Unit Detection from American Football TV Broadcast using Multimodal Content Analysis

Guy Ravitz and Mei-Ling Shyu
Department of Electrical and Computer Engineering
University of Miami, Coral Gables, FL 33124, USA
{ravitz, shyu}@miami.edu

Abstract

In this paper, a multimodal unit detection framework to detect and extract units, a novel concept towards event detection and extraction in sports TV broadcasts, is proposed. The proposed unit is defined to be a segment of a sports TV broadcast that describes a potentially interesting event, which possesses the potential of attracting the attention of the observer and satisfy his/her need of viewing the more interesting segments of the broadcast. A number of events that are considered as the unit target events are game impacting events such as score, missed score, penalties, and special game inserts, such as highlights and statistics clips. The proposed framework serves as an efficient data preprocessing procedure that can reduce the amount of data by ridding off the irrelevant data and prepare the remaining data in an efficient way for future event detection and extraction. Several experiments are conducted on various football games from different TV broadcasts, including college football and professional football. The experimental results demonstrate that the proposed framework effectively achieves the goal of data reduction, which is expected to increase the accuracy of event detection and extraction from the American football TV broadcast.

1 Introduction

Recent advances in technology, such as higher computer storage capacity, the availability of affordable digital capturing devices (e.g., digital cameras and digital audio recorders), and accessibility of broadband communication networks have made digital media more popular and in larger volumes. As digital media slowly fused into our lives, various approaches were developed to serve the basic needs of producing, viewing/listening, editing, storing, and broadcasting the information. Especially, the amount of sports-related data that is broadcasted over radio and television is simply overwhelming, which makes it impossible to see or hear it all. This has created a need among fans and professionals to have the ability of viewing or listening to mainly the interesting parts of all these available broadcasts. These needs were answered by applications such as video summarization and highlight/interesting events detection and extraction.

In response to such demands, content-based analysis to extract the information about the content conveyed by the data has emerged. The goal of content-based analysis is to analyze the content of the data in order to understand the semantics and meanings of the occurrences that the data describes. The first generation of applications that have utilized content-based analysis did so in a uni-modal manner, namely audio, video, or text [2–5, 16, 18]. More recently, a second generation of these applications came to life in the form of multimodal content analysis systems [1, 7–9, 14]. These systems processed the content of more than one media type, and used some combination method to arrive at the desired conclusion.

In sports video event detection, two major cinematic concepts adopted are *scenes* and *shots*. Many of the existing systems designed to detect events in sports have adopted video and audio processing approaches from the produced movies [5, 10, 12, 15, 19, 22] using scene and shot based processing to achieve the desired goal. This process generally involves an initial segmentation stage in which the data is segmented to scenes and/or shots, which is followed by some scene or shot based processing stage where the desired events are detected. However, the difficulty of using cinematic based processing arises from the fact that the sports domain is not as nearly structured as the movie and TV show domains. In the sports domain, the cuts, different camera angles, and camera's movement are not predetermined for the most part and depend on the flow of the game. Therefore, the meaning of each of these mentioned actions is not always clear.

In addition, as mentioned earlier, the process of digitizing data has created a huge amount of data, which makes the task of automatically detecting events from long digitized data a complicated task. This task can be simplified by using a pre-processing approach for data reduction and at the same time avoiding the loss of important data. To address these issues, a multimodal content-based preprocessing framework for event detection is proposed whose main purpose is to locate the potentially interesting event segments in a sports TV broadcast, and separate them from the rest of the broadcast. Such a potentially interesting event segment is defined as the "*unit*."

We have identified that one of the most desired type of event for a unit to describe in the area of sports event detection, mainly in the domain of American Football, is some variety of a potentially interesting play. A play segment was defined in [1, 17] as the time during a game when the ball is in bounds and in some kind of motion. At the same time, the players of both teams are in the progress of executing a set of actions whose nature is defensive for the defending team, and offensive for the offending team. For the scope of our study, an interesting play segment is defined to be a play segment which describes a significant occurrence in the game. Some examples could include scoring events such as touchdowns and field goals, forced change of possession events such as fumbles and interceptions, and well executed plays such as long completed pass, long run, first down conversions, to name a few. Finally, a potentially interesting play segment is defined to be a segment which describes a play which possesses the potential to be an interesting play. A set of potentially interesting events will ideally consist of segments of plays, excluding non-interesting segments such as short plays, incomplete short passes, and breaks.

