Sports Video Summarization using Highlights and Play-Breaks

Dian Tjondronegoro
Centre for Information Technology Innovation
Queensland University of Technology
Brisbane, Australia
d.tjondronegoro@qut.edu.au

Yi-Ping Phoebe Chen
Centre for Information Technology Innovation
Queensland University of Technology
Brisbane, Australia
p.chen@qut.edu.au

Binh Pham
Centre for Information Technology Innovation
Queensland University of Technology
Brisbane, Australia
b.pham@qut.edu.au

ABSTRACT

To manage the massive growth of sport videos, we need to summarize the contents into a more compact and interesting representation. Unlike previous work which summarized either highlights or play scenes, we propose a unified summarization scheme which integrates both highlights and play-break scenes. For automation of the process, combination of audio and visual features provides more accurate detection. We will present fast detection algorithms of whistle and excitement to take advantage of the fact that audio features are computationally cheaper than visual features. However, due to the amount of noises in sport audio, fast text-display detection will be used for verification of the detected highlights. The performance of these algorithms has been tested against one hour of soccer and swimming videos.

Categories and Subject Descriptors

H.3.1. [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Abstracting Methods*.

General Terms

Algorithm, Experimentation.

Keywords

Video Summaries, Content Analysis.

1. INTRODUCTION

Sport videos need to be summarised for effective data management and delivery. Most viewers prefer to select particular segments which are interesting and suitable for their purposes. Researchers have identified that each type of sports have a typical and predictable temporal structure, recurrent events, consistent features and fixed number of views [1]. Hence, most of the current summarization techniques have been focused on one type of sport video by detecting the specific highlights or key events.

One approach for generating highlights is by optimizing the use of visual features. Gong et al [2] highlighted soccer games into penalty, midfield, in between midfield, corner kick, and shot at goal, while Zhou et al [3] categorized basketball into left- or right- fast-break, dunk, and close up shots. They used inference engine and tree learning rules to analyze specific visual features, such as line-mark recognition, motion detection, and color analysis of players' uniform and ball. The main benefit of this approach is to enable very specific queries, such as 'Show (video) shots where team A scored from the left-side'. However, combination with other features, such as audio, can potentially detect highlights more accurately. We should therefore generate highlights by analyzing the temporal structure of audio-visual features. With this approach, Nepal et al [4] modeled the temporal syntax of goal highlights in basketball videos in terms of the occurrence of high-energy audio segments, text display and change in motion direction. Similarly, Babaguchi et al [5] used a collaboration of shots similarity, keywords analysis from text display and closed caption in order to highlight events that change the score in American football, such as touchdown and goal.

Although highlights are very effective and compact summary of sport video, different users and applications may require varying amount of information. Thus, some recent approaches proposed a more generic sports summarization which is based on the classification of play-and-break scenes. In particular, Li&Sezan [6] and Xie et al [7] used HMM (Hidden Markov Model) for analyzing the temporal variation in camera views to distinguish play from break, and the transitions between them. Different camera views were detected by measuring grass portion and player activity. However, they have not demonstrated the use of some other supportive features, such as whistle and text display to detect play-break sequences.

We will present in this paper a unified summarization scheme which integrates highlights and play-break scenes. During the automated detection, users should be allowed to select whether they prefer more accuracy or faster processing time [8]. Based on their selection, the system can customize which features to be analyzed. Generally, audio features are computationally cheaper than visual features. Hence, we will show in this paper that highlights and play-break scenes can be localized using fast detection of whistle and excitement sounds. However, due to the amount of noise in sports audio, the results from audio-based

MIR'03, November 7, 2003, Berkeley, California, USA.
Copyright 2003 ACM 1-58113-778-8/03/00011…$5.00

detection can be verified and annotated by detecting text display when users are willing to have a longer processing time.

The rest of this paper is structured as follows. Section 2 describes our summarization framework; Section 3, 4, and 5 describes the algorithms for whistle, excitement and text display detection. Section 6 will present the experimental results while Section 7 concludes this paper and suggest the future work. Moreover, soccer games will be used as our examples throughout this paper.

2. SUMMARIZATION FRAMEWORK

Play-scenes based sports summary is potentially effective for browsing purposes because viewers will not miss any important events although they skip most of the break scenes. It is due to the fact that most highlights are contained within play scenes. However, we should still retain some break scenes, especially if they contain important information which users may benefit later. For example, preparation of a set piece kick, such as free kick and corner kick, shows how the offensive and defensive teams manage their formations for best results (i.e. defensive team try to avoid conceding a goal while offensive team try to maximize their chance to score a goal). Moreover, exciting events often happen during the transitions between play and break. For instance, penalty kick is how a play is resumed after stopped due to a foul which is committed inside penalty area. Thus, a highlight should include all these play-break-play scenes to ensure that the scene contains all the necessary details.

