Integrating highlights for more complete sports video summarization

Integrating Highlights for More Complete Sports Video Summarization

Dian Tjondronegoro and Yi-Ping Phoebe Chen
Deakin University

Binh Pham
Queensland University of Technology

Summarization is an essential requirement for achieving a more compact and interesting representation of sports video contents. We propose a framework that integrates highlights into play segments and reveal why we should still retain breaks. Experimental results show that fast detections of whistle sounds, crowd excitement, and text boxes can complement existing techniques for play-breaks and highlights localization.

The multimedia community needs an effective management scheme for handling the increasing amount of sports videos from TV broadcasts. One of the most important requirements in managing video is compressing its long sequence into a more compact representation through a summarization process. Researchers have proposed many techniques to take full advantage of the fact that sports videos have typical and predictable temporal structures, recurrent events, consistent features, and a fixed number of camera views.[1] To date, most summarization techniques have focused on one type of sports video by detecting specific highlights (or key events) using

- specific features such as slow-motion replay,
- keywords analysis from closed-caption and speech recognition,[2] and
- rule-based analysis of object and motion recognition (for example, using the hidden Markov model).[3]

Although domain-specific highlights can satisfy most requirements, we must realize that different users and applications often require a varying amount of information. For example, some users might need to query "What happens just before or after a specific key event?" If the system stores key events by themselves, it won't be able to answer the query. To fill in this gap, some researchers claim that *play* sequences in sports videos are "self-consumable" because most users naturally focus their attention on events that happen within plays. A play scene is generic because it can contain a sequence of shots where the ball is being played in a soccer match or capture a swimming race. Generally, a long global shot usually corresponds to play events, while frequent and/or long close-up shots indicate break events that cause a game to stop momentarily (such as a foul, celebrating a goal, or the end of a playing period).[4]

Unlike previous work that categorizes sports videos into either highlights or play sequences, we aim to present a unifying summarization framework that integrates highlights into plays as well as reveal why we should still retain breaks. Our main purpose in this article is to construct a more complete sports video summary that can support a broader scope of users and applications. Our approach is more complete in the sense that the generated summaries contain almost all the important scenes that can support a wider range of user and application requirements. Thus, when we choose to store only the summary for compression purposes, we won't lose any important information from the full-length sports video.

To complement current available techniques to detect highlights and play-breaks, the second goal of this article is to demonstrate that most play-breaks and highlights in certain sports can be localized using fast detection of whistle and excitement sounds. In particular, whistle detection can replace visual-based play-break detection in many sports that use whistles, such as soccer, rugby, swimming, and basketball. Excitement in sports audio tracks corresponds to key events. However, because of the amount of noise in sports audio, the results from audio-based detection can be verified and annotated by detecting text display. Moreover, text occurrences in sports video can also detect some additional highlights.

Despite the fact that our algorithms are processed offline, we still prefer fast and cheap computation to support "summaries on request."

1070-986X/04/$20.00 © 2004 IEEE

Published by the IEEE Computer Society

For example, after users input a sports video into our system, they should be able to select whether they prefer fast-but-less-accurate or slow-but-more-accurate processing depending on how long they're willing to wait.[5] Based on their selection, the system can customize which features to analyze. We expect that by providing more audiovisual features—which are computationally cheaper—we can improve detection accuracy as well as detect more summaries.

More complete summarization scheme

The main purpose of summarizing sports videos is to compress unimportant contents for efficient storage because most sports viewers prefer to focus their attention on events within play segments. However, most sport videos contain many events that cause a game to stop. Even sports fans don't want to spend their time waiting for the game to resume again. Thus, a play-based summary is effective for browsing purposes because most highlights are contained within a play. We consider plays to be self-consumable because viewers won't miss any important events although they skip most of the break scenes. Play segments are also generic because they can be an individual performance in gymnastics, an offensive/defensive attempt in soccer and basketball, or a race in swimming. Moreover, in a frame–shot–event–video hierarchy, a play is at the same level as an event since a play contains complete actions within multiple video shots.

Break sequences, however, should still be retained. They're just as important as play, especially if they contain highlights that can be useful for certain users and applications. For example, a player preparing for a direct free kick or penalty kick in soccer videos shows the strategy and positioning of the offensive and defensive teams. This type of highlight can be crucial for sports coaches and training purposes.

A break can also contain slow-motion replay and full-screen texts, which the broadcaster usually inserts when the game becomes less intense or at the end of a playing period. Slow-motion scenes usually replay a key event from different angles; therefore, they can be useful for analysis. For example, viewers can verify doubtful events after being replayed slower or view a goal from different perspectives. On the other hand, texts that are displayed during a break are usually informative to keep the viewers' attention, such as number of fouls committed by a player and game statistics. Moreover, certain highlights often happen during the transitions between plays and breaks. For example, a free kick in soccer indicates how a play is resumed after a foul.

While play and break sequences are good summaries, most compact and high-level summaries of sports videos should contain only a few keyframes that represent highlights or important events. This is because plays aren't necessarily short enough for users to continue watching until they can find interesting events. For example, a match sometimes can have only a few breaks because of a rare goal, foul, or ball out of play. In this case, play segments can become too long for a summary. Play-breaks also can't support users who need a precise highlight summary. In particular, sports fans often need to browse or search a particular highlight in which their favorite team and/or players appear. Similarly, sports professionals, such as coaches, often use key events to analyze the performance and tactics of their team and/or opponents. Moreover, play-break analysis of sports videos is inadequate because users aren't interested in the ratio of the match being played and being stopped. On the other hand, users can benefit more from statistics based on highlight events. For instance, coaches could analyze the percentage of fouls committed by their teams in a game to determine the aggressiveness of their defensive tactics.

Based on these reasons, we've demonstrated the importance of integrating highlights in their corresponding plays or breaks to construct a more complete sports video summary. This approach lets users browse a different level of summary details depending on their individual needs. Furthermore, attaching highlights to their corresponding plays or breaks is also useful in generating the most exciting plays or breaks that can be easily constructed by setting a minimum number of highlights per play or break. Thus, the system can potentially achieve further compression with more confidence.

Unified summarization framework

We've applied a hierarchical structure (see Figure 1, next page) to organize a sports video summary that consists of integrated plays, breaks, and highlights. Each play and break can contain one to many highlights that we can organize into a highlights collection. For example, if users are interested in storing a highlight collection from team A, the system will compile the corresponding highlights that belong to team A into a highlight collection.

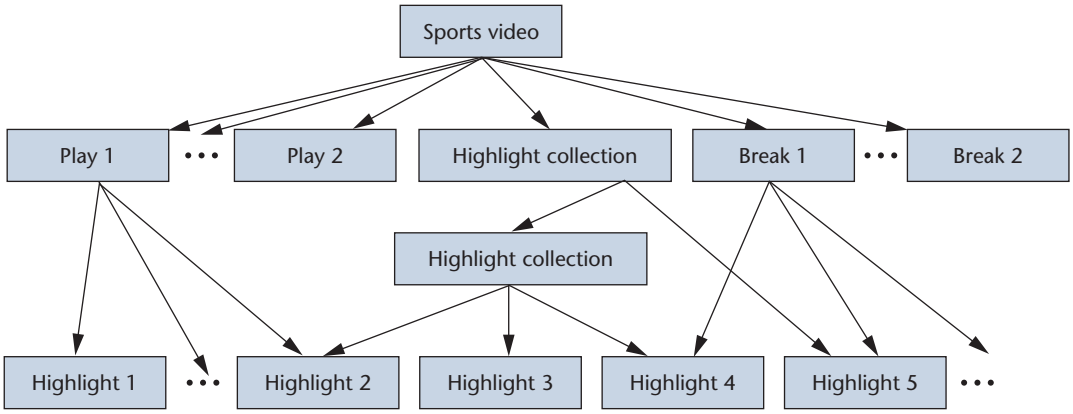


Figure 1. Hierarchy model of sports video in terms of plays, breaks, and highlights.

