SOCCER REPLAY DETECTION USING SCENE TRANSITION STRUCTURE ANALYSIS

Jinjun Wang[2,1]*, Engsiong Chng*[2]*, Changsheng Xu*[1]

[1] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{stuwj2, xucs}@i2r.a-star.edu.sg
[2] CeMNet, SCE, Nanyang Technological University, Singapore 639798
jjwang@pmail.ntu.edu.sg, aseschng@ntu.edu.sg

ABSTRACT

The replay scene detection is an useful technique for content based sports video analysis. Most current researchers try to find suitable visual and/or compressed domain features to detect the replay scene from a broadcast video. In this paper, we present a novel approach using context information from the concurrence of replay and other types of shots to detect the replay scenes. We first perform a shot classification and then a scene transition structure analysis on the generated shot label sequence to extract the replay scene. The proposed model is computationally fast and some promising results were obtained.

1. INTRODUCTION

In recent years there has been increasing research attentions in sports domain [1]. Researchers are focusing on identifying interesting sports events based on multimodal analysis techniques to assist automatic video summarization, skimming, highlight generation, etc tasks. In previous reported work, the videos examined are usually the broadcast sports videos and not raw unedited video [2]. This is because broadcast sports video contains rich post-production information, e.g. appearing of replay, transition of different shots, superimposed caption, etc, that could be utilized for semantic analysis, and promising results were reported by [1, 3].

In this paper, we examined replay scene detection of broadcast soccer video. A broadcast soccer video is typically intertwined with different shots such as close-up view, far view and replay. Of all the shot types, the replay scene is an excellent locator of semantically important segments as the broadcast director would normally launch a replay for interesting events. Hence the replay detection can greatly facilitate the semantic analysis of soccer video. Literature has reported some techniques to detect the replay scene. To give a few examples, Pan *et al* [4] proposed a replay detection method by the detection of editing effect, e.g. flying logos, before and after the replay segments; Kobla *et al* [5] used the macroblock types information in B-frames, along with vector flow information and the number of bits used for encoding each frame, to detect slow-motion replay. Pan *et al* [6] applied the zero crossing measure to evaluate the amplitude of the fluctuations in the frame differences within a sliding window to detect the presence of slow-motion. These replay detection techniques can be categorized into two classes: 1) detecting the editing effect, and 2) detecting the slow-motion, which is technologically more generic. However, as not all replay segments are intertwined with editing effect or displayed at a much slower than real-time rate, the current replay detection accuracy remain unsatisfactory. Lastly, even for replay scene which meets the assumptions of the above techniques, the high computation-cost hampers the development of real-time analysis systems. These reasons have made replay scene detection a difficult problem.

In this paper, we proposed a novel replay scene detection method based on the analysis of the scene transition structure in the broadcast sports video. The rest of the paper is organized as following: Section 2 describes our proposed method and the required algorithms; Section 3 gives our experimental results and Section 4 draws the conclusion and raised some future work.

2. OUR APPROACH

2.1. Method

As introduced in Section 1, the current editing effect detection and slow-motion detection techniques have their limitations. In our approach, we utilize the context information to identify the replay scenes.

The context information comes from the scene change during the transition of different shots. As introduced in section 1, a video consists of consecutive shots, and each shot is associated with a different type of scene. These shots give the broadcast soccer video a well defined structure and by identifying such temporal scene pattern, the appearance of replay can be detected. Our proposed method

0-7803-8874-7/05/$20.00 ©2005 IEEE

adopts a two stage approach, as illustrated in Fig.1: In the first stage, shot boundary detection and frame classification are performed and the shots are classified into three classes, namely Far view(Fig.2a), Medium view (Fig.2b) and Close-up view (Fig.2c). The obtained shot label sequence is then sent to the second stage where the scene transition structure containing the replay shots are detected using string matching algorithm, and the replay shots are additionally extracted.

