# Identification of Narrative Peaks in Video Clips: Text Features Perform Best

Joep J.M. Kierkels[1,2], Mohammad Soleymani[2], Thierry Pun[2]

[1] Department of medical physics, TweeSteden hospital,
5042AD Tilburg, the Netherlands
[2] Computer vision and multimedia laboratory (CVML)
Computer Science Department, University of Geneva,
Battelle Campus, Building A, 7 Route de Drize
CH – 1227 Carouge, Geneva, Switzerland
jkierkels@tsz.nl, { mohammad.soleymani, thierry.pun }@unige.ch

**Abstract.** A methodology is proposed to identify narrative peaks in video clips. Three basic clip properties are evaluated which reflect on video, audio and text related features in the clip. Furthermore, the expected distribution of narrative peaks throughout the clip is determined and exploited for future predictions. Results show that only the text related feature, related to the usage of distinct words throughout the clip, and the expected peak-distribution are of use when finding the peaks. On the training set, our best detector had an accuracy of 47% in finding narrative peaks. On the test set, this accuracy dropped to 24%.

## 1 Introduction

A challenging issue in content-based video analysis techniques is the detection of sections that evoke increased levels of interest or attention in viewers of videos. Once such sections are detected, a summary of a clip can be created which allows for faster browsing through relevant sections. This will save valuable time of any viewer who merely wants to see an overview of the clip. Past studies on highlight detection often focus on analyzing sports-videos [1], in which highlights usually show abrupt changes in content features. Although clips usually contain audio, video, and spoken text content, many existing approaches focus on merely one of these [2;3]. In the current paper, we will attempt to compare and show results for all three modalities.

The proposed methodology to identify narrative peaks in video clips was presented at VideoCLEF 2009 subtask on "Affect and Appeal" [4]. The clips that were given in this subtask were all taken from a Dutch program called "Beeldenstorm". They were in Dutch, had durations between seven and nine minutes, consisted of video and audio, and had speech transcripts available. Detection accuracy was determined by comparison against manual annotations on narrative peaks provided by three annotators. The annotators were either native Dutch speakers or fluent in Dutch. Each annotator chose the three highest affective peaks of each episode.

While viewing the clips, finding clear indicators as to which specific audiovisual features could be used to identify narrative peaks was not straightforward, even by

looking at the annotations that were provided with the training set. Furthermore we noticed that there was little consistency among the annotators because more than three narrative peaks were indicated for all clips. This led to the conclusion that tailoring any detection method to a single person's view on narrative peaks would not be fruitful and hence we decided to work only with basic features. These features are expected to be indicators of narrative peaks that are common to most observers, including the annotators.

Our approach for detecting peaks consists of a top-down search for relevant features, e.g., first we computed possibly relevant features and secondly we investigated which of these features really enhanced detection accuracy. Three different modalities were separately treated.

First, video, in MPEG1 format, was used to determine at what place in the clip frames showed the largest change compared to a preceding key frame. Second, Audio, in MPEG layer 3 format, was used to determine at what place in the clip the speaker has an elevated pitch or has an increased speech volume. Third, ext, taken from the available metadata xml files in MPEG 7 format, was used to determine at what place in the clip the speaker introduced a new topic. Next to this, the expected distribution of narrative peaks over clips was considered. Details on how all these steps were implemented are given in Section 2, followed by results of our approach on the given training data in Section 3. Discussions over the obtained results and evaluations are given in Section 4. In Section 5 several conclusions are drawn from these results.

In the VideoCLEF subtask, the focus of detecting segments of increased interest is on the data part, e.g., we analyze parts of the shown video-clip to predict their impact on a viewer. Even though there exists a second approach to identify segments of increased interest. This second approach focuses not on the data but directly on the reactions of a viewer, e.g., by monitoring his physiological activity such as heart-rate [5] or by filming his facial expressions [6]. Based on such reactions, the affective state of a viewer can be estimated and one can estimate levels of excitation, attention and interest in a viewer [7]. By themselves, physiological activity measures can thus be used to estimate interest, but they could also be used to validate the outcomes of data-based techniques.

## 2 Feature extraction

For the different modalities, feature extraction will be described separately in the following subsections. As the topic of detecting affective peaks is quite unexplored, only basic features were implemented. This provides an initial idea of which features are useful, and future studies could focus on enhancing the relevant basic features. Feature extraction was implemented using MATLAB (Mathworks Inc).

## 2.1 Video features

Our key assumption for video features was that dramatic tension is related to big changes in video. It is a film editors' choice to include such changes along time [8], and this may be used to stress the importance of certain parts in the clip. The proposed narrative peak detector will output a 10 s window of enhanced dramatic tension from videos with the frame rate of 25 frames per second and this precision level is too large and merely slows down computations. Hence only the key frames (I-frames) are treated.

