Automatic Sports Highlights Extraction with Content Augmentation

Kongwah WAN[1], Jinjun WANG[2,1], Changsheng XU[1] and Qi TIAN[1]

[1] Institute for Infocomm Research,
21 Heng Mui Keng Terrace, Singapore 119613
[2] Nanyang Technological University, SCE, Singapore 637598
{kongwah, stuwj2, xucs, tian}@i2r.a-star.edu.sg

Abstract. We describe novel methods to automatically augment content into video highlights detected from soccer and tennis video. First, we extract generic and domain-specific features from the video to isolate key audio-visual events that we have empirically found to correlate well with the ground-truth highlights. Next, based on a set of heuristics-driven rules to minimize view disruption, spatial regions in the image frames of these video highlight segments are segmented for content augmentation. Preliminary trials from subjective viewing indicate a high level of acceptance for the content insertions.

1 Introduction

With the world-wide growth of consumer devices such as mobile phones and home set-top-boxes (STB), content providers are looking for sustainable revenue streams from innovative video applications in their distribution networks. With its global appeal, sports video is widely seen as a key driver content to launch applications such as interactive TV. Significant research effort has also been devoted to the automatic extraction of sports highlights in the past few years [1]. In particular, interesting results from our recent work in [2] points to the potential for a secondary market for game viewer-ship on mobile devices, and that replay selection and generation is no longer the exclusive purview of the game broadcasters. Issues naturally arise as to how the business case can be further enhanced. We note that the traditional model of 30-sec advertising-run for broadcast TV may not work well on mobile platform for obvious reasons of bandwidth cost and duration ratio. In contrast, in-program content augmentation appears to be more suitable. However, existing techniques such as [3] are generally hardware-driven and expensive. Our intention in this paper is to explore alternative techniques for content insertion (used interchangeably in this paper with content augmentation) and how they can be integrated with automatic sports highlights extraction for consumer video applications. We organize the rest of the paper as follow. We first describe automatic highlight extraction for soccer and tennis in Section 2. Then we provide an overview of our content insertion techniques in Section 3. Experimental results are detailed in Section 4 before concluding on some future work in Section 5.

2 Automatic Sports Highlights

As most sports events are played in constrained settings and telecasted with a small number of cameras, the limited view categories and their repetition is intuitively amenable for techniques such as shot type classification [4]. Structures in soccer [5] and tennis [6] have been capitalized for play-break classification and high level retrieval. For low end devices, compute-efficient techniques have been developed in [7,8] using a small set of generic audio-only features. However, it should be pointed out that most of these literature address the problem of event *isolation*, ie, its start-point, rather than its boundary end-points, which is more difficult as it requires a proper enclosure of the pre-event context and post-event response. In what follows, we give a brief overview on our key recent contributions [2,8,9] in event boundary detection.

2.1 Audio End-points (Soccer): Excited Commentary

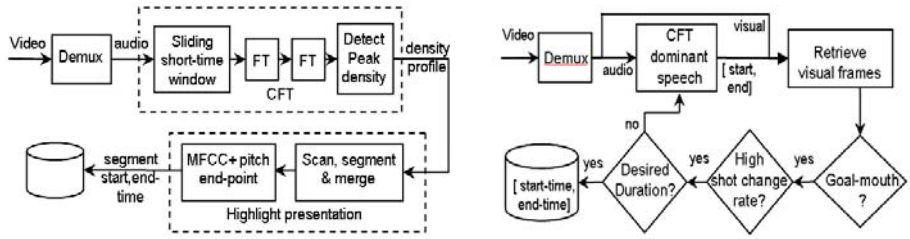

Fig. 1. Left: Composite Fourier Transform-based highlight detection on soccer audio; Right: Inclusion of visual features for end-point calculation

The difficulty in analyzing soccer audio is compounded by the large spectral variation in mixing commentary speech, field audio and noise. Figure 1 (left) replicates the system in [8], where we approach the problem as locating the "dominant speech" portions in the audio, which we model as a band-limited pulse train in the frequency domain. A novel Composite Fourier Transform (CFT) is then used to isolate the excited commentary: a first FT of the signal is applied, and the resulting magnitude spectrum is treated as another time domain signal input to a second FT. This differs from the widely-used cepstrum method, which applies an inverse FT to the log-spectrum instead. To calculate the end-points, voiced segments based on MFCC features are input to a robust pitch tracker to mark the rise and fall of excited commentary.

