Automatic Mobile Sports Highlights

Kongwah WAN, Xin YAN and Changsheng XU
Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613
{kongwah, yanxin, xucs}@i2r.a-star.edu.sg

Abstract

We report on our development of a real-time system to deliver sports video highlights of a live game to mobile videophones over existing GPRS networks. To facilitate real-time analysis, a circular buffer receives live video data from which simple audio/visual features are computed to detect for highlight-worthiness according to a priori decision scheme. A separate module runs algorithms to insert content into the highlight for mobile advertising. The system is now under trial over new 3G networks.

1. Introduction

With the onset of 3G and advent of automatic sports highlight detection technology, mobile sports video alert is becoming a reality. This is expected to generate a new and sustainable revenue stream for content owners throughout the distribution networks. Significant research results have been achieved in the analysis and automatic extraction of sports highlights [1-3]. In particular, results in [1] show that replay selection need no longer be the exclusive purview of game broadcasters and herald a secondary market for mobile devices. A natural means for advertising in such context is in-program content placement. Already, overlays are used extensively in broadcast TV: time/score/game-stats overlays, sporadic sponsor-logo "drop-down" and animation. Simple overlays are usually manually operated in standard production studios, the "pop-out" only appearing at sporadic moments corresponding to a momentary lull in the game. In contrast, sophisticated systems such as [4,5] use elaborate sensor-based tracking to blend the insertion into landmark objects in the video. Image processing-based methods in [6] generally have more image constrains and lower system throughput. But when applied properly, they can be viable alternatives.

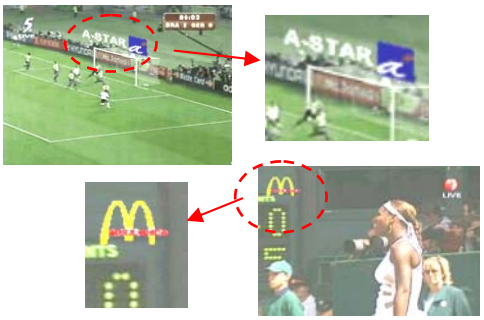

Figure 1. Examples of insertions

An attempt to place content in a video highlight is a challenge on the limits of the de-facto stand that all advertising exposures should occur during the low points of a game. To minimize the potential outrage, we introduce a 2-tier distribution model: a premium model where ad-free highlights may be sent but would be more expensive, and a (default) baseline model, where advertising exposures would be seen, but they are limited to: (a) landmark objects such as the top of the goal posts; (b) an area at the 4 corners of the screen not covering the TV station logo and other annotation icons. Figure 1 shows some examples of insertions.

Figure 2 shows the overall system work-flow. Audio and visual features are extracted to detect for highlight worthiness of the current video in the buffer. A sub-segment is output as the premium highlight video. This is routed to the content insertion module to produce a new video with content implanted (baseline). Two sets of video highlights are therefore archived at every time. Based on a given subscriber's information, one is selected for delivery. To manage bandwidth, an SMS notice is first sent, and if he opts to see the video, a new connection is then launched.

Highlight detection and content placement is described in Section 2 and 3 respectively. Section 4 details our mobile video delivery set up. Results of our experimental trials are discussed in Section 5, and we conclude in Section 6.