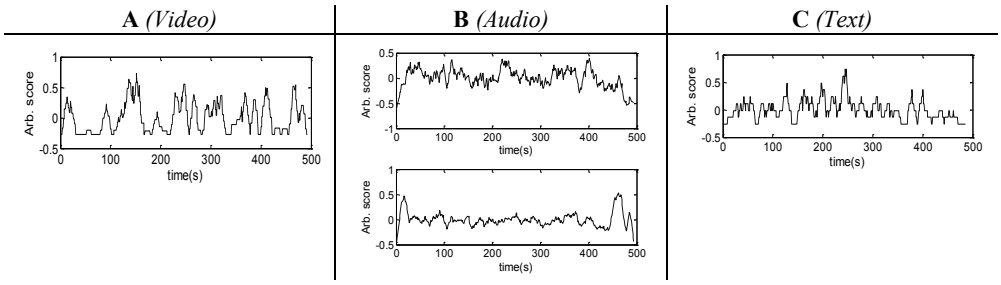


**Fig. 1.** Illustration of single modality feature values computed over time. A: Video feature, B: Audio features, C: Text feature. All figures are based on the episode with the identification code (ID), "BG_37016".

## 2.2 Audio features

The key assumption for audio was that a speaker has an elevated pitch or has an increased speech volume when applying dramatic tension, as suggested in [9;10]. The audio is encoded at 44.1 kHz sampling rate in mpeg layer 3 format. The audio signals only contain speech except a short opening and ending credits at the start and the end of each episode. The audio signal is divided in 0.5 s segments for which the average pitch of the speaker's voice is computed by imposing a Kaiser window and applying a Fast Fourier Transform. In the transformed signal, the frequency with maximum power is determined and is assumed to be the average pitch of the speaker's voice over this window. Next the difference in average pitch between subsequent segments is computed. If a segment's average pitch is less than 2.5 times as high as the pitch of the preceding segment, its pitch value is set to zero. This way, only those segments with strong increase in pitch (supposed indicator of dramatic tension) are kept.

Speech volume is determined by computing the averaged absolute value of the audio signal within the 0.5 s segment. As a final step again, the resulting signals for pitch and volume are both smoothed by averaging over a 10 s window, and the smoothed resulting signal is scaled to have a maximum absolute value of one and subsequently to have a mean of zero. Next, they are down-sampled by a factor 2, resulting in vectors *audio1* and *audio2* which both contain 1 value per second as is illustrated in Fig. 1B.

### 2.3 Text features

The main assumption for text is that dramatic tension starts by the introduction of a new topic, and hence involves the introduction of new vocabulary related to this topic. Text transcripts are obtained from the available metadata xml files. The absolute occurrence frequency for each word was computed. Words that occurred only once were considered to be non-specific and were ignored. Words that occurred more than five times were considered too general and were also ignored. The remaining set of words is considered to be topic specific. Based on this set of words, we estimated where the changes in used vocabulary are the largest. A vector $\underline{v}$ filled with zeros was initialized, having a length equal to the number of seconds in the clip. For each remaining word, its first and last appearance in the metadata container was determined and was rounded off to whole seconds, subsequently all elements in $\underline{v}$ in between the elements corresponding to these obtained timestamps are increased by one. Again, the resulting vector $\underline{v}$ is averaged over a 10 s window, scaled and set to zero mean. The resulting vector text is illustrated in Fig. 1C.

### 2.4 Distribution of narrative peaks

A clip is directed by a program director and is intended to hold the attention of the viewer. To this end, it is expected that points of dramatic tension are distributed over the duration of the whole clip, and that not all moments during a clip are equally likely to have dramatic tension.
For each dramatic tension-point as indicated by the annotators, its time of occurrence was determined (mean of start and stop timestamp) and a histogram, illustrated in Fig. 2, was created based on these occurrences. Based on this histogram, a weighting vector $\underline{w}$ was created for each recording. Vector $\underline{w}$ contains one element for each second of the clip. Each element's value is determined according to the histogram.

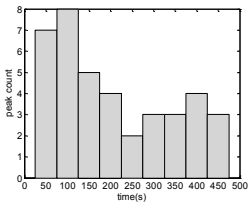


**Fig. 2.** Histogram that illustrates when dramatic tension-points occur in the clips according to the annotators. Note that during the first several seconds there is no tension-point at all.

