AUTOMATIC CONTENT PLACEMENT IN SPORTS HIGHLIGHTS

Kongwah WAN and Changsheng XU

Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613

ABSTRACT

To be viable advertising platforms, methods for in-program content placement in sports video must balance against clutter. We propose viewer relevance (VR) measures of video frames in the temporal and spatial domain. Video sub-segments with low temporal VR are first selected, within which actual content is emplaced on regions with low spatial VR. We compute VR measures using color, motion, texture and domain features, upon which spatio-temporal techniques are used to segment spatial regions for actual content placement. Results from preliminary subjective viewing trials on soccer and tennis video indicate that our approach is promising.

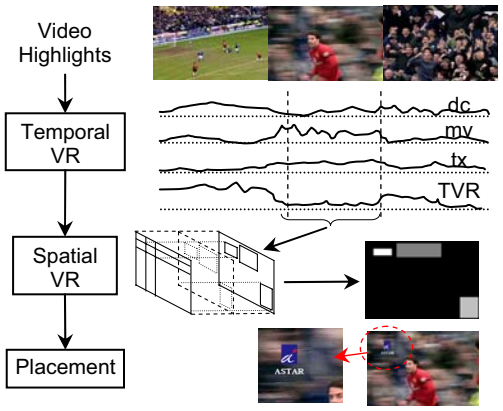

Figure1. Content Placement Work-flow

1. INTRODUCTION

Because of its global appeal, sports video is a key driver content for distribution on consumer and mobile devices. Recent progress in the automatic extraction of sports highlights [1-3] has also herald a secondary market for mobile game highlights. In such context, traditional fixed-slot advertising model may not be applicable. In-program placement are more suitable, and are already used in sports broadcast showing attendant data such as play time/score-bar or advertising effects such as trademark logos and animation overlays. There is a need to balance between maximizing exposure and minimizing clutter. Image overlay systems are simple but the logo "pop-out" must be sporadic. Realistic 3D content blending systems can support more frequent exposures but only at the expense of elaborate sensor-based tracking [5] and image processing [6].

In this paper, we extend our initial work in [4] to explore content placement in sports video, where we seek to automate the sporadic logo "pop-out". Figure 1 shows the work-flow. Short highlight clips (~30secs) in a sports video (manually segmented or automatically extracted) are input to the system. Using ideas from an attention model [7], a Temporal Viewer Relevance (TVR) curve is computed by fusing various visual features over time. The aim is for TVR to return high values on video segments that depict "relevant" game moments. Contiguous frames along a trough on the TVR curve are the supposedly "low" points in the game, and deemed to be amendable for content placement. To locate a "suitable" placement location on each of these frames, we assign Spatial Viewer Relevance (SVR) values to the local spatial regions. A spatio-temporal approach is used to define a saliency value based on the dynamic pixel variations in the frames. More precisely, image frames are first divided into fixed-size overlapping blocks, on which color, texture and motion features are extracted. Accumulating over all frames, we obtain a feature volume for each block. A SVR value is then computed for each block based on the entropy of its feature volume. Ideally, low SVR values should be assigned to blocks of crowd clutter and moving background.

In summary, TVR first tells us *when* to do content exposure. SVR then points us to *where* to do the placement. In what follows, Section 2 and 3 will provide details of the temporal and spatial VR measures we have implemented for soccer and tennis video. Results and discussions of subjective viewing trials are presented in Section 4 before concluding on some future work in Section 5.

2. TEMPORAL VR

A typical sports video highlight is likely to straddle across multiple scenes depicting pre-event play, main event and post-climax. For example, a soccer video highlight showing a goal would include short segments of the action leading

up to the goal (high VR), the climactic goal (peak VR) and finally the celebratory response (lower VR). The aim is for TVR to mirror these relevance values by a non-linear fusion of three visual features extracted from every frame: dominant color (*dc*), motion (*mv*) and texture (*tx*):