2.2 Visual End-points (Soccer): Shot-cut rate, Field positions, Goal-mouth, etc

Figure 1 (right) shows the 2 additional visual features used in [9] to compute highlight boundary end-points: shot-change rate and goal-mouth appearances. The former intuitively capture the production rule that after a significant event, views from multiple cameras are used to reproduce the action from alternative view angles or to survey the on-field emotions. In particular, computing shot cuts on video from panning cam-

eras tracking the paths of celebrating players usually result in rapid cut sequences of short duration. Together with a reliable detection of goal-mouth appearance, the number of stray outputs from an audio-only (CFT) approach is effectively reduced.

Our other recent paper in [2] further describes a novel approach to generate soccer replay segments from *only* the main panoramic video camera. When used in a multi-ple-camera-setup, the end-points can be used as time-stamp markers to collate the other video streams, and further analysis can be made as to which is a better video feed to go on air. Not only can these segments be replay candidates for decision by the broadcast TV director, they can also be distributed to a secondary channel of viewers on, say, mobile platform. A mid-level representation framework is used to create audio-visual keywords from the low level features: (Visual) F_1: Active Play Position, F_2: Ball trajectory, F_3: Goal mouth location, F_4: Motion activity; (Audio) F_5: Whistling, Acclaim. Three types of events are defined: Attack, Foul and Others. Three SVM classifiers (Gaussian kernel) are trained on an empirically derived set of mid-level keywords: Attack: F_1, F_2, F_3 and F_5; Foul: F_1, F_4 and F_5; Other: F_1 and F_4.

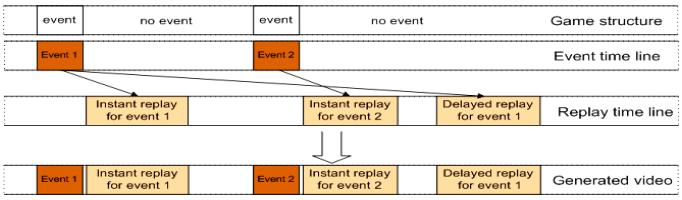

Fig. 2. Rules for replay insertion

To compute replay boundary, a search algorithm is applied to search backward and forward from the moment of event occurrence. The backward search checks whether the play position keyword F_1 has changed from t_s - D_1 to t_s - D_2, where t_s is the event moment starting time and D_1, D_2 are the minimal and maximal offset threshold respectively. A similar forward search is applied to detect the event end-time. These boundaries are input to the replay generation module (Figure 2). For any event segmented, the system attempts an *instant* replay by examining whether it can be inserted at the following "No-event" slot (instant replays for event-1 and 2 in row-2 and 3). Otherwise, it checks for delayed replay criteria: is the event important enough to be shown at a later time. If so, the system buffers the event and inserts the replay in the next available time slot (row-2 and 3 where a delayed replay for event-1 is inserted at a later time slot). Row-4 shows the generated video after replay insertion.

2.3 Audio End-points (Tennis): Applause and Ball-hits

In practice, the CFT approach works well for noisy soccer audio but would appear to be overkill for "cleaner" signals like tennis audio: the crowd is usually quiet during play and loud cheers and applause generally follow every point won. In fact, our experiments have shown that a good indicator of the quality of the game point, and therefore its highlight-worthiness, is in the *duration* of the applause/cheers. In our MFCC implementation, 40 filters evenly spaced over 50Hz to 3200Hz (6 octaves)

computes the mel-spectra on which the first 13 DCT coefficients are used as a vector input to a neural network. Training data is manually cropped and labeled as either Applause or Ball-hit. A window size of 100msec in steps of 50msec is used to generate MFCC vectors for the Back-propagation learning algorithm. On play sequences, ball-hit frames are differentiated from the silent frames using a simple ZCR feature. It is remarkable that fairly robust results can be obtained from this simple setup on 5 full-length games (3 different Grand-Slam tournaments, ~10hours). Relevant training data are only extracted from the first 5 minutes of each video and testing results are compiled on the remaining video sequence. High recall and precision rates of >70% has been obtained on segmenting game units. Ball-hit detection has also been used to identify *long rallies*, forming another criterion for highlight selection.