### 2.5 Fusion and selection

For fusion of the features, our approach merely consisted in giving equal importance to all used features. After fusion, the weights vector $\underline{w}$ can be applied and the final indicator of dramatic tension $\underline{drama}$ is derived as (shown for all three features):

$$drama = \underline{w} \cdot \left( video + \frac{(audio1 + audio2)}{2} + text \right)^T . \quad (2)$$

The estimated three points of increased dramatic tension are then obtained by selecting the three maxima from drama. The three top estimates for dramatic points are constructed by selecting the intervals starting 5s before these peaks and ending 5s afterwards. If either the second or third highest point in drama is within 10s of the highest point, the point is ignored in order to avoid having an overlap between the detected segments of increased dramatic tension. In those cases, the next highest point is used (provided that the new point is not within 10s)

**Table 1.** Schemes for feature combinations.

| Scheme number | Used features | Weights | Scheme number | Used features | Weights |
|---|---|---|---|---|---|
| 1 | Video | Yes | 5 | Video, Text | Yes |
| 2 | Audio | Yes | 6 | Audio, Text | Yes |
| 3 | Text | Yes | 7 | Video, Audio, Text | Yes |
| 4 | Video, Audio | Yes | 8 | Text | No |

## 3 Evaluation schemes and results

Different combinations of the derived features were made and subsequently evaluated against the training data. The schemes tested are listed in table 1. If no weights are used (Scheme 8) vector $\underline{w}$ contains only ones.

Scoring of evaluation results is performed based on agreement with the reviewers' annotations. Each time a peak that was detected coincides with (at least) one reviewer's annotation, a point is added. A maximum of three points can thus be scored per clip and since there are five clips in the training set, the maximum score for any scheme is 15. The obtained scores are shown in table 2.

**Table 2.** Results on the training sets. The video ID codes in the dataset start by "BG_".

| Scheme number | BG_36941 | BG_37007 | BG_37016 | BG_37036 | BG_37111 | **Total** |
|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 1 | 1 | **3** |
| 2 | 2 | 1 | 1 | 1 | 1 | **6** |
| 3 | 2 | 1 | 1 | 2 | 1 | **7** |
| 4 | 0 | 1 | 2 | 1 | 1 | **5** |
| 5 | 1 | 2 | 2 | 1 | 0 | **6** |
| 6 | 2 | 1 | 1 | 2 | 1 | **7** |
| 7 | 1 | 1 | 2 | 1 | 0 | **5** |
| 8 | 0 | 1 | 1 | 1 | 0 | **3** |

## 4 Discussion

As can be seen in table 2, the best performing schemes on training samples are scheme 3 and scheme 6 which both result in 7 accurately predicted narrative peaks and hence an accuracy of 47%. These two schemes both include the text based feature and the weights vector. Scheme 6 also contains the audio based feature but fails to achieve an increased accuracy because of this inclusion. Considering that there is also strong disagreement between annotators, an accuracy of 47% (compared against the joint annotations of three annotators) shows the potential of using the automated narrative peak detector. The fact that this best performing scheme is only based on a text based feature corresponds well to the initial observation that there is no clear audiovisual characteristic of a narrative peak when observing the clips. Five schemes have been evaluated using the test samples mainly corresponding to some of the different schemes that were previously used in table 1. The results of these five methods on the test-data, and their explanations are given in table 3. For number 5, all narrative peaks were randomly selected (for comparison with random level detection). Evaluation of these runs was performed in two ways: Peak-based (similar to the scoring system on the training data) and Point-based which can be explained as follows; If a peak that is detected coincides with annotations of more than one reviewer annotation, multiple points are added. Hence the maximum-maximum score for a clip can be nine when annotators fully agree on segments, the minimum-maximum score remains three when annotators fully disagree.
The difference between the two scoring system lies in the fact that the Point-based scoring system awards more than one point to segments which were selected by more than one annotator. If annotators agree on segments with increased dramatic tension, there will be (in total over three annotators) less annotated segments and hence the probability that by chance our automated approach selects an annotated segment will decrease. Therefore, awarding more points to the detection of these less probable segments seems logical. Moreover, a segment on which all annotators agree must be a really relevant segment of increased tension. On the other hand, this Point-based approach gives equal points to having just one correctly detected segment in a clip (annotated by all three annotators) and to detecting all three segments correctly (each of them by one annotator).
Since our runs were selected based on the results that were obtained using the Peak-based scoring system, results on the test data are mainly compared to this scoring.

First of all, it should be noted that results are never far better than random level, as can be seen by comparing to run number 5. Surprisingly, the Peak-based and Point-based scores show a distinctly different ranking of the runs. Run 1 performed the worst under the Point-based scoring, yet it performed best under the Peak-based scoring system. Based on the results obtained on the clips in the test set, it was expected that runs 1 and 3 would perform best. This is clearly reflected in the results we obtain when using the same evaluation method on the test clips, the Peak-based evaluation. However, with the Point-based scoring system this effect disappears. This may indicate that the main feature that we used, the text based feature based on the introduction of a new topic, does not reflect properly the notion of dramatic tension for all annotators, but is biased towards a single annotator.