This model has some obvious benefits. First, users can watch all play and break scenes or just the ones that have a certain number of highlights. Second, users can refer back to the whole play or break scene and thus answer, What happens before and after a highlight? or What causes a highlight? Third, the model lets viewers have one or more highlight collections for a sports video and structures them in a hierarchical scheme. Thus, users can build their own highlight collection on top of existing (or system-generated) collections.

Here we'll define various parts of our framework. A sports video summary is a 5-tuple (*PLY*, *BRK*, *COL*, *HGT*, and *ASC*), where:

- *PLY* is a set of plays,
- *BRK* is a set of breaks,
- *COL* is a set of (highlight) collections,
- *HGT* is a set of highlights, and
- *ASC* is a set of association maps (for example, λ, ϖ, and δ).

We define plays, breaks, collections, highlights, and association maps as follows:

- Play is a sequence of shots in a sports video where play flows until it's stopped.
- Break is a sequence of shots where play doesn't flow until it's resumed.
- Collection is a conceptual entity that groups similar highlights.
- Highlight is a sequence of shots containing key events, such as a goal.
- Each play and break can be associated with basic attributes $B = [Fs, Fe, Kf, Ka, Kc]$ and $[\#H]$, where *Fs* is frame start, *Fe* is frame end, and #*H* is number of highlights. *Kf*, *Ka*, and *Kc* are key (or representative) frames, audio, and clips, respectively.
- Each highlight can be associated with *B* and a set of annotations, such as type, player, and textual description. Type is the type of highlight, such as a goal, foul, and shot on a goal in soccer. Player is the actor(s) involved in the highlight. Textual description is free text or formatted text (such as XML) that further describes specific highlights, such as <Current Score Line>.
- Each play, break, and highlight may include a set of semantic and temporal links between them {**S**, **T**}. These links help users browse among plays or highlights.
- **T** is a temporal that includes [before, overlaps, during, after, and so on]. These links can be calculated automatically based on *Fs* and *Fe*.
- **S** is a semantic link that includes [caused-by (or results-in), same players, and so on]. These links can be generated manually by users or automatically, based on temporal links' and annotations' similarity.
- λ is an association map that assigns a play to each highlight where $[H_{Fs}\ H_{Fe}]$ is within $[P_{Fs}$ and $P_{Fe}]$. Thus, $H_{Fe} \leq P_{Fe}$ and $H_{Fs} \geq P_{Fs}$.

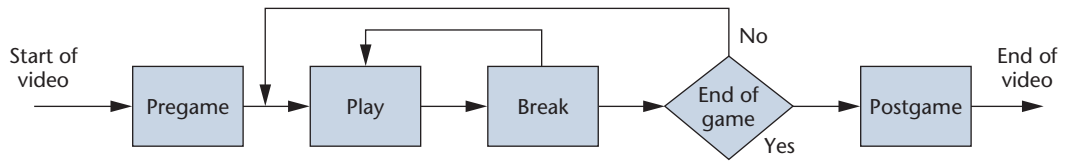


Figure 2. Generic play-break sequence model.

- ϖ is an association map that assigns a break to each highlight where $[H_{Fs}\ H_{Fe}]$ is within $[B_{Fs}$ and $B_{Fe}]$. Thus, $H_{Fe} \leq B_{Fe}$ and $H_{Fs} \geq B_{Fs}$.

- δ is an association map that assigns a collection to each highlight where $H_{\text{type}} \approx C_{\text{type}}$.

Similarity of the play-breaks structure for different sports

Generally, a sports video (without advertisements) is usually started with a pregame scene containing introductions (of the teams and/or players), commentaries (results from other games), or predictions (who's going to win). After a game starts, it's played until an event causes a break. After a play stops, it will be resumed until it's stopped again. This play-break phase is iterative until the end of the game, which is then followed by a postgame scene that has similar content to the pregame. The main difference of a postgame scene (from pregame) is that it contains commentaries about the overall game and provides some highlighted scenes. Pregame and postgame scenes are usually recorded in a studio and mark the start and end of a sports video. Figure 2 shows the generic play-break sequence model that any sport can use. However, the structure of play-break iteration in this model can be specialized for specific sports. In this article, we use three examples from three sports categories: period-, time-, and performance-based, which are distinguished based on their temporal structures' similarity.

Period-based sports are typically structured into playing periods, such as a set in tennis and basketball; half or quarter (of a match) in soccer, basketball, and Australian football; and a round in boxing. Thus, this sports category typically begins when the referee indicates the start of a playing period. After a match begins, there are some iterative play and break sequences until the end of the period. Unless we reach the end of the match, we'll see another start of the playing period after each period ends.

Note that we can predict the number of playing periods for each class of sports. For example, soccer usually has two 45-minute (normal) playing periods. In some cases, where required, there could be another two 15-minute (extended) playing periods that would end abruptly if a team scores a goal.

Time-based sports usually involve races (or competitions) that are structured around laps. Examples of this sports category are swimming, motorbike, and Formula One races. Unlike period-based sports, which are usually broadcast as an individual match, swimming is mostly broadcast as a portion of a championship or competition. For example, in day eight of the Australian National Championship live broadcast, viewers are presented with multiple races, such as "men's semifinal freestyle 50 m." Each race can be decomposed into one or more laps (which are equivalent to a play). After all the laps in a competition are finished, the winner will usually be interviewed before another race is started unless we reach the end of the program (which is marked by a postgame scene). During a lap, only little key events could happen, such as overtaking the lead and breaking a record. In Formula One or motor races, we might find accidents in an event.

Performance-based sports include gymnastics, weight lifting, golf, and track-and-field events such as the high and long jumps and throwing (for example, shot put and javelin). Performance-based sports' temporal structure is similar to time-based sports. For example, in day 21 of Olympics gymnastics, viewers will see different competitions such as the men's and women's acrobatic artistic or rhythmic semifinals. Each competition will have one or more performances (by each competitor). Similarly, the winners of each competition are usually interviewed after their performances. Due to their similarity, we could have grouped time- and performance-based sports, however, the main difference is the rarity of key events in performance-based sports. Unlike a lap, we can consider each performance a key event because there are many breaks between each performance, such as players waiting for the results (for example, points from the judges) and slow-motion replay of earlier performances.

