A MULTIMODAL SYSTEM FOR AUTOMATIC SPORTS HIGHLIGHTS GENERATION:

"Getting the good parts you want to your Mobile Platform with limited bandwidth!"

Abhijeet Sangwan, Hynek Bořil, Taufiq Hasan, John H. L. Hansen
Center for Robust Speech Systems (CRSS), Department of Electrical Engineering,
The University of Texas at Dallas, Richardson, Texas, U.S.A.

ABSTRACT

We propose a system that can automatically identify exciting events in a sports video and can use this information to automatically generate highlights. The proposed system exploits information in audio, speech and video channels to automatically segment the video to create play-by-play discrete chunks, and subsequently use a number of simple yet efficient multimodal features to assign an excitement score to every play. Using this excitement score, the new system can rank orders plays and generate highlights of desired time. Alternatively, the system can allow users to skip "boring" parts of the video and skip to the next emotional hotspot. This has specific benefits for remote viewing of multi-media content with limited bandwidth resources.

Index Terms— Multimodal Signal Processing, Sports Highlights Generation, Video Segmentation, Speech Excitability Measure

1. INTRODUCTION

We proposed a system to automatically identify exciting events in sports videos, which can be used for automatically highlight generation. The proposed technology can supports multiple applications of interest to different audiences. Content creators can use the proposed system to automatically and efficiently create various reproductions of the original sports video to suit different consumer bases. For example, a longer and shorter highlights edition can be created for viewing during weekends and daily-commute, respectively. Distributors and viewers can use the proposed system in semi-automatic or fully automatic mode to create personal video summaries that could then be shared with other viewers via social networks. For example, given the simplicity of the proposed system, the tool could be embedded as a plugin into popular multimedia websites like youtube.com. Additionally, the technology could also be used to create new user experience by allowing users to automatically navigate to hotspots. For example, multimedia streaming websites like youtube.com already support time-tagging content and the proposed system could use this ability to allow "hot-spot navigation".

2. METHOD

The proposed system workflow is shown in Figure 1. The system contains 4 major modules which accomplish (i) Semantic segmentation of the video, (ii) Video based excitement feature extraction, (iii) Audio and speech based excitement feature extraction, and (iv) Analysis and highlights generation.

The semantic segmentation of the baseball video involves detecting pitching scene that allows the system to cut the continuous video into discrete play-by-play events. Each play starts with pitching and ends at the beginning of the next pitch. A number of simple features that can be extracted at a video frame level are proposed to identify pitching scene [1]. The output of the semantic segmentation module is the start and stop time of each play. A number of video excitement measures are extracted for each play. Particularly, slow motion replay, camera motion activity, and scene cut density are computed. Presence of slow motion replay, higher camera motion and greater scene cut density are typically positively correlated with interesting events in sports videos. In order to exploit speech information, we apply a simple heuristic that the excitement in the commentators voice will be positively correlated to exciting play. Similarly, the loudness of the crowd should also be typically positively correlated with excitement in the game. Based on this assumption, we first employ speech activity detection to separate commentator's speech from pure crowd noise. Subsequently, we extract pitch, first, second and third formants, and spectral center of gravity features from the speech signal. Additionally, we compute loudness measure from the crowd noise.

In the next step, the proposed method estimates user excitability directly from low-level features using their joint-PDF (probability density function) estimated over the game videos. Even when extended videos are not available for training these models, the proposed technique can still extract highlights from a given game

