Creation of Soccer Video Highlights Using Caption Information

Seong-Yoon Shin*, Il-Ko Kang**, Yang-Won Rhee***

Abstract

A digital video is a very long data that requires large-capacity storage space. As such, prior to watching a long original video, video watchers want to watch a summarized version of the video. In the field of sports, in particular, highlights videos are frequently watched. In short, a highlights video allows a video watcher to determine whether the highlights video is well worth watching. This paper proposes a scheme for creating soccer video highlights using the structural features of captions in terms of time and space. Such structural features are used to extract caption frame intervals and caption keyframes. A highlights video is created through resetting shots for caption keyframes, by means of logical indexing, and through the use of the rule for creating highlights. Finally, highlights videos and video segments can be searched and browsed in a way that allows the video watcher to select his/her desired items from the browser.

Keyword : 하이라이트 비디오(Highlight Video), 장면전환검출(Scene Change Detection), 자막 프레임 구간(Caption Frame Interval), 자막 키 프레임(Caption Key Frame).

* 제1저자 : 신성윤
** 접수일 : 2005.10.11, 심사완료일 : 2005.11.16
***.**** 군산대학교 컴퓨터정보과학과
1. Introduction

In a video database, video contents are described as the shot or scene structure(1). A shot is a valid unit forming video information, that consists of one or more sets of successive frames, and that represents continuous motions at some point in time and space. Currently, wide attention among researchers is being paid to automated/semi-automated schemes for video shot detection and characterization.

Video indexing can be performed efficiently by means of caption text extraction and recognition. This has led to some remarkable progress, for example, in enabling the automatic conversion of hard-copy documents using optical character recognition (OCR) technology(2,3,4,5), and in simulating languages using voice recognition (VR) technology(6,7). In both cases, as exemplified above, outputs are created in ASCII text format for which indexing can be performed by conventional, not-so-perfect information retrieval techniques. From this, we can see that video sources containing a mine of information (i.e., news, ads, movies, and sports events) contain important contents in the form of voice, caption texts, and/or texts in images.

A video summary helps the user who desires to watch a particular video determine whether it is well worth watching. The video summary is classified into two types: video summary sequence and highlights video. A video summary sequence is suited to documentaries given that it provides an important overview of the whole original video. On the contrary, a highlights video highlight is suitable for video previews or sports highlights because it only contains video segments worthy of interest(8).

In this paper, we propose a scheme for generating soccer video highlights, using the structural features of captions in which appearance features in terms of time, as well as features such as spatial locations, areas, and colors are used. Such structural features of captions are used to extract caption frame blocks and caption keyframes from input frames. The caption keyframes are indexed both physically and logically. Soccer video highlights are created through a predefined rule for creating highlights. In addition, efficient retrieval and browsing techniques are provided as a means of allowing for selecting and watching videos in a fast and convenient manner.

This paper consists of the following sections. Relevant researches are described in Section 2, and structural features of captions in Section 3, extraction of caption frame intervals and caption keyframes in Section 4, creation of soccer video highlights in Section 5, experiment results in Section 6, and finally, thesis conclusions in Section 7.

II. Relevant Researches

Studies on the field of video summarization have been continuously conducted by a number of researchers. A video summary is classified into two types: video summary sequence and video highlight. Schemes for creating video summary sequences include video skimming(9), shot change graph(10), cluster validity analysis(11), and video manga(12). In addition, schemes for video highlights include movie trailer(13,14) and event-based sports summarization(15).

Christal et al.(9) proposed video skimming enabling the summarization of documentaries or news broadcasts. In this scheme, videos and their copies are assigned in order of words used, and important words in the copies are identified through language analysis. As a result, video clips are selected in order of the prioritized words. Yeung et al.(10)
proposed the shot-based change graph featuring the use of story streams. Hanjalic et al. (11) extracted keyframes and set video shots, and then created video summary sequences containing the keyframes through the use of cluster validity analysis. Uchihashi et al. (12) proposed video Manga featuring comic cartoon styles as a scheme for generating video summaries. They used video Manga to measure importance based on the scarcity and persistency of video segments.