Integrating highlights into play-breaks

We need to know which highlights should be

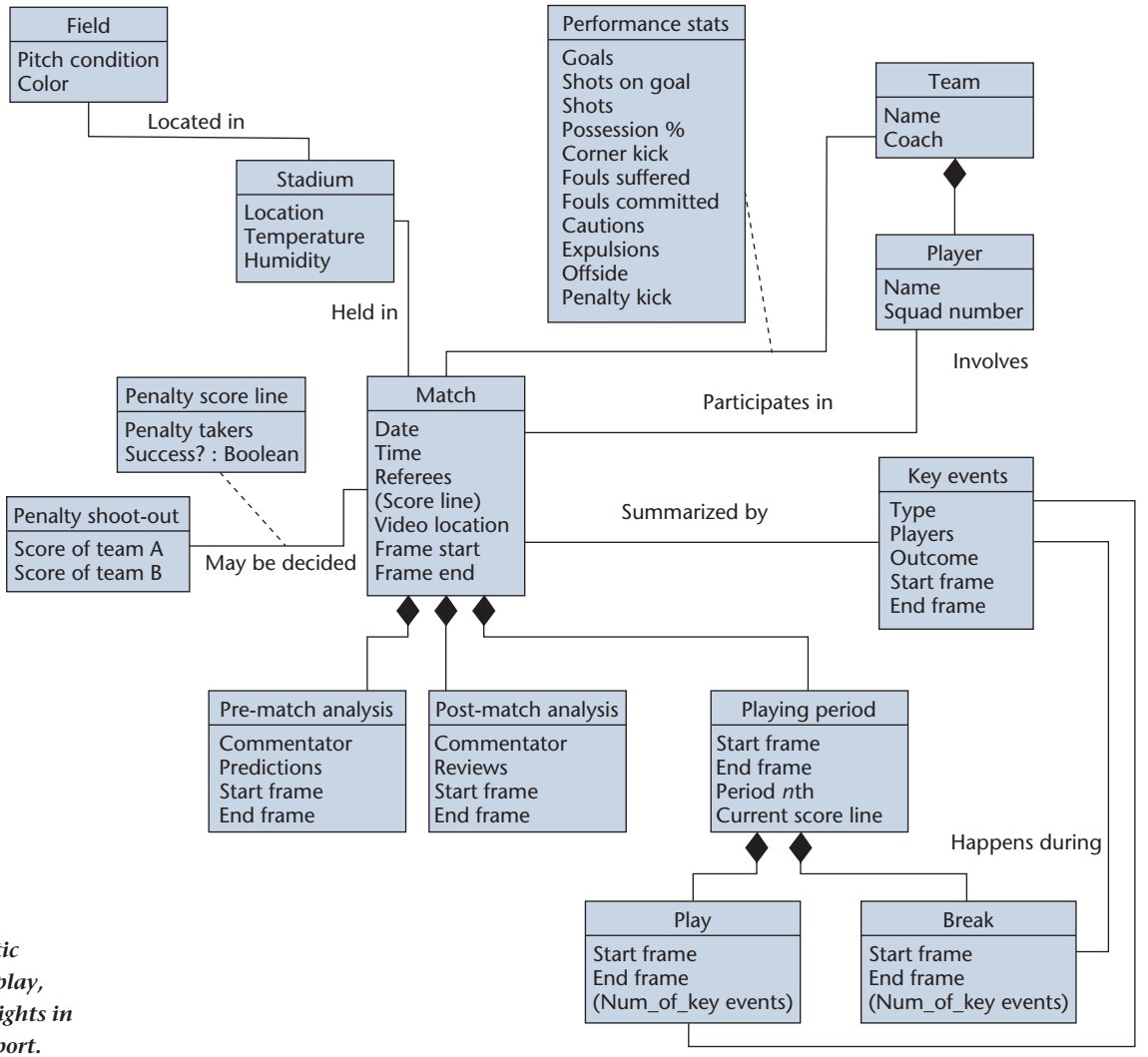


Figure 3. Semantic relationships of play, break, and highlights in a period-based sport.

attached to a play or break. For example, in soccer, when a play is stopped because of a foul, a slow-motion replay of earlier highlight(s) from different angles is usually shown along with a text display of the player's name during closeup shots. Then, the play can be resumed with a free kick. Thus, in soccer games, goals, set pieces, and fouls are the typical highlights during play-break-play transitions while good plays, goal shots, and offside highlights occur during play scenes. In addition, ceremonies, preparing for set pieces, and player substitutions are the typical highlights that happen during a break (that is, when a ball isn't in play or the camera isn't focused on players). Slow-motion replays and text displays can also be inserted during break sequences to keep viewers' attention.

Hence, a more complete sports video summary should be able to include all these highlights into plays and breaks. When all highlights are detected and attached to their corresponding play or break, the system can generate a highlights collection based on the annotation of each highlight, such as an XML tag of <Highlight Type>. During summary construction, we can either segment plays and breaks first and then localize the highlights within each of them, or segment highlights and let the system determine whether they belong to a play or break.

Figure 3 shows a diagram, based on the Unified Modeling Language, that describes the semantic relationships between a soccer match, playing periods, plays, breaks, highlights, and the participating objects. (We can easily modify this model for other sports.) The first component of this model is semantic objects. Two soccer teams are the primary participants of a soccer match. Each team's performance during the match is measured by the performance statistics, such as the number of fouls committed. Soccer players play for one of

the participating soccer teams, and they're uniquely identified by their last name and squad number. A field is where a sports match is played. However, viewers' main interests are primarily aimed toward the ball and the players since the camera mainly captures soccer players and the ball's movements during a soccer match. A stadium containing a field is where the crowds are seated. Thus, the stadium represents the soccer match's environment, which we can describe in terms of the address (location), temperature, humidity, wind speed, and so on. Environmental conditions can also be of interest because these factors can affect the players' performance. For example, if a team is playing at home, it is more likely to win since they have more support from the audience.

The second component of this diagram is the events and highlights in a sports match. A soccer match can be composed of a pre- and postmatch analysis and the (actual) playing period. Each playing period contains a continuous sequence of plays (ball in play) and breaks (ball out of play or certain events happen). A break can also indicate a match highlight. For example, the referee will stop the match when a foul occurs. A match highlight happens during a specific playing period and can be used to summarize the interesting events in a soccer match. Each match highlight is identified by the time when it happens, the type of highlight (goal, foul, and so on), the players who were involved, and the possible outcome of the event. Finally, a penalty shoot-out is sometimes needed to decide the winner of a soccer match. When a penalty shoot-out occurs, the referee records for each penalty taker the status of whether a point was gained. In the end, the difference in the total score of both teams will determine the winner.

We developed a prototype GUI (see Figure 4) to show how users can browse our proposed sports video summary structure. The interface consists of video hierarchy (of play-breaks and highlights), video player, and annotation sections. Using the drop-down lists in the hierarchy, users can select a particular game. Based on this selected game, the system will load the corresponding lists of plays, breaks, and highlight collections that (when selected) will load the particular key events attached to them. Each time a user selects a component in the hierarchy (such as a play), the annotations will appear while the video player shows the keyframe. Video controls (such as play, stop, and pause) are available from the video player.

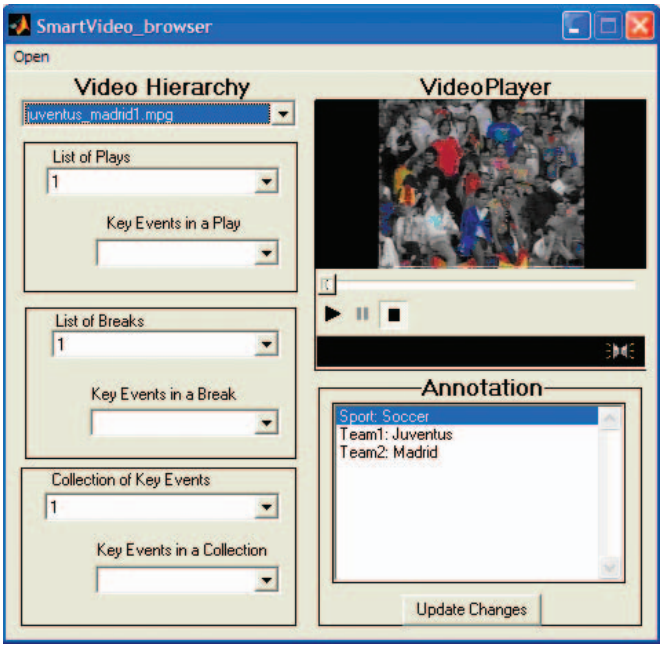


Figure 4. Browser for play-breaks and highlights.

In the future, we plan to enhance this interface to let users add or update their own highlight collections as well as write queries. Moreover, we need to let users browse on semantic and temporal links between highlights, plays, and breaks. The most challenging improvement, however, is to include a graphical hierarchy that adapts to the type of sport being browsed.

