INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Visual, Laughter, Applause and Spoken Expression Features for Predicting Engagement within TED Talks

Fasih Haider[1], *Fahim A Salim*[1], *Saturnino Luz*[2], *Carl Vogel* [1], *Owen Conlan* [1], *Nick Campbell*[1]

[1]ADAPT Centre, Trinity College Dublin, Ireland
[2]IPHSI, University of Edinburgh, UK
{haiderf,salimf,vogel,owconlan,nick}@scss.tcd.ie, S.Luz@ed.ac.uk

Abstract

There is an enormous amount of audio-visual content available on-line in the form of talks and presentations. The prospective users of the content face difficulties in finding the right content for them. However, automatic detection of interesting (engaging vs. non-engaging) content can help users to find the videos according to their preferences. It can also be helpful for a recommendation and personalised video segmentation system. This paper presents a study of engagement based on TED talks (1338 videos) which are rated by on-line viewers (users). It proposes novel models to predict the user's (on-line viewers) engagement using high-level visual features (camera angles), the audience's laughter and applause, and the presenter's speech expressions. The results show that these features contribute towards the prediction of user engagement in these talks. However, finding the engaging speech expressions can also help a system in making summaries of TED Talks (video summarization) and creating feedback to presenters about their speech expressions during talks.

Index Terms: presentation quality, video summarization, user engagement detection, expressive speech analysis, non-verbal behaviour analysis, TED talks

1. Introduction

In terms of multimodality, videos are one of the most versatile forms of content which people consume on a regular basis. Take YouTube as an example: over a billion hours of video content are watched daily[1]. Because of the amount of video content available, it is becoming increasingly difficult for users to find desired content. A recent study reports that an average American would spend more than a year over a lifetime looking for something to watch on TV[2]. One criterion for filtering content is how engaging the video is. Hence, we believe that a model to detect user perceptions of engaging vs. non-engaging videos would be beneficial for any number of applications, including video recommendation and video search. The set of videos available to viewers is very diverse, and each kind of video engages users differently or to put it differently, users watch different types of videos for various reasons, i.e. engagement with content is context dependent [1].

Here we focus on one video genre which is video presentations such as TED talks. To distinguish between engaging versus non-engaging TED talks, we first need to define the meaning of user engagement within the context of TED talks and how to quantify it. In the literature, the quality of user (human) experience with a system is called user engagement [2, 3] and a six-factor based matrix is proposed by OBrien et al. [4] for user engagement. In terms of video content, researchers have described user engagement with video content in a variety of ways, e.g. duration for which a user watches a video [5, 6] and subjective evaluation of user engagement through questionnaires [7, 8]. It can be seen that there is not much agreement in measuring engagement due to its highly context dependent nature. For this study, we describe engagement using the elaborate feedback system (described in detail in section 3).

In this paper, we propose a novel approach to detect user engagement with TED talks. The proposed approach is based on the hypothesis that multimodal features can be extracted automatically from TED videos and be correlated to user engagement criterion for a variety of interesting applications. The experimental results discussed in this paper are an extension of the experiment reported by Salim et al. [21], in which we extracted high-level features such as close up and distant shots of the speaker, and the number of instances of laughter and applause in the video. That previous study worked only on a full video; i.e., it could not identify which segments within the video were more engaging compared to others. In the present study, in addition to the high-level feature set of Salim et al. [21], we are also using spoken expressions of the presenter. We first extracted segments of TED talks using speech segmentation using Lium Toolkit [10]. Clustering was performed on the resulting dataset of segments. After clustering the dataset, classification test using LDA and statistical analysis was performed to identify the relationship between the clusters and user engagement.

The experimentation shows promising results (A-weighted F-score for engagement detection and Kruskal–Wallis test for speech expressions evaluation) in terms of:

- Identifying engaging vs non-engaging TED talks.
- Within a TED talk, we are able to identify speech segments which contributed towards that engagement.

In terms of contribution, this paper proposes the following

- A classification model based on multimodal features to identify engagement in TED talks.
- A method to identify segments within a talk that are more engaging than others.

The proposed system architecture is depicted in the Figure 1, where the user (a presenter, potential viewer or a video summarization tool) obtains feedback about a talk. The feedback is in the form of video segments (engaging and non-engaging parts of talk) and the predicted label. As the audio-visual information are correlated, we also provide users visual information during a speech segment.

2. State of the Art

Wernicke conducted statistical analysis on TED talks and proposed a metric for creating an optimal TED talk based on user

[1]https://www.engadget.com/2017/02/27/youtube-one-billion-hours-watched-daily/ – last verified: March 2017.

