

A MULTI-MODAL HIGHLIGHT EXTRACTION SCHEME FOR SPORTS VIDEOS USING AN INFORMATION-THEORETIC EXCITABILITY MEASURE

Taufiq Hasan, Hynek Bořil, Abhijeet Sangwan and John H. L. Hansen

Center for Robust Speech Systems (CRSS), University of Texas at Dallas, Richardson, TX-75080, USA.

ABSTRACT

A generic method for sports video highlight selection is presented in this study. Processing begins where the video is divided into short segments and several multi-modal features are extracted from each video segment. Excitability is computed based on the likelihood of the features lying in certain regions of their probability density functions that are exciting and rare. The proposed measure is used to rank order the partitioned segment stream to compress the overall video sequence and produce a contiguous set of highlights. Experiments are performed on baseball videos using excitement in the commentators' speech, audio energy, slow motion replay, scene cut density, and motion activity as features. Subjective evaluation of excitability and ranking of video segments yield a higher correlation with the proposed measure compared to well-established techniques indicating the effectiveness of the approach.

1. INTRODUCTION

Automatic video summarization has a wide range of applications in domains such as sports, movies, security, news and on-line video streaming. Application in sports videos is especially important for commercial purposes. Many sports viewers are unable to watch entire matches and are only interested to see the highlights. Automatically generating such highlights reliably can thus be very useful. Many approaches towards automatic event detection and summarization in sports videos have been reported in the literature. Past methods use information from a single modality [1], or combine multiple modalities [2, 3, 4, 5]. Many techniques depend on specific sports [2, 6], video effects [1], or environments. Methods such as [6] attempt to annotate the full game automatically using sophisticated machine learning and domain knowledge, whereas other methods tend to be more generic [7, 3]. In simpler methods such as in [8] for baseball games, the probability of a baseball hit and excited speech is combined to estimate the excitability of a video segment. In general, the generic highlight extraction methods aim at constructing temporal features from different modality that are proportional to or indicative of user excitability. Later some kind of fusion technique is used to generate a single excitement curve providing estimated affective state of the viewer at each point in time/video segment.

In this study we propose a probabilistic feature fusion technique for estimating the user excitability for sports videos. The method is based on a simple but powerful principle of information theory: the less likely an event, the more information it contains. We assume that

This project was funded by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen

interesting parts in a video are rare [4] and therefore have high self-information (also known as the "surprisal"). [9]. This can be intuitively understood as follows: if for a given sports video the ambient noise is always high, then audio energy as an excitement indicator [7] would be unreliable. In this game there is nothing "surprising" in a high audio energy. The proposed method aims at estimating the user excitability directly from low level features using the estimated feature PDFs over the game videos. Even if extended training videos are not available, the proposed technique can still extract the highlights from a single game video by estimating the feature PDFs, provided that the features used are generally related to user excitement. Using the proposed excitability measure, the video segments can be rank ordered and highlights automatically generated.

2. PROPOSED METHOD

The video is first segmented into small blocks for feature extraction. Several features (scalar parameters) are extracted from each segment that are modeled to be generally proportional to the excitement level of the given segment. These features represent long/short term (cumulative) characteristics from different modalities, and used to form higher level descriptors (features) such as, duration of excited speech, average motion activity, etc.

Let the random variable X_i be the i th feature and $x_i(k)$ be an observation of that feature in the k th segment. Since, X_i is in general proportional to the excitability of the video segment, $p(X_i \geq x_i(k))$ will be very low for highly exciting video segments, i.e., they will be rare outcomes for the random event $\{X_i \geq x_i(k)\}$. Therefore, the self-information measure (in bits) associated with the random event $\{X_i \geq x_i(k)\}$ given by [9],

