Replay Scene Classification in Soccer Video Using Web Broadcast Text

Jinhui Dai[1]*, Lingyu Duan[2,3], Xiaofeng Tong[1], Changsheng Xu[2], Qi Tian[2], Hanqing Lu[1], Jesse S. Jin[3]

[1]National Lab of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing China 100080
{jhdai,xftong,luhq}@nlpr.ia.ac.cn
[2]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{lingyu,xucs,tian}@i2r.a-star.edu.sg
[3]School of Design, Communication, and Information Technology, University of Newcastle, NSW 2308,
Australia, {jesse.jin}@newcastle.edu.au

Abstract

The automatic extraction of sports video highlights is a typical kind of personalized media production process. Many ways have been studied from the viewpoints of low-level audio/visual processing (e.g. detection of excited commentator speech), event detection (e.g. goal detection), etc. However, the subjectivity of highlights is an unavoidable bottleneck. The replay scene is an effective clue for highlights in broad- cast sports video due to the incorporation of video production knowledge. Most related work deals with the replay detection and/or a simple composition of all detected replays to generate highlights. Different from previous work, our work considers different flavors of different people in terms of highlight content or type through replay scenes classification. The main contributions include: 1) proposing a multi-modal (visual+textual) approach for refined replay classification; 2) employing the sources of Broadcast Web Text (BWT) to facilitate replay content analysis. An overall accuracy of 79.9% has been achieved on seven soccer matches over seven replay categories

1. Introduction

Replay is a reliable indicator of sports highlight due to the incorporation of video production knowledge. Different people have different flavors in terms of highlight content or type. Beyond replay detection and a simple composition of replay scenes, we focus on the refined classification of replay scenes. With refined replay categories, it is easy to output personalized highlights with different compositions.

There are many methods for highlights generation. Rui et al. [1] employed audio features to detect highlights in base-ball video. Peker et al. [2] utilized motion activity to generation highlights. But it is not easy to use low-level visual and audio feature to accurately locate the highlight segments. In [3], Assfalg et al. tried to extract soccer highlight with event detection. However, an event does not mean highlight as it does not directly reflect the subjective evaluation. Moreover it is difficult to determine the boundary of highlight segments even if an event can be accurately detected. Thus we resort to the replay scenes inserted into the broadcast sports video.

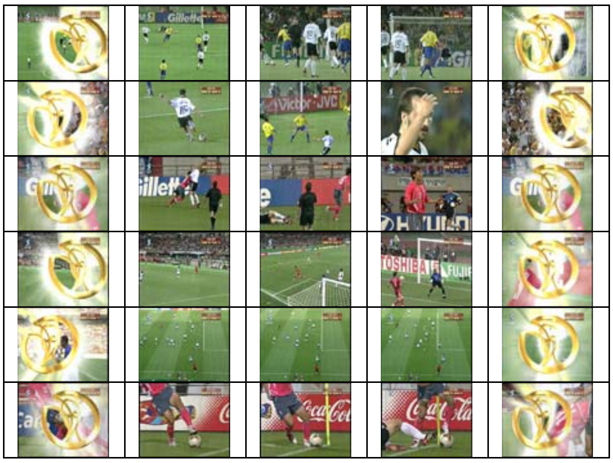


Figure 1: Sample images from different replay categories.(1st row: goal, 2nd row: shoot, 3rd row: foul, 4th row: attack, 5th row: offside, 6th row: out of bound)

The insertion of replay is based on an expert's understanding. It carries sufficient semantics for highlights. Pan et al [4] heuristically concatenated replay scenes to generate sports highlights. Ekin el al [5] heuristically linked sports summary to the presence of a replay scene. Li et al. [6] combined audio analysis and replay detection to locate the highlights. Neither did they perform further semantic analysis within replay scenes. Our work is focused on the replay classification. We want to automatically detect the replay categories to help deliver personalized highlights.

