HIGH-LEVEL SOCCER INDEXING ON LOW-LEVEL FEATURE SPACE

Masaru Sugano[†], *Koichi Uemura*[‡], *Yasuyuki Nakajima*[†], *and Hiromasa Yanagihara*[†]

[†]KDDI R&D Laboratories Inc., Japan

[‡]Science University of Tokyo, Japan

ABSTRACT

In this paper, we propose an efficient scene clustering algorithm for soccer videos. Such scene clustering in sports videos is one of the semantic indexing techniques that can be applied to editing, summarizing, and content structuring. Our proposed method exploits a small set of audio and visual low level features that can be easily extracted from an MPEG compressed video, and classifies a soccer video into five predefined scene classes, from lower importance to higher importance. Based on very simple determination criteria, simulation results have shown that our method successfully performs semantic indexing for soccer videos at low computational cost.

1. INTRODUCTION

Developments of computer and network technologies and technology standards such as digital video compression enable us to enjoy more audio visual contents at low cost in public and personal environments. To achieve efficient search and retrieval of the vast amount of audio visual content, indexing techniques play an important role. Most previous researches on indexing techniques investigate *structural indexing*, such as shot boundary detection and silence detection. These are methods of determining non-hierarchical content structure, and a brief overview of the content can be obtained through the browsing applications. However, such structural indexing is not sufficient for advanced search and retrieval, such as object-based and event-based retrieval, or higher level handling such as filtering and summarization for browsing and editing purposes, for example. Towards advanced searches, *semantic indexing* such as genre classification [1] and highlight extraction [2] is necessary.

In this paper, we propose a scene clustering method in broadcasted soccer videos by using low level features such as color, motion, and audio. These features are easily extracted from MPEG compressed video. This kind of classification makes it possible to efficiently search and retrieve particular events and to hierarchically structure the content according to significance. Hierarchical structure is very useful especially for annotating and summarizing, which can be described by MPEG-7 *Hierarchical Summary DS* [3], for example. Previously we proposed a summarization method for baseball and news, which have clear context, based on a simple shot transition model derived from the recurrent shot detection [2]. For other sports such as soccer and basketball, a lack of canonical scenes makes it difficult to identify the event boundaries and furthermore, significant events are not always aligned with event boundaries. For these sports, scene clustering is the key for semantic structuring.

This paper is organized as follows. Section 2 briefly overviews the related works and our proposal. In Section 3, audio-visual features used for clustering scenes in soccer videos are explained. Section 4 describes proposed clustering methods for each class of scene. Finally, the experimental results are shown and discussed in Section 5.

2. SEMANTIC INDEXING FOR SOCCER VIDEO

For higher level indexing of a soccer video, a number of algorithms have been proposed, which are based on an excitement component related to motion, density of cuts, and sound energy [4], dominant color and referee/player detection [5], grass area ratio and grass orientation classification [6], playfield zone classification [7], and goal event detection using controlled Markov chains [8]. For other sports, scene clustering methods for basketball [9], baseball [10], tennis [11] and so forth, have been presented. Although they provide successful results to a certain extent, most require complex preprocessing and/or computationally intensive decision criteria.

Our method analyzes spatio-temporal characteristics of audio and video, and all the features used are directly obtained from MPEG compressed video stream with few additional calculations. Here only four low-level features and five simple criteria are used for clustering a soccer video into the following five scene classes: (i) Goal and shoot scene, (ii) Free kick scene, (iii) Corner kick scene, (iv) Offensive scene, and (v) Close-up scene of players-of-interest or a referee. In general, the semantic level illustrated in Fig. 1 is addressed to these classes, in terms of browsing, summarizing, editing, etc. Using motion and color features, the soccer video is at first roughly classified into three categories: *dynamic*, *static with large grass area*, and *static with small/no grass area*. Then finer clustering is performed to identify either significant class ((i), (ii), or (iii)) or insignificant class ((iv) or (v)).

0-7803-8554-3/04/$20.00 ©2004 IEEE.

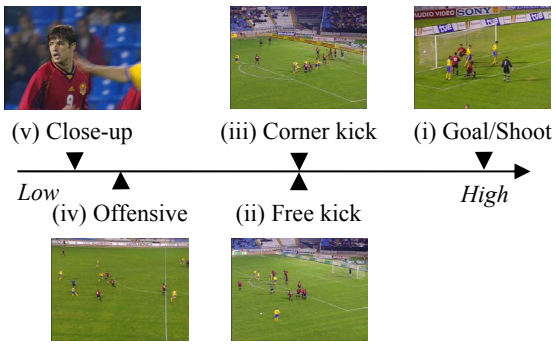

Fig. 1. Semantic level for soccer scenes

3. AUDIO VISUAL FEATURES

In our method, incoming MPEG compressed video is at first segmented into shots by applying a shot detection algorithm [12]. For each shot, several audio and visual features are extracted from the compressed audio/video stream, which are further described in this section.

