Examining Multimodal Characteristics of Video to Understand User Engagement

Fahim A. Salim, Killian Levacher, Owen Conlan, and Nick Campbell

CNGL/ADAPT Centre, Trinity College Dublin, Ireland
salimf@tcd.ie,killian.levacher@scss.tcd.ie,owen.conlan@scss.tcd.ie,
nick@tcd.ie

Abstract. Video content is being produced in ever increasing quantities and offers a potentially highly diverse source for personalizable content. A key characteristic of quality video content is the engaging experience it offers for end users. This paper explores how different characteristics of a video, e.g. face detection, paralinguistic features in the audio track, extracted from different modalities in the video can impact how users rate and thereby engage with the video. These characteristics can further be used to help segment videos in a personalized and contextually aware manner. Initial experimental results from the study presented in this paper provide encouraging results.

Keywords: Personalization Multimodality Video Analysis Paralinguistic User Engagement

1 Introduction

Videos are one of the most versatile forms of content in terms of multimodality we consume on a regular basis. They are available in ever increasing quantities. It is therefore useful to identify engagement in videos automatically for variety of applications. E.g. in [8] the presenter did statistical analysis on TED talks to come up with a metric for creating an optimum TED talk based on user ratings, while [3] and [7] try to create video recommender systems based on users viewing habits and commenting patterns.

Each kind of video engages users differently, i.e. engagement with content is context dependent [1]. For our context, by engagement we mean the elaborate feedback system used by the raters of TED videos described in detail in section 2.

We believe that it is possible to extract quantifiable multimodal features from a video presentation automatically and correlate them with user engagement criterion for variety of interesting applications. Additionally extracting and indexing those features and their correlation to engagement could impact content adaptability and contextualizing search and non-sequential video slicing. This paper discusses an initial multimodal analysis experiment. In order to test our hypothesis, we extracted features from TED videos and correlated them with user feedback scores available on the TED website.

2 Current Study

TED website asks viewers to describe the video in terms of particular words in-stead of a simple like or dislike option. A user can choose up to three words from a choice of 14 words (listed in table 1) to rate a video. This makes our problem more interesting for now we do not have a crisp binary feedback to learn from, but a rather fuzzy description of what viewers thought about a particular video. The rating system for user feedback gives us a detailed insight of user engagement with the video presentation. Since it is voluntarily information by users in terms of semantically positive and negative words, it provides good basis to analyze relevant factors of engagement described in [6].

Among the 14 rating criterion in table 1 provided to the user, we could identify 9 of them as being positive words, 4 of them being negative (shown in bold) words and 1 as neutral (shown in Italic).

Table 1. Average number of user ratings per each rating criteria for 1340 Ted videos across different topics.

Rating	**Avg. (Count)**	**Avg.(%)**	**Rating**	**Avg.(Count)**	**Avg.(%)**
Beautiful	120	6.67	Inspiring	384	18.16
Confusing	15	1.17	Jaw-dropping	118	5.45
Courageous	122	6.08	**Longwinded**	28	2.23
Fascinating	234	12.64	**Obnoxious**	23	1.62
Funny	106	4.73	*OK*	65	4.88
Informative	246	15.24	Persuasive	188	9.70
Ingenious	134	7.64	**Unconvincing**	51	3.73

As seen in Table 1, ratings tend to be overwhelmingly positive. In order to normalize them we used the following definitions for our purpose for a video to be considered “Beautiful” or “Persuasive” etc. it must have a rating count more than average rating count for that particular rating word. With this, TED talks were categorized as “Beautiful and not Beautiful”, “Inspiring and not Inspiring”, “Persuasive and not Persuasive” etc. giving two classes for classification for each of the 14 rating words.

To perform our experiment we extracted, for how many seconds there was a close up shot of the speaker and when there was a distant shot and when the speaker was not on the screen. For non-visual features we looked at number of laughter and applauses (counted from transcribed audio track i.e. subtitle files) and laughter applause ratio within TED talks. We use HAAR cascades [4] in OpenCV library [2] to detect when the speaker is on the screen or not and a simple python script to get laughter and applause count since TED talks come with subtitles files.

For correlating features with user ratings to see some potential patterns we utilized the WEKA toolkit which allows easy access to a suite of machine learning techniques [5]. We used Machine Learning algorithm Logistics Regression, and tenfold cross-validation testing for our analysis on 1340 TED talk videos, to see how feature values affected user ratings. We tested on both percentage count and actual count for each ratings.

