Travelogue Boredom Detection with Content Features

Mohammad Soleymani
Computer Vision and Multimedia Laboratory
University of Geneva
Switzerland
mohamamd.soleymani@unige.ch

ABSTRACT

In this working note, a set of features are proposed for ranking videos according to the felt boredom in users or boredom ranking. The boredom ranking can be used as a feature for video recommendation. A series of travelogue videos was used as the dataset. The features were from different modalities, namely, video, audio, and speech transcript. The amount of information that a given episode brings and the fame of places are proposed as relevant features. The boredom scores were estimated using a linear regression and relevance vector machine (RVM.). It is shown that the amount of information a video delivers to a viewer can decrease the perceived boredom.

Keywords

Video recommendation, ranking, regression, boredom detection.

1. INTRODUCTION

Users often perceive boredom when watching online content. In the case of feeling boredom the users skip the content or stop watching videos. Having estimation about the level of boredom a user will experience watching a video helps a recommendation system to deliver more entertaining material. The goal of this work is to rank the video based on an estimation of boredom.

This work has been done as a part of the affect task of Mediaeval benchmarking forum 2010. Detailed task description, the dataset and annotation methodology are given in the paper by Soleymani and Larson [7].

Boredom detection from video content falls into the category of video affect representation. Video affective represneration requires understanding of the intensity and type of user's affect while watching a video. Hanjalic et al. [3] introduced "personalized content delivery" as a valuable tool in affective indexing and retrieval systems. In order to represent affect in video, they first selected video- and audio- content based features based on their relation to the valence-arousal space that was defined as an affect model (for the definition of affect model, see Section 1.3) [3]. Then, arising emotions were estimated in this space by combining these features.

Soleymani et. al proposed a scene affective characterization using a Bayesian framework [6]. Arousal and valence of each shot were first determined using a linear regression. Then arousal and valence values in addition to content features of each scene were used to classify every scene into three classes. The three emotional classes were calm, excited positive and excited negative. The Bayesian framework was able to incorporate the movie genre and the predicted emotion from last scene or temporal information to improve the classification accuracy.

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

2. METHODS AND MATERIAL

2.1 Dataset

The dataset selected for the developed corpus is Bill's Travel Project, a travelogue series called "My Name is Bill" created by the film maker Bill Bowles (http://www.mynameisbill.com/). Each video is annotated by multiple annotators with boredom scores on nine point scale. The average boredom score given by participants of the preliminary study served as the ground truth for this benchmarking challenge. First 42 videos were released and used for training. The remaining 80 videos were served as the evaluation.

The dataset consists of information from different modalities, namely, visual information from video, speech transcripts, audio signals, titles and publication dates.

2.2 Content Features

The low level content features were extracted from audio and video signals, namely, key lighting, color variance, motion component zero crossing rate, audio energy. A detailed description of the content features can be found in [6]. Shot boundaries were detected using the method described in [5].

Video length, shot change rate and variation (standard deviation and skewness), number of shots and average shot length were extracted using the detected shot boundaries.

2.3 Proposed features

The speech transcripts were provided by a software implemented originally for speech recognition in meetings [8]. Using WordNet [2] the nouns and nouns which could describe a country, place or land were first extracted as location names. Each noun was checked to see if it has a Wikipedia page in English. The number of nouns indexed in Wikipedia was counted as an information related feature. This was extracted as an indicator of the amount of information each transcript carries. The sum of the length of all nouns' Wikipedia pages was also extracted to represent the information significance of the content. Location fame score was computed by averaging the Wikipedia page size of all the location nouns. The number of location nouns was another feature in this class. Based on the scores given to the development set, information transfer, fame score, video length, number of shots and shot change related features were formed the set of proposed features.

2.4 Regression and Support Vector Ranking

In order to determine the boredom score estimation, the regression weights were computed by means of a linear relevance vector machine (RVM) from the Tipping RVM toolbox [9]. A support vector ranking was used to rank the videos based on the selected features. SVMlight was used in the implementation of support vector ranking in this work [4].

3. RESULTS

First, the correlation between low level content features with the development set was studied. Then the features with significant Spearman correlation were chosen in the feature set. Finally, five different runs were generated by combining the new proposed features and the selected content features.

Table 1. Content features with significant ranking correlation with boredom scores in the development set.