3.1. Audio features

For audio features, sound characteristics such as applause and whistle which are specific to sporting events are analyzed. In sports videos, sudden, loud applause is often emitted from the audience in response to significant and interesting events. Therefore detecting applause sound in sports videos is very important for semantic indexing. For this purpose, we use audio subband energies which can be directly calculated from 32 subbands data of an MPEG audio stream [2]. Similar approaches are found based on a sound energy [4] and sound effect attention model [13], which requires additional processing. Here, we adopt subband energy analysis on compressed domain originally presented in [2] since our preliminary experiment has shown that this approach is applicable to the detection of applause and cheering in sports videos. In this paper, slight modification was further made according to [14], which indicates that subbands 2-7 well represent the commentator's excited speech in sports videos. Therefore the following Equation (1) is derived for calculating the subband energy for applause $SBE_{applause}$, where sb_k denotes the subband energy of the k-th subband (k=1-32).

$$SBE_{applause} = 0.1 \times sb_1 + 0.2 \times \sum_{k=2}^{7} sb_k + 0.7 \times \sum_{k=8}^{32} sb_k \qquad (1)$$

Our preliminary experiments have also shown that a referee's whistle sound can be obtained from 6th and 7th subband energies since they are relatively high among other subbands, as shown in Fig. 2, which shows a sample of subband energy distribution (sb_1 through sb_8, for one second) for whistle scenes. In Fig. 2, W1-W3 (light-grayed) indicate distributions for a whistle sound, while A1-A3 (dark-grayed) are those for loud applause only. We calculate subband audio energy for a whistle sound $SBE_{whistle}$ ($=sb_6+sb_7$) since a whistle generally indicates significant events, such as free kicks, infringements, etc.

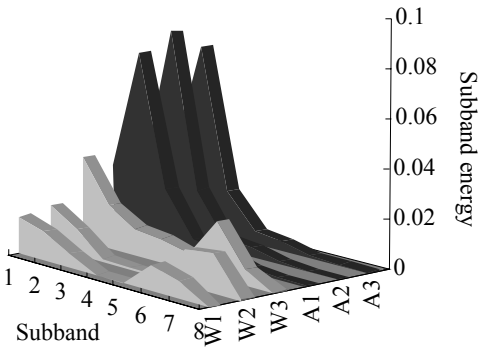

Fig. 2. Distribution of MPEG subband energies

3.2. Color feature

In global scenes, which capture a playfield largely, grass color becomes dominant as shown in Fig. 1. Many works exploit grass color detection for indexing sports videos [5][6][10]. Here, as a measure of color characteristics, we use the color histograms of Y, Cb, and Cr components of a downscaled frame (8×8) of the first frame in each shot. Although some approaches exploit DC images of I-frames [6] or 3×3 divided blocks for grass and/or sand colors detection [10], the reason why 8×8 scaled frames is used is that histogram calculations can be greatly simplified while maintaining sufficient accuracy for grass color detection. In addition, our histogram-based approach is robust to variations in grass color such as lighting conditions and weather since no predefined threshold for grass color values is specified.

3.3. Motion feature

For motion features, several works focus on camera motion analysis for scene clustering [9][10]. Our method focuses on that in close-up scenes, which capture players-of-interest or a referee, in which motion vectors become large, while in global scenes which capture playfield, motion vectors become small. This is because objects (e.g. players) are dominant in close-up scenes and the motion of these objects becomes relatively large even with slight motion. To estimate whether motion is large or small, we define shot-based motion activity MA as follows. At first, an absolute sum of all the motion vector amplitudes within a P-frame is calculated, and then it is divided by the number of macroblocks regarding motion vectors. The resulting value is accumulated through each shot, and finally, the accumulated value is divided by the number of P-frames in each shot that produces MA.

4. PROPOSED SOCCER SCENE CLUSTERING

In this section we describe the scene clustering algorithm in detail, which integrates the set of audio visual features shown in Section 3. Our proposed method is a bottom-up approach based on heuristic thresholding. The processing flow is illustrated in Fig. 3. At first the entire soccer video compressed by MPEG-1/-2 is classified into three categories, *dynamic*, *static with dominant grass area*, and *static with less dominant grass area*, by evaluating motion activity and grass color dominance. Then audio features (applause and whistle) and motion features are evaluated for those scenes, which further determine (i) Goal/Shoot (4.3), (ii) Free kick (4.4), or (iii) Corner kick (4.5). Scenes not qualifying for the above scene classes are finally identified as (v) Close-up (4.1) or (iv) Offensive (4.2).