However, only visual processing cannot suffice for the refined replay classification (See Figure. 1) due to the semantic gap between low-level visual feature and high-level semantic concepts. We thus propose a multi-modal approach for re-fined replay classification. The WEB Broadcast Text (WBT) is employed to facilitate the semantic analysis of replay scenes. The textual information is becoming an important modality for semantic multimedia computing. Zhang et al [7] detected semantic events by recognizing superimposed caption. Babaguchi et al [8] attempted to combine

* This work was performed when the author was visiting Institute for Inforcomm Research, Singapore as an intern

0-7803-9332-5/05/$20.00 ©2005 IEEE

video, audio, and close-caption text to improve the reliability and efficiency of detecting semantic events in sports video. Xu et al. [9] fused audio-visual features and text information available in match reports and game logs to detect events. Different from their work, we utilize the source of WBT. WBT is a textual record of live commentary on sports game in Web. Section 3.2 gives a detailed description of WBT. WBT is widely available from web sites and can be acquired more easily than superimposed captions, close-captions and game logs. And WBT provides more details than match report.

In soccer video, we classify replay scenes into seven categories: 1) goal replay (GR), 2) shoot replay (SR), 3) attack replay (AR), 4) foul replay (FR), 5) offside replay (OR), 6) out of bound (OBR), and 7) others (OTR). Section 3 gives a description of these seven categories of replays. Six categories are illustrated in Figure 1. The overall framework is illustrated in Figure 2. After detecting replay scenes, we extract the visual features within a replay scene and its neighborhood shots. The textual features are extracted from WBT within an associated window. Our visual features consist of several visual concepts (e.g. referee, goal net, goal view) instead of low-level features. For the textual features, we try to spot a set of keywords (See Table 2) from WBT. We calculate the term frequencies of these keywords to construct textual features. Finally we combine the visual concepts and the term frequencies to classify a replay scene into one of 7 predefined categories. As shown in Figure 2, the replay categories can be used to deliver personalized highlights.

Our main contribution consists of 1) delicate replay scene classification and 2) the introduction of BT towards a multi-modal approach.

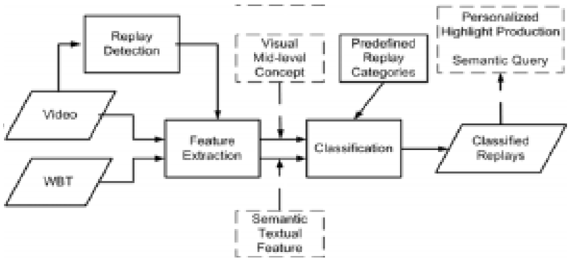


Figure2: The framework of our replay scene classification approach.

2. Replay scenes Detection

To detect replay scenes, we want to represent and identify the special digital video effects (SDVE) [10] inserted at the beginning and the end of a replay scene.

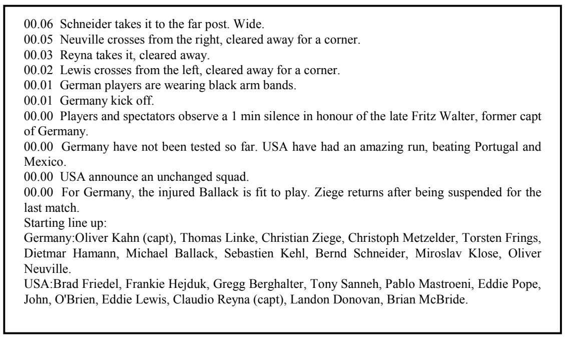
Figure3: One set of SDVE Indicating the replay scenes

As shown in Figure 3, the overlapped 'flying graphics' is a typical kind of SDVE. We choose a set of SDVE video segments as training data, and employ the spatial-temporal mode seeking to capture dominant colors that best describe the overall color appearance. And then we employ a sliding window to perform mode-based similarity matching between the resulting color modes and the segments within a window over the whole video data. A promising performance over four matches from 2002 FIFA World Cup, Recall 90~97% and Precision 85~95%, was reported in our previous work[10]

3. Replay scenes classification

As introduced above, we have predefined 7 replay categories: GR, SR, AR, FR, OR, OBR, and OTR in soccer video. GR is the replay of "goal". SR is the replay of "shooting". AR is the replay of exciting "offence" but without "shooting". FR is the replay of "foul" including severe "injure". OR and OBR are replays of disputable "off side" and "out of bound" respectively. OTR includes more than one category within one replay scene. The terms refer to the glossary [11]. Different subjects might concentrate on different categories in terms of sports highlights. Fans will prefer GR, SR and AR to others. However, coaches not only give different priority to GR, SR and AR, but also give FR and OR valuable considerations.

