First Approaches to Automatic Boredom Detection: DMIR tackles the MediaEval 2010 Affect Task

Yue Shi and Martha Larson
Multimedia Information Retrieval Lab
Delft University of Technology
{y.shi, m.a.larson}@tudelft.nl

ABSTRACT

We propose five simple approaches to automatically predict the level of boredom/engagement experienced by a viewer when watching a video. The approaches are based on characteristics of the video that reflect humor, "cuteness", dynamism, interactivity and popularity. Our results suggest that the task is feasible, but difficult, with the best results achieved by the method that uses humor in the form of the count of laughter events automatically detected in the video.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Experimentation

Keywords

Boredom detection, Affect, Video Annotation, Benchmark

1. INTRODUCTION

A video search engine returns ranked lists of video results in response to a user information need expressed as a query. When the user is looking for specific subject matter (i.e., a video on a particular topic or sub-topic), the ranking function of the search engine is designed to target material treating that subject. However, when multiple videos would fit the user need in this respect, it becomes interesting to try to improve user satisfaction with the output of the video search engine by incorporating information that is "orthogonal to topic" into the ranking function. The MediaEval 2010 Affect Task focuses on one sort of information of this kind, namely viewer-reported boredom. In this paper, we propose five simple approaches that can be used to predict the level of boredom that users report while watching video.

Our assumption is that some videos are inherently more boring or more engaging than others. We conjecture that viewer perceived boredom is related to the style and the quality of the scripting, editing and production of the video. Although personal preferences vary, without doubt, from viewer to viewer, we are interested in those aspects of a video that contribute to wide consensus among viewers as to whether a video is engaging or boring. We carry out the work reported here within the larger framework of our video recommendation and retrieval research, with an eye to future work that will incorporate reported-boredom and other related scores into algorithms that improve user satisfaction with video recommendation and retrieval results.

The MediaEval 2010 Affect Task involves automatic prediction of viewer-reported boredom for a data set consisting of a series of short travelogue episodes that were created by filmmaker Bill Bowles (http://www.mynameisbill.com/) and published on the internet as a video journal. This material was chosen because of the broad universal appeal of the topic of travel. Further, limiting the video corpus to a single topical domain means that it is easier to factor out the influence of the topic of the video and to isolate the reaction of the viewer to aspects of the video that are "orthogonal to topic." The videos in the data set were annotated by viewers who reported the level or boredom/engagement that they felt while watching the video by assigning it a score. Each video is only 3-5 minutes in length, which is advantageous because it is feasible for annotators to watch each video in its entirety before assigning the boredom score. The annotations were created by making use of the crowdsourcing platform Mechanical Turk (http://www.mturk.com). We used a procedure we designate "high commitment crowdsourcing" [3] that is designed in order to make it possible to recruit a well-qualified, demographically balance worker set from the larger worker pool and motivate the recruited workers to change their crowdsourcing behavior from the standard piecemeal-work patterns to a committed-work pattern that is extended over time and is focused on a single task necessary in order to annotate the data set.

The videos in the data set are divided into a development set containing 42 videos and a test set containing 82 videos. Each video is accompanied by speech recognition transcripts, generated by the the speech recognition system described in [4], and by shot segmentation information, generated by the system described in [2]. Further details on the data set and how it was developed are available in [3]. We made use of the videos in the development set in order to design and optimize our boredom prediction algorithms. Results are reported on the videos in the test set. The boredom scores that are assigned by the annotators are used in order to create a ranking of the test set videos, our approaches have the goal of reproducing this ground-truth ranking. The approaches are evaluated by measuring the correlation between the ranking that we generate and the ground-truth ranking. Similar evaluation methodology is applied in [5].

We first describe each of our algorithms in turn, then we report results and finally we offer a conclusion and outlook.

MediaEval 2010 Workshop, October 24, 2010, Pisa, Italy

2. APPROACHES

We propose five approaches, each based on a conjecture concerning which characteristics of a video make it engaging or boring for viewers. Note that we do not claim that these characteristics are completely unrelated to the topic of the video, but rather that the disassociation of the characteristic and the topic is large enough in order to make the characteristics potentially helpful for refining the ranking of the output of a video recommendation or retrieval system.