In contrast to the work done on play and break detection, the proposed "*unit*" is not as exact as a play segment. As mentioned in [13], it is not necessary to know the exact beginning and exact ending of the segment. In fact, in many cases, the information before the beginning or after the end of a play can provide some hints regarding the nature of the play. Some examples could be a celebration that might indicate a score or score prevention, and the slow motion segments that might indicate some interestingness level. The proposed "*unit*" is defined to begin shortly before an interesting play starts, which in American Football is usually at the time before the ball is snapped when the teams line up to execute the play, and ends shortly after the play ends. A detailed illustration of a "good" unit and a "bad" unit is provided in Section 3. It is also important to mention that the proposed "*unit*" detection process does not spend the effort of modeling, detecting, and extracting the break segments, and does not require an initialization and training stages such as those in [1, 17]. This results in a process which involves a reduced amount of computational complexity.

There are several advantages of this novel unit concept. First, it serves as a data preprocessing step for event detection via the proposed multimodal framework. It offers an improvement for the initial segmentation stage used by most of the applications by considering the interestingness in the segmentation process, making it an effective data preparation and filtering process. Filtering out unnecessary information before applying more specific event detection, such as touchdown event detection for example, can improve the performance of the event detection framework. In [7], for example, we showed that in the domain of soccer games, the ratio between goal and non-goal events was too small, and therefore the data had to be pre-processed to increase this ratio so that a data mining technique can be applied to detect the goal events successfully. In the domain of American Football, we are facing a similar problem where the ratio between highlights/interesting events, and non interesting events is small as well. The proposed "*unit*" detection framework achieves a desired ratio by separating the non-interesting data from the potentially interesting data.

Second, it releases the system from the cinematic constraints that the other systems have. The unit concept differs from the scene and shot concepts as follows.

- In [11], a scene is defined as one of the subdivisions of a play in which the setting is fixed, or when it presents continuous actions in one place. However, in most sports TV broadcast the location never changes. All of the different events take place on the field or court. Adopting this definition will yield an inefficient segmentation. Another definition presets the concept of a semantic scene. A semantic scene is defined as a collection of shots that are consistent in respect to a certain semantic [6]. According to this definition, the unit is similar to the semantic scene in the sense that it is a segment of consecutive shots that are similar in semantics. The difference is that due to the nature of the multimodal framework, some units can consist of shots that have common visual semantics, common audio semantics, and some common audio-visual semantics. However, the proposed unit extraction is not tied to the requirement of the existence of one specific semantic.

- A video shot is defined as a video sequence that consists of continuous video frames for one camera action [20]. Similarly in [21], a shot is defined to be a sequence of frames captured between a 'start' and a 'stop' camera operations. Based on such a definition, the process of shot detection and extraction in sports TV broadcast will result in many segments due to the use of many cameras and camera angles, a large amount of camera movement in sports TV broadcasts, and the fact that a unit can consist of partial shots.

However, a unit is a collection of consecutive shots. The proposed framework avoids segmenting the data into shots first and then gluing them back to a meaningful unit to avoid complexity.

In this paper, several experiments are conducted on the video data recorded and digitized from five American football games from both college and professional leagues. These games included different quarters, and were broadcasted over three different major networks. Such a data collection can demonstrate the adaptivity and robustness of the proposed framework. Our experimental results show that all the above mentioned problems can be addressed by using the proposed multimodal unit detection framework.

This paper is organized as follows. Section 2 describes our proposed framework in details. The experimental results are presented and analyzed in Section 3. We conclude our study in Section 4.

2 Multimodal Unit Detection Framework

A "unit" is defined as a segment of consecutive frames which describe a potentially interesting event from the raw video such as scores and penalties in the sports domain. Figure 1 presents the proposed multimodal framework that consists of three steps, namely audio-based unit detection, dominant intensity learning, and visual-based unit detection.