Extracting events

Figure 5 (next page) shows the two main approaches that we can take to detect play-breaks and highlights. In the bottom-up approach, we need to first apply generic feature extraction techniques to identify important shots, such as using a color histogram comparison to detect shot boundaries.[5] Second, we need to analyze the shot contents to identify the semantic objects, their activities, and the relationships that form an event. When the system detects an event within a sequence of shots, it needs to group or classify the shots into a scene. For example, the system needs to detect a sequence of shots that contains object activities including "player A runs then passes to player B," "player B scores a goal," and "player C fails to keep the ball out of goal" to identify a goal event. The main disadvantage of this approach is the complexity and long processing time required to apply event detection algorithms for each shot. In a sports game, shots are repeated many times without necessarily leading to a (key) event.

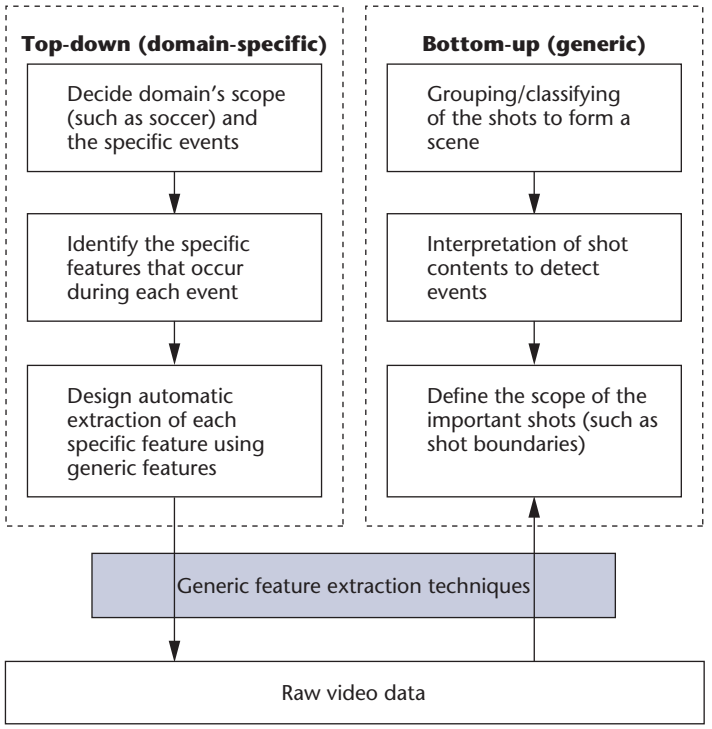


Figure 5. Top-down versus bottom-up approach for detecting play-breaks and highlights during a game.

In contrast, the top-down (or domain-specific) approach is more effective because we can use faster detection algorithms to detect specific features that occur distinctively during each highlight event. For example, when the position of play is near a goal, if the crowd's cheer and commentator's speech is excited for a certain period of time, and then a text display shows an updated score, there's a big chance that a goal event has occurred. Thus, we need to first decide the scope of domains we want to summarize (in our case, it's sports video). We then need to decide the specific events that can be used to summarize the video content (such as plays, breaks, and key events).

The system can detect each event automatically using specific features. Hence, we need to automatically extract these specific features using generic feature extraction techniques such as audio analysis (loudness and pitch) for detecting excitement, visual analysis (edge and shape) to detect text display, and closed-caption analysis (detection of key words). We can achieve a faster and more efficient highlight detection process by maximizing the use of audio features, which are generally computationally cheaper than using visual features analysis. Thus, our approach uses audio features to localize the highlights and play-break events. We localize the text displays to verify the detection and assist in the annotation process.

When applying the top-down approach for designing solutions for sports video summarization, some features are more effective for detecting different events in specific sports. For example, a crowd and commentator's excitement usually indicates key events in soccer whereas excitement is detected constantly (for each lap) during a swimming race. On the other hand, crowd noise (mostly applause) in tennis indicates the end of each play because the crowd has to stay quiet until a point is awarded after each play; comments are usually made during a short break (before another play is resumed by a ball serve).

Here's how our summarization processing approach works. The system detects highlight, play, and break scenes from sports games and stores them as raw video data. The detected highlights are stored as links to the raw video in the database after the system detects the text display and verifies the highlights. Annotators then use the text display to verify the type of highlight and annotate it with some information for retrieval. For example, a goal can be described in terms of the updated scores between the competing teams, details of the goal scorer (player name, team name, and squad number), and the time in which the goal is scored. Moreover, the annotators can also recheck the highlight scene to ensure that it is consumable (that is, it can be viewed) by itself since this process is subjective and almost impossible to automate. To assist annotators, information about a sports game and its highlights are often shareable with other games, especially if they're the same type of sport. Thus, we can achieve a faster highlight construction by storing the most common information during an offline process.

In the following sections, we briefly describe algorithms for detecting play-breaks, highlights, and text annotation. You can find more details on these algorithms elsewhere,[6] including the thresholds we used.

Detecting play-breaks

Play-break transitions during most sports videos can be generically detected using a camera-views classification. Generally, a long global shot with some interleaving short zoom-in and close-up shots usually correspond to a play. On the other hand, long zoom-in or close-up shots indicate a break. A common approach to classifying the main shots in a sports video is to use the grass (or a dominant color) ratio, which measures the amount of grass pixels in a frame. The basic con-

cept is that global shots contain the highest grass ratio, while zoom-in contains less, and close-up contains the lowest (or none).[4] Moreover, slow-motion replay scenes or full-screen texts should also be regarded as breaks. To detect slow-motion replay, we can use the hidden Markov model to capture the possible transitions of features—such as editing effects, visual slow-motion effects, still fields, and normal motion replay—that occur during slow-motion replays.[7]

To complement or replace these detection methods, we can use whistle detection to localize play-breaks in specific sports that use whistles, such as soccer, rugby, swimming, and basketball. The main benefit of using a whistle as a play-break indicator is that whistle occurrences in sports video is distinctive and can be detected quickly. A whistle sound can overcome the complexity of tracking audio in sports videos because it's distinguishable from human voices and other background sounds. In soccer, whistle sounds indicate the start and end of the match and playing period; play stops, such as a foul or offside; and play resumes (after being stopped). In swimming, a long continuous whistle tells swimmers to get ready for a race.

Detecting whistles requires calculating the spectral energy within the whistle's frequency range (for example, 3500 to 4500 Hz for soccer[8]) as follows:

$$PSD_W = \sum_{WL}^{WU} \left| S(n) * \mathrm{conj}\left(S(n)\right) \right|$$

where WU and WL are the upper and lower bounds of the whistle frequency range (respectively), and $S(n)$ is the spectrum (produced by the fast Fourier transform) of the audio signal at frequency n Hz.

Because users enter WL and WU in terms of Hz, we applied the following equation:

$$WX = \mathrm{round}\ \{(WX_{\mathrm{Hz}}/fs) * N\}$$

where WX_{Hz} is WL or WU in terms of its Hz value, fs is the sample frequency, and N is the n-point fast Fourier transform performed. Table 1 shows the whistle ranges for various sports.

Within each video clip, our system marks a frame as (potentially) containing a whistle sound if it contains a PSD_W value greater than the minimum value for $PSDw$ that we regard as a potential whistle (threshold1). We then consider this current value of PSD_W as the current significant value. Finally, the system determines that a clip has a whistle sound if it finds at least n neighboring frames containing a PSD_W value of at least 80 percent of the current significant value. Thus n is threshold2, which specifies the minimum number of frames required to confirm whistle existence.

Table 1. Whistle ranges for various sports.