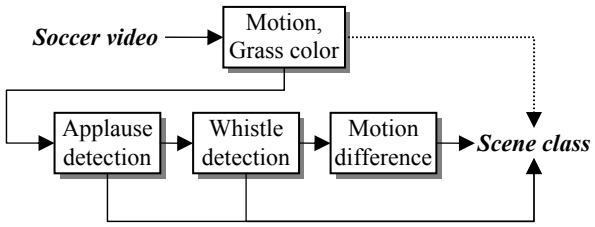

Fig. 3. Soccer scene clustering flow

4.1. Close-up scenes

As mentioned in 3.3, in a close-up scene of players or a referee, the following cases may apply; either large motion exists or grass color is not dominant. Thus close-up scenes can be determined using motion and/or color features. At first, screening is conducted by evaluating motion feature where shots with motion activity MA larger than TH_1 are determined as close-up scene candidates. Note that semantically important close-up scenes following a goal or shot on goal will be included in Goal/ Shoot scenes. Then the color feature is evaluated for further determination. Here, we use color histograms of Y, Cb, and Cr components in downscaled representative frames, and the values in the most frequent set of bins are regarded as grass color. In order to reduce the total amount of calculations, downscaled images are created only if $MA<TH_1$. Then all the pixel values are evaluated and if the number of pixels with grass color N_{gr} is smaller than TH_2, corresponding shots are determined as close-up.

4.2. Offensive scenes

In global scenes capturing offensive plays, grass color is dominant. After thresholding motion activities MA as described in 4.1, if the number of pixels with grass color N_{gr} derived from YCbCr histograms is larger than TH_2, corresponding shots are determined as offensive scenes.

4.3. Goal/Shoot scenes

As described in 3.1, in the case of goal scenes or shoot scenes, very loud applause is emitted from the audience. Here we exploit audio energies defined in 3.1 for determining whether loud applause exists. Therefore, if the subband energy for applause $SBE_{applause}$ is larger than TH_3, it is determined that there exists a shot accompanied with goal or shoot. Since a set of shots is generally important for these interesting events, a few shots before and after the featured shot are aggregated into a scene.

4.4. Free kick scenes

Free kick is taken when serious infringement occurs and it accompanies the referee's whistle. In our method, a free kick is determined by detecting the whistle sound. That is, if the subband energy for a whistle sound $SBE_{whistle}$ is larger than TH_4, a free kick scene is identified.

4.5. Corner kick scenes

In corner kick scenes, the camera view typically changes from a "close-up" capturing a kicker to a "zoom-out" capturing around the penalty area. As mentioned in 3.3, motion in a close-up scene is larger than that in a zoom-out scene, thus the shot transition described above may accompany large changes of motion. Therefore if ΔMA, a difference between motion activities of two adjacent shots is larger than TH_5, a corner kick scene is located.

5. EXPERIMENTAL RESULTS

5.1. Experimentation setup

To evaluate our proposed method on Windows PC implementation, we used three soccer programs (*Soccer I*, *II*, and *III*) compressed by MPEG-2 (720×480, Video: 4Mbps, Audio: MPEG-1 Layer II 224kbps). Each video is 45-minutes duration (not including commercials), i.e. half of the game. All the thresholds, shown in Table 1, were determined using a different 45-minutes soccer video, so that the maximum recall value is obtained.

Table 1. Determined thresholds by training data

Symbol	Description	Scene class	Value
TH_1	Motion activity MA	Close-up	average
TH_2	Number of grass color pixels N_{gr}	Offensive, Close-up	24
TH_3	Subband energy for applause $SBE_{applause}$	Goal/Shoot	average+0.01
TH_4	Subband energy for whistle $SBE_{whistle}$	Free Kick	average×2
TH_5	Difference of motion activities ΔMA	Corner Kick	35

Table 2. Scene clustering results for *Soccer I*, *II*, and *III*

No	Scene class	Detected	Correct	Precision	Recall
I	Goal/Shoot (4/10)	16	14 (4/10)	87.5%	100%
	Free kick (7)	8	6	75.0%	85.7%
	Corner kick (6)	5	3	60.0%	50.0%
	Offensive (82)	80	70	87.5%	85.4%
	Close-up (124)	126	114	90.5%	91.9%
II	Goal/Shoot (3/9)	16	12 (3/9)	75.0%	100%
	Free kick (10)	5	4	80.0%	40.0%
	Corner kick (2)	2	1	50.0%	50.0%
	Offensive (68)	60	58	96.7%	85.3%
	Close-up (108)	116	106	91.4%	98.1%
III	Goal/Shoot (2/9)	11	10 (2/8)	90.9%	90.9%
	Free kick (9)	11	5	45.5%	55.6%
	Corner kick (3)	7	3	42.9%	100%
	Offensive (69)	64	58	90.6%	84.1%
	Close-up (124)	117	112	93.2%	90.3%