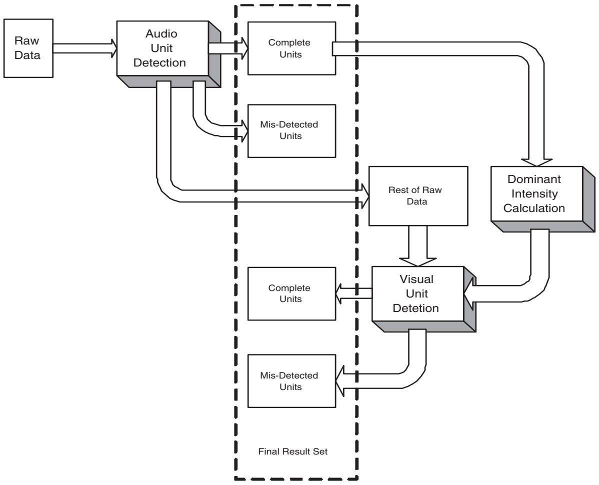

Figure 1. The proposed multimodal unit detection framework.

The proposed framework operates in the following sequence. First, the raw data file is passed through the audio-based unit detection step, where the units are detected and extracted based on the average energy of the audio track. This results in three sets: (i) a set of the complete units, S_1. This set includes those units which were correctly detected and extracted by the audio based algorithm. In other words, this is a set of complete segments which each one of them describes some variation of a potentially interesting event; (ii) a set of mis-detected units, S_2, includes those non-unit segments that were erroneously classified and extracted as units. In other words, this set consists of segments which describe events such as breaks and non-related broadcast interrupts; and (iii) a set of the rest of the raw data, S_3. This last mentioned set is a reduced version of the original raw data set. This set includes break information and potentially interesting events which were missed by the audio based step.

Next, in the dominant intensity learning step, this algorithm uses the good units in S_1 to calculate the median of the dominant intensity of the visual frames. Then, the result is passed to the visual unit detection stage together with the set S_3 to generate a set of the complete units V_1 and a set of mis-detected units V_2. Finally, the output from this framework includes the sets S_1, V_1, S_2, and V_2 obtained from the audio-based step and the visual-based step. The rest of the data is discarded due to the conclusion that it does not contain any information whose content describes a potentially interesting event.

2.1 Audio-based Unit Detection Algorithm

The use of the energy information enables the algorithm to compute whether a segment is potentially interesting or not, based on the crowd and/or commentator's excitement. As mentioned in our previous work [13], interesting events on the American Football field stimulate the crowd and commentators. This stimulation results in a loud and prolonged crowd cheer and louder, faster, and a more excited commentators speech throughout the event. These segments of excitement are represented by segments of high energy levels in the audio track. In order to detect these segments, our audio-based algorithm uses the Average Energy Level feature. The relationship used to calculate the average energy of the audio track is:

$$E = \sum_{n=1}^{L-1} \frac{X(n)^2}{L}, \tag{1}$$

where $X(n)$ is the amplitude value of the n^{th} sample of audio signal X, and L is the total number of samples used to calculate the average energy. As we mentioned in our previous work [13], when extracting features from a continuous audio signal, it is common to segment the signal to smaller units called frames or bins and then extract the features for each of the bins, rather than to extract the features from the entire signal once. The bin size of 50 milliseconds was selected for this purpose of extracting the feature from the audio tracks, which is obtained from some empirical studies. Hence, L is equivalent to 50 ms (2,205 samples) when the audio is sampled at 44.1 kHz. The average energy for each 50ms bin is calculated and all of these results are stored in the Average Energy Vector (AEV). Each index of the vector is populated with the average energy value for the 50 milliseconds it represents. For example, AEV[0] contains the average energy value for the first segment of 50 milliseconds of the audio track, AEV[1] contains the average energy value for the second segment of 50 milliseconds of the audio track, and so on until AEV[n] which contains the average energy value for the n^th (and the last) segment of 50 milliseconds of the audio track.

In this algorithm, three different thresholds are defined. The first threshold, *thresh1*, is an amplitude test. Its value is calculated from the data. After many observations, it is set to about 50% of the central tendency of the AEV vector, i.e., $thresh1$ = 0.53 × median(AEV). The second threshold *thresh2* is used as a duration test whose value is determined using domain knowledge. That is, an important segment (play) cannot be shorter than 5 seconds, which are represented by 100 bins. The third threshold, *thresh3*, is used to find the end of the unit to avoid over-segmenting a unit. Its value is also determined using domain knowledge that it can have up to 500 milliseconds (10 bins) where the value of the bins is less than *thresh1* at any point, and still be considered as a member of the unit. More detailed discussion on this algorithm can be found in [13].

