the development of television on demand. For these purposes, it should be interesting to have an automatic annotation of events happened in soccer games. One solution consists in analyzing the audio soundtrack associated to the soccer video and to detect the interesting frames.

In this paper we use the adaptive time-frequency decomposition of the soundtrack as a feature extraction procedure. This decomposition is based on the Matching Pursuit concept and a dictionary composed of Gabor functions. The parameters provided by these transformations constitute the input of the classification stage. The results provided for real soccer video will prove the efficiency of the adaptive time-frequency representation as a feature extraction stage.

1. INTRODUCTION

Soccer video highlights retrieval is not only a subject of research but also a need considering the huge amount of data that we can find on the internet. Most of the video archives are not indexed and an automatic parsing approach is a marketable issue. The development of television on demand has also this need of video index. Viewers could have access to the information they need, without having to watch hours and hours of videos.

Several methods have been proposed, based on the information provided by video frames, such as camera motion, court lines detection, motion vectors, location and movements of the players as in [1], others on audio/video features extraction [2] or only on audio features, for instance dominant and excited speech [3].

In this paper, we propose a method based on audio feature extraction. There are typically a tremendous amount of crowd activity, which differs depending on the type of highlight in a game. For instance, when a goal is scored, the crowd cheers are increasing progressively before it, and continues for a few seconds after. For a penalty or free-kick goal, the crowd cheers are sudden, whereas when a goal is missed, crowd cheers begin and stop soon after. Finally, during a normal game sequence, crowd activity is usually not particularly intense. Considering these observations, we assert that if the human ear is able to distinguish the crowd reaction, signal processing tools would be able to do so.
The idea behind this work is to use an adaptive time frequency decomposition (ATFD) [4,5] on the audio soundtrack of the sequences as a starting point for the feature extraction and classification.

The paper is organized as follows. In Section 2 a brief presentation of adaptive time-frequency decomposition concept is done. The classification of the soccer sequences, based on the parameters provided by the ATFD, is described in Section 3. The efficiency of the proposed method is analyzed trough the results in the Section 4. We conclude our discussions in Section 5.

2. ADAPTIVE TIME-FREQUENCY TECHNIQUES

Most of the natural signals are non-stationary. Since their structure is generally complex, some transformations in a more intuitive representation spaces are usually well suited. Linear expansions in a single parameter basis, whether it is a Fourier, wavelet, or any other basis are not flexible enough. A Fourier basis provides a poor representation of functions well localized in time, and wavelet basis are not well adapted to represent functions whose Fourier transforms have a narrow high frequency support. Thus, a flexible decomposition technique can be considered for representing signal components whose localization in time and frequency vary widely.

Matching pursuit (MP), introduced in [4], is a technique that decomposes a signal into a linear expansion
of waveforms that belong to a redundant dictionary of functions. These waveforms, selected in order to best match the signal structure are selected among a dictionary of time-frequency atoms. The aim of the algorithm is to obtain a parsimonious description in order to estimate the original signal with as fewer coefficients as possible. Generally, considering a signal x and a dictionary

$D = \{g_\gamma | \gamma \in \Gamma, \|g_\gamma\| = 1\}$, the signal decomposition is expressed as

$$x = \sum_{\gamma \in \Gamma} \lambda_\gamma g_\gamma \,, \tag{1}$$

where the decomposition coefficients λ_γ are obtained by the inner product between the signal *x* and the function g_γ : $\lambda_\gamma = \langle x, g_\gamma \rangle$.

The MP builds up the signal decomposition one element at a time, picking up the most energy dominant component first. The MP begins by projecting the signal *x* on a function $g_{\gamma_0} \in D$ and computes the residue $Rx = x - \langle x, g_{\gamma_0} \rangle g_{\gamma_0}$. Thus, the *Rx* is orthogonal to g_{γ_0}. The MP algorithm chooses $g_{\gamma_0} \in D$ such that $|\langle x, g_{\gamma_0} \rangle|$ is maximum:

$$|\langle x, g_{\gamma_0} \rangle| \geq \alpha \sup_{\gamma \in \Gamma} |\langle x, g_\gamma \rangle|, \tag{2}$$

where $\alpha \in (0,1]$ is an optimal factor. The MP iterates this procedure by decomposing the residue. If we suppose the m^{th} order residue $R^m x$ has been computed, the next iteration chooses $g_{\gamma_m} \in D$ such that :