2.1. Dominant Color

Sports video is frequently characterized by the presence of a dominant color, eg, green soccer field. In a sports video highlight, the *non*-dominant color frames are usually the post-climax frames depicting player close-up or celebratory response. Ignoring low intensity pixels, denote $H_H[h], H_S[s]$ and $H_I[i]$, $h \in [0..63], s \in [0..63]$, $i \in [0..127]$ as the histogram of the HSI-color component of all pixels in a frame *f*. The histogram entropy for H component P_H is: $p_H = -\sum_i p_h(i) log(p_h(i))$, where $p_h(i) = \frac{H_H[i]}{\sum_i H_H[i]}$. P_S and P_I are similarly computed. P_H, P_S and P_I will be low whenever there is a dominant component in their histograms. The final *dc* is computed as a reciprocal:

$$dc_f = (p_H * p_S * p_I)^{-1}$$

2.2. Motion

To obtain mv_f of a frame *f*, each pixel is first computed for its optical flow using [8]. The magnitude of flow vectors and their direction uniformity provide an indication of the global motion. Define $mf_{i,j,f}$ as the magnitude of pixel (*i,j*) with flow vectors (*dx,dy*) normalized over *all* frames as:

$$mf_{i,j,f} = \frac{\sqrt{dx_{i,j}^2 + dy_{i,j}^2}}{\max_{i,j,f}(\sqrt{dx_{i,j}^2 + dy_{i,j}^2})}$$

Normalizing by the frame dimension *N* give the mean flow magnitude for frame *f*: $\overline{mf_f} = \frac{1}{N}\sum_{i,j} mf_{i,j,f}$.

For flow direction, we quantize every flow vector in frame f into *D=8* direction-bins and accumulate into histogram H_M. The direction entropy mp_f is computed as:

$mp_f = -\sum_d^D p_M(d) log(p_M(d))$, where $p_M(d) = \frac{H_M[d]}{\sum_d H_M[d]}$

mp_f is low when the flows are aligned. The final mv_f is:

$$mv_f = \overline{mf_f}(1 - \overline{mf_f} * mp_f) \ ,$$

2.3. Texture

Texture measures based on the Gray Level Co-occurrence Matrix (GLCM) are used. Let $p(i,j,d,\theta)$ be the normalized GLCM of pixels (*i,j*) separated by distance *d* in orientation θ. The Contrast th_w is the moment of inertia around the matrix's diagonal and indicates the smoothness of pixel variation within a window *w*:

$$th_w = \sum_{i,j \in w} \frac{p(i,j,d,\theta)}{(1 + (i-j)^2)}$$

The final texture tx_f is computed by summing th_w over all *w* and normalizing by the maximum over the video.

$$tx_f = \frac{\sum_{w \in f} th_w}{\max_f \sum_{w \in f} th_w}$$

2.4. Fusion

To obtain the final TVR curve, we use a priori formulation:

$$TVR_f = \frac{dc_f}{mv_f * tx_f}$$

This encodes the observation that frames characterized by a high dominant color *dc* (usually pre- and main event frames) are intuitively more important. Frames showing high motion *mv* and background clutter *tx* (usually post-climax frames) will return lower TVR values. Content placement, if any, ought to be performed on the latter. Low TVR segments can now be selected using a straightforward duration-based scheme. A sliding window, corresponding to the desired exposure duration, accumulates the TVR values under the curve. The window with the least cumulative TVR sum can then be used as the starting frame for content exposure. An empirical threshold *T* is applied to ensure that the TVR cumulative sum is sufficiently low.

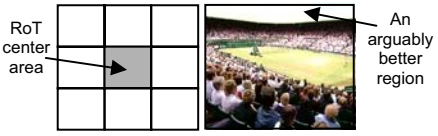

Figure 2. "Problem" with arbitrary placement

3. SPATIAL VR

Once TVR isolates the post-climax frames, content can arguably be placed anywhere outside the "Rule of Third" (RoT) center area. However, as Figure 2 shows, arbitrary placement may not yield the best effect. The wide area at the top is a better choice as it has the most background uniformity and contrast. Its elongated geometry would also fit a horizontal text ticker well. These considerations lead us to develop a block-based spatio-temporal approach.

3.1. Block-based Spatio-Temporal Entropy

In this method, overlapping fixed-size spatial blocks are first defined over each image frame, from which visual features are extracted and quantized. If we think of the quantized features in a block as *its* states along the temporal

axis, the distribution of the state (feature) transitions may be obtained by a temporal histogram, and the corresponding probability distribution function (*pdf*) approximated by normalization. Accumulating features over a block also encodes into the *pdf* the collective spatial relationship of pixels within their local neighborhood. Once a feature *pdf* of a block is obtained, its entropy can be used as a measure of feature consistency.