[2]http://gizmodo.com/is-it-a-bad-thing-that-we-spend-1-3-years-of-our-lives-1788632578 – – last verified: May 2017.

http://dx.doi.org/10.21437/Interspeech.2017-1633

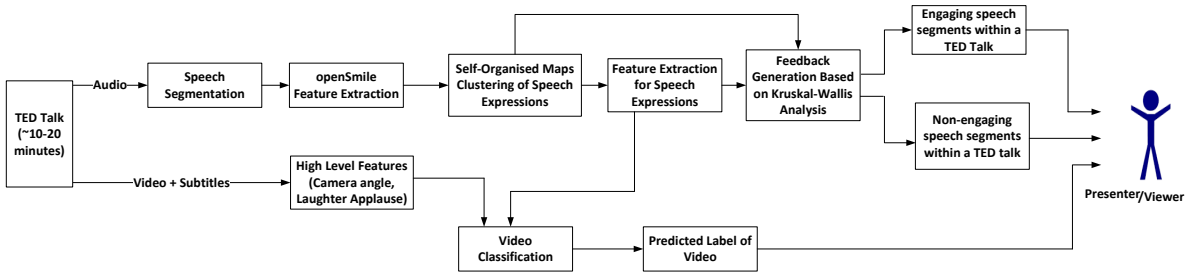


Figure 1: *System Architecture.*

ratings [11]. A major difference between his study and our work is that the TED user ratings on which the former study is based were considerably simpler i.e. viewers could simply 'like' or 'dislike' a particular TED talk. We deploy a more comprehensive rating system (detailed in section 3). Recommender system development for viewers based on their viewing/listening preferences and commenting patterns has attracted considerable interest. For example, Tan et al. use heterogeneous data from different sources to create a recommender system based on user video preferences [12]. Brezeale et al. use movie subtitles and low-level visual descriptors to cluster the data (videos), and then use Hidden Markov Model (HMM) to learn the sequence of clusters to predict the users' preferences [13]. Anwar et al. proposed to file videos into different categories using the caption text and visual features [14].

A significant amount of research is currently being conducted within the field of video summarization. In video summarization, importance is mostly attributed to visual features [7, 15]. However, multimodal features are also receiving considerable attention due to the added value they bring in terms of identifying important chunks. An example of comprehensive multimodal feature extraction is the work of Evangelopoulos et al. who take advantage of all three visual, audio and linguistic modalities to create video summaries [16]. Other interesting examples of multimodal feature extraction are work of Dong et al. [17] and Haesen et al. [8]. Extracting all these multimodal features and indexing them would make videos more searchable and would also help correlate videos with user feedback, and thereby user engagement. In terms of assessing the quality of a presentation, there have been many studies which focus on the analysis of speech utterances and body gestures [18, 19]. The Multimodal Learning Analytics (MLA) dataset contains student presentations graded by teachers in terms of body language, self-confidence, loudness level, eye contact and content [20]. While presentation rating in the educational context is a well-researched area, the factors affecting presentation ratings by ordinary viewers and listener have hardly been investigated.

In our work, TED talk data and user feedback provide us with the opportunity to evaluate the presentation from a viewer's perspective, not a teacher's perspective. In previous work, TED talk data and ratings were used to detect the engagement level using low-level acoustic features and some high-level visual features (camera shots durations and angle, the numbers of laughter and applause) [9, 21, 22]. However, those approaches are not able to generate feedback for presenters. Although it is found that high-level features are correlated with user ratings, this is based on the assumption that the professional camera operators focus on facial expressions and body language using those camera angles. We ultimately aim to develop a system that will support our objective of personalised and contextually aware video slices. The video summarization literature is worth investigating to this end. Although the objective of this study is not to create video summaries, the process of creating a video summary involves similar steps to those we needed for this study. This study, in addition to detecting user engagement with a presentation, provides us with a basis to evaluate the emotions and speech expressions of presenters. We assume that different expression styles engage the audience differently, which can be used to detect engagement level of a talk and helps us in summarization of a talk/video by identifying engaging segments.

3. Data Set

The TED website reports what percentage of viewers rated the video as "Informative", "Unconvincing" etc. However, relying on a rating given by a single user is not sufficient as different viewers may perceive a video differently, and may not rate the same video consistently across the different categories provided. Most of the ratings shown in Table 1 are positive, and we can see that the positive ratings have a higher average count and percentage than negative and neutral ratings. The minimum average count and percentage for a positive rating ("Funny") are 106 and 4.73% respectively, and the maximum average count and percentage for negative rating ("Unconvincing") are 51 and 3.73% respectively.

Table 1: *The average number of user ratings per each rating criteria for 1340 Ted videos across different topics. Avg(Count) is the average number of votes for each rating word, while Avg(%) refers to the average percentage of votes against each rating word.*

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

Therefore, if only the ratings of an individual video were

considered, it would seem like all the videos engage the viewers mainly positively. We did not see a video to which a negative rating word got the highest count by the viewers. So to deduce which video is found to be "Obnoxious" or "Longwinded" by viewers, some kind of normalisation is required. To do that we used the following definitions: 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 categorised as "Beautiful and not Beautiful", "Inspiring and not Inspiring", "Persuasive and not Persuasive", and so on, giving two classes for classification for each of the 14 rating words. The details of user rating distribution (Yes, No) for videos is depicted in Figure 2.

3.1. Data Pre-Processing and Feature Extraction

Speech segmentation is performed on all the audio files of TED videos using the Lium toolkit [10]. The duration of chunks is between a few seconds to 20 seconds maximum. As a result, we have 120,382 chunks of audio from 1338 videos for experimentation (clustering).