There are many textual resources of sports game widely in the web. They provide high level semantics and more details of a sports game. All these information could be leveraged to support the semantic analysis of sports video. As an important textual resource, WBT is available widely in many sports-websites (e.g. www.soccernet.com, www.rediff.com, etc.) and club-websites (e.g. www.inter.it, www.manutd.com, etc.). Figure4 gives an example of WBT.



00.06 Schneider takes it to the far post. Wide.
00.05 Neuville crosses from the right, cleared away for a corner.
00.03 Reyna takes it, cleared away.
00.02 Lewis crosses from the left, cleared away for a corner.
00.01 German players are wearing black arm bands.
00.01 Germany kick off.
00.00 Players and spectators observe a 1 min silence in honour of the late Fritz Walter, former capt of Germany.
00.00 Germany have not been tested so far. USA have had an amazing run, beating Portugal and Mexico.
00.00 USA announce an unchanged squad.
00.00 For Germany, the injured Ballack is fit to play. Ziege returns after being suspended for the last match.
Starting line up:
Germany:Oliver Kahn (capt), Thomas Linke, Christian Ziege, Christoph Metzelder, Torsten Frings, Dietmar Hamann, Michael Ballack, Sebastien Kehl, Bernd Schneider, Miroslav Klose, Oliver Neuville.
USA:Brad Friedel, Frankie Hejduk, Gregg Berghalter, Tony Sanneh, Pablo Mastroeni, Eddie Pope, John, O'Brien, Eddie Lewis, Claudio Reyna (capt), Landon Donovan, Brian McBride.

Figure4: An example of Web Broadcast Text

Using visual features only cannot suffice for our target. We combine textual semantic features from WBT and visual concept features to construct a 37-dimensional feature vector (add the number of sentence as a dimension of the vector) for training a classifier. Many machine learning algorithms can be used. In our experiment, the decision tree C4.5 is employed.

Textual retrieval is effective from the semantics level queries point of view. With the keywords in WBT associated with a replay, we may provide rich queries towards personalized highlights. For example, the fans may want to retrieve the highlights where their "super star" appears.

3.1 Visual concept feature extraction

Currently we develop three visual concepts, i.e. “goal view”, “referee”, and “goal net” for replay scenes classification. (See Figure 5)

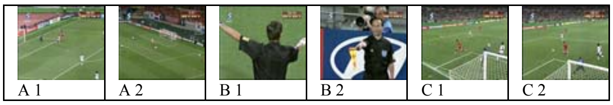


Figure 5: Sample images of visual concepts (A1, A2: “goal-view”; B1, B2: “referee”; C1 C2: “goal-net”)

The concept of “goal view” usually appears at the latter end of a “Field view” shot previous to the scenes of GR, SR and AR as illustrated in Table 1. The “goal view” is identified by the shape of homogeneous playfield [12]

The concept of “referee” lies in a close-up shot previous to the scenes of FR or OR as shown in Table 1. This concept is based on the observation that the referee’s clothing color is uniform and different from players’ clothing color for a match. We establish the color model of a referee to identify the “referee” by using a color matching method [13].

The concept of “goal-net” lies in the shots within the scenes of GR, SR and AR as illustrated in Table 1. A “goal net” view is recognized through combining the features of color and texture [13].

Table 1: The associations between visual concepts and replay scenes. √: this class contains this feature; ×: this class does not contain this feature; N.A: this feature is useless for this class.

Visual Concept	GR	SR	AR	FR	OR	OBR	OTR
“Goal view”	√	√	√	NA	NA	NA	NA
“Referee”	×	×	×	√	√	×	NA
“Goal-net”	√	√	√	×	×	×	NA

3.2 Textual semantic feature extraction

As shown in Figure 4, a WBT of a soccer game is usually composed of two parts: introduction and game action. In the introduction part, there is useful preliminary information about this game such as the line-up of two teams. We can use it to establish the lists of player names. In the game action, a complete description of the playing game is instantly produced in sentences and archived for later use. Each sentence can be represented as : <sentence time, sentence content>. The resolution of each sentence is roughly in one minute.