Another type of video summarization is highlights extraction. Lienhart et al. (13) and Pfeiffer et al. (14) proposed a scheme for automatically creating movie trailers by searching low-level visual/audio features, motion information, and color information. In this paper, the empirical features of the basic physical parameters for digital videos were used as a means of selecting important objects, persons, actions, dialogues, title texts, and title music clips. Babauchi et al. (15) proposed a sports video summarization scheme using event-based video indexing. Despite its usefulness in the summarization of videos, this scheme loses a considerable number of important features that represent meanings.

III. Structural Features of Captions

3.1 Caption Region Analysis

Captions appearing in a soccer video are an important element that allows the user to see how the soccer game develops. The caption region of a soccer video differs in the following features from that of a drama or documentary:

1) Position: The positions of different caption regions are fixed by shape.
2) Area: Each caption region has its own area.
3) Existence: (Once an event occurs), every caption region immediately appears, exists for a while, and disappears.
4) Change: The position of texts changing within a caption region is fixed.
5) Color: Each caption region has its own color.
6) Point in Time of Appearance: The caption region of an event appears directly following occurrence of the event.
7) Order of Appearance: The order of appearance of a caption can differ depending on each individual game.

Based on the caption features described above, caption keyframes are extracted and video segments are indexed. This process plays an important role in creating soccer video highlights.

3.2 Caption Scene Analysis

In a soccer video, a limited number of captions occur. In this paper, captions are divided by semantics of video contents into:

1) Team(Cteam): Both team names (Each regional/ team name)
2) Stadium(Cplc): Stadium name (Stadium and its region)
3) Broadcast Booth(Cbo): Sportscaster names (Announcer and commentator)
4) Referee(Cref): Referee names (Chief referee and sub referee)
5) Player List(Clst): Player list of both teams;
6) Game Begin(Cbg): Beginning of the game (First/second half of the game and team names)
7) Scoreboard(Csco): Game scores (First/second half of the game, time frame, team names, and scores)
8) Bench(Cbch): Benches for both teams (Head coach/assistant coach/bench player names)
9) Player(Cpl): Names of players who had shoot, assists and free kicks, and who committed general fouls;
10) Player exchange(Cchg): Names of players who are exchanged
11) Goal(Cgo): Names of players who scored goals:
12) Foul(Cfol): Names of players who were given red cards such as warning or exit for committing critical fouls.
13) Game End(Cend): Termination of the game (Names of and scores of both teams).

In addition to the caption shots described above, captions with little relevance to the current game (i.e., sports, game type, tournament, team ranking, weather, other stadiums, and records) occur in a soccer video. In this thesis, no consideration was given to the categorization and extraction of such additional captions.

3.3 Caption Features in Terms of Time

As shown in (Table 1), captions are classified by the order of appearance into start caption, event caption, and end caption. In general, those captions have some structures in terms of time as shown in (Fig. 1).

<table>
<thead>
<tr>
<th>Categorization as a Function of Time</th>
<th>Caption Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Caption</td>
<td>Ctem, Cplc, Cbct, Cref, Clst, Cbgn</td>
</tr>
<tr>
<td>Event Caption</td>
<td>Csoo, Cbch, Cplr, Cchg, Cgol, Cfol</td>
</tr>
<tr>
<td>End Caption</td>
<td>Cend</td>
</tr>
</tbody>
</table>

As illustrated in (Fig 1), event captions representing events include Csoo, Cbch, Cplr, Cchg, Cgol, and Cfol. There are some differences in the order of appearance of the event captions. Some event captions appear directly following occurrence of a particular event while the soccer game goes on. The caption Cend indicating the termination of the game displays the result of the game.

〈Table 2〉 shows the appearance features of captions.