$$\zeta_i(k) = I[X_i \geq x_i(k)] = -\log_2 [p(X_i \geq x_i(k))], \quad (1)$$

will be proportional to excitability. For D feature parameters, we define the random vector $\mathbf{X} = (X_1, X_2, \dots, X_D)$ as the feature vector, and $\mathbf{x}(k)$ as an observation vector in the k th video segment. We can now refine (1) for D dimensions as:

$$\zeta(k) = -\log_2 [p(\mathbf{X} \geq \mathbf{x}(k))], \quad (2)$$

where $\zeta(k)$ is a measure excitability in segment k from D features. Assuming that X_1, X_2, \dots, X_D are independent, we have

$$\zeta(k) = -\log_2 \prod_{i=1}^D p(X_i \geq x_i(k)) = -\log_2 \prod_{i=1}^D \int_{x_i(k)}^{\infty} f_{X_i}(\lambda) d\lambda, \quad (3)$$

where f_{X_i} is the PDF of the i th feature. The idea is illustrated in Fig. 1 with two features X_1 and X_2 . For an observation $\mathbf{x}(k) =$

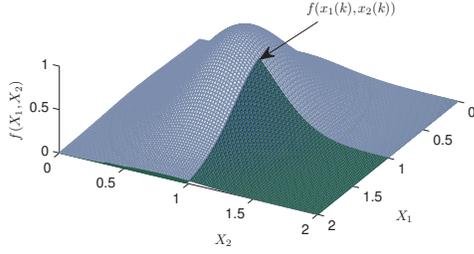


Fig. 1. Conceptual joint PDF of two multimodal features X_1 and X_2 extracted from video segments. Shaded area reveals the high tail region indicating exciting events.

$(x_1(k), x_2(k))$ obtained from the k th segment, the area under the shaded region determines how likely it is that other segments would have higher feature values compared to this observation. The advantage of using the proposed measure is that it not only considers the value of the observation $x_i(k)$ in the k th segment, but also takes into account how likely it is that this feature yields a higher value than $x_i(k)$. Thus $\zeta(k)$ can be used to rank video segments from high to low excitement level.

3. SYSTEM DESCRIPTION

We use six baseball game videos from the 1975 World Series to evaluate the proposed highlight generation scheme. Our highlights video generation depends on a *semantic video segmentation*, though other method of segmentation can also be utilized. We define *semantic segments* as short self explanatory video segments that can be used as building blocks for the highlights video. Examples of such segments can be: play times in soccer games; time interval between each bowling in cricket; times between each pitching in baseball game, *etc.* For our experiments we perform segmentation at the pitching scenes. This is the only part of the implementation which is game dependent.

The notation used from this point forward are as follows: t, k and i denote video frame, video segment, and feature index, respectively. For the i -th feature, $\Phi_i(t)$, $x_i(k)$ and $G_i(t)$ indicate feature value at time t , feature parameter extracted from segment k , and viewer arousal curve at time t estimate as in [7], respectively. The multimodal events/features used for excitability measure: (1) slow motion replay, (2) camera motion activity, (3) scene cut density, (4) commentators' speech in high and (5) low excitement levels, and (6) audio energy. For comparison we also implemented the highlight detection scheme presented in [7].

3.1. Video processing

3.1.1. Slow motion detection

Pixel-wise mean square distance (PWMSD) feature is used for detecting slow motion regions using the zero-crossing method as described in [1]. Since slow motion replay is displayed after some interesting event in sports, we assume that duration of a slow motion shot in the k -th *semantic segment* is proportional to excitability (given the segment is sufficiently long) and thus we use this measure