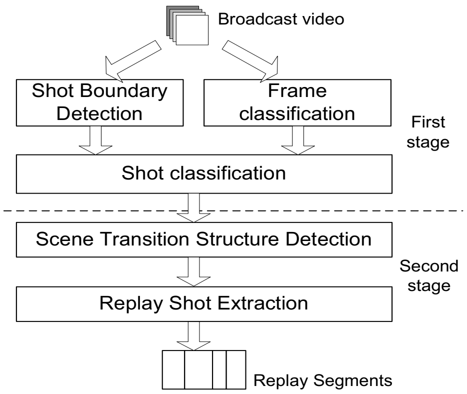


Fig. 1. Diagram of structure analysis

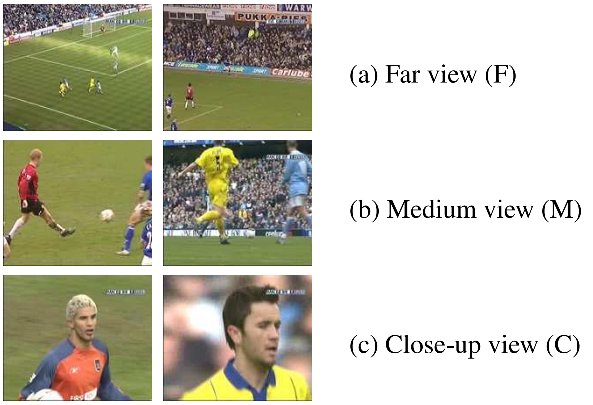
We make use of visual information from soccer video to obtain and analyze the shot transition structure. As the low-level operations to obtain visual frames are straightforward, we proceed immediately to present the analysis parts, starting from the first stage of the model.

2.2. Implementation of Stage 1

2.2.1. *Shot boundary detection*

In broadcast video the shot can be regarded as a basic analysis unit [7]. In our model the first stage output is a sequence of shot classification label from the original broadcast video, which requires firstly the shot boundaries to be detected .

We use a commercial software M2-Edit Pro [8] to do this shot boundary detection task. By experimentally setting the only related threshold in the software, namely “Sensitivity”, around 80% shot boundary could be automatically detected. We then manually refine the result by deleting any obviously unnecessary boundaries to combine the neighboring shots.



(a) Far view (F)

(b) Medium view (M)

(c) Close-up view (C)

Fig. 2. Shot type examples

2.2.2. *Frame classification*

In addition, frame classification is also performed in this stage. Three classes of possible view type labels is given to each frame for our model, specifically: Far view (F) (Fig.2a), Medium view (M) (Fig.2b), or Close-up view (C) (Fig.2c).

Figure 3 summarizes the algorithms to classify each frame into respective view type.

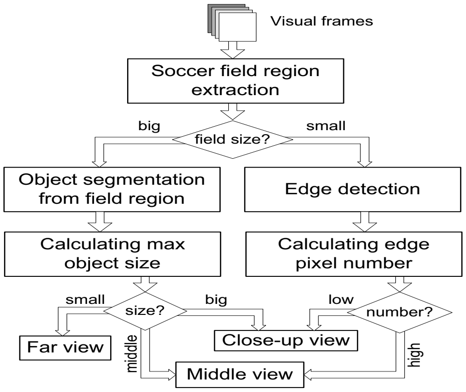

Visual frames
Soccer field region extraction
big
field size?
small
Object segmentation from field region
Edge detection
Calculating max object size
Calculating edge pixel number
small
size?
big
low
number?
Far view
middle
Close-up view
high
Middle view

Fig. 3. Frame classification algorithms

2.2.3. *Shot classification*

The shot classification is then performed using a simple weighted majority voting of frames identified for a single shot. After this processing, the output is a sequence of shots labeled as Far, Medium or Close-up shot. This sequence is processed by the next stage to detect replay scenes.

2.3. Implementation of Stage 2

2.3.1. *Scene Transition Structure Detection*

From our analysis, we observed that a replay scene has some definitive scene transitions. E.g, in soccer game replays, the Close-up views are often launched before and after each replay. Hence, if the scene transition structures that are related with replay can be identified, it could be used to search for replay scenes, and such transition structures can be detected using string matching algorithms.

In this paper, Dynamic Programming Algorithm (DPA) [9] is applied to detect any matched segments from the shot label sequence with desired scene transition structure. The DPA algorithm works in the following manner: The predefined scene transition structure can be regarded as a substring $S = \{s_1, s_2, s_3, ..., s_i\}$, $i = 1...U$ and the shot label sequence as a string $L = \{l_1, l_2, l_3, ..., l_j\}$, $j = 1...V$ where s_i and l_j are the individual symbols of S and L, respectively. The DPA computes the Edit Distance (ED) [9] matrix $D = DPA(S, L)$, U by V in size. The component of D, known as the ED between symbol s_i and l_j, is denoted as $d_{i,j}$ where

for $i = 1...U, j = 1$ (the first column of D)

$$d_{i,j} = i \tag{1}$$

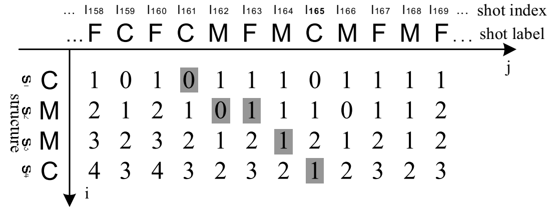
and for $i = 1...U, j = 2...V$

$$d_{i,j} = min \left\{ \begin{array}{c} d_{i-1,j-1} + matchcost(s_i, l_j) \\ d_{i-1,j} + c \\ d_{i,i-1} + c \end{array} \right\} \tag{2}$$