Sport	Whistle Range (Hz)
Soccer	3500 to 4500
Swimming	2800 to 3200
Rugby	2500 to 2700
Basketball	3500 to 3700
Netball	2300 to 2600

Detecting highlights

We can localize highlights in sports videos by detecting slow-motion replay scenes.[9] This approach is robust because nearly all sports videos use slow-motion replay to indicate interesting events. The main disadvantage is that if there is no slow-motion scene after an exciting event, we miss the highlight. Moreover, not all slow-motion replay scenes are displayed after the key event. In most cases, broadcasters wait until the game becomes less intense to replay earlier highlights. In this case, the replay scene usually doesn't contain all the necessary details to form a comprehensive highlight, such as text annotation and audio key words. Hence, we should extract specific features to detect the key events.

Many approaches already exist for detecting key events based on visual features.[10] For example, Gong et al.[3] and Zhou et al.[11] summarized videos of soccer and basketball (respectively) games according to the playing position, such as midfield and penalty area. They used an inference engine or tree-learning rules (that is, if-then-else) to analyze the lines detected from the soccer playing ground, motion detection, and color analysis of the balls and players' uniforms.

Similarly, Chang et al.[12] used statistical models to analyze the spatial-temporal characteristics of typical camera views in baseball videos to detect more specific highlights, such as home-runs and nice hits, in addition to pitch and batting that Rui et al.[13] have detected. The main benefit of this approach is that broadcasters always prepare the visual component of a sports video in a consistent manner to help viewers understand the content. For example, only certain camera operations can capture specific objects during a sports game. Moreover, we can support visual-related queries, such as "show (video) shots where team A scored from the left side of the field." However, combined with other features, this approach can potentially detect highlights more accurately being used alone. In particular, we can model the temporal syntax of

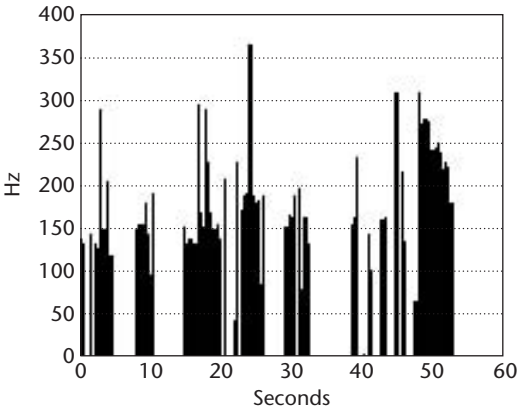


Figure 6. Pitch determination and silence detection applied to an audio clip containing a goal attempt.

different sports highlights based on the occurrence of specific features such as high-energy audio segments, text displays, closed captions, specific camera views, and motion direction.[2]

Unlike play-breaks, the scope and detection methods for key events are more specific for different sports. For example, goals, fouls, and kicks are common key events for period-based sports such as soccer and rugby. Overtaking the lead, nearing the finish line, and record-breaking times are more common for race-based sports such as swimming. Thus, we generalized key events as the segments in which our system could detect excitement from the crowd and/or commentators. When exciting events occur, generally the crowd's cheer and commentator's speech become louder, faster, and higher (in pitch) and less pauses occur.

To localize louder clips, we used the whistle-detection method, but replaced the calculation of volume for that of PSD_W. We used this equation to calculate the volume of each audio frame:

$$\text{Volume} = \frac{1}{N} * \sum_{n=1}^{N} \left| s(n) \right|$$

where N is the number of frames in a clip and $s(n)$ is the sample value of the nth frame.

To calculate pitch and silence, we applied the subharmonic-to-harmonic ratio-based pitch determination from Sun[14] for its reliability (see Figure 6).

Based on the pitch values, we can calculate the high-frequency and pause rates in a clip using dual-fashioned equations:

$$\text{PauseRate} = \#P_f/N * 100 \text{ percent}$$

$$\text{HighpitchRate} = \#HP_f/N * 100 \text{ percent}$$

where $\#P_f$ is the number of frames containing speech pauses 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.

Detecting text for annotation

During or after key events in most sport videos, broadcasters insert a text box on screen to draw users' interest to some important information, such as player details and the updated score. Full-screen texts display team members' names as well as statistics after each playing period. Moreover, smaller-sized texts are usually constantly displayed to keep viewers up to date with the current progress of the game, such as current elapsed time and score.

Our text display detection method is based on an assumption that, in most cases, sports videos use horizontal text to display important information.[15] Thus, if we can detect a prominent horizontal line in a frame that corresponds to the text box, we can locate a text area's starting point. For this purpose, we used the Hough (or Radon) transform on gradient images (produced by the Sobel filter) to detect prominent lines in video frames. The main benefit of this method is that most text displays in sports videos are surrounded by a large rectangle box to distinguish them from the background. However, some sports videos use vertical text, which can be detected using methods presented elsewhere.[16]

Our text display detection method works as follows. First, we segment the video track into a 1-minute clip. The system preprocesses each of the frames within a 1-second gap of the video track (by converting the color scheme to grayscale and reducing the clip's size to a preset smaller size) to optimize performance. Second, we apply the Sobel filter to calculate the edge (gradient) of the current frame and then apply the Radon transform on the gradient image to detect line spaces (R) that are in between a 180-degree angle. We applied threshold1 (the minimum value that R usually corresponds to a prominent line) on these R values to detect the potential candidates of prominent lines usually formed by the box surrounding the text display.

After our system detects these lines, it calculates 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 the theta (t) value, which indicates the line's angle.

To verify that the detected lines are the candidates of a text display region, the system only retains the lines that follow these criteria:

- the absolute value of r is less than n percent of the maximum y-axis, and
- the corresponding t is equal to 90 (horizontal). This n-value represents threshold2, the maximum possible location of the horizontal line in terms of the y-axis.

The first check is important to ensure that the location of the lines is within the usual location for a text display. The second check is to ensure that the line is horizontal because there are potentially other prominent horizontal lines that can be detected from other areas besides the text display, such as the boundary between a field and a crowd. Finally, for each of the lines detected, the system checks that their location (the r values) is consistent for at least m seconds (that is, if the video frame rate is 25, 2 seconds is equal to 50 frames). We consider this m-value as threshold3, the minimum period (in terms of seconds) that the lines must stay in a consistent location. The purpose of this check is to ensure that the lines' location is consistent for the next frames because text displays always appear for at least 2 seconds to give viewers ample reading time. 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. Figure 7 illustrates how the Sobel filter and Hough transform detect text.

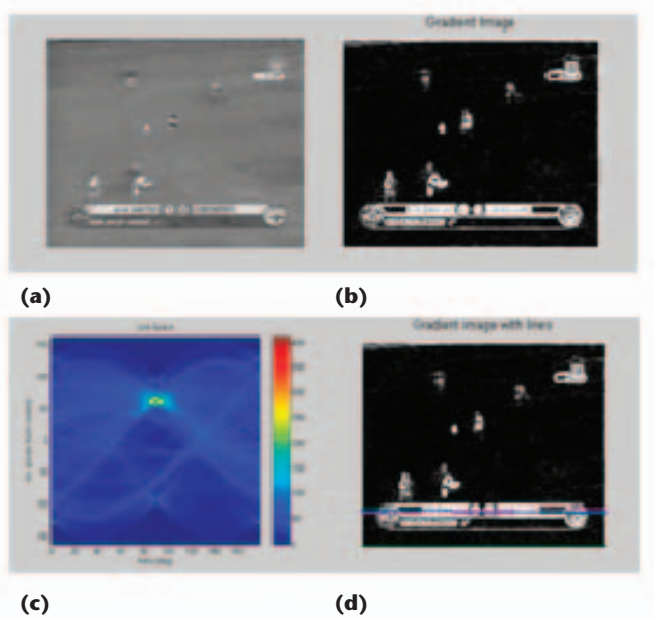


Figure 7. How the Sobel filter and Hough transform detect text: (a) Grayscaled and resized frame. (b) Gradient image reveals the lines. (c) Peaks in Hough transform. (d) Horizontal lines detected.