0-7803-9332-5/05/$20.00 ©2005 IEEE

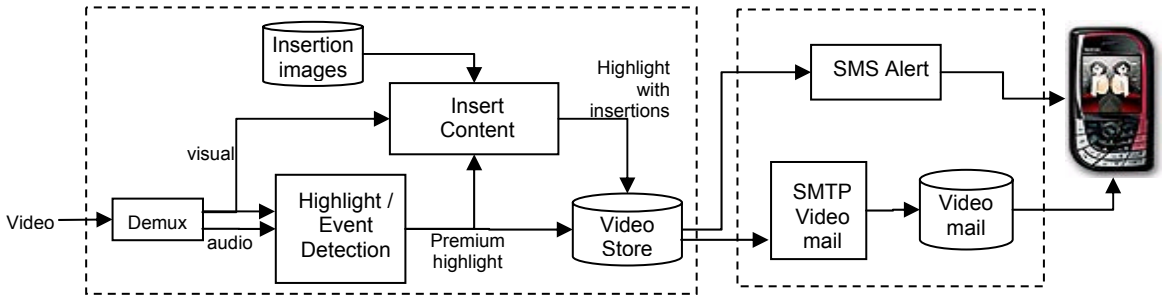

Figure 2. System Architecture of Mobile Sports Alert

2. Detecting Highlight-Worthiness

A MPEG-1 video stream is input to a circular buffer capable of storing 1 minute's worth of data. Audio and visual frames (PCM, RGB) are first decoded in real-time. The techniques in [3] are extended to assess the Highlight-Worthiness of the buffer content. Denote this by the vector $HW_{B,f}$ where f is the frame index within the buffer. $HW_{A,f}$ and $HW_{V,f}$ are vectors defined similarly for the audio and visual domain.

The PCM frames feed into the rear of a local circular audio FIFO buffer of 200msec. When 100msec of new data has arrived, the entire audio buffer is processed to compute for mid-level audio *keywords* [8] such as pitch and energy. The new values enter into another circular buffer maintaining a fixed number of the most recent values. The median values provide an indication of the level of audio energy over the recent past. A high value should correlate with $HW_{A,f}$. Domain-specific features may be also used, eg an applause detector for tennis audio.

RGB frames are similarly buffered and computed for their color histogram (CH_{frame}). A small buffer of 1 sec is kept, wherein a mean CH_{bufer} is maintained by averaging over successive CH_{frame}. An abrupt difference of CH_{frame} with CH_{buffer} indicates a potential shot boundary and CH_{buffer} is reset. The number of shot changes over the fixed-length recent window indicates the shot change rate, which generally will be higher during an important scene.

We compute $HW_{B,f}$ as a simple linear combination of the $HW_{A,f}$ and $HW_{V,f}$. Let $F = \arg\max_f HW_{B,f}$. If F is the *center* index in the buffer, and $\max_f HW_{B,f}$ exceed a priori threshold, then a highlight event has been deemed to occur. (Note the delay of *half* buffer time). We then search for the start and end boundary of the final highlight segment by searching back and forth the center index respectively. One strategy is to assign the end boundary as the point where $HW_{V,f}$ drops to the same level before F. However, in a mobile context, video highlights must be short (<15sec). We have therefore use a fixed offset where pre-event duration is twice the post-event duration.

3. Automatic Content Placement

Given a video highlight, the intent is to locate a region for content placement that would generally be acceptable and not be seen as intrusive. The real-time requirement prevents us from using compute-intensive methods such as optical flow-based motion modeling and object tracking. Hence we look for video entities (eg, tennis white lines and soccer goal mouth) that are easy to detect and track. The reader can refer to [7] for more details. We also segment regions with static overlay (eg, TV station logo, time/score annotation) by detecting pixels with low temporal variation.


Figure 3. Static regions and their location mask

3.1. Static Region Location

In general, static graphical insertions in TV are opaque or semi-transparent. To obtain a static region mask, we use a gradient-based approach by computing edge change over successsive frames. A preliminary static mask S_i is obtained via time-averaging:

$$S_i = G_i + \alpha S_{i-1}$$

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, arising for instance from slight changes from the time-ticker, morphological operations are applied using a horizontal kernel crafted to work well on the station logos and graphical annotations on video data from local broadcast. Figure 3 shows examples of location masks obtained. Using the location mask, we deduce the positions of existing overlays. Adhering to the "Rule of Third" norm, these positions are usually found in the four corners of the frame. Our placement can now be safely over the "unused" corners.

3.2. Locating Post-Event Frames

Even if we have located a spatial region for content placement, we need to time their exposure. It is intuitive that video frames showing the pre- and main event are more important than the post-event frames. The climax has passed and what follows are frames depicting the human emotions, eg, celebrations, agony, crowd applause. The general mood is relax, and content placement, if any, would best happen here.