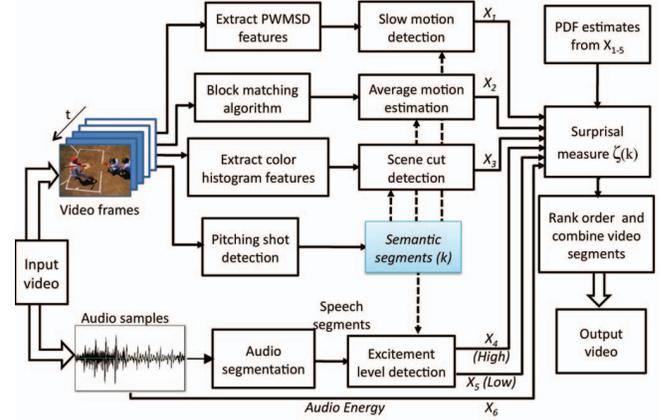


Fig. 2. Proposed system block diagram.

as the feature parameter $x_1(k)$. To obtain $G_1(t)$, we first define the slow motion function as, $\Phi_1(t) = 1$ if slow motion is detected at time t ; or 0 otherwise (Fig. 4 (a)). Next, we filter $\Phi_1(t)$ to obtain $G_1(t)$ to fulfill the “smoothness” criteria required for the method presented in [7]. In general, for the i th feature we use the following filter:

$$G_i(t) = \Phi_i(t) * K(l, \beta) \quad (4)$$

where, $K(l, \beta)$ indicates a Kaiser window of length l and scale parameter β ($l = 500$ and $\beta = 5$ is used).

3.1.2. Camera motion estimation

To detect camera motion, we use a block-matching algorithm similar to [10] for motion vector estimation between successive video frames. The raw motion values are normalized and stored in $\Phi_2(t)$, then smoothed using (4) as before to obtain $G_2(t)$. Segmental feature $x_2(k)$ is computed by averaging $G_2(t)$ across the k -th segment.

3.1.3. Scene-cut density

Scene cut density measure is known to be correlated with excitement in sports videos and is extracted as is [7]. At each video frame index t we compute: $\Phi_3(t) = e^{(1-n(t)-p(t))/\delta}$, where $n(t)$ and $p(t)$ are frame indices of the two nearest scene-cuts to the left and right of the frame t , respectively. Again, we use (4) to obtain $G_3(t)$ from $\Phi_3(t)$ and average $G_3(t)$ over the k -th segment to compute $x_3(k)$.

3.1.4. Pitching scene detection

The pitching times in baseball are very well suited locations for *semantic segmentation*. To detect the pitching scenes, the following operations are performed on each video frame: the field pixels are detected using the HSV color space condition as in [11] and a binary image is formed. Fig. 3 (b) shows an example binary image from a pitching scene. A video frame has to pass four condition tests to be eligible as a pitching scene: (i) area ratio [11] lies in the range of 25% – 45%, (ii) lower half of the image contains more field pixels [11], (iii) strong valley found in the vertical profile of field pixels [11] (as in Fig. 3 (c)) and (iv) higher intensity occurs in specific block regions (7, 11, 10, 14 as shown in Fig.3 (d)) due to the presence of the pitcher and batter [6]. In the evaluation of this algorithm on manually labelled data an accuracy of 80.6% was reached.

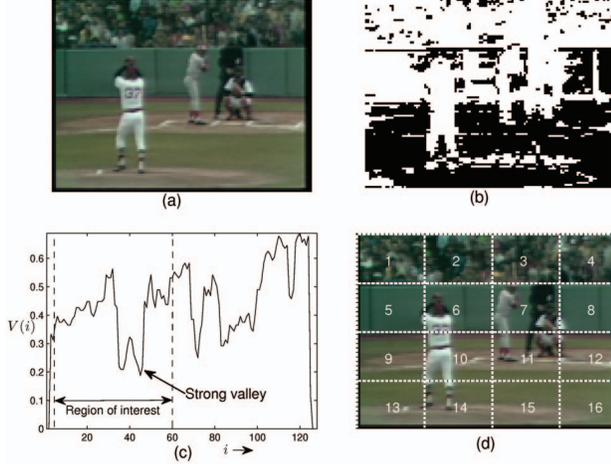


Fig. 3. Pitching scene detection: (a) a sample pitching scene, (b) detected field pixels (shown in black), (c) vertical profile of the field pixels, and (d) 16 blocks dividing the image.