5.2. Results and discussions

Table 2 shows the scene clustering results for each soccer video. The values in parentheses in the 'Scene class' column indicate the actual number of scenes. As shown in the table, the most significant scenes, i.e. Goal/Shoot scenes, are successfully determined for all the videos. Other significant scenes, i.e. Offensive scenes and Close-up scenes are classified at high precision and recall. The misclassified scenes as Goal/Shoot scenes include offsides taken after a goal, and offensive and defensive plays inside/around the penalty area, which are regarded as highlights. Missed detections for Free kick scenes are caused since the whistle sound is biased, or buried in applause or other noises in Soccer *II* and *III*, which results in the difference in recall rates between Soccer *I* and other two. The results of Corner kick scenes also show low precision and recall in all videos since its threshold is not optimally chosen due to a lack of training data. The false detections caused by goal kick scenes, which are beyond our scope but have the same shot transition described in 4.5, also incurred low precision for Corner kick scenes.

Misclassifications between Offensive and Close-up scenes are complimentary; that is, Offensive scenes with large motion are determined as Close-up scenes. Similarly, some Close-up scenes are classified into Offensive scenes because they may have a dominant grass area. These errors are caused at the initial stage of determination by thresholding motion activities. Therefore, the order of feature evaluation process should be considered. In addition, since some scene classes may have similar characteristics in low level feature space, more detailed analysis of these characteristics and incorporation of other features need to be considered to improve accuracy.

As for other criteria, our algorithm requires far less computational cost than real-time playback, only about 20% even for MPEG-2 video, depending on context.

6. CONCLUSIONS

This paper proposes an efficient scene clustering algorithm for MPEG compressed soccer videos. Although greatly simplified, our method successfully classifies a soccer video into a predefined set of scene classes. It exploits low level audio visual features easily obtained from MPEG compressed bitstream. Future studies include more detailed and accurate clustering and building a general framework for structuring sports videos.

The authors thank Dr. T. Asami, Dr. S. Matsumoto and Dr. M. Wada for their continuous support. This research is organized by NICT, National Institute of Information and Communications Technology of Japan.

7. REFERENCES

[1] M. Sugano, et al., "Shot Genre Classification using Compressed Audio-visual Features," *IEEE ICIP2003*, Vol.2, pp.17-20, Sep. 2003.

[2] M. Sugano, et al., "MPEG Content Summarization base on Compressed Domain Feature Analysis," *SPIE ITCom2003*, Vol. 5242, pp. 280-288, Sep. 2003.

[3] ISO/IEC 15938-5, Information Technology – Multimedia Content Description Interface – Part 5: Multimedia Description Scheme, Jul. 2003.

[4] A. Hanjalic, "Generic Approach to Highlight Extraction from a Sports Video," *IEEE ICIP2003*, Vol.1, pp.1-4, Sep. 2003.

[5] A. Ekin and A.M. Tekalp, "Robust Dominant Color Region Detection and Color-based Applications for Sports Video," *IEEE ICIP2003*, Vol.1, pp.21-24, Sep. 2003.

[6] P. Xu, et al., "Algorithm and System for Segmentation and Structure Analysis in Soccer Video," *IEEE ICME2001*, pp.721-724, Aug. 2001.

[7] J. Assfalg, "Automatic Extraction and Annotation of Soccer Video Highlights," *IEEE ICIP2003*, Vol.2, pp.527-530, Sep. 2003.

[8] R. Leonardi, et al., "Semantic Indexing of Soccer Audio-Visual Sequences: A Multimodal Approach Based on Controlled Markov Chains," *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 14, No. 5, pp. 634-643, May 2004.

[9] Y.-P. Tan, et al., "Rapid estimation of camera motion from compressed video with application to video annotation," *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 10, No. 1, pp. 133-146, Feb. 2000.

[10] W. Hua, et al., "Baseball Scene Classification using Multimedia Features," *IEEE ICME2002*, Vol.1, pp.821-824, Aug. 2002.

[11] E. Kijak, et al., "HMM Based Structuring of Tennis Videos using Visual and Audio Cues," *IEEE ICME2003*, Vol. 3, pp.309-312, Jul. 2003.

[12] Y. Nakajima, et al., "Universal Scene Change Detection on MPEG Coded Data Domain," *SPIE VCIP97*, Vol. 3024, pp. 992-1003, Feb. 1997.

[13] R. Cai, et al., "Highlight Sound Effects Detection in Audio Stream," *IEEE ICME2003*, Vol. 3, pp.37-40, Jul. 2003.

[14] D.A. Sadlier, et al., "MPEG Audio Bitstream Processing Towards the Automatic Generation of Sports Programme Summaries," *IEEE ICME2002*, Vol. 2, pp. 77-80, Aug. 2002.