Sports games are usually played on a field/court with a distinct color tone for easy spectatorship. This may of course vary from venues, and to weather and lighting conditions. We cannot specify a value for the dominant color but assume the existence of one. We compute the statistics of the dominant color in the HSI space by taking the mean value of each color component around their respective histogram peaks, i_{peak}. An interval $[i_{min}, i_{max}]$ is defined around each i_{peak}, where

$$\sum_{i=i\min}^{ipeak} H[i] <= 2H[ipeak] \text{ and } \sum_{i=i\min-1}^{ipeak} H[i] > 2H[ipeak]$$

$$\sum_{i=ipeak}^{imax} H[i] <= 2H[ipeak] \text{ and } \sum_{i=ipeak}^{imax+1} H[i] > 2H[ipeak]$$

$$\text{Color mean} = \sum_{i=imin}^{imax} H[i]*i \Big/ \sum_{i=imin}^{imax} H[i]$$

Color pixels are quantized into 64 hue, 64 saturation and 128 intensity bins. The color mean is converted back into 24-bit RGB (R_{peak}, G_{peak}, B_{peak}) and to determine a given pixel *I(x,y)* is a field color:

$$G(x,y) = \begin{cases} 1, \text{ if} \begin{cases} I_G(x,y) > I_R(x,y) + K(G_{peak} - R_{peak}) \text{ and} \\ I_G(x,y) > I_B(x,y) + K(G_{peak} - B_{peak}) \text{ and} \\ \left|I_R(x,y) - R_{peak}\right| < R_t \text{ and} \\ \left|I_G(x,y) - G_{peak}\right| < G_t \text{ and} \\ \left|I_B(x,y) - B_{peak}\right| < B_t \text{ and} \\ I_G(x,y) > G_{th} \end{cases} \\ 0, \text{otherwise} \end{cases}$$

R_t=10, G_t=20, B_t=10, K=0.9 and G_{th}=80.

Frames with less than a third of *G(x,y)* are deemed to be non-field frames, and hence also post-event frames. We can afford to be strict in imposing consecutive frames in the buffer be consistently classified before insertion. It also has the appealing effect that insertion occurs at the moment of shot change.

4. Mobile Video Highlight Delivery

Amidst its hype and optimism, 3G is arguably still in its infancy. What is an appropriate bit-rate to encode our video highlights and what is the best mode of delivery? While the publicized rate of 3G streaming at 50-64kbps is generally acceptable for talking head, it is only marginal for high motion sports video. Live streaming incurs additional transport overhead and is liable to network breakage. MMS has size limitation at 100KB for 2.5G GPRS, and 300KB for 3G. Encoding at 50kbps, this corresponds to a maximum video length of 16-secs and 48-secs respectively. Email attachment has no such limitations, but we need to watch out for long download time that will kill interest before any viewing gratification can start. Our best bet is that a user can at most wait 30-secs. Over a 2.5 GPRS network, it is the time to download a 90KB 50kbps video attachment to a phone.

The standard video format for mobile network is 3GPP with MPEG-4 or H.263 codec. Every video enabled mobile phone supports this format. There are several free or inexpensive GUI-based encoding tools available for 3GPP format. However we need a command line tool for the communication server to send sport video alert automatically. Command line tools which support 3GPP format are rather expensive now. The price is beyond our budget.

RealPlayer is the default media player on many video enabled mobile phones. It supports 3GPP and the proprietary Real format. We choose Real Producer 10 Plus as our mobile video encoding tool. It has a command line executable and supports 3 video formats: RV8, RV9 and RV10. The video quality of RV9 and RV10 format is better than RV8 for the same video encoding rate, but they require more CPU power. The RV8 format is suitable for a mobile phone with CPU speed of around 100MHz. One such phone is the Nokia-7610, which we are using for our trial.

The emerging H.264 codec is very promising for mobile video applications. For the same encoding rate, H.264 gives a much better video quality. Our next trial is to exploit H.264 on 3G networks.