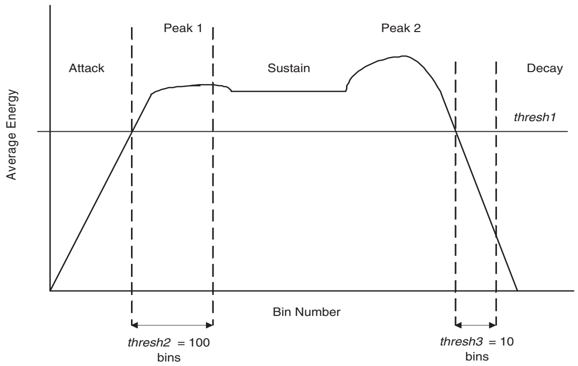

Figure 2. Envelope used to model the average energy audio characteristics of a unit

The three mentioned above thresholds were designed to detect the audio envelope presented in Figure 2. This envelope was observed to describe the typical audio characteristics of a unit. It can be seen from the figure that a typical audio pattern of a unit begins with a sharp raise in the audio level (attack) causing a sharp increase in the audio average energy which reaches a certain peak level (peak 1) usually due to an exciting burst at the snap of the ball. Next, the level is observed to slightly drop, and remain fairly constant for the duration of the event (Sustain). This sustain describes a continuous crowd and commentators excitement during the interesting event. The sustain ends with another peak, usually due to another excitement burst at the end of an event, which is followed by a drop in the average energy (decay) which notes the end of the event. Figure 2 describes how the three mentioned above thresholds were used to model this envelop.

As we have learned from our study, this envelope unfortunately describes the characteristics of part but not all of the units in the American Football broadcast domain. In [13], we have reported a high precision performance of (96%) but a lower recall of (79%) when using the audio average energy only. This translates to a higher ability of the algorithm to extract units when they are detected, but a lower ability to detect the majority of the units. We have addressed this issue by adding a visual feature to the framework in order to improve the recall performance of the proposed framework. This visual feature is described in more details in the following subsection.

2.2 Dominant Intensity Learning

In the proposed framework, the dominant intensity in the video frames is adopted as the visual feature in the visual-based unit detection algorithm. It was observed that during the interesting play segments, the dominant presence of pixels was those pixels representing the background, which is the field. Hence, it was concluded that identifying sequences of video frames which are dominated by the grass (field) pixels will result in unit detection. In the literature, a common feature used to identify such dominance is called the dominant-color ratio. In [1, 9, 17], some variation of the dominant color ratio feature extraction was used and demonstrated high levels of success. However, this feature extraction requires calculations which involve some level of computational complexity, since it is extracted from the color information of the video frames. Keeping in mind that the proposed unit detection in this paper is designed to perform as a pre-processing framework, some simplifications are introduced.

First, the processing is all done using the gray level images, and therefore dominant intensity is calculated instead of the dominant color. Second, the proposed algorithm spends no effort on ratio calculation. Figure 3 illustrates that determining the difference between frames which include a large number of field pixels to ones that do not, can be simply done by comparing the index of the maximum intensity in the histogram for each frame. This index represents the dominant intensity of each frame. At the top of Figure 3, a sequence of frames representing a play segment is presented. Below that sequence, four frames which are members of four different non-interesting events are presented. The number below each of these frames denotes the index value of the maximum intensity of the histogram of each of these frames. Hence, the dominant intensity was calculated to identify the video frames that describe the presence of a large amount of field pixels.

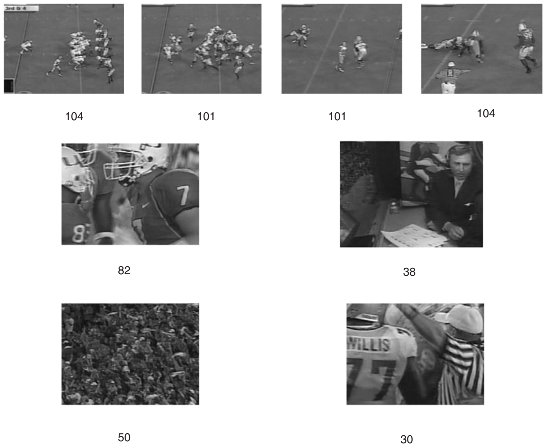

Figure 3. Examples of representative frames and their respective index of dominant intensity.