For each replay, we locate a window along time axis. Let T_{replay} denote the ending time of a replay. T_l, T_r denote the left boundary and the right boundary of the associated window respectively. For each replay scene, we select those sentences within the window $<T_{replay} - T_l, T_{replay} - T_r>$ for textual feature extraction. The term-frequency vector Xi of a replay is defined as:

$$X_i = [x_{1i}, x_{2i}, ..., x_{ji} ..., x_{mi}]^T$$

where x_{ji} denotes the frequency of the term $f_j \in W$ in the related sentences of the i th replay $W=\{f_1, f_2, f_3, ... f_m\}$, where W denotes the complete related vocabulary set of selected semantic event terms and their synonymous terms.

In WBT, there exist some special terms associated with a semantic event, such as shoot, off side and so on. For some events, there are some synonymous terms. For example, for the event of “shoot”, such words as “smash, shoot, shot, crack, head” have the same meaning as “shoot”. Table 2 illustrates the correlativity between selected terms and replay categories. There are some overlaps between some categories’ correlative terms

Table2: Selected semantic event terms and synonymous terms

Replay Category	Correlative terms
GR	G-O-A-L, goal
SR	shoot, smash, blast, chip, fire, crack, ranger, head, shot,
AR	corner, punch, collect, save, clear, cross, goalkeeper, box, goal,
FR	free kick, yellow card, book, red card, medical, injured, treatment, stretcher, go off, return, foul,
OR	off side
OBR	throw in, goal kick, corner
OTR	N.A.

4. Experiments

Our experimental dataset consists of seven full matches from 2002 FIFA World Cup as listed in Table 3.

Table 3: Table 3: Experimental Dataset

Match	Replay Number	Match	Replay Number
GER-BRA(30/6//02)	33	GER-KOR(25/6/02)	71
BRA-ENG(21/6/02)	32	KOR-TUR(29/6/02)	55
SEN-TUR(22/6/02)	46	GER-USA(21/6/02)	67
KOR-SPA(22/6/02)	69		

The WBT comes from www.rediff.com. Table 4 lists the statistics about replay categories. The SR, AR and FR are three major categories of replay scene. The GR and OR are two minor categories but have significant meanings for users

Table 4: Statistics about replay categories.

GR	SR	AR	FR	OR	OBR	OTR	Total
18	100	62	134	18	17	24	373
4.8%	26.8%	16.6%	35.9%	4.8%	4.5%	6.4%	100%

We employ the decision tree C4.5. The tool WeKa [14] is used. Table 5 lists the classification results using visual feature, textual features, and their combination. Figure 6 illustrates the performance differences to show the advantages of the combination of visual features and textual features. We exploit the overall F_measure to evaluate the performance of classification. F_measure is defined as:

$$F\_measure = \frac{2 * precision * recall}{precision + recall}$$

It can be seen from Table 5 that using the visual feature only, we can get more or less satisfactory results of two categories SR and FR. But they are unable to distinguish the replay scenes of GR and OR. However, they are meaningful for personalized highlight production.

Table 5: Replay classification result.(T: only textual features. V: only visual concepts. T&V: the combination of textual features and visual concepts) (P: Precision. R: Recall. F: F-measure)

		GR	SR	AR	FR	OR	OBR	OTR
V	P	0	0.589	0.417	0.725	0	0	0
	R	0	0.930	0.323	0.903	0	0	0
	F	0	0.721	0.364	0.804	0	0	0
T	P	0.667	0.672	0.719	0.630	0.643	0.750	0.600
	R	0.667	0.780	0.371	0.866	0.500	0.176	0.125
	F	0.667	0.722	0.489	0.730	0.563	0.286	0.207
T & V	P	0.857	0.772	0.776	0.839	0.750	0.692	0.615
	R	0.667	0.950	0.613	0.933	0.500	0.529	0.333
	F	0.750	0.852	0.685	0.883	0.600	0.600	0.432

From Table 5, we notice the performance of each category has been improved with the combination of visual concepts and textual features. Two major categories of "Shoot" and "foul" have achieved promising result. For FR, we can attempt to perform further classification on the replay scenes of "yellow card", "red card", etc. We can also notice prominent improvements for other minor categories. Figure 6 gives a clear comparison.

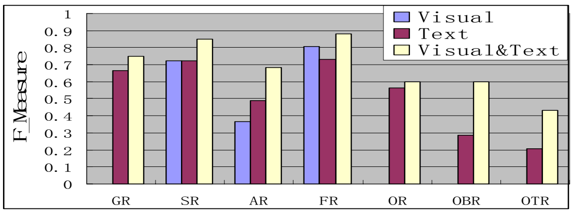


Figure 6: An illustration of performance comparisons in the cases of visual only, textual only, and visual plus textual.