<table>
<thead>
<tr>
<th>Captions</th>
<th>Caption Configuration</th>
<th>Appearance Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Captions</td>
<td>Ctem, Cplc, Cbct, Cref, Clst, Cbgn</td>
<td>Convey an overview of the game; Appear independently.</td>
</tr>
<tr>
<td>Event Captions</td>
<td>Csoo</td>
<td>Appear in a short-/long-term manner throughout the soccer game; Appear at any time; Appear independently or in an overlapping manner with other captions.</td>
</tr>
<tr>
<td>End Captions</td>
<td>Cend</td>
<td>Conveys the result of the game; Broadcast terminated.</td>
</tr>
</tbody>
</table>
3.4 Caption Features in Terms of Space

The spatial caption structure is the information of the position region in a soccer video where captions appear. As illustrated in Fig 2, the captions appear in a limited number of regions 6.

As shown in Fig 2, the caption regions such as Cso, C1st, Cbgm, Cend, and Cref overlap with each other. However, they have their own area. The captions such as Ctem, Cplc, Cbct, Cbch, Cptr, Cgol, Cflo, and Cchg differ in size little from each other, and appear in the middle of the lower part of the frame.

\[ |x_h - x| < T_x \quad AND \quad |y_h - y| < T_y \]  

\[ \text{Eq. (1)} \]

In this thesis, caption keyframes are extracted from 13 caption scenes, based on the features of caption regions. The keyframes are indexed to search video segments for different subjects. Therefore, the extraction and indexing of caption regions plays an important role in developing a soccer video database.

IV. Extraction of Caption Frame Intervals and Keyframes

4.1 Extraction of Caption Frame Intervals

Every caption in a soccer video has its own position, area, and color value. Based on such features, caption frame intervals are extracted by means of similarity measurement.

Caption frame intervals are extracted by performing a comparison between the structural features predefined for caption positions, areas and colors, and the features of input frames. If any similarity between those features exists, the input frames become caption frame candidates. Such similarity measurement is continuously performed across the whole frame. The successive frames meeting similarity requirements are specified as caption frame blocks. The method for similarity measurement is addressed as follows:

4.1.1 Similarity Measurement of Caption Positions

\[ \frac{A_h}{A_s} > T_d \quad AND \quad \frac{A_s}{A_s} > T_d \]

\[ \text{Eq. (2)} \]

In Equation (1), \( x_i \) and \( y_i \) are coordinate values for input frames; \( x_t \) and \( y_t \) are reference values obtained by means of structural features; and \( T_x \) and \( T_y \) are critical values for positions.

4.1.2 Similarity Measurement of Caption Areas
In Equation (2), Ami refers to the area of the caption region overlapping between input frames and reference frames. Afi and At refer to the area of caption regions for input frames and the area of caption regions for reference frames, respectively. In addition, TA refers to the critical value for areas.

4.1.3 Similarity Measurement of Caption Colors

\[
C^R_{fi} - C^R_i < T_C \quad AND \\
C^G_{fi} - C^G_i < T_C \quad AND \\
C^B_{fi} - C^B_i < T_C \\
\text{Eq. (3)}
\]

In Equation (3), \(C^R_{fi}\) and \(C^R_i\) refer to the maximum number of RGB pixels in the caption region for input frames and the maximum number of RGB pixels in the caption region for reference frames, respectively. TC refers to the critical value for colors.

4.2 Extraction of Caption Keyframes

Caption frame blocks extracted indicate important events. However, all frames aren’t necessarily keyframes. As such, one typical frame representing the whole caption frame block becomes the caption keyframe.

As illustrated in (Fig. 3), the first of the caption frame blocks becomes the keyframe. The caption keyframe plays an important role in performing video retrieval and browsing, and in creating highlights.

![Caption Keyframe Extraction](image)

V. Highlight Creation

5.1 Scene Reset

5.1.1 Necessity of Resetting Scenes

Event captions like Cpl, Cch, Cgo, and Cft appear directly following occurrence of an event. Therefore, real events exist in the preceding segments of the caption keyframe. Scene reset refers to the process of resetting event scenes in such a way as to allow scenes to have real events.

In the soccer video, event captions appear immediately following occurrence of an event, with the exception of the start/end captions and intermittent captions appearing on the scoreboard. As described in (Table 3), scene reset is implemented for the scenes that need to be reset.