The dominant intensity is learned by calculating the dominant intensity of the units that were detected and extracted in the first audio-based detection stage of the proposed framework. Then the median of the intensities is used as the dominant intensity. We believe such a design can yield a more robust learning stage due to the fact that in an American football broadcast, the ratio between the grass dominated frames and non-grass dominated frames is much smaller, and learning the grass pixel intensity from randomly selected frames cannot reach the same level of confidence. This high confidence level was made possible since most of the units extracted by this first detection stage describe play events, in which the dominant intensity will be of the pixels representing the field.

2.3 Visual-based Unit Detection Algorithm

The main purpose of this stage is to detect and extract the units that the first audio-based algorithm missed, and hence it takes the set S_3 and the calculated dominant intensity values as its input. It works almost identically to the audio-based algorithm, except it uses the dominant intensity feature instead of the average energy vector and uses some different threshold values.

In this algorithm, the content of each bin is the average dominant intensity. The relationship used to calculate this value is:

$$Di = \frac{\sum_{n=1}^{L-1} MAX(Hist(M(n)))}{L}, \qquad (2)$$

where L is the number of frames in a bin (nine in this case) and $M(n)$ is the n^{th} frame of the movie file. The bin size is chosen to be 9 frames, representing about 1/3 of a second for video sampled at 30 frames per second. Again, as in audio processing, the goal is to extract features from smaller units rather than from the entire video clip.

The value of *thresh1* is set to the dominant intensity index value learned in the Dominant Intensity Learning stage, and the algorithm tests whether the bin value is within + or - 10% of the dominant intensity value. It is important to notice that the dominant intensity is calculated only from a specified region of the frame but not the entire frame, which eliminates, as much as possible, the phenomena of sideline noise. This noise is caused by the appearance and disappearance of the sidelines and the crowd in the frames due to the large amount of camera movement. This noise can affect the dominant intensity value of the video frames. This region can be observed in Figure 4, where this mentioned region is bounded by the rectangle surrounding the center of the frame.

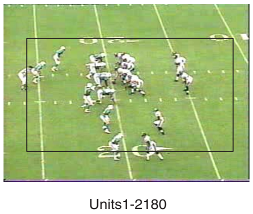

Figure 4. Region for which dominant intensity is calculated to avoid sideline noise

The *thresh2* value is set to the same value of five seconds of *thresh2* of the audio based algorithm. As mentioned earlier, this threshold gives the unit a minimum duration, under which it will not be counted as a potentially interesting segment. Finally, the *thresh3* value is set to 2 seconds, which is about 6 bins or 54 frames. This threshold was set mainly due to the fact that sometimes, just before a play would start, the broadcast will show a close-up of one of the players, which will change the dominant intensity of the video frames. In case that the algorithm counts a consecutive number of bins that are in the + or - 10% of the *thresh1* range, and the count is larger than *thresh2* but there is a 1 to 2 seconds drastic change in the dominant intensity value, this will over-segment the segments to two incomplete units. Therefore, similarly to the audio based algorithm, *thresh3* was introduced to avoid the over-segmentation of the units.

The addition of the visual-based algorithm which used the result of the audio-based algorithm to learn the dominant intensity value helped improve the performance of the overall framework by reducing the amount of missed units, which will be demonstrated in Section 3.

3 Experiments and Analysis

As mentioned earlier, in this study, an interesting event in the American football includes the plays which resulted in touchdowns, interceptions, or field goals; well executed plays (defense or offense) such as getting or preventing a first down; and an explanation of a penalty by a referee, inserted highlight clips from production, etc.

3.1 Data Preparation

In data preparation, about 140 minutes of American football video was recorded and digitized. This data was of game-time only, and the game-time is referred to as the broadcast segments where the content of the video describes only the game related events. Such events include plays, crowd shots, player/referee/coach shots, and commentator shots. This game-time does not include commercials or interrupts such as news breaks. There exists some technology to detect those types of interrupts, which is beyond the scope of this paper. The data collected was from five different games, in both college and professional leagues, including different quarters, and was broadcasted over three different major networks. Such a data collection represents various game broadcasting styles, which can be used to demonstrate the adaptivity and robustness of the proposed framework.