$$|\langle R^m x, g_{\gamma_0} \rangle| \geq \alpha \sup_{\gamma \in \Gamma} |\langle R^m x, g_\gamma \rangle|, \tag{3}$$

and deduces $R^{m+1}x$ by :

$$R^{m+1}x = R^m x - \langle R^m x, g_{\gamma_m} \rangle g_{\gamma_m} \tag{4}$$

Summing for *m* between 0 and M-1 yields

$$x = \sum_{m=1}^{M-1} \langle R^m x, g_{\gamma_m} \rangle g_{\gamma_m} + R^M x = \sum_{m=1}^{M-1} a_m g_{\gamma_m} + R^M x \tag{5}$$

where a_m (m=0,..,*M-1*) are the decomposition coefficients.

There are two major factors which guarantee the success of a MP algorithm. The first one is the choice of stop criteria. Since the MP is an iterative decomposition, establishing the number of iterations, *M*, is very important for the considered application. One of the most used criteria is the choice of *M* such that the residual energy is smaller than a percentage of the signal energy:

$$M \left| \varepsilon^2 \|x\|^2 - \|R^{M+1}\|^2 - \text{minimal} \right. \tag{6}$$

This criterion is not well adapted when a signal to noise ratio (SNR) is relatively small [5]. In this case, the signal energy contains the noise contribution and, consequently the correct setup of *M* is almost impossible.

However, since in our application the signals of interest are the soccer soundtracks we can assume that the noise is relatively small and, more importantly, its level and properties are almost the same for all signals. In these conditions, we can use the criterion (6) whose ratio ε is empirically setup.

The second factor which guaranties the efficiency of the MP is the choice of the elementary functions g_γ. Intuitively, the parameters of these functions should ensure a good matching on the signal's time-frequency structures. A common choice is to design a function with as much degrees of freedom (i.e., control parameters) as possible. On the other hand, the time-frequency resolution is another property of interest, especially for feature extraction applications. According to these requirements, we consider for our application an elementary function defined as

$$g_{\gamma_m}(t) = \frac{1}{\sqrt{s_m}} g\left(\frac{t - u_m}{s_m}\right) e^{j(2\pi f_m t + \phi_m)} \tag{7}$$

These atoms are issued by dilatation (s_m), modulation (f_m) and translation (u_m) of the Gaussian window

$$g(t) = 2^{1/4} e^{-\pi t^2} \tag{8}$$

The fourth parameter, ϕ_m, stands for the initial phase. According to this definition, inspired from [4], the atoms (called Gabor functions) are characterized by three parameters : u_m, s_m, f_m, ϕ_m. This type of elementary functions allows to define an adaptive time-frequency tilling, unless the Gabor or wavelet transform (figure 1).

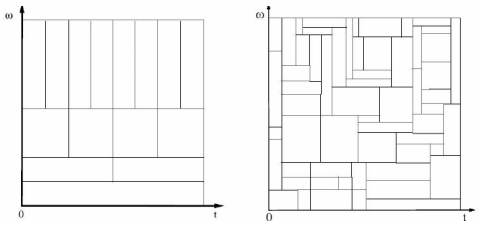

Fig. 1. T-F tilling : MP versus wavelet decompositions

The choice from an arbitrary time-frequency tilling constitutes the main property interesting for characterization purposes. It will be used in the next section for the separation of soccer events based on the MP analysis of the corresponding soundtracks.

3. CLASSIFICATION OF SOCCER HIGHLIGHTS

Most of the time, what interests a soccer watcher are the highlights, such as goals, of course, but also missed goals and scored free-kicks and penalties are of interest. Hence, we have chosen sequences of theses three types, adding a "normal game" class, which is relevant in order to differentiate an interesting sequence (from the 3 classes above) from an "uninteresting" one, in terms of highlight retrieval. This defines 4 classes as illustrated in Fig. 2 : *goals, missed goals, penalties/free-kicks, normal game*.