For example, in the case of color, we extend the *HSI* color component histograms defined in Section 2.1 to accumulate colors from pixels in the same spatial blocks over *all* frames in the sequence. Denote $CSTH_b$ as the new spatio-temporal color histogram for the b^{th} block. Normalizing, we obtain the block color-*pdf*:

$$PDF_{color,b}(h,s,i) = \frac{CSTH_b[h,s,i]}{\sum_{(h,s,i)\in\Omega} CSTH_b[h,s,i]},$$

where Ω is the *HSI* quantization space. The color spatio-temporal entropy at block b is then given by:

$$CSTE_b = -\sum_{(h,s,i)\in\Omega} PDF_{color,b}(h,s,i)\log(PDF_{color,b}(h,s,i))$$

Without further elaboration, the block-based feature entropy for texture can similarly be built. The feature entropy values can be projected onto an energy map. Figure 3 shows some examples. Given a modality, a high feature consistency within the temporal accumulation window will induce a low *STE* value. For example, the central blocks in the soccer frames (top-row) show a greater color uniformity compared to the background clutter. This is a result of the camera tracking the celebrating player in his red jersey throughout the shot. Similarly, the texture *STE* map in the tennis frames (bottom-row) registers a higher texture consistency in the middle blocks covering the spectator area. In contrast, the blocks covering the stadium top and tennis court area show more changes as a result of camera panning.

3.2. Spatial Segmentation and Placement

Additional steps are needed to segment each STE map into candidate regions. A Gaussian smoothing filter is first applied to remove high frequency noise before collating all blocks with STE below a threshold. This threshold is iteratively raised so that the cumulative area of all blocks covers one-tenth of the image area. Morphological closing and opening with a horizontal kernel is then used to join neighboring blocks. Examples of the final regions are shown in red bounding boxes in figure 3.

There are several considerations as to which candidate regions to use and the appropriate *type* of content placement. Firstly, all regions overlapping with the center RoT area are rejected. Secondly, candidate regions from the Color-STE maps tend to have greater background uniformity and hence are suitable for a text/image overlay using alpha keying. This is a relatively unobtrusive placement. Enhanced text visibility can also be achieved by

(a) Key-frame (b) Color STE (c) Texture STE

Figure 3. Spatio-Temporal Entropy maps

selecting a foreground color with the best contrast to the mean background. On the other hand, candidate regions in Texture-STE maps tend to be clutter areas and moving background. An overlay needs to have its own background for better contrast. This type of placement is visually more intrusive, but is arguably acceptable as it only occludes "irrelevant" video data. Lastly, the geometry of the candidate regions can be used to decide whether a text, image or animation is more suitable.

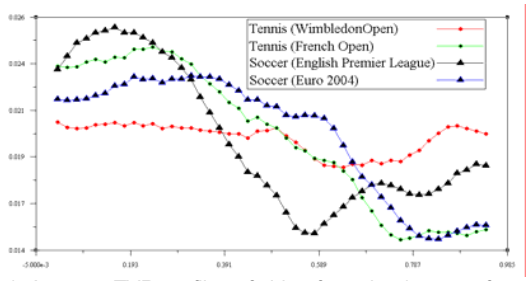

Figure 4. Average TVR profiles of video from the 4 types of test data, normalized to unit duration and unit cumulative TVR area.

4. EXPERIMENTAL RESULTS

We examine the validity of our priori VR formulation by running the content placement algorithm on broadcast sports editorial video. Data from two sports domain are tested: soccer from the weekly highlights of the English Premier League/Euro-2004, and tennis from the daily highlights of the French Open/Wimbledon Open. We manually segment the video into individual short highlight clips, removing all non-game segments such as interviews and commercials. Because our VR is formulated to model the boundaries of a single game event, we also removed the replays. A total of 53 soccer highlights and 42 tennis highlights are segmented. Figure 4 shows the average TVR profiles of video from the 4 tournaments normalized to unit duration and unit cumulative TVR area. The troughs are clearly seen. Usually, these are the transition to the post-climax scenes. The slight variations among the 4 profiles are due to the difference in broadcast style and venue ambience. For instance, the EPL highlights tend to show a lot more of the post-climax scenes, and hence their troughs are generally earlier.

Of the 95 input clips, 13 (throughput=86%) fail to detect their TVR troughs below the cumulative threshold *T* (Section 2.4; empirically set to 0.019 in our experiments). This happens when the post-climax scenes are too short. An example is when play resumes very quickly and the camera switches back to the far-view focus on the dominant color field. The TVR would then treat that as part of the main event segment, and fail to register a trough. In general, our TVR model works best when there is a clear demarcation of pre-event, main-event and post-event boundaries. This is also why when applying TVR on automatic video highlights generated by [3,4], throughput drops to only ~30%.