Play and break scenes are however not sufficient to support users who want to query specific highlights. In particular, sport fans usually like to view all highlights which belong to their favorite team. Play segments are also not necessarily short enough for users to keep on watching until they can find interesting events. For example, a match sometimes can have only a few breaks due to the rarity of highlights which causes the game stopped, such as goal, foul, or ball out of play. In this scenario, play scene can become too long to be a summary. In addition, users are hardly interested in the ratio of the match being played and being stopped. On the other hand, users can benefit more from statistics which are based on highlight events. For instance, sport coaches could analyse the percentage of fouls committed by their teams in a game in order to determine the aggressiveness of their defensive tactics. Moreover, not all play and break sequences are interesting. For instance, play can be paused shortly due to ball out of play and resumed by throw-in in soccer. This event can happen many times in a sports game and hardly can become an interesting highlight.

Hence, the main benefit of integrating play-break and highlights scenes is to combine their strengths and to achieve a more complete summary. For example, play scenes are generic because they can be an individual performance in gymnastics, an offensive/defensive attempt in soccer and basketball, or a race in swimming. However, the most compact summary of sport videos should contain only a few key frames which highlight most important events. In this case, detection of highlights (or key events) is needed in addition to play-break detection.

The process of sport summarization involves detecting and localising the start and end (frame) of each event, verifying that the resulting scene is self-adequate (i.e. it contains every detail that viewers need to fully understand the content, and annotating the scene for retrieval. These steps should be done (semi-) automatically because manual detection is subjective, very time-consuming and often incomplete. Hence our summarization framework is presented in Figure 1.

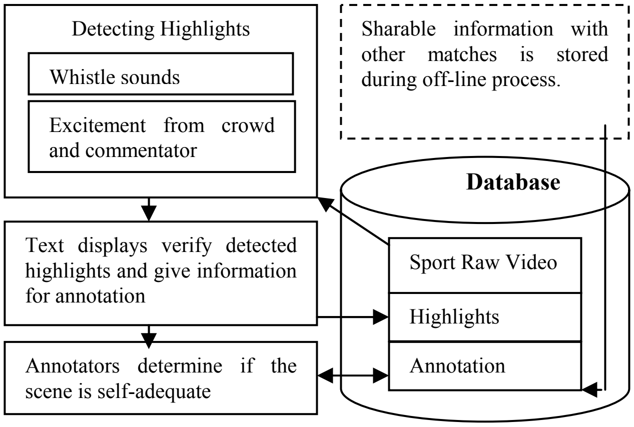

Detecting Highlights
Whistle sounds
Excitement from crowd and commentator
Sharable information with other matches is stored during off-line process.
Database
Text displays verify detected highlights and give information for annotation
Sport Raw Video
Highlights
Annotation
Annotators determine if the scene is self-adequate

Figure 1. Sports Video Summarization Framework.

Sport games are stored as raw video data in the database. From this raw data, whistle detection is used to identify the frames in which the game is being stopped (i.e. distinguishing play and break scenes). In addition excitement from crowd and commentator are localized to detect highlights which are caused by interesting events. Each detected highlight is stored as the position (i.e. start and end frame of the scene) to the raw video data. The text display is then used to confirm the highlights and as well as giving information for the annotators. For example, a goal can be described using the description in scoreboard display which includes scorer's details (e.g. name, team name, and squad number), the updated score-line, and the time in which the goal is scored. At the same time, the annotators can also manually re-check the highlight scene to ensure that it is consumable by itself since this process is very subjective and almost impossible to be done automatically. In order to assist annotators, information about sport game and its highlights are often shareable with other games, especially if they are the same type of sports. Thus, a faster or rapid highlight construction can be achieved by storing the most common information during off-line process.

We used a hierarchical structure to organize play, break and highlight scenes as described in Figure 2. Each play, break or combination of play-and-break can contain one to many highlight scene(s) which can be organised into a highlight collection. For example, if users are interested in storing highlight collection from team A, the corresponding highlights which belong to team A will be compiled into a highlight collection. We have shown the utilization of MPEG-7 standard descriptions to annotate and query highlights collection in [8]. Based on this structure, users can select to watch all play and/or break scenes or just the ones which have a certain number of highlights (i.e. interesting play and break scenes).Users can also refer back to the whole play or break scene if they found a highlight is not adequate.