3.2. Audio processing

3.2.1. Speech/non-speech classification

We use an unsupervised non-stationarity measure based speech /non-speech classification scheme as presented in [5]. The approach is based on a long term standard deviation calculated on Mel-filter bank energies of the audio frames, which is sensitive to non-stationarity. Using this measure, speech/background classification is performed using a two mixture Gaussian mixture model. This technique yields a classification accuracy of 80.1% across all 6 baseball game audio.

3.2.2. Excitement measurement in speech

The excitement-level classification in our current system follows the one proposed in [5] based on the analysis of the correlation between speech production parameters and subjectively perceived excitement levels in commentators' speech. In that study, recordings of commentators' speech from six baseball games were labeled by an expert annotator into four subjective perceived excitement levels (*level 1* – no excitement, *level 4* – maximum excitement) and used to train and evaluate a binary excitement level classifier, where excitement levels 1 and 2 were grouped into a *moderate excitement* class and levels 3 and 4 into a *high excitement* class. Similarly as in [5], the current classifier uses mean conversational turn fundamental frequency F_0 , spectral center of gravity SCG , and first three formant center frequencies F_{1-3} normalized by the mean and variance of their distributions found in the training set. The performance of the Gaussian mixture model (GMM) based classifier when trained on 4 baseball games and scored on two open test set games against the perceptual labels reaches equal error rate (EER) of 21–22% in a 3-turn round robin experiment.

To estimate the $G_i(t)$ functions for high and moderate/low excitement in speech, we use the same principle used for slow motion feature. First, we form the function $\Phi_4(t)$, such that $\Phi_4(t) = 1$ if high excitement class was detected at time instant t , and 0, otherwise. Similarly, we form $\Phi_5(t)$ for detected moderate/low excitement class. These functions are shown in Fig. 4 (d) and (e). Next, the corresponding $G_i(t)$ functions are computed following (4). The

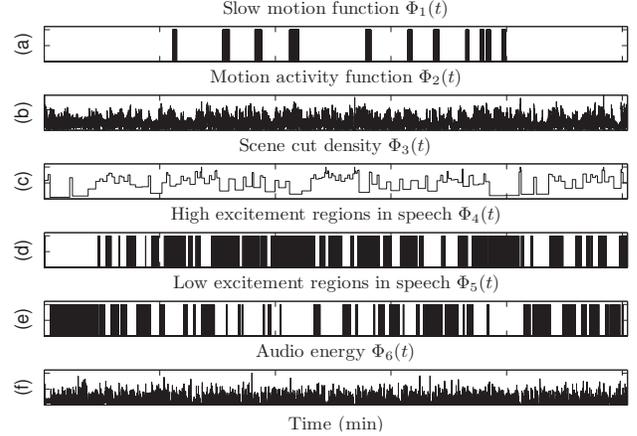


Fig. 4. A time-line view of the detected events/feature functions $\Phi_i(t)$ across a 25 minute segment of a Baseball game video.

only difference here is that the function $\Phi_5(t)$ is inverted before filtering, following the fact that low excitement in speech is inversely proportional to the viewer arousal.

3.2.3. Audio energy measure

For computing audio energy, a fixed audio segment size of 267 samples is used which is equivalent to our video frame-rate of 29.97 frames/sec. For each frame t , audio energy $\Phi_6(t)$ is extracted and filtered using (4) to obtain $G_6(t)$. Averaged value of $G_6(t)$ in the k -th segment is used to compute the segmental features $x_i(k)$.