4.1. Communications Server Architecture

Figure 4 shows our video communications server. We send the video clip as email attachment using

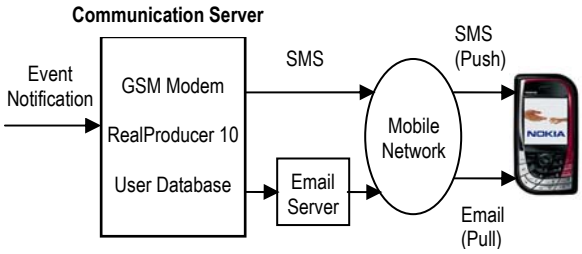

Figure 4. Communications Server

SMTP through a high speed wired network to the email server of the mobile operator. Then the registered user can check out (pull) the email message with video clip attachment. We send (push) a SMS message first through a GSM modem to notify the user that a new video highlight is available. The user is required to check his email box to download the video clip. Normally the SMTP server is free to access.

The tasks of the server are as follow:

a) Wait for new event notification from the Highlight Detection Server.
b) Send a notification SMS to the registered user. This usually reaches the user in 5-10 secs.
c) Re-compress the event video clip from MPEG-1 format (frame-size 352x288, 1150kbps, 25fps) to RV8 format (frame-size 176x144, 50kbps, 6.25fps). Re-compression takes about 1 times real-time on a 1.8GHz CPU.
d) SMTP-send the RV8 video clip to the email-box by a high speed wired network.

5. Experimental results

We conduct subjective viewing tests to evaluate the three key elements in our system: the quality of the automatic highlights produced by our algorithms, visual quality of the 50kbps bit-rate used to encode the highlights, and the overall system usability. 17 subjects from our institute (researchers, engineers and students) are invited for the trial. These are tech-savvy people used to high-speed internet connection at work and in school, and who are also aware of the limitations of mobile network. Each subject is first asked to watch the last 15-mins of the Wimbledon Ladies 2004 Final, and asked whether the automatic highlights are truly representative of the game. At the same time, the video highlights are sent to their mobile phones. After the game viewing, they proceeded to download the video emails and are then asked to evaluate the quality of the 50kbps video playback on the mobile phone.

The cumulative results over 17 viewers are tabulated in Table 1. Most gave a high opinion of the quality of our highlights, and think the 50kbps video is acceptable on the mobile phone. Over all, the subjects accept this prototype system. The minority opinion of opposition came in 2 forms: (1). The duration of 12-secs for highlight is not long enough; (2). Some suggest to improve the video quality further.

Table 1. Experimental results

	Bad	Acceptable	Good
Quality of highlights	11.8%	58.8%	29.4%
Video quality (50kbps)	11.8%	76.5%	11.8%
Overall usability	11.8%	52.9%	35.3%

6. Conclusions

A mobile sports highlight delivery system that integrates advertising content placement is presented. Methods are proposed to minimize viewing disruption caused by the placement. With the onset of 3G and mass appeal of premium sports content, there is now a clear opportunity for new revenue in secondary market for mobile sports highlights.

7. References

[1] J. Wang, C. Xu, E. Chng, K. Wan and Q. Tian, "Automatic Highlight Detection and Replay Generation for Soccer Video" *Proc ACM Multimedia 2004,* pp 31-38.

[2] L. Xie, et.al, "Structure Analysis of Soccer Video with Domain Knowledge and Hidden Markov Models", *Pattern Recognition Letters*, Vol 25, Issue 7, May 2004, pp 767-775.

[3] K. Wan and C. Xu, "Efficient Multimodal Features for Automatic Soccer Highlight Generation", *Proc ICPR 2004*, Vol 3, Aug 23-26, 2004, pp 973-976.

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

[5] Sportsvision. 1st and Ten. http://www.sportvision.com/

[6] G. Medioni, et.al, "Real-time billboard substitution in a videostream", *Proc 10th Tyrrhenian International Workshop on Digital Communications*, Italy, 1998, pp 71-84.

[7] K. Wan, X. Yan, X. Yu and C. Xu, "Real-time goal-mouth detection in MPEG soccer video", *Proc ACM Multimedia 2003*, pp 311-314.

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