3 Content Augmentation

Fig. 3. Landmark detection in soccer video and content insertion therein

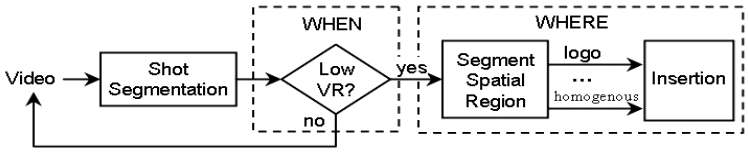

Fig. 4. A framework for Content Augmentation

Advances in multimedia communications have made it possible for real-time computer-aided digital effects to be introduced into video presentations. For example, PVI's famous first-down line in American football is blended onto the field using clever chroma-keying [3], achieving a realistic implant that appears to be part of the field. However, the hardware-based method is labor-intensive and expensive. In [10], a software method is reported for inserting advertising images into selected target areas in a video broadcast. Insertion areas are manually selected and simple edge/color features are extracted for matching using geometric hashing. Video frames are buffered for smoothly aggregated insertions that do not appear abruptly. Apart from billboards, reliable segmentation of other generic landmark targets in soccer (Figure 3) is also addressed in [11]. While these methods do not need prior labeling, their dependence on domain features, eg, the soccer center ellipse, is undoubtedly a limiting factor. In the remainder of this section, we explore novel methods to overcome this limitation. Figure 4 encapsulates the essential ideas in a conceptual framework.

3.1 Temporal Segments with Low Viewer-Relevance

The basic unit of computation in the framework remains at the shot level. Collated frames in the shot are computed to obtain a viewer-relevance (VR) measure. It is not hard to conjecture a fairly accurate VR of a specific game. For instance, in soccer, the general opinion is likely to correlate a play build-up in or towards the goal area as exciting/relevant. The equivalent to tennis is probably the game point unit, commencing from serve to out-of-bound. Basing an insertion decision upon the VR of a video shot segment facilitates the intuitively appealing notion of minimal (or none at all) disruption to viewer's enjoyment of the game. The underlying question is to ask WHEN to insert. In this regard, the play-break classification in [5] may be usable for soccer insertion, while the applause detection methods (Section 2.3) are applicable for tennis insertion. Every image frame within each video shot that passes the criteria for low VR subsequently undergoes spatial region segmentation for content insertion.

3.2 Spatial Regions with Low Viewer-Relevance

In general, several heuristics guide the design of a spatial region segmenter for insertion. Region attributes such as homogeneity, texture similarity, motion and clutter are all factors to assign a degree of relevance to a region. For instance, the watermark indicia and time/score annotation used by many sports content distributors/broadcasters occupy small screen space, and are placed at the 2 upper corners to be of least visual disruption. One would then naturally project that these positions would also offer the "best" locations for any new content to be inserted further downstream.

On the other hand, a general characterization of a region VR appears plausible. A viable postulate is regions in a fast motion scene, eg, from a panning camera tracking a player, should have high VR. This means that even if the current shot qualifies for insertion with a low VR in the first WHEN-decision, it should arguably still fail the WHERE-decision. In contrast, regions that are uniform in terms of certain visual homogeneity metric qualitatively obtainable from the images are arguably better candidates because they contain less information and would appear to be less prominent to the human eye. The above considerations motivate the following techniques.

3.2.1 Static region (TV logo/graphical annotation) segmentation

In general, static graphical insertions in TV are opaque or semi-transparent. The first intuition is to apply variance analysis on pixel intensity to obtain a static region mask. But this approach is sensitive to the fluctuation due to video digitization noise. Instead, we adopt a gradient-based approach by computing edge change over successsive frames. A preliminary static mask S_i is obtained via time-averaging:

$$S_i = G_i + \alpha S_{i-1} , \quad (1)$$

where G_i is the gradient image of current frame i, α denotes a decay factor set to 0.7. To cope with spurious holes, eg, from the timer-ticker, morphological operations are

then applied based on an elongated kernel designed to work with most TV logos and graphical annotations. Figure 5 shows examples of static masks obtained.

