A Mid-level Visual Concept Generation Framework for Sports Analysis

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

[1]*National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China 100080*
{xftong, luhq} @ nlpr.ia.ac.cn
[2]*Institute for Infocomm Research, 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 development of mid-level concepts helps to bridge the gap between low-level feature and high-level semantics in video analysis. Most existing work combines the customized mid-level concepts and statistical models to detect particular events. Based on broadcast sports video production knowledge, we extend our previous work to present a unified framework for mid-level concept generation in this paper. A video segment is characterized via three essential aspects: camera shot size, an object appearing in a scene, and video production technology. These three aspects clearly summarize the primary concerns in terms of a generic concept generation. Within this framework, we can flexibly and clearly define meaningful mid-level concepts towards comprehensive video content analysis, such as replay classification and the detection of events (e.g. goal, shoot, attack, foul, offside, and out of bound, etc.).

1. Introduction

As a popular video genre, sports video has attracted much research attention. Towards semantic sports video analysis, a three-level framework [6] has been proposed and experimentally proved sound. At the low layer, low-level features are directly extracted from raw video data. In previous work, visual features [1, 2], audio features [3, 4], textual features [5] and multi-modality features [6, 7] have been extensively developed. The mid-level representations can be constructed based on those low-level features with clustering or classification methods. For the inference of high-level events, advanced statistical learning models have been investigated, such as Dynamic Bayesian network (DBN) [8, 9], hidden Markov model (HMM) [10], hierarchical HMM (HHMM) [11], etc. These methods have been proved to be effective in multivariate time series analysis. However, between efficient low-level feature extraction and effective high-level reasoning algorithms, a so-called semantic gap usually exists, which makes semantic analysis more difficult and less flexible. Intuitively, a mid-level layer may be indispensable to bridge this gap. Driven by sports domain knowledge and general broadcast sports video production knowledge, we may define a set of mid-level concepts. In this paper we present a mid-level visual concept generation framework for sports video analysis.

A mid-level concept can be considered as a descriptor computed from low-level feature. Its definition is usually motivated by high-level semantic inference. In visual domain, it can be semantic shot classes. Within commonly used three-layer event detection architecture, the implementations of mid-level concepts are the major modules at the mid-layer. There exist some mid-level concepts in previous work. Those concepts were usually customized and developed according to the visual/audio characteristics of predefined events. We have some general mid-level concepts, such as replay, field-view, player medium, player following, close-up, goal-view, excited audience, motion patterns, etc. These concepts were task-dependent, and do not clearly delineate how to extend them. That is, they are lack of a general viewpoint and unified explanation which restricted its ability in terms of generic and comprehensive semantics analysis.

In this paper, we focus on the mid-level visual concepts in field-ball sports video. The framework studies the generation of visual concepts from three aspects: camera shot size, an object appearing in the scene, and video production technology. It compactly characterizes a segment in most scenarios. With the various combinations of instances of the three factors, we can deduce many meaningful mid-level concepts, which have almost covered the shot descriptors in existing work. Comprehensive semantics analysis can be performed according to these concepts.

The remainder of this paper is organized as follows. Section 2 introduces the concept generation model. A concept detection method is discussed in Section 3. Applications and

* Part of this work was performed when the author was visiting the Institute for Infocomm Research as an intern.

0-7803-9332-5/05/$20.00 ©2005 IEEE

discussion are given in Section 4. Finally, we draw conclusions in Section 5.

2. Concept generation model

As introduced above, we characterize a segment in broadcast sports video from three aspects: camera shot size, an object appearing in the scene and video production technology. It is summarized according to the generation procedure of a concept in existing work. This framework indicates three significant considerations when one shot or a series of shots are produced to describe a scenario. The structure of our proposed model is illustrated in Figure 1.

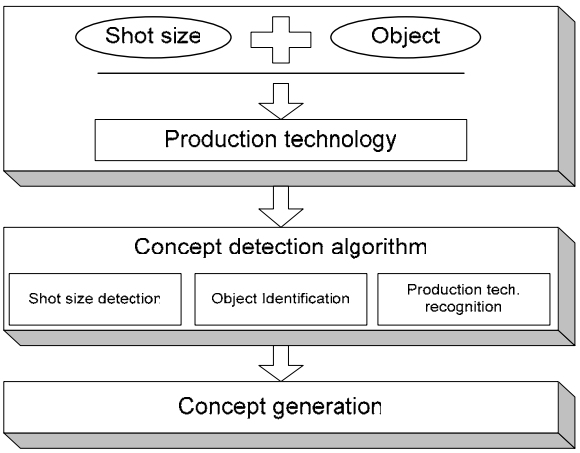

Figure 1. Concept generation architecture

The combination of the three factors is denoted by a three-dimension vector $c_i = <s_i, o_i, p_i>$, where s_i is the shot size, o_i is the object type, and p_i is the production way of the i^{th} segment. Through the combination of these variables, we can generate different concepts. Note that not any arbitrary combination can produce a meaningful descriptor.

