INDEXING OF NFL VIDEO USING MPEG-7 DESCRIPTORS AND MFCC FEATURES

Syed G. Quadri, Sridhar Krishnan and Ling Guan

Dept. of Electrical and Computer Engineering, Ryerson University
Toronto, Canada, M5B 2K3
{squadri,krishnan,lguan}@ee.ryerson.ca

ABSTRACT

In this paper, we propose an application system to classify American Football (NFL) Video shots into 4 categories, namely: Pass plays, Run plays, Field Goal/Extra Point plays (FG/XP) and Kickoff/Punt plays (K/P). The proposed system consists of two stages. The first stage is responsible for play event localization and the latter stage is responsible for feature mapping and classification. For play event localization we have proposed an algorithm that uses MPEG-7 motion activity descriptor and mean of the magnitudes of motion vectors, in a collaborative manner to detect the starting point of a play event within a video shot with 83% accuracy. The indexing and classification stage uses MPEG-7 motion and audio descriptors along with Mel Frequency Cepstrum Coefficients (MFCC) features to classify the events into 4 categories using Fisher's LDA. We obtain indexing accuracy of 92.5% by using leave-one-out classification technique on a database of 200 video shots taken from 4 different games obtained from 4 different networks.

1. INTRODUCTION

The concept of On-Demand entertainment and programming is fast becoming a reality with the popularity of digital TV channels. Nearly every professional sports league and team in North America has a digital channel boasting of On-Demand programming and statistics. But the reality is that it takes nearly three to four hours in post-production work to prepare the highlights for a game. For example, on NFL Sunday Ticket you get Highlights-On-Demand on Monday morning for the games played on Sunday. In order to minimize this delay, we need a system that can analyze the contents of the broadcast and derive the semantics from the input. These semantics can be made available to the users for querying in order to create a true On-Demand experience. Recently a lot of research has been conducted on automating the process of indexing and annotating the sports video streams. Nearly all the major sports have been used to test the indexing and retrieval systems. One of the major projects working in generating semantic sports video annotations is the ASSAVID project. As detailed in [1], this project focuses on developing a system that can categorize different types of sports and provide users with an interface to query events in a particular sport.

In [2] Miyauchi et. al., used audio, textual and visual information to classify NFL video into events like touchdowns and field goals. In [3] Lazarescu et. al., classified different types of formations within NFL games using the natural language commentary from the game, the geometrical information about the play and the domain knowledge. In [4], Nitta et. al. used closed caption text and audio visual information to classify plays into 3 categories namely: scrimmage, FG/XP and K/P.

All of the works mentioned above rely on domain knowledge to classify different high level concepts within American football. On the other hand, we propose a system that classifies recurring events of the game without using any domain knowledge. These recurring events are the most basic components of the game. By classifying these basic components first we can look for higher concepts contained within each of the basic events and thus generate a hierarchical graph of concepts which varies from low level to high level. In this work we focus on utilizing existing standard descriptors of MEPG-7 as the basic feature set. In [5], the authors have proposed applications for generating summary highlights in sports domain using MPEG-7 motion descriptor, but to our knowledge no one has used MPEG-7 audio and motion descriptors to index recurring events in the American football domain.

Section 2, will detail the algorithm proposed for localization of play events within NFL video shots along with an analysis on the performance of the algorithm. Section 3, provides details on the features set used for indexing and the classification scheme utilized. Section 4, presents the results of the classification scheme and Section 5, provides the concluding remarks and future directions.

2. LOCALIZATION OF PLAY EVENT

Sports have a very well defined structure. They have a set of rules that must be followed in order for the game to be