Figure 8. Various text displays in our data set.

Experimental results

Here we describe the reliability and robustness of our events detection algorithm for various sports genres.

We used a data set of around three hours of sports videos that included a wide range of sports: soccer, swimming, tennis, rugby, a bike race, a horse race, basketball, and netball (a popular sport in Australia that's similar to basketball). Table 2 (next page) provides the details of each video, including its commentators' characteristics. Figure 8 shows examples of the videos' text displays. We took these samples from different broadcasters and events so that we could test the robustness of our detection algorithms. To import video and audio streams to MathWorks' Matlab software, we performed some preprocessing (as illustrated in Figure 9 on p. 33) on each video using a combination of available software.

For our first set of ground truths, we first performed manual detection on the occurrences of whistle sound, excitement, and text displays. However, the manual perception of excitement

Table 2. Characteristics of our data set.

Video	Event	Excitement
Soccer 1 Manchester United vs. Deportivo (20 minutes)	Champions league 2002	Generally talkative and very descriptive. Crowd was emotional following the match.
Soccer 2* Juve vs. Madrid (20 minutes)	Champions league 2003	Same as the Soccer 1 video, but generally more excited because this match is more important than Soccer 1.
Soccer 3** Brazil vs. Germany (20 minutes)	Federation Internationale de Football Association (FIFA) World Cup 2002	Less talkative and less descriptive. Less emotional when the match gets exciting.
Soccer 4 Milan vs. Inter Milan (20 minutes)	Champions league 2003	Same as Soccer 2 video.
Swimming 1* Women 100-m freestyle (5 minutes)	Australian national competition	Have male and female commentator. Event was indoors and had less background noise than an outdoor game.
Swimming 2* Women 200-m breaststroke (5 minutes)	Australian national competition	Same as the Swimming 1 video, but only had a male commentator.
Tennis Martina Hingis vs. Jennifer Capriati (20 minutes)	Australian Open 2002	Two females. They only make comments after a play is stopped.
Rugby 1* Australia vs. Romania (20 minutes)	World Cup 2003	Constantly describing the match. Crowd was emotional following the match.
Rugby 2* France vs. Japan (17 minutes)	World Cup 2003	Less talkative and less descriptive than Rugby 1 but almost the same amount of excitement during key events.
Bike race (race 2)*** 8 cars naked-bikes race (8 minutes)	Australasian FX-pro twins 2003 championships	Two male commentators are constantly talking about the race progress with no significant excitement when key events occur.
Horse race*** (10 minutes)	Carlton Draught Caulfield Cup	During the race, commentator(s) become very talkative and excited.
Basketball** (14 minutes)	Australia's Womens National League 2003–2004	There is one female and one male commentator. They are both descriptive and follow the crowd's emotions well.
Netball** (9 minutes)	Australian National Championships 2003	Same as the basketball video. However, the commentators are generally more excited because it's a more important event than the usual league games because it's a final.

*Video starts from the beginning of the game (including ceremonies or players' introduction).
**Video finishes after a playing period is stopped.
***Video features a full-length race (start to the end) as well as the interview of the winner.

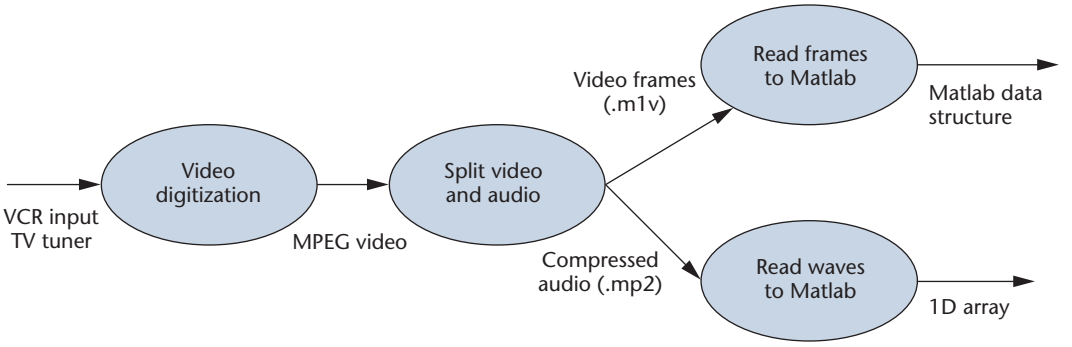


Figure 9. Preprocessing steps of videos.

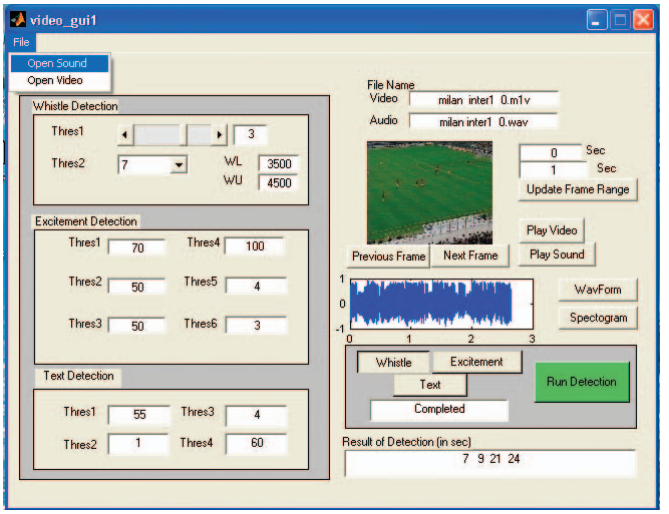


Figure 10. User interface for running and checking the detection algorithms.

can be subjective because of individual sensitivity to loudness and noise. Moreover, the level of excitement varies for different sports, such as soccer versus swimming, as well as from one video to another video due to different crowds, commentators, and recording characteristics. Since our excitement detection is based on detecting clips containing a higher number of louder, high-pitched, and fewer pause frames, we combined subjective hearing (of excitement) with manual observation of these features. In particular, we used a waveform in addition to a graph of volume to locate louder clips and pauses. We used a diagram plotting pitch values against time to locate the clips containing higher-than-average pitches to confirm manual hearing on high-pitched speech. We manually mimicked the algorithm to combine these features before we checked final excitement candidates, ensuring that they actually represented excitement and/or key events.

For the second set of ground truths, we manually localized the occurrences of highlights and play-breaks for each sports video. Then, we checked whether each play-break transition and highlight could be localized by whistle, excitement, or text.

Results and discussion

We tested the detection algorithms that we developed in MathWorks' Matlab 6.5 using a Pentium 4 1.5-GHz PC with 512 Mbytes of memory in the Windows XP professional platform. We constructed a GUI (see Figure 10) for testing the detection algorithms for whistle, excitement, and text. The main purpose of this GUI is to help us modify the best thresholds used for each algorithm.

Moreover, the correctness of the detection results can be checked directly by viewing the particular video frames or playing the particular sound clip.

To evaluate the performance of our detection algorithms, we used these measures:

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

- Precision rate (PR): 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 tradeoff for achieving minimum misdetections. This is because the lower thresholds we use, the fewer of missing events we have, but at the same time we'll get more false detections.

Table 3. Performance measures of whistle, excitement, and text detection.