The recorded material was later digitized using an ATI All-In Wonder XT video card. This card is installed in a computer with an Intel Pentium 4 processor and running in the Windows XP Professional environment. Virtualdub and TMPGEnc were used to generate the audio-only and video-only files, respectively. The audio was sampled using a rate of 44.1 KHz, at 16 bit. The video files were sampled at a rate of 30 frames per second. Finally, the processing was done using the MathWorks Matlab software.

3.2 Performance Evaluation

In our earlier study, we have reported the precision (96%) and recall (79%) of the audio-based unit detection algorithm from three games with a total of 120 minutes [13]. Though this audio-based algorithm alone achieved exceptionally good extraction performance, there were a large number of units that were completely missed by it. In order to improve the performance, the proposed multimodal framework is developed.

In order to evaluate the proposed framework, it was necessary to define a good unit in a measurable manner. An illustration of a "good" unit and a "bad" unit is provided in Figure 5. The top of this figure describes a typical sequence of events in an American Football TV broadcast, and the below two segments present the possible outcome of the unit detection process. The first, on the left, is labeled as a good unit. This segment includes an entire potentially interesting event, with some extra information before and after the play segment. This unit is considered a good unit as it follows the two guidelines of a good unit, which are (i) a segment which includes a potentially interesting event; and (ii) a segment which is complete. A complete segment is a segment which includes the beginning and the and of an event. The amount of extra information a good unit includes is limited for evaluation purposes, which is discussed in more details in our ground truth rules for performance evaluation. An example of a bad unit is provided on the right, where it can be seen that the segment is not complete. These considerations enable us to evaluate our algorithm based on two performance metrics, namely the correct detection of potentially interesting events and the complete extraction of these events.

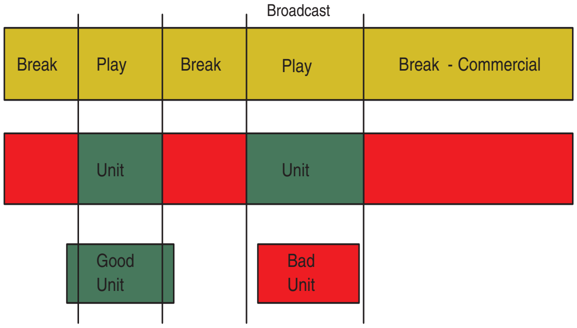

Figure 5. A good unit vs. a bad unit

As a preparation for the experiments, all the examined video clips were viewed, to log the start-time and end-time for each potentially interesting event segment. The start time is logged as the time before the ball is snapped, and the end time is logged as one or two seconds after the action related to the play segment ended. The goal is for the proposed framework to detect and extract these segments as units. The unit start and end times generated by the algorithm are compared to the list in the log file in the following manner (the ground truth rules):

- A detection of a unit was declared as a success, as long as the start time was earlier than the start time from the log, and the end time was later than the one in the log for the respective interesting event segment. Since one of the goals of this process was data reduction, a limit of 2 seconds was set on how much extra information the unit can contain. If a unit does not pass this rule, meaning it is too short or too long, it is counted as a missed segment.

- If an interesting event segment in the log is not included in any unit detected by the algorithm, it is counted as a missed segment.
- A unit generated by the algorithm that did not exist in the log is considered as a mis-detected unit.

Table 1. Experimental results

G	Q	Clip	IE	D	M	M-D	P	R
j_m	1	clip_1	7	6	1	0	1	0.83
j_m	1	clip_2	7	6	1	0	1	0.85
c_f	3	clip_2	5	5	0	0	1	1
c_f	3	clip_3	3	2	1	0	1	0.6
c_f	3	clip_4	5	5	0	1	0.83	1
c_f	3	clip_5	7	6	1	1	0.85	0.85
c_f	3	clip_6	4	4	0	0	1	1
c_f	4	clip_1	8	8	0	1	0.88	1
c_f	4	clip_2	6	6	0	0	1	1
c_f	4	clip_4	6	6	0	1	0.85	1
c_f	4	clip_5	5	4	1	1	0.75	0.8
c_f	4	clip_6	3	3	0	0	1	1
b_g	1	clip_1	5	5	0	0	1	1
b_g	1	clip_2	8	8	0	2	0.67	1
n_c	3	clip_1	8	8	0	0	1	1
n_c	3	clip_2	6	5	1	1	0.83	0.83
n_c	3	clip_3	8	6	2	0	1	0.75
n_c	4	clip_1	7	7	0	0	1	1
n_c	4	clip_2	5	4	1	2	0.66	0.8
n_c	4	clip_3	7	6	1	0	1	0.86
n_c	4	clip_4	6	6	0	0	1	1
n_c	4	clip_5	6	6	0	1	0.86	1
n_c	4	clip_6	8	8	0	1	0.88	1
w_d	3	clip_1	8	6	2	0	1	0.75
w_d	3	clip_2	7	6	1	1	0.86	0.86
w_d	3	clip_3	8	7	1	1	0.88	0.88
w_d	3	clip_4	8	6	2	1	0.86	0.75
O			**171**	**155**	**16**	**15**	**0.91**	**0.91**