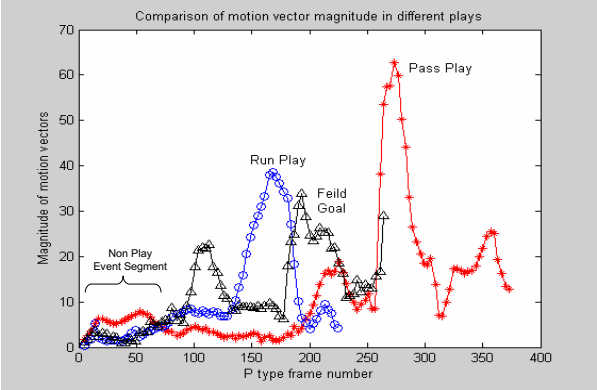

Figure. 1. Motion Vector magnitudes for various plays

played properly. Many sports such as golf, baseball, bowling and American football have a requirement that the team or players must be in a distinctive position before each play. In Golf, the player positions himself by the ball in order to hit it in a certain direction. Likewise in American football the two teams first line up face to face before the ball is snapped to begin the play. The common theme among all these sports is that before the play starts, the level of motion activity in the video is lower compared to when the play has started. This distinction in the motion activity is utilized in the proposed algorithm to segment play events from non-play events. Figure 1, shows the magnitude of motion vectors in different types of NFL plays.

2.1. Proposed Play Event Detection Algorithm

The primary objective of the algorithm is to detect the key frame that can be used as the starting point of the play event in the shot. The end point of the play event is not extracted, as in most American football video shots containing play events, the shot usually terminates at the end of the play.

In order to extract the intensity of motion descriptor, MPEG-1 video motion vectors are used. Only the motion vectors from the P frames are analyzed in order to speed up the processing time. In MPEG-7 the motion activity descriptor represents the Standard Deviation of motion vector magnitudes within a frame. The intensity of motion activity descriptor along with the mean of the motion vector magnitudes is used collaboratively in the algorithm to detect the starting point of the play event. An analysis of 20 video shots selected from each category was conducted to estimate the thresholds for the mean and standard deviation of motion vectors. The following steps detail the algorithm:

Step 1: Find a P frame with a mean value of 4 or higher

Step 2: Determine the gradient of the mean values within a window (3 or 4 adjacent frames)

Step 3: If gradients are all positive mark the frame as possible starting point, else go back to Step 1.

Step 4: If the intensity of motion descriptor has a value of 2 or higher, return frame number as the starting point

Step 5: Otherwise, determine the gradient of the standard deviation values within a window (3 or 4 adjacent frames)

Step 6: If the gradients are all positive return the frame number as the starting point, else go back to Step 1.

2.2. Play Event Detection Algorithm Evaluation

The above algorithm was tested on the American football video shot database which consists of 200 video shots taken from 4 different games and 4 different networks. In order to measure the performance of the algorithm, we have to establish the ground truth about the starting point of the play event within each video shot. This was accomplished by having an observer manually index the frame number which best represented the start point of the play event.

Comparison of results was done by getting the delta between the ground truth frame number and the frame number estimated by the algorithm. The results still needed to be evaluated in terms of what this delta meant in actual time domain. That is we need to evaluate if the algorithm is estimating a starting point too early or if it is estimating the starting point after a certain amount of delay.

Since MPEG-1 video has a frame rate of 30 frames/sec, building a histogram whose bin size was 30 frames would give a general idea of how apart the estimated frame numbers were from the ground truth in actual time domain. Figure 2, shows the histogram of the number of shots within each time unit. Negative time units represents early detection and positive time units represents a delayed detection.

From Figure 2 we can see that the algorithm detects the starting points of the play with 83% accuracy. That is 166 of the 200 video shots in the database had the starting points detected within ± 1 seconds of the original starting point. The accuracy of the algorithm can be increased to 86.5% by increasing the window size from 3 frames to 4 frames. But this change in window size has its own side effect. By increasing the window size we are looking for motion activity being sustained for a longer period of time, which means we will get more shots with delayed detection.

3. INDEXING AND CLASSIFICATION

One of the biggest application areas for MPEG-7 is multimedia indexing and retrieval. Since the introduction of MPEG-7 standard, there has been significant research effort put in developing applications based on descriptors from MPEG-7, but to date there has been only a few applications that utilize MPEG-7 descriptors for sports video indexing and retrieval. The application we are proposing is a first

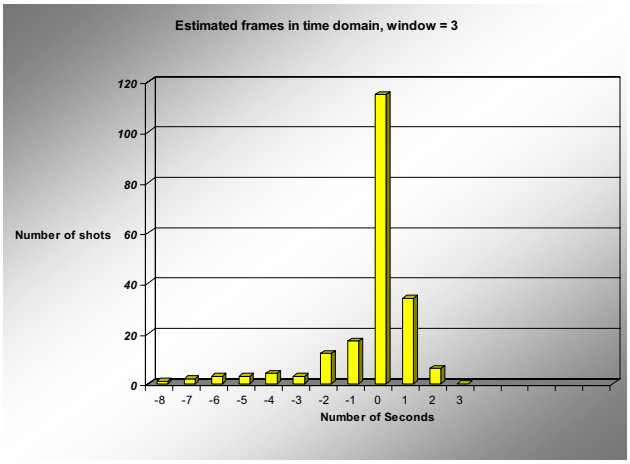

Figure. 2. Performance of algorithm in actual time

in the American football domain, which utilizes MPEG-7 motion and audio descriptors along with MFCC features.