Sample Video	Whistle					Excitement					Text				
				RR	PR				RR	PR				RR	PR
	Nc	*Nm*	*Nf*	%	%	*Nc*	*Nm*	*Nf*	%	%	*Nc*	*Nm*	*Nf*	%	%
Soccer 1	13	9	9	60	60	50	12	10	81	83	9	3	3	75	75
Soccer 2	7	2	2	77	78	21	1	14	96	63	11	4	10	71	52
Soccer 3	11	5	2	69	84	41	0	26	100	61	6	0	4	100	60
Soccer 4	2	1	0	67	100	18	4	8	82	69	9	2	3	82	75
Swimming 1	1	0	0	100	100	8	0	3	100	64	19	3	5	86	73
Swimming 2	1	0	0	100	100	13	3	2	81	87	6	3	4	67	60
Tennis	Whistle isn't used				35	6	3	85	92	Algorithm not applicable					
Rugby 1	17	2	5	90	77	8	5	6	62	57	15	2	5	88	75
Rugby 2	9	0	0	100	100	12	2	6	86	67	14	3	7	82	67
Bike race	Whistle isn't used				2	1	1	67	67	6	6	1	50	86	
Horse race	Whistle isn't used				4	0	1	100	80	9	2	4	82	69	
Basketball	9	6	3	60	75	25	2	9	93	74	30	1	2	97	94
Netball	36	2	4	95	90	16	1	6	94	73	14	0	2	100	88
Average				82	86				87	72				83	80

We used the following equations for these indicators:

$$RR = Nc/(Nc + Nm) * 100 \text{ percent}$$

$$PR = Nc/(Nc + Nf) * 100 \text{ percent}$$

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

Table 3 shows the performance measures. Our goal was to achieve a precision rate of 70 percent or greater. The lowest acceptable precision rate should not be less than 60 percent. Based on the average statistics, we can justify that the detection algorithms are overall robust and reliable.

Table 4 shows the comparisons of detected play-breaks and highlights using whistle detection only, whistle-text detection, whistle-excitement detection, and finally whistle-excitement-text detection. This table demonstrates the advantage of using the three features in terms of the number of highlights that each algorithm can detect and the time each algorithm spends on detection. This table shows that whistle detection is fast, but it can only localize 20 to 30 percent of the total highlights, which are mostly caused by fouls and offsides (that is, when plays are stopped).

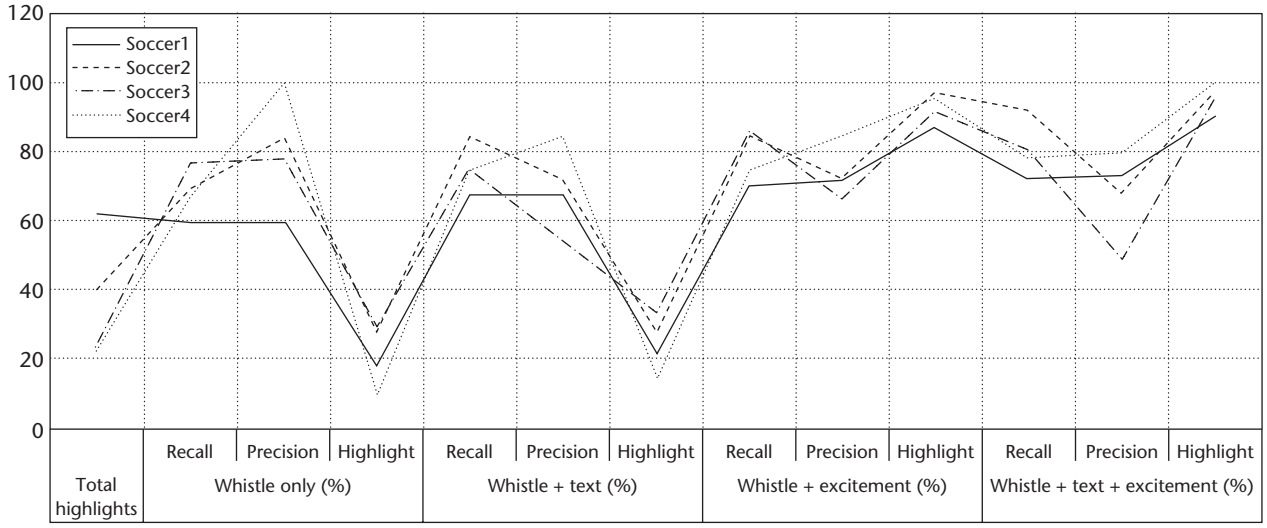
In most cases, however, a whistle isn't really used to indicate play being resumed again with a free kick, unless there's a substantial pause during the break. For example, a direct free kick taken near the penalty area will be indicated by a whistle after the period of time in which the teams are preparing their formation. In contrast, with a free kick taken from the defensive to midfield area, the whistle is only blown to indicate that there's a foul or offside without indicating the free kick itself. Hence, we need to adopt the camera-views-based method[4] so that we can define the play-break transition more precisely.

By combining whistle and excitement noises, users only need to wait slightly longer to detect 80 to 90 percent of the highlights since the excitement algorithm can locate most exciting events, such as good attacks or defensive plays, goals, free kicks, and sometimes even fouls. In addition, excitement detection is effective for localizing goal highlights because of the massive amount of excitement during the start of a good attack, which often leads to the goal itself. Moreover, the excitement will still be sustained during the goal celebration and slow-motion replays, especially when the commentator and crowd are excited about the goal.

When we combine whistle and text detection, the number of highlights detected only slightly increases and the processing period is longer than using excitement. This is because visual features are generally more expensive computationally than audio features. We need text detection for localizing the start of a match, a goal, and shot on goal, as well as confirming offside and foul events. Large (to full-screen) text is usually displayed before a match begins to show the starting lineup of each team and the formation they use for the

Table 4. Statistics of highlight detection using various combinations of features.

		Automatically Detected Highlights							
		Using Whistle Only		Using Whistle + Text		Using Whistle + Excitement		Using Whistle + Excitement + Text	
Sample Video	Total Highlights	Number of Highlights	Time (Minutes)	Number of Highlights	Time (Minutes)	Number of Highlights	Time (Minutes)	Number of Highlights	Time (Minutes)
Soccer 1 (40 minutes)	62	11	1.7	13	37.1	54	22.9	56	58.2
Soccer 2 (20 minutes)	24	7	0.7	8	24.8	22	10.6	23	35.4
Soccer 3 (20 minutes)	40	11	0.7	11	26.7	39	8.8	39	35.5
Soccer 4 (20 minutes)	22	2	0.9	3	18.1	21	8.9	22	19
Swimming 1 (5 minutes)	3	1	0.2	3	5.1	1	3.2	3	8.1
Swimming 2 (5 minutes)	3	1	0.2	3	5.2	1	3.3	3	8.3
Tennis (20 minutes)	40	0	0	0	0	33	9.9	33	28.8
Rugby 1 (20 minutes)	34	18	0.9	20	20.6	25	10.9	27	29.9
Rugby 2 (17 minutes)	21	8	0.7	9	16.9	18	9.6	19	18.5
Bike race (8 minutes)	9	0	0	5	6.5	2	3.5	7	9.9
Horse race (10 minutes)	2	0	0	2	8.4	2	4.5	2	12.9
Basketball (15 minutes)	37	7	0.8	12	14.6	30	7.9	35	21.9
Netball (9 minutes)	43	36	0.4	39	8.8	38	4.9	41	13.4
Average time spent for 1-minute segment			0.04		1.06		0.52		1.49



match. Since these text displays are large and contain a lot of information, they're usually displayed for the whole 1- or 2-minute time span. After a goal is scored, a text box shows the updated score. Similarly, after a shot on a goal, usually the text will confirm that there's no change in score or show the details of the player(s) involved (such as the forward player and goalkeeper.

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

Figures 11 and 12 show the scalability of our detection algorithms for soccer and rugby videos, respectively. For different amounts of total highlights, we can achieve almost the same performance measures for different combinations of whistle, excitement, and text detections. These figures include the highlight ratio, which is the

Figure 11. Scalability of the highlights detection algorithms for soccer videos.