Table 6: The confusion matrix of the classification with the combination of visual and textual.

classified_as>	GR	SR	AR	FR	OR	OBR	OTR
GR	12	4	0	1	0	0	1
SR	0	95	2	0	1	1	1
AR	1	19	38	1	0	1	2
FR	1	3	3	125	2	0	0
OR	0	0	7	0	10	0	1
OBR	0	1	3	4	1	8	0
OTR	0	1	11	3	0	1	8

Table 6 illustrates a confusion matrix of classification with the combination of visual and textual. OBR and OTR's classification results are not satisfied because usually they are not mentioned in WBT. Some replays of GR OR and AR are missed because there are no associated textual sentences in WBT. Especially for GR, some replay scenes are repeatly inserted at "calm" periods when there are no significant events. Thus those delayed GR replays are missed. The overlap of replays' associated windows is generally the source of the confusion of classification. Especially for AR and SR, some sentences related with SR fall in the AR' associated window and some replays of AR are falsely classified as SR.

In our experiments, we set $T_l = 0$ (refer to the window size in Section 3.2) due to the time-lag of textual sentences. Table 7 compares the results with different window size T_r. The best accuracy of 79.9% is achieved when we set $T_r = 1.5$ minutes.

Table 7: A comparison of results with different T_r

Tr(mins)	1.0	1.5	2.0	2.5
F_Measure	70.2	79.9	78.5	76.6

5. Conclusion

We have proposed a multi-modal approach to classify replay scenes in broadcast soccer videos. We have studied the feasibility of using a new textual source of WBT to facilitate video content analysis. As the replay scenes condense the sports highlights by video production techniques, we can use automatically classified replay categories to deliver a useful digest system of broadcast sports video. Moreover, rich keywords in WBT provide us a way to retrieve video segments while an alignment problem has to be further studied. Our future work includes the extension of replay classification to other sports games and the combination of visual, audio, and textual (WBT) to detect events.

6. References

1. Yong Rui, Anoop Gupta, Alex Acero: "Automatically extracting highlights for TV baseball programs", Proc ACM Multimedia, Oct 2000, Los Angeles USA,pp.105-115
2. KA Peker, R. Cabasson and A. Divakaran, "Rapid generation of sports highlights using the MPEG-7 motion activity descriptor," SPIE Conference on Storage and Retrieval from Media Databases, San Jose, CA, USA, January 2002
3. J. Assfalg, M. Bertini, A.Del Bimbo, W. Nunziati and P. Pala: "Soccer highlight detection and recognition using hmms" In: IEEE International Conference on Multimedia and Expo. (2002)
4. H.Pan, P. Van Beek, and M.I. Sezan: "Detection of slow-motion replay segments in sports video for highlights generation" In Proc. of ICASSP'01, 2001.
5. A. Ekin, A.M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization", IEEE Transactions on Image Processing 12(7): 796-807, 2003.
6. B.Li H.Pan, and I. Sezan, "A general framework for sports video summarization with its application to soccer", In Proc. of ICASSP'03, 2003.
7. Dongqing Zhang and Shih-Fu Chang: "Event detection in baseball video using superimposed caption recognition". ACM Multimedia 2002: 315-318
8. N.Babaguchi, Y.Kawai, and T.Kitahashi: "Event based indexing of broadcasted sports video by intermodal collaboration". IEEE Trans-action on Multimedia, Vol .4, No.1, March 2002
9. Huaxin Xu, Tat-Seng Chua: "The fusion of audio-visual features and external knowledge for event detection in team sports video", Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval, 2004.
10. L.-Y.Duan, et al, "Mean shift based video segment representation and applications to replay detection" In Proc. of ICASSP'04, 2004.
11.glossary:http://www.firstbasesports.com/soccer_glossary.html
12. L-Y. Duan, M. Xu, Qi, Tian, C.-S. Xu, J. S. Jin, "A unified framework for semantic shot classification in sports video" to appear on IEEE Transactions on Multimedia
13. Xiaofeng Tong, Lingyu Duan, Chengsheng Xu, Qi Tian, and Hanqing Lu, "Mid-level visual concept detection for semantics analy-sis in sports video", Submitted to ICME2005.
14. Weka3. http://www.cs.waikato.ac.nz/~ml/weka/