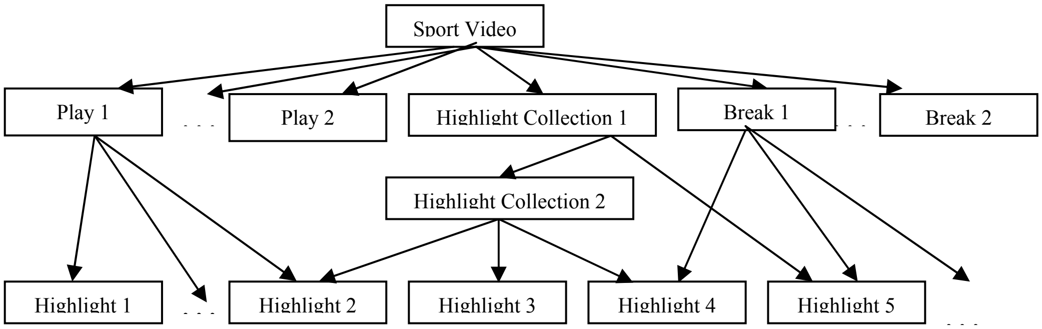
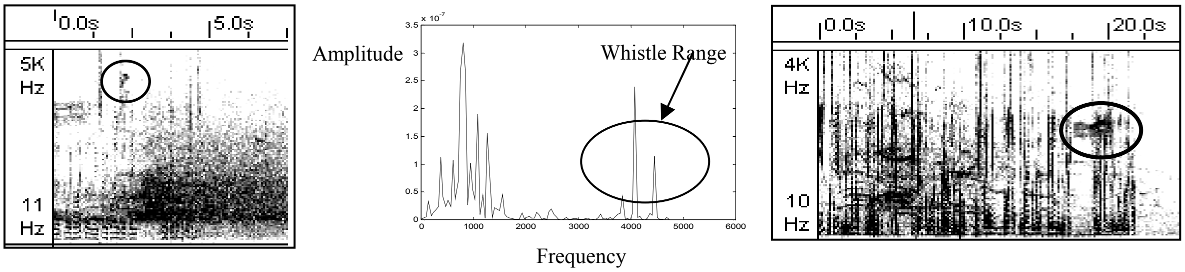


Figure 2. Hierarchy model of Sport Video Summary based on Play, Break, and Highlight.



(a) *(b)* *(c)*

Figure 3. a) Spectrogram Containing Short Whistle, b) Peak Energy in Soccer Audio is not Within Whistle Frequency Range, c) Whistle in Swimming Audio.

In addition, users can build their own highlight collection on top of existing (or system generated) collections. For example, users can have a highlight collection called "my favourite highlight collections" which contains the existing highlight collections, such as goal, free kick and shot on goal.

The next three sections will present our methods for whistle, excitement and text display detection. The following pre-processing of audio track was performed before whistle and excitement detection: Audio track of sport video is normalized using its maximum absolute sample-value; The channel is converted into mono if the audio is stereo; Each audio track is segmented into one-second clips while each clip is then further segmented into 40 ms frames with half overlaps (i.e. 1 clip contains 50 frames).

3. DETECTING WHISTLE

The audio track in sport videos is very complex due to noises from human voices and background sounds However, we noticed that whistle sound occurrences during sport videos are very distinctive as shown by the spectrogram in Figure 4.

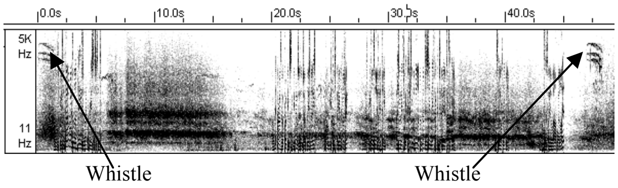


Figure 4. Spectrogram of Soccer Whistle.

During soccer matches, whistle sounds indicate:

- The start and the end of the match and playing period
- Play stops, such as foul or offside which may lead to yellow or red card is given to a player for punishment;
- Play resumes (after being stopped). Depending on the outcomes, play can be resumed with set piece (or dead-ball) kick, such as penalty kick, free kick, goal kick or corner kick.

Moreover, in swimming videos, a long-continuous whistle is used to tell the swimmers to get ready for a race.

Zhou et al [9] detected sports event boundary by identifying the whistle sound from a referee which usually indicates the start or end of an event. They found that whistle sounds have high frequency and strong spectrum with frequency range from 3500 – 4500 Hz. Hence, the peaks of whistle spectrums would be in that frequency range. Based on these assumptions, they suggested that a whistle sound can be detected if there is a longer than 1s window of peak frequencies which fall into the range between 3500-4500 Hz (let' call it whistle frequency range). While this technique's performance has not been reported by any experiment work, we have predicted three potential limitations. Firstly, if an audio clip only contains a very small portion of whistle (i.e. when the whistle is blown very short as shown in Figure 3a), the energy ratio within the whistle frequency range might not be the peak due to the other dominant sounds or noises. Secondly, for most sport video sounds, the peak is often within lower frequency range than the whistle range especially when the commentator's speech and crowd is loud (as shown in Figure 3b). Thirdly, whistle sound is not always within 3500-4500Hz frequency range, such as in swimming video which is shown in Figure 3c.

To overcome these problems, we have developed another method to detect whistle sound in sports video. We performed an *N*-point Fast Fourier Transform (FFT) to calculate the spectrum (i.e. the histogram of frequencies) of each (audio) frame. We then calculated the Power Spectral Density (PSD_W) of the signal within whistle's frequency range using this formula:

$$PSD_W = \sum_{WL}^{WU} | S(n) * conj(S(n)) | \qquad (1)$$

Where, WU is the upper bound of frequency range, WL is the lower bound, N is the n-point FFT, and *S(n)* is the spectrum of the audio signal at frequency *n* Hz. Complex conjugation (Conj) is required because there are imaginary and real components of spectrum as a result of Fourier Transform.