In American football domain visual or motion features play a significantly dominant role in discriminating between different types of plays as evident from Figure 1. Therefore first we evaluate the efficacy of using motion descriptors for an American football video indexing system and then we evaluate the changes in system performance by adding audio descriptors and MFCC features.

3.1. MPEG-7 motion features

According to MPEG-7 description [6], the standard deviation of the magnitude of motion vectors formulate the intensity of motion descriptor. The descriptor takes on the value of 1 through 5, low value meaning low intensity of motion. Experiments done by using 5 levels showed that most of the motion descriptors were quantized into 2 or 3 levels. Thus to provide better motion activity resolution the descriptor was quantized into 12 levels. Similarly according to MEPG-7 description the dominant direction descriptor is calculated by quantizing the angles of the motion vectors into 8 levels. In this work the same 8 quantization levels were used to define the dominant direction descriptor.

A 2D feature map was created by combining the two motion activity descriptors. The motivation behind this was to create a feature set that can model both the intensity of motion and the direction of motion, thus discriminating between high intensity motion in upward direction versus high intensity motion in lateral direction. As can be seen from Figure 3, the feature map provides a unique representation of only 12×8 dimensions for both the intensity and direction of motion. In the feature map, blue colour corresponds to low values and red colour corresponds to high values.

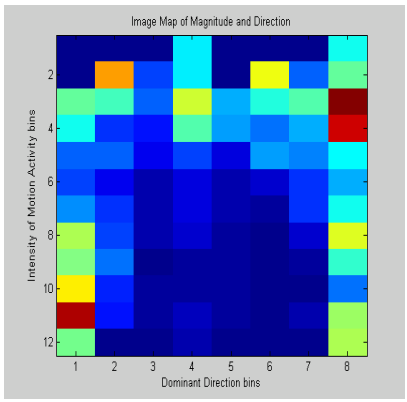

Figure. 3. Motion feature map

3.2. MPEG-7 audio features

The motivation behind using audio descriptors is due to the fact that most sports have a certain vocabulary associated with each event. Almost all the announcers will utilize some of the vocabulary to describe similar events. Therefore we wanted a compact representation of audio characteristics to describe the general tone and pitch of the announcer. The objective is to analyze the similarity in the spoken sound between similar events.

We used 3 MPEG-7 basic spectral audio features, namely: Audio Spectrum Envelope (ASE), Audio Spectrum Centroid (ASC) and Audio Spectrum Flatness (ASF) to achieve our objective. The ASE descriptor represents the power spectrum of an audio signal and can be calculated by taking the Fourier transform (FFT) of the audio signal which is windowed using a Hamming window with an overlap of 50% between adjacent windows.

The ASC descriptor represents the center of gravity of the power spectrum. This is calculated by adding the energy in each frequency bin in the FFT spectrum and dividing it by the total energy in the frame as shown below:

$$\mathbf{ASC}(l) = \frac{\Sigma_{k=0}^{K-1} k.|P(l,k)|^2}{\Sigma_{k=0}^{K-1}|P(l,k)|^2} \ , \qquad (1)$$

where k is the frequency bins index. The descriptor shows which frequencies are dominated in the spectrum.

The ASF descriptor represents the overall tonal component in the power spectrum of the audio signal. It is calculated by calculating the geometric mean of the audio frame and dividing it by the arithmetic mean of the audio frame as shown by the equation:

$$\mathbf{ASF}(l) = \frac{(\Pi_{k=0}^{K-1}|P(l,k)|^2)^{\frac{1}{N}}}{\frac{1}{N}\Sigma_{k=0}^{K-1}|P(l,k)|^2} \ , \qquad (2)$$

where k is the frequency bins index and N is the size of the short time fourier transform window.