3.3. Feature fusion and excitement estimation

To generate the highlights time curve in [7], for each t , the functions $G_i(t)$ are filtered using a weighting function

$$w(t) = \frac{1}{2} \left(1 + \operatorname{erf} \left(\frac{\bar{G}_M(t) - d}{\sigma} \right) \right) \quad (5)$$

to obtain $G'_i(t) = G_i(t)w(t)$, $i = 1, \dots, D$, where $D = 6$ is the number of features, the choice of $d = 40$ and $\sigma = 5$ follows [7], and

$$\bar{G}_M(t) = \min_i \left[\operatorname{sort}_i^M (\{G_i(t) | i = 1 \dots D\}) \right]. \quad (6)$$

Here sort_i^M provides top M values of $G_i(t)$ at time stamp t . We use $M = 3$. Finally the highlights time curve is generated as

$$H_M(t) = \left[\frac{\max_t(a(t))}{\max_t(\tilde{a}(t))} \right] \tilde{a}(t) \quad (7)$$

where $a(t) = \sum_i \eta_i G'_i(t)$, $\tilde{a}(t) = K(l, \beta) * a(t)$ and $\sum_i \eta_i = 1$. To estimate excitability in a segment k , we use the averaged values of $H_M(t)$ in that segment to obtain $\bar{H}_M(k)$.

For the proposed excitability measure, first the multi-modal feature vector $\mathbf{x}(k) = (x_1(k), x_2(k) \dots x_D(k))$ is computed for each segment k . Next, the histogram of the $x_i(k)$ values across all segments of the video is used to estimate the PDFs $f_{X_i}(\lambda)$. The excitability measure $\zeta(k)$ from each segment can then be computed using (3). For generating highlight videos, segmentation is performed using the detected pitching shot locations. The proposed

measure can then be used to rank order and combine the video segments according to the user defined highlights duration. Sample highlights generated using the proposed technique can be found in <http://sites.google.com/site/icassp2012highlights/>.

4. SUBJECTIVE EVALUATION

In order to evaluate the effectiveness of the proposed measure of excitability, we conducted a subjective evaluation involving five viewers familiar with the game of baseball. To make the evaluation feasible, we selected 100 pitching scenes (semantic segments) of about 15 seconds duration from a baseball game. The subjects were asked to watch the videos and rank the excitability of the scene in a scale of 0-30. The rubrics used are: boring (0-10), moderately exciting (10-20), and very exciting (20-30). Since the dynamic range of the ranking scores may vary across subjects, the scores obtained from each subject are normalized to zero mean and unity variance, and then averaged to obtain the final subjective measure. Note that the normalization here does not affect the ranking order of the segments decided by the subject, i.e., order from lowest ranking to highest ranking ones will be preserved. Next, we extract excitement estimation in the selected 100 video segments obtained from (i) $\bar{H}_M(k)$ computed using [7], (ii) $\gamma(k)$ computed from a simple addition and normalization of the feature parameters $x_i(k)$ [5], and (iii) the proposed measure $\zeta(k)$. The measures are evaluated by comparing their correlation and ranking agreement with the subjective measure.

From Fig. 5, we observe that all of the selected feature parameters $x_i(k)$ are correlated with the subjective scores with positive values, i.e., they are proportional to excitability. The proposed excitability measure $\zeta(k)$ is seen to have the highest correlation coefficient of 0.801, which outperforms the other two excitability measures $\gamma(k)$ and $\bar{H}_M(k)$, yielding correlations of 0.606 and 0.702, respectively. In each correlation value in Fig. 5, the probability of the null hypothesis is $p < 0.01$, i.e., the correlation is significant. To evaluate objective rank ordering of the segments, we choose N top subjectively ranked video segments as a ground truth for highlights. The excitement ranking measures $\bar{H}_M(k)$, $\gamma(k)$ and $\zeta(k)$ are then evaluated with respect to true positive rate (TPR) and false positive rate (FPR). From the results summarized in Fig. 6, again the proposed measure provides superior performance compared to [7] and [5] for most values of N . As an example, for $N = 20$ the proposed technique selects 70% of the top 20 subjectively ranked segments with a false positive rate of 7.5%, while for [7] the TPR and FPR values are 55% and 11.2%, respectively.