The detection (to localize whistle) starts from the first until the last clip of an audio track to check whether the clip contains whistle sound. Within each clip, a frame is marked as (potentially) containing whistle if it contains a PSD_W value which is greater than *threshold1* (this current value of PSD_W is then regarded as *current significant value*). Finally, a clip is determined to have whistle sound if we can find at least *n* neighboring frames which contain PSD_W value of at least 80% of the current significant value. Thus *n* can be regarded as the minimum number of frames required to be confirmed for whistle existence which is *threshold2*. The starting time of a clip is attached to the output array if it is found to have a whistle sound. The values of thresholds and whistle range (WU and WL) are experimental and should not be static because of the variations in whistle sounds. For example, whistle sound is affected by the type of whistle being used, the whistle blower (which usually affect the length), as well as the environment, such as the amount of background noise, and the volume of recording.

During our experiment, whistle frequency range was set as 3500-4500 Hz for soccer and 2800-3200 Hz for swimming, while threshold 1 was set to 1.5 to 3 which is adjustable according to how noisy or loud the overall signal. Threshold 2 was set as either 5 or 7 depending on the average length of whistle being blown in the video.

4. DETECTING EXCITEMENT

We noticed the following changes in sports audio track when exciting event occurs: *1.* crowd's cheer and commentator's speech become louder; *2.* Commentator's voice has a slight raise of pitch; and *3.* Commentator's talk becomes more rapid (i.e. more talkative) thus has less pauses. Based on this concept, the essence of our excitement detection algorithm is the use of three main excitement *candidates* which are based on three features: lower pause rate (*candidate1*), higher pitch rate (*candidate2*), and louder volume (*candidate3*).

Based on the three primary candidates, we combine and filter out some of them to obtain a set of (final) candidates which are most likely to contain excitement. For this purpose, we performed some steps as depicted in Figure 5. Firstly, *candidate1* and *candidate2* are combined into *candidate4* to verify *candidate3*. It is due to the fact that loudness-based excitement is less reliable since they can be detected from loud crowd cheer and background noise which does not always correspond to exciting events. For example, crowd cheer can get louder to give more support for their team, particularly when the team is playing at home and have conceded a goal. Secondly, *candidate5* is formed by combining all three features (which is most likely to have excitement), while *candidate6* contains loud clips which does not have low pause and high pitch (thus less likely to contain excitement). Thirdly, before combining *candidate5* and *candidate6* to produce the final candidates for excitement clips, we discard loud clips (*candidate6*) which does not last for at least 3 seconds. This step is to eliminate loud clips which are too short since they are less likely to contain exciting events. Finally, we group excitement clips which have less than 2 second gaps and check if the length is longer than a certain threshold. This step is important to produce excitement segments which are significant enough to contain highlights. The process for detecting excitement with the particular thresholds used for each of the steps is based on this figure.

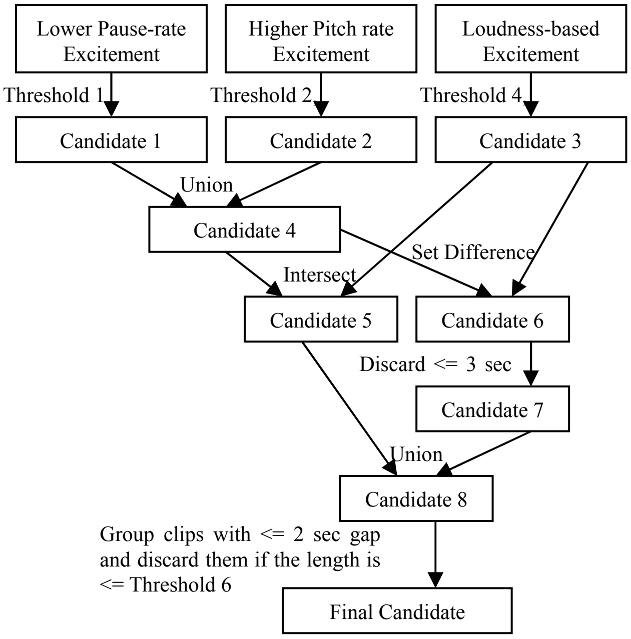


Figure 5. The Excitement Detection Framework.

To localize louder clips (i.e. *candidate3*), the method for whistle detection can be reused, but replacing calculation of volume for that of PSD_W. We used this equation to calculate the volume of each audio frame:

$$Volume = \frac{1}{N} * \sum_{n=1}^{N} | x(n) | \qquad (2)$$

Where, N is the number of samples in a frame and x(n) is the sample value of the n^{th} frame.

Detecting pitch and silence of the voiced speech components from sports audio is very difficult due to the complex background noise which often mixes with the speech component, as well as the unavailability of a standard method to calculate pitch. Thus, for our purpose, we have used the sub-harmonic-to-harmonic ratio based pitch determination in [10] to determine pitch and detect silence. This method produces reliable result and their algorithm (called ***shrp***) is available online and is well documented. Figure 6 shows that *shrp* predict the pitch value of the speech component, and whether it is a silence (i.e. pause) or non-voiced speech which is marked by pitch value equals to zero.

