Identification of Narrative Peaks in Clips:
Text Features Perform Best.

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

[1] University of Geneva, Computer Science Department
Battelle Building A, 7 Route de Drize
CH – 1227 Carouge, Geneva, Switzerland
{Joep.Kierkels, 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%.

Keywords: Feature detection, Video analysis, Attention.

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 video-clips or documentaries. Once such sections are detected, other less relevant sections may be removed from a recording. By doing so, 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 specific patterns related to ball or player movement. 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 show results for all three modalities and try to identify which modality is actually the most valuable in the detection of segments which supposedly include narrative peaks.

For our participation in the VideoCLEF 2009 subtask on "Affect and Appeal" [4], we propose in this paper a methodology to identify narrative peaks in video clips. The clips that were used in this subtask were all taken from a Dutch program called "Beeldenstorm". They were in Dutch language, had duration 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 (Dutch speaking) annotators.

While viewing the clips, we failed to see any clear indicators as to which specific audiovisual features could be used to identify narrative peaks, even when looking at the annotations that were provided. 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 our belief 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. We expect these features 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 establish which of these features really enhance detection accuracy. We separately treated three different modalities.

- Video, taken from the available .mpg files, was used to determine at what place in the clip frames showed the largest change compared to a preceding frame.
- Audio, taken from an mp3 conversion of the .mpg frames, was used to determine at what place in the clip the speaker has an elevated pitch or has an increased speech volume.
- Text, taken from the available .mp7 files, was used to determine at what place in the clip the speaker introduced a new topic.

Next to this, we considered the expected distribution of narrative peaks over clips. 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. In Section 4 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 it is outside the scope of the subtask, it is worth to mention that there exists a second approach to identifying 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. Because the evaluation in the VideoCLEF task will be against the explicit annotations provided by the three annotators we did not include recordings of physiological activity in our VideoCLEF contribution.

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, we decided to implement only basic features. 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.
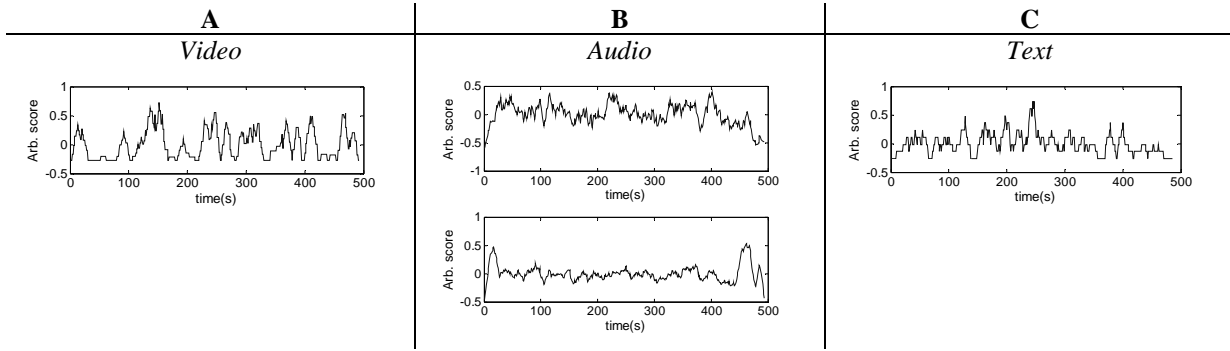
2.1 Video features

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 change between different frames [8], and we believe that this may be used to stress the importance of certain parts in the clip. Video is recorded at 25 frames/second. Because our narrative peak detector will output a 10 s window of enhanced dramatic tension, this precision level which is too large and merely slows down computations, furthermore frame changes are often not too obvious on subsequent frames. Our treatment of video-frames starts at frame 1 and subsequently jumps about 0.5 s to frame 13, frame 25, and so on.

All frames are converted to grayscale levels. The difference between subsequent frames is computed as:

$$\Delta_{vid} = \sum_{n=1}^{N} \sum_{m=1}^{M} \left| F_{13}(n,m) - F_1(n,m) \right|, \tag{1.1}$$

in which N and M are the width and the height of the frames and F_{13} indicates the matrix containing pixels values of the 13th frame. As a next step, the change in Δ_{vid}, indicated as $d\Delta_{vid}$, over subsequent observations is determined and compared against a threshold value that reflects how much change in this value is observed when a scene changes. This threshold was determined on the training-set and set to $5*10^5$. If $d\Delta_{vid}$ is below this threshold, it is set to zero. As a final step, the resulting $d\Delta_{vid}$ is 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, it is down-sampled again by a factor 2, resulting in vector _video_ which contains 1 value per second as is illustrated in Fig. 1A.