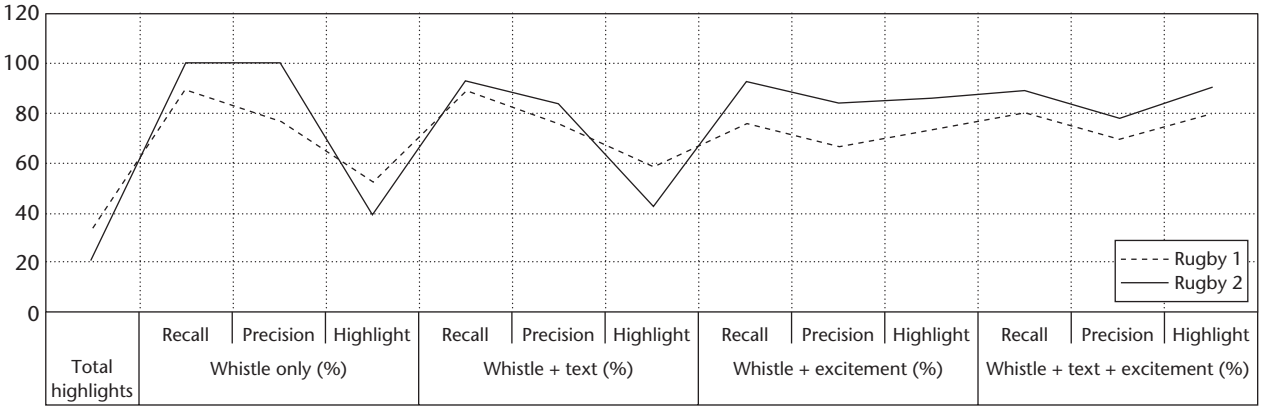


Figure 12. Scalability of the highlights detection algorithms for rugby videos.

percentage of highlights the algorithms correctly identify. The HR indicates that for any number of total highlights, whistle-only detection can only detect 9 to 29 percent, while whistle and text detect 13 to 33 percent of the highlights. The whistle and excitement algorithm is a good combination because it can detect 87 to 95 percent of the highlights. Finally, when we combine whistle, text, and excitement, the algorithm detects 90 to 100 percent of the highlights.

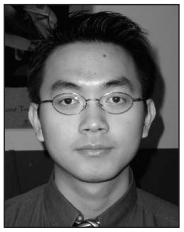
Future work

We've used some slightly adjustable thresholds when applying our algorithms to different videos to avoid misdetections and reduce false detections. For future work, we'll need to design an automated method for deciding the thresholds, so that these algorithms become fully automated and less biased by subjective decisions. We'll then adapt current video optical character recognition techniques to extend our text detection method to achieve a fully automated verification and annotation of highlight and play-break sequences. We hope to also extend our video summarization scheme to include methods for indexing and retrieval, so that we can show its benefits in terms of meeting user and application requirements. In addition, because mobile devices that can play video have become more common, we aim to integrate these devices with our algorithms to provide sports fans with a summarized version of sports videos from anywhere. **MM**

References

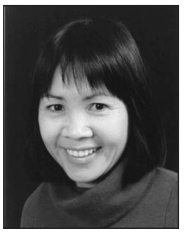
1. D. Zhong and S.-F. Chang, "Structure Analysis of Sports Video Using Domain Models," *Proc. IEEE Int'l Conf. Multimedia and Expo* (ICME 2001), IEEE CS Press, 2001, p. 182.
2. N. Babaguchi, Y. Kawai, and T. Kitahashi, "Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration," *IEEE Trans. Multimedia*, vol. 4, no. 1, 2002, pp. 68-75.
3. Y. Gong et al., "Automatic Parsing of TV Soccer Programs," *Proc. Int'l Conf. Multimedia Computing and Systems*, ACM Press, 1995, pp. 167-174.
4. A. Ekin and A.M. Tekalp, "Generic Play-Break Event Detection for Summarization and Hierarchical Sports Video Analysis," *Proc. Int'l Conf. Multimedia and Expo 2003* (ICME 03), IEEE Press, 2003, pp. 169-172.
5. H.J. Zhang, "Content-Based Video Browsing and Retrieval," *Handbook of Internet and Multimedia Systems and Applications*, B. Furht, ed., CRC Press, 1999, pp. 255-280.
6. D. Tjondronegoro, Y.-P.P. Chen, and B. Pham, "Sports Video Summarization Using Highlights and Play-Breaks," *Proc. ACM SIGMM Int'l Workshop Multimedia Information Retrieval*, ACM Press, 2003, pp. 201-208.
7. H. Pan, P. van Beek, and M.I. Sezan, "Detection of Slow-Motion Replay Segments in Sports Video for Highlights Generation," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing* (ICASSP 01), IEEE Press, 2001, pp. 1649-1652.
8. W. Zhou, S. Dao, and C.-C. Jay Kuo, "On-Line Knowledge- and Rule-Based Video Classification System for Video Indexing and Dissemination," *Information Systems*, vol. 27, no. 8, 2002, pp. 559-586.
9. N. Babaguchi et al., "Linking Live and Replay Scenes in Broadcasted Sports Video," *Proc. ACM Workshop Multimedia*, ACM Press, 2000, pp. 205-208.
10. A. Ekin and M. Tekalp, "Automatic Soccer Video Analysis and Summarization," *IEEE Trans. Image Processing*, vol. 12, no. 7, 2003, pp. 796-807.
11. W. Zhou, A. Vellaikal, and C.C.J. Kuo, "Rule-Based Video Classification System for Basketball Video

Indexing," *Proc. ACM Workshop Multimedia*, ACM Press, 2002, pp. 404-441.
12. S.-F. Chang, "The Holy Grail of Content-Based Media Analysis," *IEEE MultiMedia*, vol. 9, no. 2, Apr.–June 2002, pp. 6-10.
13. Y. Rui, A. Gupta, and A. Acero, "Automatically Extracting Highlights for TV Baseball Programs," *Proc. ACM Int'l Conf. Multimedia*, 2000, ACM Press, pp. 105-115.
14. X. Sun, "Pitch Determination and Voice Quality Analysis Using Subharmonic-to-Harmonic Ratio," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing* (ICASSP 02), IEEE Press, 2002, pp. 333-336.
15. R. Lienhart and A. Wernicke, "Localizing and Segmenting Text in Images and Videos," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 12, no. 4, 2002, pp. 256-268.
16. H. Li, D. Doermann, and O. Kia, "Automatic Text Detection and Tracking in Digital Video," *IEEE Trans. Image Processing*, vol. 9, no. 1, 2000, pp. 147-156.



Dian Tjondronegoro is a PhD candidate at Deakin University, Melbourne, Australia. His research interests include video content analysis and retrieval, XML databases, MPEG-7, and XQuery. Tjondronegoro received a BS in information technology (first honors) from Queensland University of Technology.

Yi-Ping Phoebe Chen is an associate professor and head of the Multimedia Streams and the Bioinformatics Research Labs in the School of Information Technology, Faculty of Science and Technology at Deakin University, Melbourne, Australia. Her current research interests are multimedia and bioinformatics. Chen received a BS in information technology (honors) and a PhD in computer science from the University of Queensland. She is the founder of the Asia-Pacific Bioinformatics Conference and is the steering committee chair for the Multimedia Modeling Conference and Asia-Pacific Bioinformatics Conference.



Binh Pham is a professor and director of research in the Faculty of Information Technology at the Queensland University of Technology, Brisbane, Australia. Her research interests include computer graphics, multimedia, image analysis, and intelligent systems and their applications in diverse domains. Pham has a PhD in numerical computing from University of Tasmania and a DipEd from Monash University.

Readers may contact Yi-Ping Phoebe Chen at the School of Information Technology, Deakin Univ., 221 Burwood Hwy., VIC 3125, Australia; phoebe@deakin.edu.au.