3 Experiment Results and Discussion

Our aim is to see the value in the multimodality of video content. We experimented by removing our visual features to see if this will affect the correct classification of video for the ratings. Figure 1 shows that the accuracy of correctly classified instances increased with the inclusion of visual features for the majority of rating words, 7 to be precise. While for 3 it remained equal but for 4 rating words it actually decreased.

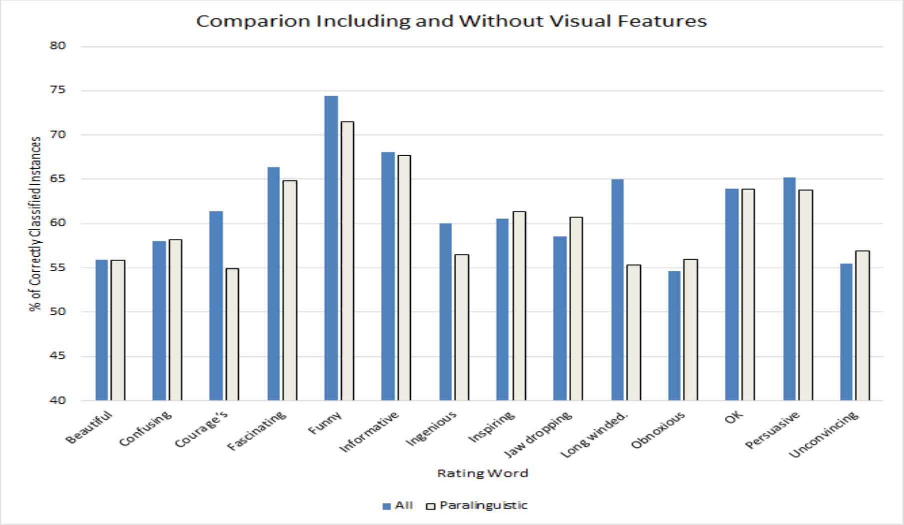

Fig. 1. Comparison with and without visual features for classification of Ratings.

Results of our study are interesting in many regards. Firstly, it is the preliminary step towards our thesis about the value of different modalities within a video stream. Another interesting aspect of this study is that all the features were automatically extracted, i.e. no manual annotation was performed. So any model based on our feature set could be easily used for new content and any advancement in computer vision and paralinguistic analysis technology would help in making our model better. This model would become a component of personalization systems to enhance contextual quires.

Our current approach however is not without any limitations. The biggest of them is that it cannot be used for all type of videos such as movies and sports videos etc. Another limitation of our approach is that we have analyzed the video as a whole unit i.e. we simply do not have information on which portion of video was more “Funny” or “Beautiful” or “Long-winded” compared to other portions.

For further investigations we would like to extend our signal extraction to a focused set of features. We are planning to extract more visual features to see what impact they have on user ratings. In addition to visual features we would also like to introduce paralinguistic features from the audio stream to the fold. Most importantly we are planning to see the correlation between linguistic features of TED talks and their corresponding user ratings. This extracted meta-data and engagement analysis will feed into a model to create multimodal segments in a personalized and contextually aware manner.

4 Acknowledgement

This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at School of Computer Science and Statistics, Trinity College, Dublin.

References

1. Attfield, S., Piwowarski, B., Kazai, G.:Towards a science of user engagement. WSDM Workshop on User Modelling for Web Applications, Hong Kong (2011).
2. Bradski, G.:The OpenCV Library. Dr. Dobb. J. Soft. Tool. (2000).
3. Brezeale, D., Cook, D.J.: Learning video preferences using visual features and closed captions. IEEE Multimed. 16, 3947 (2009).
4. Lienhart, R., Maydt, J., Lienhartintelcom, R.: An extended set of Haar-like features for rapid object detection. Proceedings. Int. Conf. Image Process. 1, 900903 (2002).
5. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software. ACM SIGKDD Explor. 11, 1018 (2009).
6. OBrien, H.L., Toms, E.G.: Examining the generalizability of the User Engagement Scale (UES) in exploratory search. Inf. Process. Manag. 49, 10921107 (2013).
7. Tan, S., Bu, J., Qin, X., Chen, C., Cai, D.: Cross domain recommendation based on multi-type media fusion. Neurocomputing. 127, 124134 (2014).
8. Wernicke, S.:Lies, damned lies and statistics (about TEDTalks).(2010).