<table>
<thead>
<tr>
<th>Category</th>
<th>Relevant Scenes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenes that need no resetting</td>
<td>Cmr, Cplc, Cbct, Cref, Cst, Cbn, Csoo, Cbch, Cend</td>
</tr>
<tr>
<td>Scenes that need to be reset</td>
<td>Cpl, Cch, Cgo, Cft</td>
</tr>
</tbody>
</table>

In (Table 3), the scenes that need no resetting are left intact with no scene reset.

5.1.2 Scene Reset Technique

The scenes that need to be reset must include caption frame intervals and preceding events. (Fig 4) shows how to reset the scenes that need to be reset. The caption frame intervals include all frames containing their captions. The number of preceding events should ample enough to represent important events.
Generally, the amount of time required to represent the preceding event is between 10 seconds and 20 seconds. Hence, in this paper, we set the amount of time required for the preceding event at 20 seconds. Accordingly, the preceding event has a total 600 frames based on the standard 30 frames per second. As a result, a caption scene comprises a preceding event (600 frames/20 seconds) and a caption frame interval.

5.2 Highlight Creation

Prior to creating highlights, the number of highlight-based scenes must be determined. The most important event in the soccer video is a goal. Therefore, the number of highlight-based scenes needs to be determined in a flexible manner with the number of goals taken into account. (Fig 5) shows a flow diagram of highlight generation and the rule for creating highlights is described below:

1) A highlight consists of the following scenes:
   Ctem, Cgnd, Cbct, Cref, Clst, Cbgn, Cplr, Cchg, Cgal, Cfol, and Cend. The more detailed rule for creating highlights is as follows:

A) Start Scene
   (1) The initial scene consists of Ctem.
   (2) The next scene consists of Cgnd-Cbct or Cbct-Cgnd in order of the original video.
   (3) The next scene consists of Cref-Clst or Clst-Cref in order of the original video.
   (4) The final scene consists of Cbgn.

B) End Scene
   (1) The end scene consists of Cend.

C) Event Scene
   Event scenes are included in the highlight in order of prioritized next events. The process of comparing between the total number of highlight scenes created for each step and the number of highlight-based scenes is implemented until the two values are equal or the final event is included in the highlight.
   (1) The initially appearing event consists of Cgals in order of the original video.
   (2) The second event consists of Cplrs in order of the original video.
   (3) The third event consists of Cfol in order of the original video.
   (4) The finally appearing event consists of Cchgs in order of the original video.

2) If the total number of highlight scenes created so far is less than that of reference scenes, the final highlight video becomes the current highlight.
During the 2004 Korean professional football league (K-League), an average of zero-to-six goals were scored per game. In that case, a highlight has 0 or up to 6 goals. Other scenes are included in the highlight according to the rule for creating highlights.

5.3 Retrieval and Browsing

An indexed video and a highlights video need to be browsed by the user who wants to search video segments. As shown in (Fig 6), the video browser has retrieval and VCR functions.

![Browser diagram](image)

**그림 6. 브라우저의 가능**
**Fig. 6 Browser Functions.**

It is desirable that the user should implement the search function. As shown in (Fig 6), the user can select the date, stadium, and event to search his/her desired video segments. Search results are then displayed on the screen. In addition, the VCR function enables selected video segments to be played, stopped, and moved.

![Browser UI](image)

**그림 7. 자막 키 프레임 추출 과정의 화면**
**Fig. 7. The Screen Showing a Caption Keyframe Extraction Process.**

Extracted caption keyframes are important elements for creating soccer video highlights. (Table 4) shows the result of extracting caption keyframes.

<table>
<thead>
<tr>
<th>Video</th>
<th>Number of Caption Keyframes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video A</td>
<td>62</td>
</tr>
<tr>
<td>Video B</td>
<td>54</td>
</tr>
<tr>
<td>Video C</td>
<td>67</td>
</tr>
<tr>
<td>Video D</td>
<td>58</td>
</tr>
</tbody>
</table>

**Table 4 The Result of Extracting Caption Keyframes.**

VI. Highlight Creation

For the purpose of this paper, we used Visual C++ 6.0 in Pentium 1.3Ghz and Windows 2000.
(Fig 8) shows the process of creating highlights in the soccer video according to the rule for creating highlights. On the upper part of the screen are a video player, simple operational buttons and a frame indicator. On the lower part of the screen are caption keyframes included in the highlight.