2.1 Cuteness (`cute`)

We conjecture that viewers are more engaged by playful or charming videos. We build the `cute` approach on the idea that cuteness is related to the presence of children and animals. A video receives two points for each shot it contains that depicts a child (person under age 12) and an additional one point if that shot also contains an animal. Videos with the most points are ranked the least boring. Videos with no children or animals are ranked randomly. We used the MTurk crowdsourcing platform in order to determine whether a shot contains a child or an animal. In order to automatize the method without any human intervention, it would be necessary to use an automatic concept detector. Annotating concepts with the crowdsourcing platform gives us an estimation of whether our definition of "cute" is viable.

2.2 Dynamism (`dynam`)

We conjecture that fast-paced videos are more engaging. We take sense of motion to be reflected by the relative number of shots. Videos are ranked by the the number of shots they contain, normalized by their length in seconds.

2.3 Humor (`humor`)

We conjecture that if people depicted in the video are laughing, then there is situation depicted in the video that viewers might also find humorous, heightening their sense of engagement. Videos are ranked by the raw count of laugher events that they contain. Laughter is detected by the speech recognition system [4].

2.4 Interactivity (`interact`)

We conjecture that the video is engaging based on how much the narrator interacts with the audience. A similar conjecture is made in [1]. Videos are ranked by a score that summarizes how much Bill appears in the video, and a selection of properties describing *how* he appears, including whether he is up close speaking to the camera or whether he appears further away. As with `cute`, the annotation of the shots for `interact` was carried out using MTurk.

2.5 Popularity (`pop`)

We conjecture that videos about popular locations will be more engaging for viewers. Popularity is taken to be reflected in the amount of information related to the video's location that is available on the Internet. In order to estimate this amount, we took the location description string of the video (made available in the metadata) as being representative of its location and the number of hits returned when this location was submitted as a query to Google as being representative of the amount of information available on the topic. Note that `pop` is not "orthogonal to topic", rather it helps us gauge the extent to which we can safely ignore topic for this task.

Table 1: Correlation on the test set between the ground-truth ranking and the ranking generated by the proposed approaches

Run	Kendall τ	p-value
`cute`	0.04	0.6
`dynam`	0.16	0.04
`humor`	0.19	0.01
`interact`	0.16	0.04
`pop`	0.1	0.2

3. RESULTS AND CONCLUSION

Our results (cf. Table 1) are reported in terms of the correlation between the videos in the test set as ranked by the ground truth scores and as ranked by our aproaches. No approach yields a strong correlation with the ground-truth ranking, however the fact that we do achieve weak correlations suggests that it is indeed feasible to automate the prediction of viewer-reported boredom, at least within the context of this data set. The best performing approach is `humor`, a result that is particularly interesting in light of the fact that the laughter detection of the speech recognizer was very good, but not perfect. This result supports the view that it is not necessary for a feature to be perfect in order to be useful to refine video results lists. Future work will involve further exploration of `dynam` and `interact` focused on refining the mechanism used to combine the evidence and generate the score by which the video was ranked. Optimizing this process might yield a better ranking. Finally, `pop` may also hold future promise. Most location description strings contained a proper name of a place, but others are simply descriptive of the type of place, e.g., "in the office." Treating place names differently from commons nouns holds potential to improve this approach.

Acknowledgments The research leading to these results has received funding from the European Commission's 7th Framework Programme (FP7) under grant agreement no. 216444 (EU PetaMedia Network of Excellence).

4. REFERENCES

[1] B. Jochems, M. Larson, R. Ordelman, R. Poppe, and K. Truong. Towards affective state modeling in narrative and conversational settings. In *Interspeech*, pages 490 – 493, 2010.

[2] P. Kelm, S. Schmiedeke, and T. Sikora. Feature-based video key frame extraction for low quality video sequences. In *Image Analysis for Multimedia Interactive Services, 2009. WIAMIS '09. 10th Workshop on*, pages 25 –28, 2009.

[3] M. Soleymani and M. Larson. Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*, 2010.

[4] A. Stolcke, X. Anguera, K. Boakye, Ö. Çetin, A. Janin, M. Magimai-Doss, C. Wooters, and J. Zheng. The SRI-ICSI Spring 2007 meeting and lecture recognition system. In *Multimodal Technologies for Perception of Humans*, Lecture Notes in Computer Science, pages 450–463. Springer, 2008.

[5] Y.-H. Yang and H. Chen. Music emotion ranking. In *ICASSP*, pages 1657 –1660, 2009.