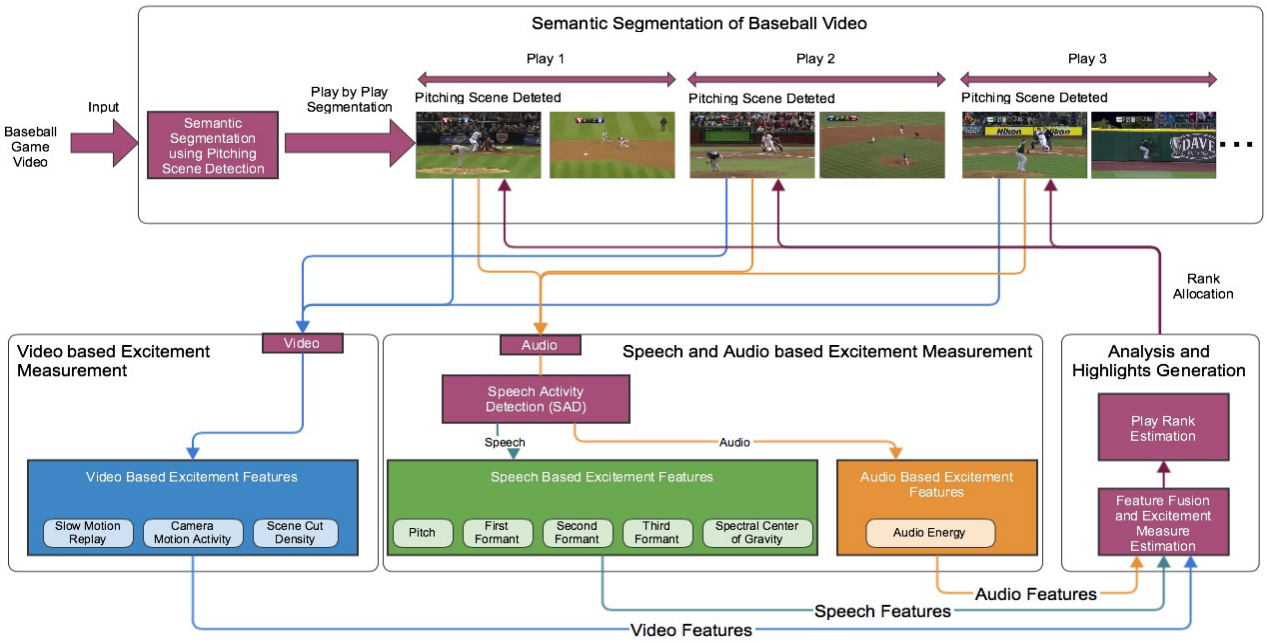

Figure 1: Proposed System for Automatic Sports Highlights Generation

video by estimating the feature PDFs from itself in an unsupervised fashion. An advantage of the proposed method is that it is less affected by extreme values of a single feature due to off-field distractions since the joint behavior of the features is considered in a probabilistic framework. Using the proposed excitability measure, the video segments can be rank-ordered to automatically generate highlights. The technique can also be used to estimate an excitement-time curve to demonstrate user-affective states over a time sequence of the video stream. Interested readers can review [1] for a more detailed description of the underlying algorithms.

3. RESULTS

We use six baseball game videos from the 1975 World Series to evaluate the proposed highlight generation method. In order to evaluate the effectiveness of the proposed measure of excitability, we conducted an independent subjective evaluation involving five viewers familiar with the game of baseball. The subjects were asked to watch the videos and rank the excitability of the scene on a scale from 0 to 30. The rubrics used are boring (0 to 10), moderately exciting (10 to 20), and very exciting (20 to 30). Using the average human ratings, correlation between human and machine scores were computed. The proposed system showed a correlation of 0.8, which was significant at $p < 0.05$. This result demonstrates the power of the proposed system.

4. DEMONSTRATION

The system demonstration will showcase highlights generation and hotspot navigation. The user will first select a baseball video from a menu of options. In highlights generation, the user will specify duration and the system will automatically generate a personal highlight for viewing. In this manner, users will be able to generate multiple highlights that will allow them to compare and contrast the quality of the videos. In hotspot navigation, the user will be able to skip "boring" sections of video using by turning on "skip" functionality. This will allow users to evaluate the benefit of such functionality within their viewing experience.

5. CONCLUSION

A simple yet effective video highlights generation scheme that utilizes audio/speech excitement and low-level video features has been presented. In our experimental studies, the proposed scheme has been shown to outperform state-of-the-art generic excitability ranking methods [1].

6. REFERENCES

[1] Hasan *et al.*: Multi-modal highlight generation for sports videos using an information-theoretic excitability measure. *EURASIP Jrl on Advances in Sig. Proc.* 2013 **2013**:173.

[2] H Pan, P Van Beek, M Sezan, Detection of slow-motion replay segments in sports video for highlights generation, in *Proc. IEEE ICASSP* 7–11 May 2001.

[3] M Delakis et. al.: Audiovisual integration with Segment Models for tennis video parsing. Comput. Vis. Image Underst. **111**(2), 142–154 (2008).