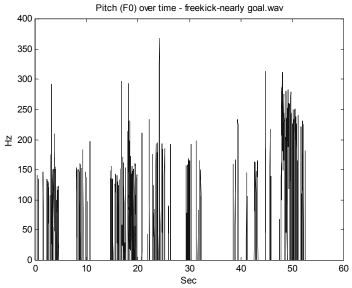


Figure 6. Pitch Determination and Silence Detection

In order to localize clips with more frames containing high-pitch and less pauses (*candidate1, 2*) we run *shrp* to calculate pitch values of each frame in an audio clip. Based on the pitch values, high-frequency and pause rate in a clip are calculated using dual-fashioned equations:

$$PauseRate = \# P_f \;/\; N * 100\% \qquad (3)$$

$$HighpitchRate = \# HP_f \;/\; N * 100\% \qquad (4)$$

Where, $\#P_f$ is the number of frames containing speech pause in a clip, $\#HP_f$ is the number of frames containing high pitch speech in a clip and N is the number of frames in a clip.

High pitch is determined if the pitch value is greater than *threshold3* while pause (or silence) is determined if the pitch value is equal to zero. Based on these rates, the algorithm determines the values and location of local maximums in high-frequency rates and local minimums in low-pause rates. Each of the local minimum of pause rate (which represents the clips containing less pause frames) is compared to *threshold1* and if it is less than that value, the location of the clip is added to *candidate1*. Similarly, when local maximum of high-frequency rate is greater than *threshold2*, the location of the clip is added to *candidate2*.

During experiment, threshold1 (i.e. minimum duration of more rapid speech) was set to the mean (or average) value if there is a large difference between the highest and lowest values of the local minimum pause-rate values. Otherwise, it was set as 50%. Similarly, if there is a large difference between the highest and lowest values of the local maximum high-frequency rate values, threshold2 (i.e. minimum duration of higher pitch speech) was set to the mean value. Otherwise it was set as 70%. Threshold3 (i.e. how much higher pitch difference should an excitement has) was set to 50% of the frequency bandwidth of the current audio clip which is calculated as max (pitch value) – min (pitch value). However, since silence is indicated with pitch value equals to 0, we specified that min (pitch value) is 50 Hz (i.e. lowest pitch for typical male speaker) because female's lowest pitch is generally higher than male. Threshold4 (i.e. how much louder should an excitement be) was set as 100% or 150% of the standard deviation value from the absolute values of samples in the current clip. Threshold5 (i.e. minimum duration of louder clips to be defined as excitement) was set to 4 (frames) and finally threshold6 (minimum duration of an excitement sequence) was set to 3 (clips).

5. DETECTING TEXT DISPLAY

To detect text in video, Wernicke and Lienhart [11] used the gradient of color image to calculate the complex-values from edge orientation image which is defined to map all edge orientation between 0 and 90 degrees and thus distinguishing horizontal, diagonal and vertical lines. Similarly, Mita and Hori [12] localized character regions by extracting strong still edges and pixels with a stable intensity for two seconds. Strong edges are detected by *Sobel* filtering and verified to be standstill by comparing four consecutive gradient images. Thus, current text detection techniques attempt to detect the edges of the rectangle box formed text regions in color video-frames as well as checking if the edge will stay for more than 2 seconds, allowing viewers to read and understand the content.

Based on this concept, the essence of our text display detection method is based on an assumption that in 99% of the cases, sport videos only use horizontal text as shown by examples in Figure 7. Thus, if we can detect strong horizontal line in a frame, we can locate the starting point of a text region. For this purpose, we have used *Radon Transform* on gradient image (which is produced by *Sobel Filter*) to detect strong lines in video frames. The main benefit of this method is that most text displays in sport (and news) video are in most cases surrounded by a large rectangle box in order to distinguish them from the background. The main steps in text display detection are described in Figure 9.

Firstly, video (frames) track is segmented into one minute clip. Each of the frames within 1 second gap (which is currently to be analysed) is pre-processed to optimize performance by converting the color scheme to grayscale and the size is reduced to a preset smaller size. Secondly, we apply *Sobel Filter* to calculate the edge (gradient) of the current frame and then apply *Radon Transform* on the gradient image to detect *line spaces (R)* which are in between 0-180 ^{0} angle. We then apply a threshold on these R values to detect potential candidates of strong lines which are usually formed by the box surrounding text display. After these lines are detected, we calculate the rho (r) value of the peak coordinates to indicate the location of the line in terms of the number of pixels from the center, and *t* (theta) which indicate the angle of the line. In order to verify that the detected lines are candidates of text display region, we only retain the lines which follow these criteria: the absolute value of r ≤ 80% of the maximum y-axis and the corresponding t is equal to 90 (horizontal).

The first check is important to ensure that the location of the lines is within the usual location for text display. The second check is to ensure that the line is horizontal because there are potentially other strong horizontal lines which can be detected from other than text display, such as the boundary between field and crowd.