Each video clip in the dataset was only annotated for its top three narrative peaks. The lack of a fully annotated dataset with all possible narrative peaks, made it difficult to study the effect of narrative peaks on low level content features. Having all the narrative peaks at different levels on a larger dataset, the correlation between the corresponding different low level content features could have been computed. The significance of these features for estimating narrative peaks could therefore have been further investigated.

**Table 3.** Results on the test set.

| run number | (scheme nr) | Score (Peak-based) | Score (Point-based) |
|---|---|---|---|
| 1 | 3 | 33 | 39 |
| 2 | 7 | 30 | 41 |
| 3 | 6 | 33 | 42 |
| 4 | 8 | 32 | 43 |
| 5 | -- | 32 | 43 |

## 5 Conclusions

The narrative peak detection subtask described in the VideoCLEF 2009 Benchmark Evaluation has proven to be a challenging and difficult one. Failing to see obvious features when viewing the clips and only seeing a mild connection between new topics and dramatic tension peaks, we resorted to the detection of the start of new topics in the text annotations of the provided video clips and the use of some basic video- and audio-based features. In our initial evaluation based on the training clips, the text based feature proved to be the most relevant one and hence our submitted evaluation-runs were centered on this feature. When using a consistent evaluation of training and test clips, the text based feature also led to our best results on the test data. The overall detection accuracy based on the text-based feature dropped from 47% correct detection on the training data to 24% on the test data. It should be stated that results on the test data were just mildly above random level. The randomly drawn results by chance performed better than random level. The simulated random level results are 40 for the point based and 30 for the peak based scoring schemes.
The reported results based on the Point-based scoring differed strongly from the results obtained using the scoring system that was employed on the training data. It was shown that although using the peaks distribution as a data driven method enhanced the results on the training data the same approach cannot be generalized due to its bias toward the annotations on the training samples.

In fact, the number of narrative peaks is unknown for any given video. The most precise annotation of such documentary clips can be obtained from the original script writer and the narrator himself. Not having access to these resources, more annotators should annotate the videos. These annotators should be able choose freely any number of narrative peaks. To improve the peak detection, a larger dataset is needed to compute the significance of correlations between features and narrative peaks.

Given the challenging task that was given, it is our strong belief that the indication that text based features (related to the introduction of new topics) perform well, is a valuable contribution in the search for an improved dramatic tension detector.

**Acknowledgments.** The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2011] under grant agreement n° 216444 (see Article II.30. of the Grant Agreement), NoE PetaMedia. The work of Soleymani and Pun is supported in part by the Swiss National Science Foundation.

## References

1. Hanjalic A.: Adaptive extraction of highlights from a sport video based on excitement modeling. IEEE Transactions on Multimedia. 7(6), 114--1122 (2005)
2. Gao, Y., Wang, W. B., Yong, J. H., Gu, H. J.: Dynamic video summarization using two-level redundancy detection. Multimedia Tools and Applications 42(2), 233--250 (2009)
3. Otsuka, I., Nakane, K., Divakaran, A., Hatanaka, K., Ogawa, M.: A highlight scene detection and video summarization system using audio feature for a Personal Video Recorder. IEEE Transactions on Consumer Electronics. 51(1), 112--116 (2005)
4. Larson, M., Newman, E. and Jones, G.J.F.: Overview of VideoCLEF 2009: New Perspectives on Speech-based Multimedia Content Enrichment. In: Multilingual Information Access Evaluation Vol. II Multimedia Experiments. LNCS, Springer (2010)
5. Soleymani, M., Chanel, G., Kierkels, J. J. M., Pun, T.: Affective Characterization of Movie Scenes Based on Multimedia Content Analysis and User's Physiological Emotional Responses. In: IEEE International Symposium on Multimedia, (2008)
6. Valstar, M. F., Gunes, H., Pantic, M.: How to Distinguish Posed from Spontaneous Smiles using Geometric Features. In: ACM Int'l Conf. Multimodal Interfaces (ICMI'07), (2007)
7. Kierkels, J. J. M., Pun, T.: Towards detection of interest during movie scenes. In: PetaMedia Workshop on Implicit, Human-Centered Tagging (HCT'08), Abstract only, (2008)
8. May, J., Dean, M. P., Barnard, P. J.: Using film cutting techniques in interface design. Human-Computer Interaction. 18(4), 325--372 (2003)
9. Alku, P., Vintturi, J., Vilkman, E.: Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation. Speech Communication. 38(3--4), 321-334 (2002)
10. Wennerstrom, A.: Intonation and evaluation in oral narratives. Journal of Pragmatics. 33(8), 1183-1206 (2001)