Table 1 gives the experimental results, where the five games are denoted by j_m, c_f, b_g, n_c, and w_d. Each broadcast video was divided into quarters, and each quarter was divided into several clips. Each clip was about five minutes long, and was named clip_x, where 'x' stands for the chronological order of the clips within the specific game and quarter. The data was divided into shorter clips to avoid memory related issues involving Matlab and the environment the experiment was conducted. Table 1 reads as follows: the columns from left to right represent the game (G), quarter (Q), clip (Clip), number of interesting events in the clip (IE), detected units (D), missed units (M), mis-detected units (M-D), precision (P), and recall (R). The last row of Table 1 provides overall (O) performance details. The total number of units logged is **171**, where **155** units were detected correctly, **16** units were missed, and **15** units were mis-detected. That is, our proposed multimodal framework achieves **91%** in both recall and precision, which demonstrates that use of the combination of audiovisual features improves the recall performance. Those segments that were completely missed by the audio-based processing step failed to pass the different threshold tests of the audio-based algorithm. On the other hand, these missed segments were then detected when they passed through the video-based processing step.

The results clearly present the superiority of the proposed framework in detecting and extracting units from American Football TV broadcasts. The audio-based algorithm can achieve a very good precision value; while the visual-based feature boosts the recall value. Moreover, the proposed multimodal framework effectively reduces the data from **8,163** to **3,441** seconds, a reduction of approximately **58%** of the data. This demonstrates that the proposed framework can serve as a data preprocessing procedure that reduces the time complexity and at the same time assists in improving the performance of event detection and extraction.

4 Conclusions

A new multimodal unit detection framework toward event detection and extraction from sports TV broadcasts, using multimodal content analysis, is presented in this paper. The proposed multimodal framework will serve as a data preprocessing procedure that filters out the irrelevant data by detecting and extracting the potentially interesting segments, called the *units*, from the sports broadcast video. In the proposed framework, the average energy was first used to detect and extract the units from the raw data in the audio-based algorithm, and then the dominant intensity feature was used to detect the units from the remaining raw data that was not extracted by the audio-based algorithm. It is shown that it performs with satisfactory overall results of 91% in precision and recall. In addition, it can significantly filter out 58% of the irrelevant data to reduce the data set. It is expected that the chance of correct classification by the framework will increase due to the reduction in the amount of irrelevant data, and a more accurate set of units can lead to a better result in event detection and extraction.

5. Acknowledgment

For Mei-Ling Shyu, this research was supported in part by NSF ITR (Medium) IIS-0325260 and National Oceanic and Atmospheric Administration (NOAA). This research was carried out in part under the auspices of the Cooperative Institute for Marine and Atmospheric Studies (CIMAS), a Joint Institute of the University of Miami and NOAA, cooperative agreement #NA17RJ1226. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of the funding agency.

References

[1] M. Abdel-Mottaleb and G. Ravitz. Detection of plays and breaks in football games using audiovisual features and hmm. In *Proceedings of the 9th International Conference on Distributed Multimedia Systems*, pages 154–159, Miami, Florida, September 2003.

[2] K. Abe, T. Taketa, and H. Nunokawa. An efficient information retrieval method in www using genetic algorithms. In *Proceedings of International Workshop on Parallel Processing*, pages 522–527, Aizu-Wakamatsu City, Japan, September 1999.

[3] Y. Avrithis, N. Tsapatsoulis, and S. Kollias. Broadcast news parsing using visual cues: a robust face detection approach. In *Proceedings of IEEE International Conference on Multimedia and Expo*, volume 3, pages 1469–1472, New York City, New York, July-August 2000.