All the above descriptor were quantized into 10 levels, thus providing a feature set of 30 dimensions.

3.3. MFCC features

Due to the fact that most of the video shots contain a lot of crowd noise, and we want to extract the perceived rhythm and sound of the spoken content, we needed a feature that can model the human hearing and also works well under noisy condition. MFCC has been used extensively in the speech recognition systems as it tries to emphasize the frequencies that are more perceptive to the human ear.

First the audio file is pre-processed in order to remove the silent segments. Then 13 MFCC coefficients are extracted for each segment. Each of the segments have 50% overlap, and thus there is lot of redundancy between adjacent MFCC values. In order to reduce the dimension of the matrix, the MFCC values are passed to a feature reduction stage. The MFCC features are reduced to a 12×64 matrix.

4. EXPERIMENTAL RESULTS

In order to evaluate the efficacy of the feature set, we used simple classification scheme such as Fisher's Linear Discriminant Analysis (LDA). In a specific sense LDA also commonly refers to techniques in which a transformation is done in order to maximize between-class separability and minimize with-in class variability. LDA works on the feature set with no prior assumptions about the nature of the data set. It tries to compute a weight vector $\boldsymbol{w}$, which when multiplied by the input feature vector $\boldsymbol{x}$ would generate discriminant functions $g_i(\boldsymbol{x})$. For C classes problem we define C discriminant functions $g_1(\boldsymbol{x})...g_C(\boldsymbol{x})$. The feature vector $\boldsymbol{x}$ is assigned to a class whose discriminant function is the largest value of $\boldsymbol{x}$.

The test database consists of 200 video shots with durations varying from 5 seconds to about 25 seconds. In the database there are 88 pass plays, 67 run plays and 45 kicking plays. A total of 8 different teams were used to create the database from 4 different networks. This variety in the database ensured that the sample space of our work was diverse and included all major broadcasters.

Table 1, shows the indexing results of using MPEG-7 motion and audio descriptors along with MFCC features.

5. CONCLUSIONS

In this paper we have proposed a system with two main components. The first component finds the starting points of play events within a video shot. The second component is responsible for indexing and classification of events in the American football domain. Both the components of the system utilize MPEG-7 motion descriptors, while MPEG-7 audio and MFCC features are added to enhance the indexing capabilities of the system.

Play Events	MPEG-7 motion	MPEG-7 motion+audio	MPEG-7 motion audio+MFCC
Pass	79.5%	85.2%	94.3%
Run	92.5%	91.0%	89.6%
FG/XP	87.5%	87.5%	93.8%
K/P	65.5%	82.8%	93.1%
Overall	82.5%	87.0%	92.5%

Table 1. Classification Performance Summary

Although there is no baseline to compare our results with, but somewhat similar works reported in indexing and retrieval of American football events [2] [3] [4], have shown indexing accuracy of 84%, 81% and 84% respectively. In this work the we obtained classification accuracy of 82.5% by using a MPEG-7 motion features alone, while the above mentioned works used multiple modalities. *By using multiple modalities, our system is able to index the events into 4 categories with 92.5% accuracy*.

6. REFERENCES

[1] W.J. Christmas B. Levienaise-Obadia J. Kittler, K. Messer and D. Koubaroulis, "Generation of semantic cues for sports video annotation," in *Proc. of IEEE Intl. Conf. on Image Processing*.

[2] N. Babguchi S. Miyauchi, A. Hirano and T. Kitahashi, "Collaborative multimedia analysis for detecting semantical events from broadcasted sports video," in *Proc. of IEEE 16th Intl. Conf. on Pattern recognition*.

[3] G. West M. Lazarescu, S. Venkatesh and T. Caelli, "On the automated interpretation and indexing of american football," in *IEEE Intl. Conf. on Multimedia Computing and Systems*.

[4] N. Babaguchi N. Nitta and T. Kitahashi, "Extracting actors, actions and events from sports video - a fundamental approach to story tracking," in *Proc. of IEEE Intl.Conf. on Pattern recognition*.

[5] R. Radhakrishnan Z. Xioing and A. Divakaran, "Generation of sports highlights using motion activity in combination with a common audio feature extraction framework," in *Proc. of IEEE Intl. Conf. on Image Processing*.

[6] P. Salembier B.S. Manjunath and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*, John Wiley and Sons, England, UK, 2002.