5. CONCLUSION

A generic video summarization scheme based on an information theoretic measure of excitability was presented. The proposed measure was shown to effectively combine multi-modal features and estimate the self-information measure in each video segment, which is proportional to its excitableness. The measure was also found to be highly correlated with a subjective evaluation of excitability in individual video segments and outperformed well-established excitability ranking methods. The proposed scheme was applied on baseball videos to generate meaningful highlights.

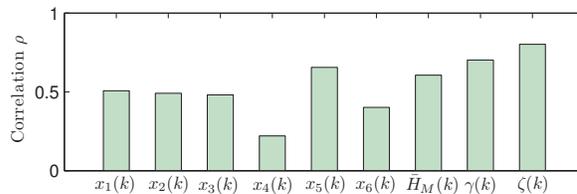


Fig. 5. Correlation coefficients of different feature parameters/objective excitement measures and subjective evaluation scores

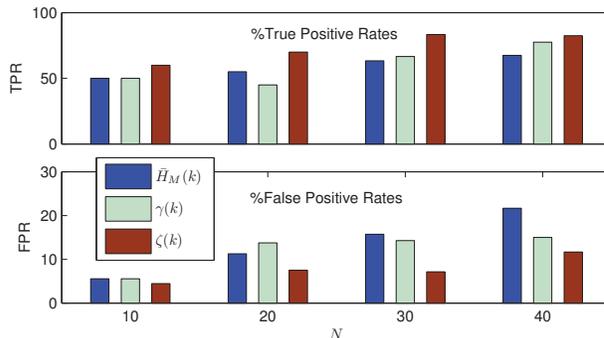


Fig. 6. Performance evaluation by comparing top N highlight segments selected using subjective and objective measures. True positive and false positive rates are shown.

6. REFERENCES

- [1] H. Pan, P. Van Beek, and M. Sezan, "Detection of slow-motion replay segments in sports video for highlights generation," in *Proc. IEEE ICASSP*, 2001.
- [2] M. Delakis, G. Gravier, and P. Gros, "Audiovisual integration with Segment Models for tennis video parsing," *Computer vision and image understanding*, vol. 111, no. 2, pp. 142–154, 2008.
- [3] M. Fleischman, B. Roy, and D. Roy, "Temporal feature induction for Baseball highlight classification," in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007, pp. 333–336.
- [4] Z. Xiong, R. Radhakrishnan, and A. Divakaran, "Generation of sports highlights using motion activity in combination with a common audio feature extraction framework," in *Proc. IEEE ICIP*, 2003.
- [5] H. Bofil, A. Sangwan, T. Hasan, and J. H. L. Hansen, "Automatic excitement-level detection for sports highlights generation," in *Proc. Interspeech'10*, Makuhari, Chiba, Japan, September 2010, pp. 2202–2205.
- [6] C. Lien, C. Chiang, and C. Lee, "Scene-based event detection for baseball videos," *J. of Visual Comm. and Image Representation*, vol. 18, no. 1, pp. 1–14, 2007.
- [7] A. Hanjalic, "Adaptive extraction of highlights from a sport video based on excitement modeling," *Multimedia, IEEE Transactions on*, vol. 7, no. 6, pp. 1114 – 1122, 2005.
- [8] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," in *Proceedings of the eighth ACM international conference on Multimedia*. ACM, 2000, p. 115.
- [9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience, 1991.
- [10] H. Liu, W.-j. Zhang, and J. Cai, "A fast block-matching algorithm based on variable shape search," *Journal of Zhejiang University - Science A*, vol. 7, pp. 194–198, 2006, 10.1631/jzus.2006.A0194.
- [11] W. Chu and J. Wu, "Explicit semantic events detection and development of realistic applications for broadcasting baseball videos," *Multimedia Tools and Applications*, vol. 38, no. 1, pp. 27–50, 2008.