In sports video, three types of shot size are usually used: long shot, medium shot and close-up. A long shot captures the global view of a scene. It is usually photographed with a wide angle lens. A medium shot has less view coverage than a long shot. It is zoomed in to a specific part of a view. A close-up shot gives the details of an even smaller part of a subject or view. Generally, an above-waist view of a person is captured in a close-up once the event of a shoot or a foul happens.

Different objects are concerned at different scenarios. The appearance of different objects is usually associated with different semantic meanings. For examples, in soccer video, a player close-up often appears after a shoot; a referee close-up is often present with the occurrence of a severe foul; an excited audience view is usually an indicator of an interesting shoot, etc. Therefore, further object identification is useful for subtle event detection.

Broadcasting technologies (particularly post-editing rules) are often utilized to make the program more attractive. We designate them as video production technology. It includes camera movement, replay scene, superimposed caption, the composition of shots, etc. For examples, camera pan or tilting are often used to track player. Camera switching is applied to capture a scene from different angles. A superimposed caption is used to display the score or player's information. As a significant video editing technique, a replay scene highlights an interesting or important segment once or several times with a slow motion pattern.

3. Concept detection method

Without loss of generality, we take soccer video as an example in this paper for it is the most popular and has been widely studied. The value sets of those three factors are shown in Figure 2. The set of camera shot sizes includes long shot, medium shot and close-up. For the object set, eight types of object are concerned: player (two teams), referee, goalkeeper (two teams), coach, audience, field, ball, and net. For the production technology, replay is particularly concerned.

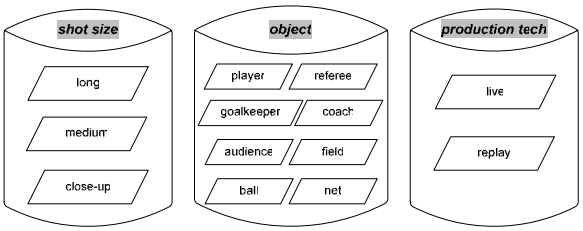

Figure 2. Value set of variable in the concept model

With difference instances of the feature vector $c_i = <s_i, o_i, p_i>$, we can deduce many useful mid-level concepts. Some examples are listed in Table 1.

TABLE 1. Mid-level concepts in soccer

No.	Concept	value of <si, oi, bi>
1	field-view	<long, field, live>
2	audience	<long, audience, live>
3	goal-net	<long, goal-net, live>
4	goal-view	<long, goalmouth, live>
5	medium-view	<medium, field, live>
6	player close-up	<close-up, player, live>
7	Referee	<close-up, referee, live>
8	Goalkeeper	<close-up, goalkeeper, live>
9	replay player close-up	<close-up, player, replay>
⋮	⋮	⋮

3.1. Shot size detection

Generally speaking, the detection of camera shot size depends on the estimation of the size of an object in a view. Accurate object segmentation is difficult. We thus roughly estimate the size of objects which is able to indicate the shot size in the view. The task of object size estimation can be reduced to two sub-problems:

(1) If the ratio of playing field is high, we can directly estimate the object size via in-field object extraction. The procedure consists of dominant color detection, playfield extrac-

tion and object segmentation [12]. Small object size indicates a field-view; larger size corresponds to a medium view; a close-up is declared by the largest object size derived by special head size detection (See Figure 3 (a), (b) and (c)).