Fig. 5. Static mask examples on soccer and tennis

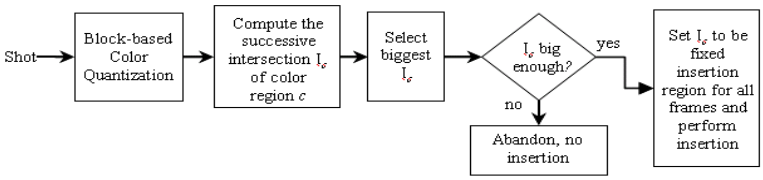

Fig. 6. Homogeneous region segmentation

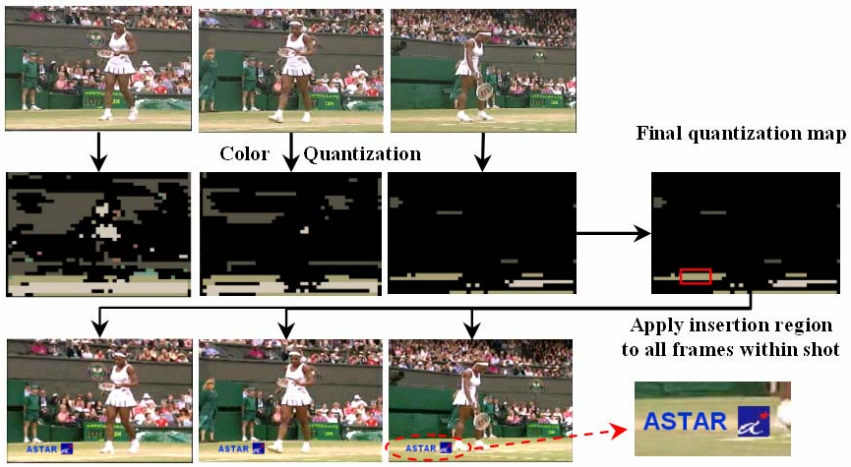

Fig. 7. Top-row: 3 original images successively spaced 1-sec apart. Row-2: evolution of the color quantization map; Row-3: Insertions (logo) onto all frames within shot

3.2.2 Homogeneous region segmentation

Figure 6 shows an exemplary flow-chart for segmenting a color-homogeneous region for insertion. Image frames in the shot are first divided into 32x32 non-overlapping blocks. These are then collated to be quantized [12] into a small number of colors (eg, 8). The quantized color indices are then used to isolate regions with color consistency. This is done by computing the boundary extent of each color-coded region using simple connected component. As the frame content changes, the boundary extent of these regions will also change. By taking the intersection of all regions computed

within the shot, we derive a rectangular region which has maintained its color integrity for the duration of the shot. A decision is then made as to whether the region is usable for insertion. This may be based on the match of its dimension to the geometry of the content to be inserted: a small square region is useful for a logo insertion (Figure 7), while an elongated one can be used for an animation sequence (Figure 8).