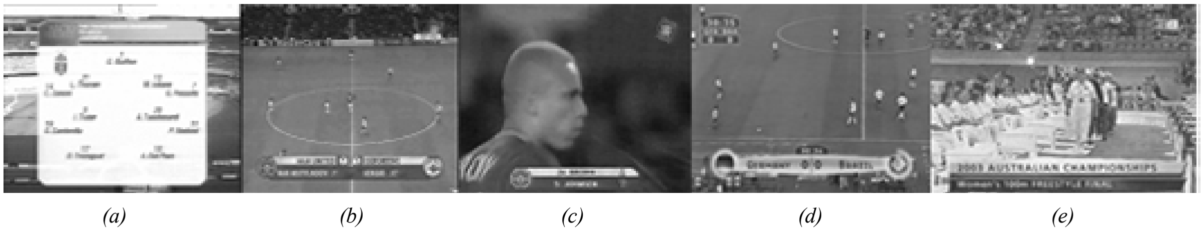


Figure 7. ***a)*** **Starting line-up of soccer team,** ***b)*** **Current score-line,** ***c)*** **Player substitution,** ***d)*** **Static text appearing for the whole match,** ***e)*** **Text in swimming**

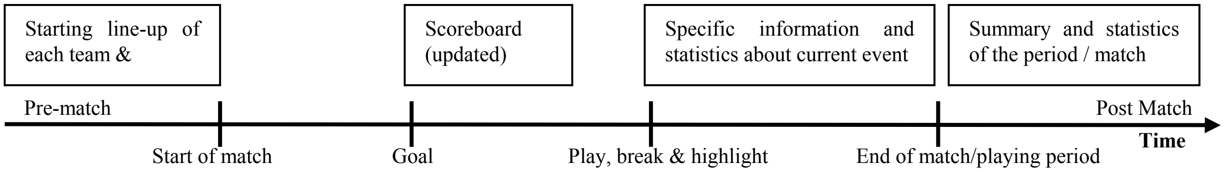

Starting line-up of each team &
Scoreboard (updated)
Specific information and statistics about current event
Summary and statistics of the period / match
Pre-match
Post Match
Time
Start of match
Goal
Play, break & highlight
End of match/playing period

Figure 8. Text displays occurrences during a soccer match.

Finally, for each of the lines, we check that their location is consistent for at least 2 seconds (i.e. if the video frame rate is 25, 2 seconds is equal to 50 frames).The purpose of this check is to ensure that the location of the lines are consistent for the next and/or previous frames since text display always appear for at least 2 seconds to give viewers time to read. Moreover, when the text display size is large and contains lots of information, it will be displayed even longer to give viewers enough time to read.

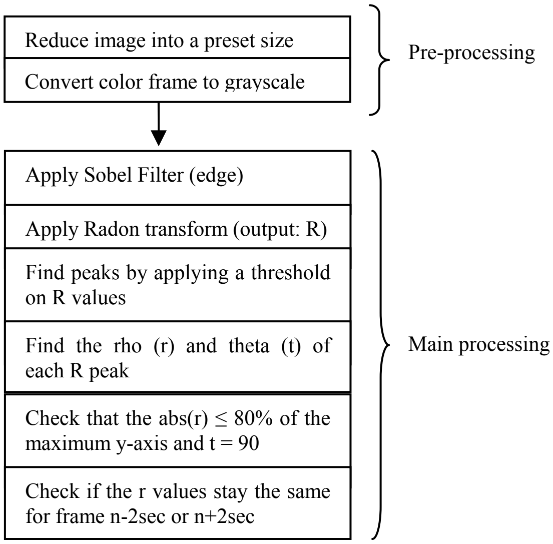

Reduce image into a preset size
Convert color frame to grayscale
Pre-processing
Apply Sobel Filter (edge)
Apply Radon transform (output: R)
Find peaks by applying a threshold on R values
Find the rho (r) and theta (t) of each R peak
Check that the abs(r) ≤ 80% of the maximum y-axis and t = 90
Check if the r values stay the same for frame n-2sec or n+2sec
Main processing

Figure 9. Text Detection using Radon Transform

During the experiment, we noticed that it takes longer to check text display in all video frames since there are not so many texts during a sport match (otherwise they can be distractive). Hence, specific domain knowledge should be used to predict text appearances in soccer videos which are quite typical from one match to another. Figure 8 depicts our concept on predicting text occurrences based on specific domain knowledge for soccer video. However, in this paper we still had to check all frames since text display can detect some events which sometimes cannot be detected by whistle and excitement sounds. For example, whistle does not always exist during start of a match.

6. EXPERIMENTAL RESULTS

We tested the detection algorithms which were developed in MATLAB 6.5 using a Pentium 4- 1.5 GHz PC with 512MB memory in Windows XP pro. platform. We used five samples: three soccer matches and two swimming competitions. The first and second soccer video were recorded from the first 20 minutes of UEFA Champions League games (which belong to the same broadcaster and commentator), but the later was noisier because the crowd constantly gave supports for their home-team and the commentator is more excited because the match is semi-final. The third soccer video was selected from the last 20 minutes of FIFA world cup final thus it belongs to different competition, broadcaster and commentator. Using this video, we can check the robustness of our algorithms for different characteristics of audio and formats of text displays. In addition, the two swimming videos were recorded from Australian National Championship competitions. Each of them is 5 minutes long and they are selected to show that our algorithms can support different sports with little modification on the parameters (including thresholds). To evaluate the performance of our detection algorithms, we used these measures:

- **Recall rate (RR)** is the percentage of true detection performed by the automated detection algorithm with respect to the actual events in the video (which is calculated as total of correct and missed detections). This indicator is important to show that our algorithm can detect most of the events while achieving as little misdetections as possible.