As $d_{i-1,j-1}$ will be out of the border of D when $i = 1$, we initialize a virtual row with all components to be 0 at $i = 0$ just for computational purpose. c is a constant and

$$matchcost(s_i, l_j) = \left\{ \begin{array}{ll} c & s_i \neq l_j \\ 0 & s_i = l_j \end{array} \right. \tag{3}$$

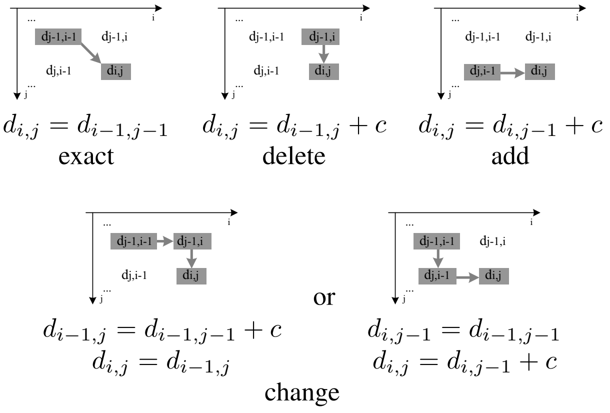
The last row of D, $d_{U,j}, j = 1...V$, records the total matching cost between S and L. The smaller the value of $d_{U,j}$, the smaller the difference between S and a segment of L ending at index position j. A zero value means an exact match. By setting a suitable threshold to $d_{U,j}$, the segments that match the predefined scene transition structure can be identified from the label sequence L. A threshold value that is greater than zero makes the system tolerant to a few unmatched symbols that either come from the shot classification error or from the variance in the actual video. To give an example, in Fig 4, a total matching cost that is below the threshold has been extracted at shot index position $j = 165$ ($d_{U,165} = 1$), and the matched label segment is "C-M-F-M-C" from $j = 161 \sim 165$. The related matching path is rendered in gray color. The substring to be extracted is "C-M-M-C" as the j-axis, and the i-axis shows a segment of the shot label sequence.



		...	l158	l159	l160	l161	l162	l163	l164	l165	l166	l167	l168	l169	... shot index
		...	F	C	F	C	M	F	M	C	M	F	M	F	... shot label
structure	s1 C		1	0	1	0	1	1	1	0	1	1	1	1	j
	s2 M		2	1	2	1	0	1	1	1	0	1	1	2	
	s3 M		3	2	3	2	1	2	1	2	1	2	1	2	
	s4 C		4	3	4	3	2	3	2	1	2	3	2	3	
	i														

Fig. 4. Example of Edit Distance matrix

This best matched label segment (i.e. "C-M-F-M-C" in the above example) can be traced by analyzing the spacial matching patterns between $d_{i,j}$ and its three possible ancestor points, either $d_{i-1,j}$, $d_{i,j-1}$ or $d_{i-1,j-1}$. In string matching tasks there are 4 possibilities of matches, namely "exact", "add", "delete" and "change", resulting in the following four patterns (Fig5):



$d_{i,j} = d_{i-1,j-1}$
exact

$d_{i,j} = d_{i-1,j} + c$
delete

$d_{i,j} = d_{i,j-1} + c$
add

$d_{i-1,j} = d_{i-1,j-1} + c$
$d_{i,j} = d_{i-1,j}$

or

$d_{i,j-1} = d_{i-1,j-1}$
$d_{i,j} = d_{i,j-1} + c$

change

Fig. 5. Matching patterns

By first examining the total matching cost and then tracing back using the above mentioned spacial patterns, the best matched segments from L can be extracted. Note that in this paper, the predefined scene transition structures contains not only replay shots but also non-replay shots. Hence, the next processing extract the precise replay scenes from the detected segments.