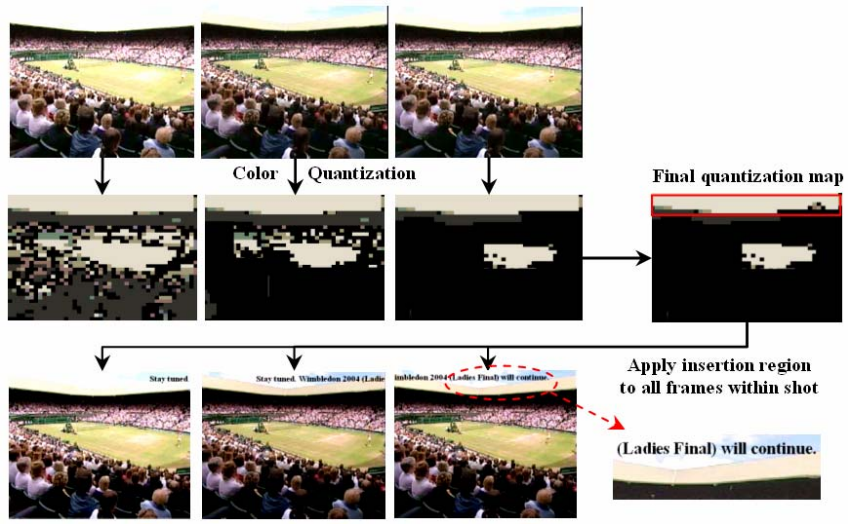

Fig. 8. Text insertion on elongated regions in original images 2-sec apart

4 Experimental Results

To verify the validity of our approach, we conduct subjective viewing tests using highlights that are automatically generated from a soccer video and a tennis video. Content augmentation described in Section 3 is then applied to these highlights. Since not every highlight segment has a successful insertion, we arbitrarily chose 9 soccer highlights of which 8 has insertions, and 10 tennis highlights of which 8 has insertions. This makes a total of 16 insertions on the 19 highlights. Only static logo insertions are used, and these are uniquely taken from famous trade-marks such as Nike, Mastercard, etc. To throw the subjects off guard, they are first asked to identify some trivia in the video before viewing starts. After the viewing, they are then asked to recall any logos they have noticed (subtlety), and whether these exposures have reduced their viewing experience (acceptability). The cumulative results over 8 viewers are tabulated in Table 1. Most gave a high opinion of the quality of our highlights, and did not notice the pattern of advertising insertions only after quite a few of them have appeared. We also have a general concurrence on our design of making advertising insertions occur during play-break, in order to minimize interference with the game proper. The minority opinion of opposition came in 3 forms: (1). No insertions

at all; (2). Restrict the insertions to a single place; (3) Legal concern on insertions over the static watermark logo. On the whole, Column 2 and 3 show a high level of acceptance and high recall of the exposure. This must be good news for advertisers.

Table 1. Subtlety and Acceptability of Insertions

Sports	Subtlety	Acceptability
Soccer	50%	50%
Tennis	40%	60%

5 Conclusion and Future Work

We expect that automatic sports highlight technology will facilitate alternative channels for sports content distribution and broadcasting. A framework for augmenting video highlights with attendant content is developed in this paper. The obvious use of this is in product branding/advertising, a traditional financial pillar for broadcast media. Based on the viewing patterns, it is also easy to generalize the framework to incorporate means for mining purchasing preferences for target advertising.

References

1. Adami, N., Leonardi, R., Migliorati, P.: An Overview of Multi-modal Techniques for the Characterization of Sport Programmes. In: Proc. SPIE – VCIP 2003, pp. 1296-1306
2. Wang, J., Xu, C., Chng, E., Wan, K., Tian, Q.: Automatic Highlight Detection and Replay Generation for Soccer Video. Full paper to appear in ACM Multimedia 2004.
3. PVI Virtual Media Services: http://www.pvimage.com/pvi/index.html
4. Duan, L., Xu, M., Chua, T., Tian, Q., Xu, C.: A mid-level representation framework for semantic sports video analysis. In: Proc ACM Multimedia 2003, pp. 33-44
5. Xie, L., Chang, S., Divakaran, A., Sun, H.: Structure Analysis of Soccer Video with Hidden Markov Models. In: Proc ICASSP 2002, Orlando, FL, USA
6. Sudhir, G., Lee, J., Jain, A.: Automatic classification of tennis video for high-level content-based retrieval. In: Proc CAIVD, 1998, pp. 81 -90
7. Xiong, Z., et.al, T.: Audio events detection based highlight extraction from baseball, golf and soccer games in a unified framework", In: Proc of ICASSP 2003, Vol V pp 632-635
8. Wan, K., Xu, C.: Robust Soccer Highlight Generation with a Novel Dominant-Speech Feature Extractor. In: Proc ICME 2004, Taiwan.
9. Wan, K., Xu, C.: Efficient Multimodal Features for Automatic Soccer Highlight Generation. To appear in: Proc ICPR 2004, Cambridge, UK.
10. Medioni, G., Guy, et.al: Real-time billboard substitution in a videostream. In: Proc 10th Tyrrhenian International Workshop on Digital Communications, Italy, 1998, pp.71-84
11. Wan, K., Yan, X., Yu, X., Xu, C.: Real-time goal-mouth detection in MPEG soccer video. In Proc ACM Multimedia 2003, pp. 311-314
12. Gervautz, M., Purgathofer, W.: A Simple Method for Color Quantization: Octree Quantization. In: Proc. CGI '88, pp. 219-231