Fig. 2. Video sequences isolated for soccer retrieval experiments

Once the classes' definition has been done rigorously, the sequences were grabbed form the Internet and from TV recordings. We retained duration of 5 seconds to analyse each highlight video sequence. Indeed, when watching these sequences, we can note that usually, the crowd cheers are growing during the 2 seconds before the ball crosses the goal line, and continues until 3 seconds after, in most cases. All the audio soundtracks of the video sequences have been extracted in a mono, 8 bits, 44,1kHz format. It corresponds to 220500 samples per sequence. The scheme for the feature extraction and classification of the audio soccer sequences shown in Fig. 3 and is as follows:

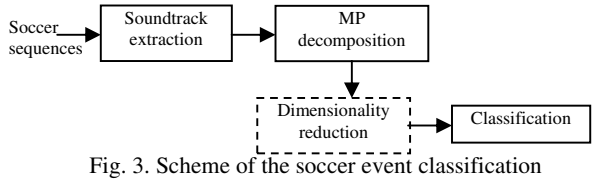

Fig. 3. Scheme of the soccer event classification

The first step is the extraction of the soccer sequence soundtrack. The extracted signals are inputs of the *MP decomposition*. Knowing that we consider N=220500 samples per sequence we setup the parameters of the elementary function dictionary as follows :

- the time parameter, u_n, ranges from 0 to 220499;
- the scale parameter, s_n, ranges from 2 to the closest integer of $log_2(N)$ (17 in our case);
- the frequency parameter, f_n, ranges from 0 to 22050 Hz (the half of sampling frequency). The number of frequency parameters is given by the required spectral resolution. In our application we consider 8192 values which corresponds to a spectral resolution 2.65 Hz.

With the parameters ranging the intervals given previously, the experimental results provided for real data show that the stop criteria (6) needs less than 2200 iterations. For this reason we will limit the number of the iterations to this value. The decomposition parameters are organized in a matrix structure as indicated in (9)

$$\begin{array}{c} \\ \text{Iteration index} \end{array} \begin{bmatrix} \text{Energy} & \text{Octave} & \text{Frequency} & \text{Time} & \text{Phase} \\ E_1 & s_1 & f_1 & t_1 & \phi_1 \\ E_2 & s_2 & f_2 & t_2 & \phi_2 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \quad (9)$$

Experimentally, we observed that these parameters have different discrimination efficiency when applied on audio signals. For example, the time parameter is difficult to be used in our case since it is impossible to synchronize the crowd reaction of all sequences at a fixed sample. Therefore, only the frequency and scale parameters will be used for the classification purpose of audio sequences. It constitutes a first step of data size reduction. Nevertheless, while we work with 2500 iterations, the number of classification parameters is about 5000. In order to reduce the dimensionality of input data, the *linear discriminant analysis* (LDA) technique is applied [6]. LDA is a supervised learning projection that uses information on the within-class scatter and between-class scatter to construct a projection matrix in the reduced space. It maximizes the ratio of between-class variance to the within-class variance in any particular data set thereby guaranteeing maximal separability. As it will be shown in the next section the LDA improves the classification performances compared to the classification in the original space.

Finally, using the feature vectors provided by LDA we used the *nearest neighbors classifier* for the classification task [6]. This operation will be performed in two phases. The first one, *learning stage,* consist in processing a training set of features with apriori known class. The second step, *testing*, is based on the computation of the distance between a new unknown feature vector and each feature vector from the training set. The short distance corresponds to the nearest neighbor. This algorithm will be more computationally intensive as the size of the training set grows but the performances will be improved.

The method proposed in this section has been used for the classification of the soccer sequences for a significant dataset. The most important results will be presented in the next section.

4. RESULTS

In this section we present the results obtained for a data set which consists of 47 sequences, composed of 10 goals sequences, 9 missed goals, 21 normal game sequences and 7 penalties. The sequences are decomposed applying MP algorithm, returning parameters as illustrated in the matrix (9).

The main idea behind the classification process is to compare the modulation frequencies of the Gabor functions with comparable scales for each sequence. This principle has been establishing by comparing the histogram of the

frequency parameters issued from the MP algorithm applied for data corresponding to each class . This is illustrated in the Fig. 4 for the scale 13.