Table 1. Performance Measures of the Detection Algorithms

Sample Video	**Whistle**		**Excitement**		**Text**	
	Recall	Precision	Recall	Precision	Recall	Precision
Soccer1	53.3 %	72.7 %	80.7 %	92.6 %	75 %	85.7 %
Soccer2	77.7 %	77.7 %	95.5 %	55.3 %	75 %	31.3%
Soccer3	68.8 %	84.2 %	100 %	61.2 %	85.7 %	16.6 %
Swimming1	100 %	100 %	100 %	63.6 %	86.4 %	73.2 %
Swimming2	100 %	16.7 %	81.3 %	86.9 %	66.7 %	60 %
Average	**80 %**	**64%**	**92 %**	**72 %**	**77 %**	**62 %**

Table 2. Detected Highlights using Combination of Whistle, Excitement and Text.

Sample Video Soccer 1,2,3 is 20 mins Swimming 1&2 is 5 mins	**Total** **Highlights and Play-breaks**	**Automatically Detected Highlights and Play-breaks**							
		Using Whistle Only		**Using Whistle +Excitement**		**Using Whistle +Text**		**Using Whistle +Excitement +Text**	
		Number of Highlight	Time (min)	Number of Highlight	Time (min)	Number of Highlight	Time (min)	Number of Highlight	Time (min)
Soccer1	**41**	**8**	0.9	**33**	10.8	**10**	19.3	**35**	29.2
Soccer2	**24**	**7**	0.7	**22**	10.6	**8**	24.8	**23**	35.4
Soccer3	**40**	**11**	0.7	**39**	8.8	**11**	26.7	**39**	35.5
Swimming1	**3**	**1**	0.2	**1**	3.2	**3**	5.1	**3**	8.1
Swimming2	**3**	**1**	0.2	**1**	3.3	**3**	5.2	**3**	8.3

- **Precision Rate (PR)** is the percentage of true detection with respect to the overall events detected by the algorithm (which is indicated by the number of correct and false detections). This percentage can indicate the trade-off from achieving minimum misdetections. It is due to the fact that the lower thresholds we use, the less are the number of missing events, but at the same time we will get more false detections.The equations that we used for these indicators are:

$$RR = Nc / (Nc+Nm) * 100\% \qquad (5)$$

$$PR = Nc / (Nc+Nf) * 100\% \qquad (6)$$

Where, Nc is the number of correctly-detected highlights, Nm is the number of misdetected highlights, and Nf is the number of false detections.

The performance measures are shown in Table 1. Our goal was to achieve a precision rate of 70% or greater while the lowest precision rate which can be compromised should not be less than 60%. Based on the average statistics, it can be justified that the detection algorithms are overall robust and reliable. However, recall rate for whistle detection in Soccer1 video is considerably low (i.e. 53 %) due to the fact that we could not sacrifice too much precision when we decided the threshold. In fact, we had detected all whistle occurrences in this video using a slightly lower threshold, but the precision rate decreased to 25%. In addition, although the precision rate for whistle detection in Swimming2 video was only 17 %, it cannot accurately justify the performance as there were only 6 false detections in the 5-minutes audio. Similarly, the precision rate for our text detection in Soccer3 was only 17% since we decided to use a lower threshold for the minimum peak value (of the detected lines) since most of the text displays (in that video) were surrounded by oval and rectangle lines (as shown in Figure 7d). In fact, we achieved more than 60% of precision rate by applying a slightly higher threshold, but some important text displays were not detected. Finally, the low precision rate of excitement detection in Soccer2 and Soccer3 videos was due to the fact that both matches were very exciting and there are a lot of loud crowd cheers during the whole match. Thus, most of the extra excitement detections were actually correct because they localizes the times where the crowd gets excited. However, we did not include some of these sequences in the ground truth when they do not correspond to exciting events.

Table 2 shows the comparisons of detected highlights using whistle detection only, whistle-text detection, whistle-excitement detection, and finally whistle-excitement-text detection. This table demonstrates the advantage of our summarization framework in terms of the number of highlights (which can be detected) and the time spent by each detection algorithms.

Based on this table, it is clear that whistle detection is very fast, but it can only localize 20 to 30% of the total highlights which are mostly caused by foul and offside (i.e. play stopped). In most cases, however, whistle is not really used to indicate play being resumed again with a *free kick*, unless if there is a substantial period of waiting during the break. For example, a *direct free kick* which is to be taken nearby penalty area will be indicated by whistle after the period of time in which the teams are preparing their formation. In contrast if a free kick is to be taken from defensive to midfield area, the whistle is only blown to indicate that there is a foul or offside without indicating the free kick itself. Hence, we need to adopt the camera-views based method [6][7], so that the play-break transition can be defined more precisely.