(2) Otherwise, if most part of a view is non-field, the size of an object is estimated through texture measurement in the view. In texture computation, the energy, entropy and contrast of the gray-level co-occurrence matrix (GLCM) are considered [12] (Figure 3 (d), (e)).

long | medium in field | close-up | long | medium out field | close-up

Figure 3. Examples of shot sizes

3.2. Object identification

Eight kinds of objects are concerned in soccer video: a player, a referee, a goalkeeper, a coach, an audience, a field, a ball, a net. The objects of "a field" and "a player" are the most frequent ones. They have derived major shot classes such as field-view (i.e. long shot + field), and players in a medium shot and close-up. The objects of "a referee", "a goalkeeper", "a coach" and "a audience" often appears within the "break" segment of a game. A "goal-net" view usually appears when a wonderful shoot or a goal happens, which is captured by the camera at the back of the net. The ball is seldom shot individually. But in field-views, the ball trajectory is useful for event detection.

For "a field", the color feature is distinct. Generally speaking, the field color is just the dominant color in sports video. The dominant color can be extracted by an accumulated color histogram. We may use the dominant color to identify the object of "a field".

For "a player" (two teams), "a referee" and "a goalkeeper" (two teams), we firstly construct their color models and then use the models to identify the object type. The color models are built up as follows: (1) Robust face detection and field detection are carried out to get the close-up views with a field as the background. As illustrated in Figure 4, for these close-ups views, five classes' objects usually appear: the player of team A (player-A), the player of team B (player-B), referee, the goalkeeper of team A (goalkeeper-A), and the goalkeeper of team B (goalkeeper-B). Their percentages in the restrained close-ups for the game of 2002 World Cup Brazil vs. England are given in Figure 4. (2) The spatial relationship constraint within a face-body region [13] is used to locate the object body region. For each body region, mean shift based color characterization is performed to extract the color modes [14]. (3) k-means (k = 5 here) clustering is applied to construct the color models of these classes of object. The clusters are sorted in a descendent order according to the size of cluster. The first two big clusters correspond to two teams' players. The third one is "referee". The last two are goalkeepers. After clustering, each object's color is modeled as a GMM. With those five GMMs, a pixel can be categorized into one of six classes (the abovementioned five classes plus an undefined class) with a naïve Bayesian classifier.

After the pixel-level color classification, we employ morphological operations and region connection to generate a uniform object region. Then we determine whether an object of interesting type appears.

player-A 38.1% | *player-B* 47.8% | *referee* 11.0% | *goalkeeper-A* 0.90% | *goalkeeper-B* 2.2%

Figure 4. Five classes of close-ups and their proportions

The "ball" object is usually the focus of the main camera. Although the ball is small, its background is uniform and can be simply modeled and filtered. The detection and tracking is thus feasible [15]. Practically, the ball trajectory in field-views is useful for event detection [16]. The experiments have reported the accuracy of over 98% for the testing frames in a game.

"Goal-net" views are captured by a camera placed at the back of a net. A "goal-net" view is defined as the scene with field background and uniform texture characteristics which can be easily represented by edge histograms.

For "audience", we only concern the audience views captured by a long shot. It is often an indicator of an excited scene. The texture in the view has a high complexity.

"Coach" views often appear in the play segment. However, it is hard to distinguish them from other close-ups such as an audience close-up, etc.

Figure 5 illustrates the five classes of objects as discussed above.

field | ball | goal-net | audience | coach

Figure 5. Another five objects

3.3. Replay detection

A replay scene may consist of several shots. If a shot belongs to a replay scene, it is called as a replay shot. It is hard to discriminate between a replay shot and a live shot without any contextual information. Regardless how many shots appear in a replay scene, we treated a replay scene as a whole. Fortunately, the transition like "flying logo" often appears at the beginning and the end of a replay scene. The replay logo can be automatically detected by template matching. A context-based verification (i.e. inter-shot transition) can be done to recognize replay scene [17].

4. Applications and discussion

4.1. Applications

For shot size detection, we have applied a hierarchical scheme to perform shot classification [12]. The shots are

categorized into field-view, medium view, close-up or audience (long shot).

Field color is generally treated as the dominant color in a whole game. Not only it can be used to extract the region of playfield, but it is a useful cue to discriminate different sports genres. For object identification, "a referee", "a goal-view" and "a goal-net" have been combined to perform replay scene classification and event detection [18]. The classified replay scenes include: (1) goal replay; (2) shoot replay; (3) attack replay; (4) foul replay; (5) offside replay; (6) out of bound; and (7) others. As the replay scene is closely related to the highlights, we can not only detect highlights but also determine the type of highlight, such as goal, shoot, foul, attack, offside and out-of-bound, etc. The capturing of a ball is basically the focus of a broadcast ball game video. In [16], we have successfully detected and tracked the ball to perform trajectory-based semantics analysis. As the presence of an excited audience view is a cue for highlights, we used in [8] the objects of "an audience" and "a close-up" to detect "shoot".