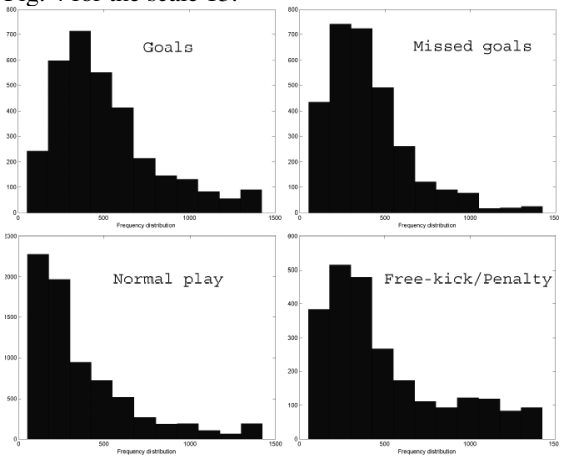

Fig. 4. Mean frequency distributions for the 4 classes

As feature parameters we use the vectors of bins centers. Empirically, we found that the best bins centers vector is of size 12. Applying the LDA algorithm for the vectors obtained for our dataset, three non-zero eigenvalues are obtained. We choose these 3 eigenvectors as a 3D projection space. The classification is then performed using the nearest neighbors (NN) method coupled to the Mahalanobis distance. The LOO (Leave-One-Out) cross-validation technique [7] has been used because of the reduced number of examples in the database. The results are shown in figure 5...

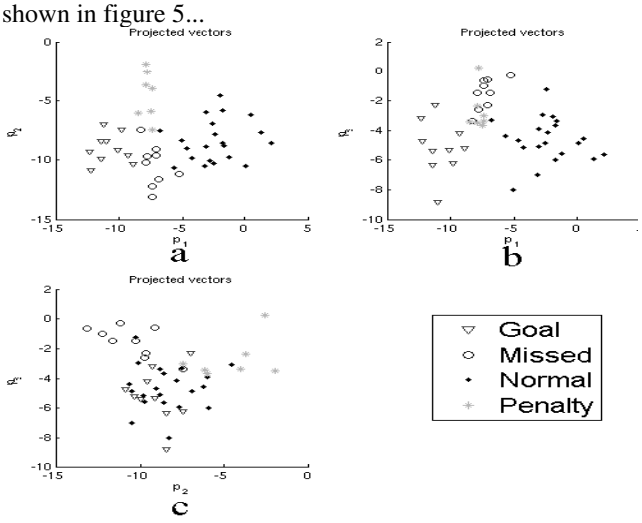

Fig. 5. Clusters given by classification in the reduced space

Note that the four classes are close but properly separated in fig. 5.a, whereas the points corresponding to penalties sequences are spread among the other classes in fig. 5.b This comes from the fact that all the penalty sequences chosen in the set are also scored penalties, so although the crowd cheers are more sudden than for a goal sequence, they are very similar. Table 1 provides the classification results with and without dimensionality reduction (provided by LDA and PCA – principal component analysis).

Table 1. Classification rates

Space	Goal	Missed	Normal	Penalty	*mean*
Initial	90%	89%	86 %	71%	84%
PCA euclidean	100%	89%	81%	43%	78%
LDA euclidean	100%	89%	90%	100%	94%
LDA mahalanobis	100%	100%	100%	100%	100%

The classification accuracies obtained clearly shows that the LDA is well adapted to transform the data provided by the MP algorithm.

5. CONCLUSION

In this paper, we have proposed a new technique for soccer events classification based on the Matching Pursuit algorithm. The dictionary of elementary functions has been manipulated according to the application in hand.

After MP decomposition the feature parameters have been projected via a dimensionality reduction stage. The LDA technique based on the nearest neighbor method yields better classification performances and improves the computational efficiency of the classification stage. In the future works, we intend to use other parameters of the functions with a larger dataset

ACKNOWLEDGEMENTS

The authors would like to thank Lastwave Software developers and Karthi Umapathy of Ryerson University for providing the Matching Pursuit routines.

6. REFERENCES

[1] Y.Gong, L.T.Sin, C.H.Chuan, H.Zhang, M.Sakauchi, “Automatic parsing of TV soccer programs”, Proc. ICMCS95, Washington DC, USA, 1995.

[2] K. Wan, C. Xu, “Efficient multimodal features for automatic soccer highlight generation”, 17th ICPR04, 2004.

[3] K. Wan, C. Xu, “Robust soccer highlight generation with a novel dominant speech feature extractor”, IEEE International Conference on Multimedia Expo ICME, 2004.

[4] S. Mallat and Z. Zhang, “Matching pursuit with time-frequency dictionaries”, IEEE Trans. Signal Processing vol. 41, pp. 3397-3415, Dec. 1993.

[5] S. Mallat, *A Wavelet Tour of signal processing*, Academic Press, 1998.

[6] R.O. Duda, P.E. Hart, D.H. Stork, *Pattern Classification* (2nd ed.), Wiley Interscience, 2000.

[7] K. Fukunaga, *Introduction to statistical pattern recognition* (2nd ed.), Academic Press Professional, Inc. San Diego, CA, USA, 1990.