[4] T. Bauer and D. Leake. Using document access sequences to recommend customized information. *IEEE Intelligent Systems*, 17(6):27–33, November-December 2002.

[5] P. Chang, M. Han, and Y. Gong. Extract highlights from baseball game video with hidden markov models. In *Proceedings of the 2002 International Conference on Image Processing ICIP 2002*, volume 1, pages 609–612, Rochester, New York, September 2002.

[6] S.-F. Chang and H. Sundaram. Structural and semantic analysis of video. In *Proceedings of IEEE International Conference on Multimedia and Expo*, volume 2, pages 687–690, New York City, New York, July-August 2000.

[7] S.-C. Chen, M.-L. Shyu, M. Chen, and C. Zhang. A decision tree-based multimodal data mining framework for soccer goal detection. In *IEEE International Conference on Multimedia and Expo (ICME 2004)*, volume 1, pages 265–268, Taipei, Taiwan, June 2004.

[8] S.-C. Chen, M.-L. Shyu, W. Liao, and C. Zhang. Scene change detection by audio and video clues. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME2002)*, pages 365–368, Lausanne, Switzerland, August 2002.

[9] S. Dagtas and M. Abdel-Mottaleb. Extraction of tv highlights using multimedia features. In *IEEE 4th workshop on Multimedia Signal Proessing*, pages 91–96, France, October 2001.

[10] A. Ekin and M. Tekalp. Generic play-break event detection for summarization and hierarchical sports video analysis. In *Proceedings of 2003 IEEE International Conference on Multimedia and Expo (ICME '03)*, volume 1, pages 169–172, Baltimore, MaryLand, July 2003.

[11] Z. Rasheed and M. Shah. Scene detection in hollywood movies and tv shows. In *IEEE Proceedings of the computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)*, volume 2, pages 343–348, Madison, Wisconsin, June 2003.

[12] D. Sadlier, N. O'Connor, S. Marlow, and N. Murphy. A combined audio-visual contribution to event detection in field sports broadcast video. case study: Gaelic football. In *Pricedding of the 3rd IEEE International Symposium on Signal Processing and Information Technology ISSPIT 2003*, pages 552–555, December 2003.

[13] M.-L. Shyu, G. Ravitz, and S.-C. Chen. Unit detection in american football tv broadcasts using average energy of audio track. In *Proceedings of the IEEE Sixth International Symposium on Multimedia Software Engineering*, volume 3, pages 193–200, Miami, Florida, December 2004.

[14] H. Sundram and S.-F. Chang. Computable scenes and structures in films. *IEEE Transactions on Multimedia*, 4(4):482–491, December 2002.

[15] X.-F. Tong, H.-Q. Lu, and Q.-S. Liu. A three-layer event detection framework and its application in soccer video. In *2004 IEEE International Conference on Multimedia and Expo (ICME)*, volume 3, pages 1551–1554, June 2004.

[16] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, July 2002.

[17] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun. Structure analysis of soccer video with hidden markov models. In *Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP '02)*, volume 4, pages 4096–4099, Orlando, Florida, May 2002.

[18] M. Xu, N. C. Maddage, C. Xu, M. Kankanhalli, and Q. Tian. Creating audio keywords for event detection in soccer video. In *Procedding of the 2003 International Conference on Multimedia and Expo (ICME'03)*, volume 2, pages 281–284, Baltimore, MaryLand, July 2003.

[19] Y.-Q. Yang, Y.-D. Lu, and W. Chen. A framework for automatic detection of soccer goal event based on cinematic template. In *Proceedings of Third International Conference on Machine Learning and Cybernetics*, volume 1, pages 3759–3764, Shanghai, August 2004.

[20] C. Zhang, S.-C. Chen, and M.-L. Shyu. Pixso: A system for video shot detection. In *Procedding of the 2003 Joint Conference of ICICS-PCM*, volume 3, pages 1320–1324, Singapore, December 2003.

[21] W. Zhao, J. Wang, D. Bhat, K. Sakiewicz, N. Nandhakumar, and W. Chang. Improving color based video shot detection. In *Proceedings of IEEE International Conference on Multimedia Computing and Systems*, volume 2, pages 752–756, Florence, Italy, June 1999.

[22] D. Zhong and S.-F. Chang. Structure analysis of sports video using domain models. In *Proceedings of International Conference on Multimedia and Expo*, pages 713–716, Tokyo, Japan, August 2001.