A replay scene is a significant indicator for highlights. It has been widely utilized in lots of work on sports video analysis [6, 8].

4.2. Discussion

Mid-level concepts play a critical role in semantics sports video analysis. With the proposed concept model, we can generate visual mid-level descriptors. In broadcast video, a scene is represented via multi-modalities. Towards a comprehensive semantics analysis, the fusion of visual, audio and text cues should be considered.

5. Conclusions

In the paper, we have proposed a unified framework for mid-level visual concept generation from three fundamental factors. Those three factors delineate the procedure of a visual concept production. Within the framework, we can flexibly generate effective visual concepts that have almost covered the descriptors in existing sports video work. Comprehensive semantic analysis can be performed based on the visual concepts generated by this framework.

This framework can be applied to basketball, volleyball and tennis videos. Although the implementation may differ in different sports, the construction of mid-level concepts remains the same. Future work will be extended to the combination of the visual, audio, and textural cues to derive more complicated concepts.

6. Acknowledgement

This work is supported by National Science Foundation of China (NSFC) under Grant No. 60475010.

7. References

[1] J.Assfalg, M.Bertini, C.Colombo, A.Del Bimbo, W.Nunziati, "Semantic annotation of soccer videos: automatic highlights identification", CVIU, November 2003.

[2] B. Li, M.I. Sezan, "Event detection and summarization in American football broadcast video", Proc. SPIE conf. on Storage and Retrieval for Media Databases, vol. 4676, 2002, pp.202-213.

[3] M. Xu, N.C. Maddage, C.S. Xu, M. Kankanhalli, and Q. Tian, "Creating audio keywords for event detection in soccer video", ICME 2003, pp.281-284.

[4] Y.Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs", ACM Multimedia, 2000, pp.105-115.

[5] D. Zhang, S-F. Chang, "Event detection in baseball video using superimposed caption recognition", ACM Multimedia, 2002, pp.315-318.

[6] L.Duan, M. Xu, T. Chua, Q. Tian, and C..Xu, "A mid-level representation framework for semantic sports video analysis", ACM Multimedia 2003, pp.33-44.

[7] J. Kittler, K. Messer, W.J. Christmas, B. Levienaise-Obadia and D. Koubaroulis, "Generation of semantic cues for sports video annotation", ICIP 2001, pp.26-29.

[8] X. Tong, Q. Liu, H. Lu, "Semantic Units Based Events Detection in Soccer Videos", ICIP 2004.

[9] F. Wang, Y. Ma, H. Zhang, and J.Li, "Dynamic Bayesian network based event detection for soccer highlight extraction", ICIP 2004.

[10] H. Pan, P. Beek, MI Sezan, "Detection of slow-motion replay segments in sports video for highlights generation", ICASSP 2001, pp.1649-1652.

[11] L. Xie, S. Chang, A. Divakaran, and H. Sun, "Unsupervised discovery of multilevel statistical video structures using hierarchical hidden Markov models", ICME 2003, pp. 29-32.

[12] X. Tong, Q. Liu, H. Lu, H. Jin, "Shot Classification in Sports Video", ICSP 2004, China, pp.1364-1367

[13] L. Wang, B. Zeng, S. Lin, G. Xu, and H.Y. Shum, "Automatic extraction of semantic colors in sports video", ICASSP 2004, pp. 617-620.

[14] L. Duan, M. Xu, Q. Tian, C. Xu, "Nonparametric color characterization using mean shift", ACM Multimedia 2003, pp.243-246.

[15] X. Tong, H. Lu, Q. Liu, "An effective and fast soccer ball detection and tracking method", ICPR 2004, pp.795-798.

[16] X. Yu, C.Xu, H. Leong, Q.Tian, Q. Tang, K. Wan, "Trajectory-based ball detection and tracking with applications to semantics analysis of broadcast soccer video", ACM Multimedia 2003, pp. 11-20.

[17] X. Tong, H. Lu, Q. Liu, H. Jin, "Replay Detection in Broadcasting Sports Videos", ICIG 2004, Hong Kong, pp.337-340

[18] J. Dai, L.Duan, X. Tong, C. Xu, Q. Tian, H. Lu, "Highlights extraction in broadcast sports video based on replay classification using visual and textural information", accepted by ICME2005.