The number of highlight-based scenes was set at 20 for creating highlights. Accordingly, one highlights video consists of 20 caption scenes. The average amount of time for an input video to be played is 58.3 minutes, and the average amount of time for a highlights video to be played is 6.7 minutes. Therefore, a video watcher can understand the whole contents of a video within 6.7 minutes by watching a highlighted version of the video.

If a video watcher selects his/her desired date or duration, a list of events satisfying such conditions appear. At that time, the video watcher can select his/her desired event, and narrow his/her search further. In addition, when a list of keyframes for selected events appears, the video watcher can select a keyframe to watch a selected shot. (Fig 9) shows video browser.

(Fig 9) Video Browser.

(7able 5) shows a comparison between the proposed scheme and existing video summarization schemes. Through the use of the structural features of captions, the
Table 5. Comparison among Video Summarization Schemes.

<table>
<thead>
<tr>
<th>Schemes</th>
<th>Features</th>
<th>Strengths</th>
<th>Problems/Difficulties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summarization Sequence</td>
<td>Video Skimming</td>
<td>-Color/captions</td>
<td>-Language analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Audio keywords</td>
<td>-Object/caption recognition</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Keyword frequency</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Scene Change Graph</td>
<td>-Story streams</td>
<td>-Setting time windows</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Time windows</td>
<td>-Overlapping stories</td>
</tr>
<tr>
<td></td>
<td>Cluster Validity Analysis</td>
<td>-Clustering</td>
<td>-Cluster configuration</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Cluster distance</td>
<td>-Criteria for validity</td>
</tr>
<tr>
<td></td>
<td>Video Manga</td>
<td>-Comic cartoon style</td>
<td>-High overhead costs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Segment scarcity</td>
<td>-Setting a matching criteria</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Segment persistency</td>
<td>-Lack of accurateness in the</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>extraction of keyframes</td>
</tr>
<tr>
<td>Highlights</td>
<td>Movie Trailer</td>
<td>-Low-level physical features</td>
<td>-Experience accumulation and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Empirical parameters</td>
<td>its application</td>
</tr>
<tr>
<td></td>
<td>Event-based Sports Highlights</td>
<td>-Event-based indexing</td>
<td>-Semantic matching</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Closed captions</td>
<td>-Changes in experience</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Time windows</td>
<td>-Event extraction</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Difference in point in time</td>
</tr>
<tr>
<td>Proposed Scheme</td>
<td></td>
<td>-Event-based/caption-based</td>
<td>-Limited domain</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Structural features (in terms of</td>
<td>-Caption focus</td>
</tr>
<tr>
<td></td>
<td></td>
<td>time and space)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Rules for creating highlights</td>
<td></td>
</tr>
</tbody>
</table>

proposed scheme creates highlights in a more efficient manner than other schemes do.

VII. Conclusions

In this paper, we proposed a scheme for creating soccer video highlights in an efficient manner through the use of the structural features of captions in terms of time and space. The structural features of captions were extracted by performing an analysis of caption information, and caption frame intervals and caption keyframes extracted through the use of those features. Also, we used an efficient rule for creating highlights so that video watchers can their desired video in a fast and convenient manner. As a result of performing experiments, it was observed that while the average amount of time for an input video to be played was 58.3 minutes, the average amount of time for a highlights video to be played was 6.7 minutes. This allows video watchers to understand the whole contents of an original video within 6.7 minutes, thereby enabling the video watchers to determine whether the original video is worth watching. In addition, the proposed scheme allows for selecting from a diverse range of videos, as well as enabling reduction in time and costs. The video browser described in this paper has been designed for efficient and fast retrieval. This paper provides
a basis for the implementation of enhanced sports video management systems.

It is believed that future research needs to be focused on enabling automatic extraction of active captions and on extracting semantic-based events. Also, an automatic indexing of video segments is considered to be achieved.

References