2.3.2. Replay Shot Extraction

The actual replay shots can be identified from the detected shot label segments using the following rules which are obtained from examination of our database:

1. *The close-up view shots are not replay scenes. The camera coverage in close-up view is too small to make close-up view shots suitable for replay;*
2. *The far view and medium view shots can be replay scenes, but any medium view shot prior to another far view shot is not replay scene;*
3. *Any far view or medium view shot can not be replay scene if it is too short, because the replay is a slow-motion display of a prior action, and the action always lasts a duration of time.*

Applying these rules to the example in Fig4, the replay shots at index position $j = 163$ and $j = 164$ are detected.

3. EXPERIMENT

In this experiment, three broadcast game videos (totally 2.5 hours) from England Premier League (EPL) were used for testing. After the first step processing, a shot classification accuracy of 94% (89 incorrect in totally 1328 shots) is achieved. In the second step processing, three structures are defined for all the three videos as below (the meanings of symbol “F”, “M” and “C” are explained in Fig2). Their respective mismatching thresholds are all set to 1:

- C-M-M-C
- C-C-F-M-C
- C-C-C-M-C

In the experiment it is observed that in the “Scene Transition Structure Detection” phase of the second stage, the detected segments from different predefined structures might overlap. For this case, the segments were merged, resulting in a longer segment. Then all the detected segments were sent for “Replay Shot Extraction”, and the final replay detection results are listed in Table1.

Table 1. Replay detection result

Game	correct	miss	false	recall	precision
Game 1	21	2	7	91.3%	75.0%
Game 2	18	5	2	78.3%	90.0%
Game 3	11	1	5	91.7%	68.8%

Game 1: Bolton vs Liverpool, first half
Game 2: Everton vs Manchester United, first half
Game 3: Manchester City-Birmingham, first half

There are some factors that caused the false or missed detection of replay scenes. One is the shot classification errors that come from either shot boundary detection error by M2-Edit Pro or the variance in game features. The shot classification errors result in incorrect shot type labels and thus affect later structure analysis. Another factor is the suitability of the defined structures. Some shot transition patterns related with replay are not distinguishing in the broadcast video, so the inclusion/exclusion of these patterns in structure definition will cause false/mised detection of replay scenes.

4. CONCLUSION AND FUTURE WORK

In this paper we introduced a novel replay detection method using scene transition structure analysis of broadcast soccer video. By shot classification and scene transition structure analysis, the proposed replay scene detection method is proved to be computationally fast and the performance is promising.

In the future, the accuracy of the model can be improved in the following ways: Firstly, the performance of shot classification could be improved to assist structure analysis. Secondly, as the scene transition structures actually reflect the temporal pattern in shot label sequence, other suitable statistical classifiers can be examined, e.g. Hidden Markov Model classifier. Thirdly, the experiment database should be enlarged to discover more appropriate transition structure definitions. And lastly, some current editing effect detection or slow-motion detection techniques can be added to improve the detection precision. These techniques are only carried on the detected shots instead of the whole video, thus the computational cost can remain low.

5. REFERENCES

[1] N Adami, R Leonardi, and P Migliorati, “An overview of multi-modal techniques for the characterization of sport programmes,” *Proc. of SPIE-VCIP’03*, pp. 1296–1306, July, 2003.

[2] Wang Jinjun, Changsheng Xu, Eng-Siong Chng, Kongwah Wan, and Qi Tian, “Automatic replay generation for soccer video broadcasting,” *To appear on ACM MM’2004*, Oct.10-16 2004.

[3] Linyu Duan, Min Xu, T. S. Chua, Qi Tian, and Changsheng Xu, “A mid-level representation framework for semantic sports video analysi,” *Proc. of ACM Multimedia’03*, pp. 33–44, Nov., 2003.

[4] H. Pan, B. Li, and M. I. Sezan, “Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transitions,” *in Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002.

[5] V. Kobla, D. DeMenthon, and D. Doermann, “Detection of slow-motion replay sequences for identifying sports videos,” *Proc. IEEE Workshop on Multimedia Signal Processing*, 1999.

[6] H. Pan, B. Li, and M. I. Sezan, “Detection of slow-motion replay segments in sports video for highlights generation,” *in Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.

[7] S. Smoliar and H. Zhang, “Content-based video indexing and retrieval,” *IEEE Multimedia*, vol. 1, pp. 62–72, 1994.

[8] USA MediaWare Solutions Pty Ltd, “M2-edit pro^{TM},” 2002.

[9] E. Ukkonen, “On approximate string matching,” *in Proc. of Int. Conf. on Foundations of Comp. Theory*, vol. Springer-Verlag, LNCS 158, pp. 487–495, 1983.