By combining whistle and excitement, users only need to wait slightly longer to detect 80% to 90% of the highlights since excitement can locate most of exciting events, such as *good display of attacking or defending, goal, free kick*, and even *foul* sometimes. In addition, excitement detection is very effective to localize goal highlights due to the massive amount of excitement during the start of good attack which often leads to the goal itself. Moreover, the excitement will still be sustained during goal celebration and slow-motion replays especially when the commentator and crowd is very excited about the goal, such as an important goal, or an elegant goal.

When whistle and text detection are combined, the number of highlights detected will only slightly increase while the waiting-period is longer than using excitement. It is due to the fact that visual features are generally more expensive computationally than audio features. Text detection is needed to localize *start of a match*, *goal* and *shot on goal*, as well as confirming offside and foul events. As shown earlier in Figure 8, there are some large texts displayed before start of a match to show the starting line-up of each team, as well as showing the formation they use for the match. Since they are large and contains a lot of information, they are usually displayed for the whole 1 or 2 minute time-span. After a goal is scored, a text is displayed to show the updated score-line. Similarly, after a shot on goal, usually the text will confirm that there is no change of score-line or showing the details of player(s) involved (e.g. forward player and goal keeper).

Finally, when whistle-, excitement-, and text- detection are all used, 85% to 100% of highlights can be detected. However, if users can afford missing some events which can only be detected by text, we recommend whistle and excitement detection to take the advantage of their fast processing time. Nevertheless, text displays which are located nearby these highlights should still be detected for annotation purposes.

7. CONCLUSION & FUTURE WORK

While most of current work on sports summarization aims to detect highlights or play scenes only, in this paper, we have proposed a more unified framework for sports video summarization. In particular, we have described the main reasons why play or highlights alone are not sufficient to support wide range of requirements. In addition, it is identified that break scenes, such as preparation of a free kick, ceremonies and commentaries should still be retained to support future queries.

For semi-automatic construction of the sports video summary, we have developed efficient detection algorithms of whistle, excitement and text which are reliable and precise (despite its simplicity and fast processing). We have also demonstrated how a full-combination of whistle, excitement and text can detect most of soccer and swimming highlights as opposed to only using one of them. However, at this stage, we have only used some slightly-adjustable thresholds for different audio characteristics in order to optimize the performance measures. Hence, for future work, we need to design an automated method for deciding the thresholds, based on the desired precision and recall, so that the algorithms can be fully automated. We also will adapt current video Optical Character Recognition techniques to extend the text detection results to enable a fully automated verification and annotation of highlight sequences. Moreover, we will extend the framework to include methods for indexing and retrieval, so that we can show its benefits in terms of meeting user and application requirements. In particular, since mobile devices which can play video have become very common, many sports fans will be able to benefit from a summarized version of sports video which is available from anywhere.

8. REFERENCES

[1] Zhong, D. and S.-F. Chang. *Structure Analysis of Sports Video Using Domain Models*. in *IEEE ICME 2001*. 2001. Tokyo, Japan.

[2] Gong, Y., et al. *Automatic parsing of TV soccer programs*. in *Multimedia Computing and Systems, 1995., Proceedings of the International Conference on*. 1995: Practical.

[3] Zhou, W., A. Vellaikal, and C.C.J. Kuo. *Rule-based video classification system for basketball video indexing*. in *ACM Workshops on Multimedia*. 2000. Los Angeles, California, United States: ACM.

[4] Nepal, S., U. Srinivasan, and G. Reynolds. *Automatic detection of 'Goal' segments in basketball videos*. in *ACM International Conference on Multimedia*. 2001. Ottawa; Canada: ACM.

[5] Babaguchi, N., Y. Kawai, and T. Kitahashi, *Event based indexing of broadcasted sports video by intermodal collaboration.* Multimedia, IEEE Transactions on, 2002. **4**(1): p. 68-75.

[6] Li, B. and M. Ibrahim Sezan. *Event detection and summarization in sports video*. in *IEEE Workshop on Content-Based Access of Image and Video Libraries, 2001(CBAIVL2001).* Sharp Labs. of America, Camas, WA, USA: Practical.

[7] Xie, L., et al. *Structure analysis of soccer video with hidden Markov models*. in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002*. Columbia University.

[8] Tjondronegoro, D. and Y.-P.P. Chen, *Content-based Indexing and Retrieval Using MPEG-7 and X-Query in Video Data Management Systems.* World Wide Web Journal, 2002. **5**(2): p. 207-228.

[9] Zhou, W., S. Dao, and C.-C. Jay Kuo, *On-line knowledge- and rule-based video classification system for video indexing and dissemination.* Information Systems, 2002. **27**(8): p. 559-586.

[10] Sun, X. *Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio*. in *ICASSP2002*. 2002. Orlando, Florida.

[11] Wernicke, A.L., R. *On the segmentation of text in videos*. in *IEEE International Conference on Multimedia and Expo ICME 2000*.

[12] Mita, T.H., O. Improvement of video text recognition by character selection. in Proceedings of Sixth International Conference on Document Analysis and Recognition, 2001.