Figure 1. Illustration of single modality feature values computed over time. A: Video feature, B: Audio features, C: Text feature. All figures are based on 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]. Audio is recorded at 44.1 kHz. 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 *audio*1 and *audio*2 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 .mp7 files. A histogram was computed for all unique words that occurred in the clip, counting their number of occurrences. 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 *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 .mp7 file was determined and was rounded off to whole seconds, subsequently all elements in *v* in between the elements corresponding to these obtained timestamps are increased by one. Again, the resulting vector *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 *w* was created for each recording. Vector *w* contains one element for each second of the clip. Each element's value is determined according to the histogram.
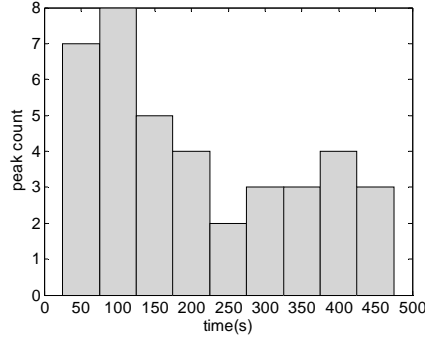
Figure 2. Histogram that illustrates when dramatic tension-points occur in the clips according to the annotators. Note that during the first several seconds there is no tension-point at all.

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):

$$drama = \underline{w} \cdot \left(video + \frac{(audio1 + audio2)}{2} + text \right)^{T}. \qquad (1.1)$$

The estimated three points of increased dramatic tension are then obtained by selecting the three maxima from $\underline{drama}$. Our estimates for the three 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 $\underline{drama}$ 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)

3 Evaluation schemes and Results

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

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

Scoring of evaluation results is performed based on agreement with the reviewers' annotations. Each time a peak that we 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 in training sets.

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**

3 Discussion

As can be seen in table 2, the best performing schemes 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 our initial observation that there is no clear audiovisual characteristic of a narrative peak when observing the clips. All non-Dutch speaking observers failed to see indicators of narrative peaks. The observation that narrative peaks seem to correspond to the introduction of new topics in the clip can only be made when an observer understands the spoken content. It is expected that this observation is also true for clips in other languages.

In our contributions to VideoCLEF, we included five runs, mainly corresponding to some of the different schemes that were previously used in table 1. The results of our runs on the test-data, and their explanations are given in table 3. For number 5, all narrative peaks were randomly selected (for comparison).

Evaluation of these runs was performed in two ways: Peak-based (similar to our scoring system on the training data) and Point-based which can be explained as follows; If a peak that we 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 then 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 to scoring 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). When considering that annotators may have different tastes and that one annotator could fully disagree with the other two annotators, it is unsatisfying that a system that would fully resemble this annotator's taste gets rewarded only the same number of points as a system that predicts merely one correct segment. In this view, a 100% correspondence with a human annotator should lead to an optimal score for this clip, since the program cannot be expected to outperform the people that are hired to evaluate it.

Because 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.

Table 3. Results on the test set.

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

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. In Fig. 3, the score is shown when calculated based on only the annotations of a single annotator. It should be noted that in this setting, the Point-based and Peak-based scoring system are identical.
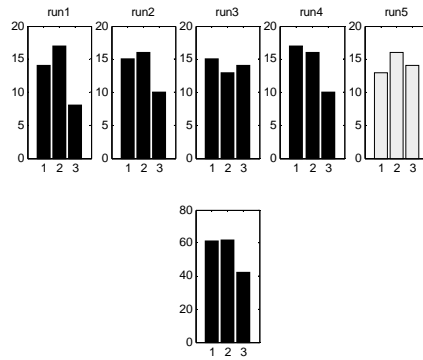
Figure 3. Histograms that illustrates scoring based on single annotators (x axis). The upper row of histograms indicate scoring (y axis) for the different runs. The lower histogram shows the joint histogram of all runs except run 5, which involved random selection of segments.

In this figure, it can be seen that scoring based only on annotator 3 performed significantly worse than scoring based only on annotator 1 or annotator 2. This indicates that our approach is biased towards the opinions of some reviewers.

Because our runs were selected based on those schemes that performed best under Peak-based metric, one should mainly compare to the Peak-based results on the test-clips. Knowing now the exact scoring system that was eventually employed in the VideoCLEF scoring, it would be recommendable to also compute the results on the training set using the Point-based metric. Possible, this would lead to a different ranking of the evaluated schemes that is more similar to the ranked results VideoCLEF reported on the Point-based evaluation. However, such a reflection on the training data was not performed because of the strict deadlines for submitting working notes (seven working-days after releasing the results).

---

[a] The Peak-based score reported here deviates slightly from the official Peak-based score reported by the VideoCLEF organizers.

In the VideoCLEF score, nearby peaks of different annotators were merged in order to create a new 10s segment. This implies that when an estimated peak is nearby a peak detected ONLY by annotator 2, this estimated peak will score points. However, if annotator 1 or 3 also selected a nearby peak, the interval in which points can be scored shifts and the same estimated peak may not score a point anymore. Since the second situation actually involves an estimated peak which is not far from two annotated peaks, it seems contradictory to the authors to not award points here. For evaluating the training data, points were awarded under similar circumstances and hence we did award points to such peaks also for the test data.

4 Conclusions

The 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 centred around 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 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 this is probably caused by a bias of our method towards the annotations given by annotators 1 and 2.

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.

Acknowledgement

References

1. Liu, C. X., Huang, Q. M., Jiang, S. Q., Xing, L. Y., Ye, Q. X., Gao, W.: A framework for flexible summarization of racquet sports video using multiple modalities. Computer Vision and Image Understanding. 113(3), 415--424 (2009)
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. Website, http://www.cdvp.dcu.ie/VideoCLEF/
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), pp. 38--45. (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)