Feature	Kendall's Tau correlation
Average of the third Mel-Frequency cepstral coefficients (MFCC)	-0.24
Average of the 10th MFCC	0.21
Average of the third MFCC	0.24
Standard deviation of the third coefficient of the autocorrelation of MFCC	0.24
Video key lighting	0.23
Average of 16th bin of the Luminance histogram (out of 20 bins)	0.25
Average of 17th bin of the Luminance histogram	0.23

Table 2. Ranking evaluation results for all the 5 submitted runs and random level.

Run	Kendall Tau ranking correlation		Spearman ρ		Kentall Tau raking distance	Spearman footrule distance
	r	p	ρ	p		
Random level	-	-	-	-	1660	27.4
1	0.05	0.45	0.09	0.41	1603	25.5
2	0.05	0.48	0.08	0.48	1558	24.9
3	**0.13**	**0.07**	**0.19**	**0.08**	**1700**	**23.2**
4	0.10	0.17	0.14	0.20	1650	25.2
5	**0.10**	**0.18**	**0.14**	**0.19**	**1439**	**24.4**

3.1 Evaluation Criteria

Four ranking distance metric were used to evaluate the boredom ranking results. The Kendall's Tau ranking correlation, Kendall's Tau ranking distance, Spearman ρ and Spearman footrule distance. The details about these metrics is available in [1].

3.2 Support Vector Ranking Results

The first two submissions was obtained using support vector ranking. The first run was done by the combination of the selected content features and proposed features. The second run only used the proposed features. The development set was used to train the support vector ranking. The results of the support vector ranking did not have any significant correlation.

3.3 Regression Results

The three following submitted runs 3,4,5 used the regression results. The third run used the proposed feature and not the content features. The fourth run used all the content features and proposed features together (217 features). Finally the last run used the combination of the selected content features and proposed features. The best results were obtained from regression using the proposed set of features and the combination of selected content features with proposed features. None of the generated ranked lists on the test set had significant ranking correlation with the ground truth p<0.05).

4. CONCLUSIONS

A set of features for the ranking of felt boredom in video are proposed. The boredom ranking results were evaluated using Kendal Tau's ranking correlation and ranking distances. The fame score, amount of information and shot segmentation information are proposed to be useful for boredom ranking.

5. ACKNOWLEDGMENTS

I acknowledge Bill Bowels for granting permission to use his videos for our research. The automatic speech recognition transcripts were kindly provided by ICSI and SRI International and I want to acknowledge their generous help for providing the speech transcripts. For the shot detection, I also thank Sebastian Schmiedeke and Pascal Kelm at the Technische Universitaet Berlin. This work is supported in part by the Swiss National Science Foundation and in part by the European Community's Seventh Framework Programme [FP7/2007-2011] under grant agreement Petamedia n° 216444.

6. REFERENCES

[1] Fagin, R.; Kumar, R.; Sivakumar, D. 2003 Comparing top k lists. Siam Journal on Discrete Mathematics 17, 1 (2003), 134-160.

[2] Fellbaum, C. 1998 WordNet: an electronic lexical database: Cambridge, MA: MIT Press.

[3] Hanjalic, A.; Xu, L. Q. 2005 Affective video content representation and modeling. IEEE Transactions on Multimedia 7, 1 (2005), 143-154.

[4] Joachims, T. 1999. Making large-scale support vector machine learning practical. In: Schölkopf, B.; Burges, C. J. C.; Smola, A. J. eds. Advances in kernel methods: support vector learning. MIT Press;169-184.

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

[6] Soleymani, M., Kierkels, J. J. M., Chanel, G., and Pun, T. 2009. A Bayesian Framework for Video Affective Representation., International Conference on Affective Computing & Intelligent Interaction (Amsterdam, the Netherlands, Sept. 11 09). ACII 2009. 267-273.

[7] Soleymani, M. and Larson, M. 2010. Crowdsourcing for Affective Annotation of Video: Development of a Viewer-reported Boredom Corpus., Workshop on CS SIGIR 2010 (Geneva, Switzerland, 2010).

[8] Stolcke, A., Anguera, X., Boakye, K., Cetin, O, Janin, A, Magimai, M., Wooters, C., and Zheng, J 2008. The SRI-ICSI Spring 2007 Meeting and Lecture Recognition System. In *Multimodal Technologies for Perception of Humans* 450-463.

[9] Tipping, M. E. 2001 Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research 1, 3 (2001), 211-244.