Once post-climax frames are segmented, SVR saliency values are computed to find a placement location. In our experiments, we used spatial blocks of size 8 by 8, each of which overlap with its adjacent blocks by 4 pixels (overlap of 50%). We found that visual features in a frame sequence increase in entropic randomness over time. As a result, many high entropy blocks are rejected (Section 3.2). On shorter temporal accumulation window (<3-secs), the STE maps show better homogeneity. Hence, all segmented TVR troughs must first be minimally 3-secs long. Then, using a fixed sliding window of 3-secs, the entire feature volume of each spatial block is divided into overlapping sub-volumes. The final STE value of the spatial block is the *least* STE value computed on *all* its sub-volumes.

To assess the suitability of placement locations, we conduct simple subjective viewing trials as follows. First, 25 short video highlight clips are randomly selected. Content placement is performed on 16 of these clips, using a short list of famous trademarks such as Nike, McDonald, etc. Some examples are shown in figure 5. The 25 clips are then merged in random order. To minimize bias, subjects are asked to watch the video to answer some trivia questions and are only told of the deliberate content placement *after* the viewing. They are then asked to write down all the trade-mark names in the video, including whether the placements are visually acceptable. From the answers, we compiled 2 subjective statistics: *Subtlety*, which measures the subjects' ability to recall the logos, and *Acceptability*, whether people can accept the "extra" content in the video presentation. Subtlety is computed as the percentage of logos correctly recalled and is indicative of the "eye-catchiness" of the placements. Acceptability is calculated by the normalized scale index.

Table 1. Subjective results from 23 subjects, 16 placements

Sports (duration, # of placement)	Subtlety	Acceptability
Soccer (2:53, 8)	57%	52%
Tennis (4:03, 8)	43%	67%
Average	50%	60%

The cumulative results over 23 viewers are shown in Table-1. Most commented that they only noticed the pattern of placements only after a few have appeared. Given that our logo exposures were only over small screen area, we consider the subtlety rate of 50% to be a fairly high recall rate, which is good news for advertisers.

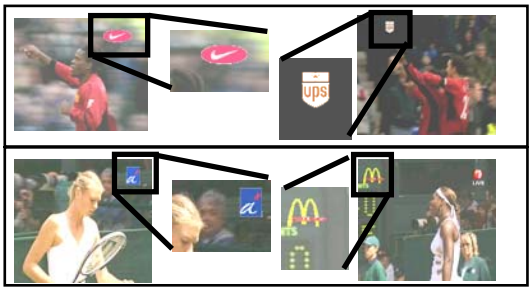

Figure 5. Examples of content placement

5. CONCLUSIONS AND FUTURE WORK

A framework for automatic content placement in sports video highlights is developed in this paper. A combination of generic visual features are computed from the video and combined to form a measure of viewer relevance in both the temporal and spatial domain. Our results on editorially-created clips show that our VR model is fairly accurate. This can be seen as domain-specific user attention modeling, where a priori knowledge such as the presence of a dominant color in sports video is factored into the VR. It is interesting to see how this can be generalized in other video domain such as movie trailers and Music TV.

6. REFERENCES

[1] J. Wang, et al, "Automatic Replay Generation for Soccer Video Broadcasting" *Proc of ACM MM 2004*, pp 32-39.

[2] G. Sudhir, J. Lee and A. Jain, "Automatic classification of tennis video for high-level content-based retrieval", *Proc CAIVD*, 1998, pp 81 -90.

[3] K. Wan and C. Xu, "Efficient Multimodal Features for Automatic Soccer Highlight Generation", *Proc of ICPR 2004*, Vol 3, pp 973-976.

[4] K. Wan, J. Wang, C. Xu and Q. Tian, "Automatic Sports Highlights Extraction with Content Augmentation", *Proc PCM 2004,* vol 2, pp 19-26.

[5] PVI Virtual Media Services: http://www.pvimage.com/

[6] G. Medioni, G. Guy, H. Rom and A. Francois, "Real-time billboard substitution in a videostream", *Proc 10th Tyrrhenian Int'l Workshop on Digital Communications*, Italy, 1998, pp 71-84.

[7] Y. Ma, L. Lu, H. Zhang and M. Li, "A User Attention Model for Video Summarization", *ACM Multimedia 2002*, pp 533-542.

[8] B. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," *Proc of DARPA Image Understanding*, 1981, pp 121–130

.