Acoustic feature extraction is performed using openSMILE toolkit [23]. The feature set is extracted using the openEAR configuration file. This set is also used for emotion and speech expression recognition [24] and consists of low-level descriptors as well as statistical functionals applied to these descriptors. We also performed a correlation test between duration of each audio chunk and its features, selecting those features which are less correlated with chunk duration ($R < 0.2$). As a result, we have 387 features in total for clustering. The feature set was further centred with mean value 0 and standard deviation 1.

Visual features are extracted using HAAR cascades [25], from the OpenCV library [26]. We calculated the time duration (seconds) of close up (detected face size $\geq 20\%$ of frame) and distance shots along with the duration (seconds) of a person not on screen (no face detected). We consider the number of instances of laughter and applause by Ted audiences as paralinguistic features. Theses features are extracted from the subtitles of Ted talks using a python script.

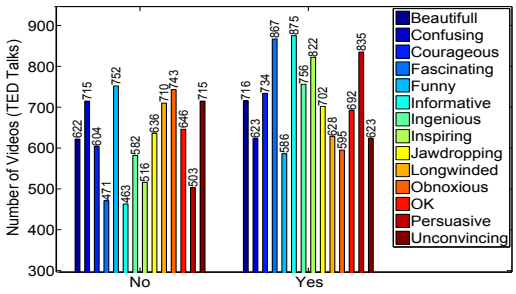


Figure 2: *Number of videos present in each class (Yes/No).*

4. Statistical Analysis

First, we employed SOM (Self Organised Map) to cluster the speech segments into 10 clusters. The motivation behind using clusters size of 10 for SOM is to separate the 6+1 universal spoken expressions (happiness, sadness, fear, surprise, disgust, anger and neutral) and any other non-speech segments (music, applause, laughter). Results of clustering are shown in Figure 3a and 3b. Then we calculated the number of speech segments in each cluster for every video and then divided it by the total number of speech segments within a video. Later, to analyse the significance of speech expressions, we used Kruskal–Wallis test and null hypothesis in the following manner:

H: The number of speech expressions in each group (e.g. beautiful and not beautiful) has the same mean value.

The Kruskal–Wallis test rejects the null hypothesis ($p < 0.05$) for many clusters (speech segments). For example, speech segments in clusters number 1,2,4,5,7 and 8 have a significant difference in their mean values for beautiful-YES and beautiful-NO. Speech segments from cluster number 1,4,5,7 and 8 have higher mean for beautiful-YES than beautiful-NO. Hence we can say that the speech segments in these clusters (that also represents speech expression) are engaging. The cluster number 2 have higher mean for beautiful-No than beautiful-YES, hence we can say that the speech segments in these clusters are non-engaging. The details for all engagement ratings are depicted in Table 2.

Table 2: *Statistical significant clusters for each rating.*

Rating	Cluster $p < 0.05$	YES	NO
Beautiful	1, 2, 4, 5, 7, 8	1, 4, 5, 7, 8	2
Confusing	7, 8	nil	7,8
Courageous	3, 4, 6, 7, 8, 9	3, 4, 6, 7, 8	9
Fascinating	6, 7, 9	7, 9	6
Funny	3 , 4, 5, 6, 7, 9	4, 5, 7, 9	3, 6
Informative	1 4 5 7 8	4	1 5 7 8
Ingenious	2 3 4 6 9	2,9	3, 4, 6
Inspiring	3 4 6 7 8	3, 4, 6,7, 8	nil
Jaw-dropping	2, 3, 7, 8, 9	2, 3, 7, 8, 9	nil
Longwinded	3 7	3	7
Obnoxious	4 6 7 9	4 6	7 9
OK	2 3 7 8 9	nil	2 3 7 8 9
Persuasive	1 3 7 8 10	3	1 7 8 10
Unconvincing	2 4 5 7 9	nil	2 4 5 7 9

5. Engagement Detection

Based on the statistical analysis results, we conclude that most of the speech expressions are statistically different for engaging and non-engaging speech segments. So models for an automatic engagement detection system are trained. We have performed three experiments using three different feature set for classification as described below:

Experiment One: we used high-level visual and paralinguistic features (camera angles, laughter, applause) extracted from video and subtitles.

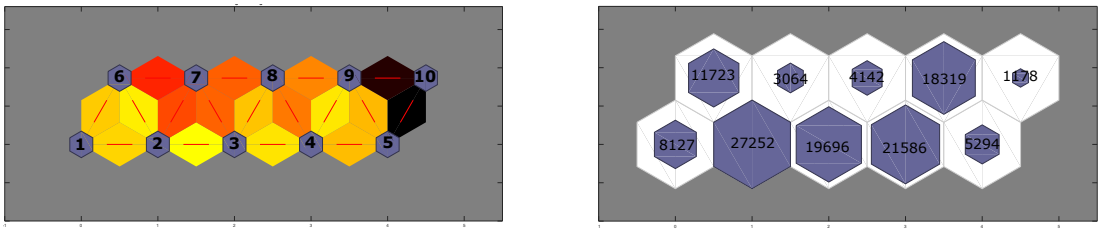
Experiment Two: we used the speech expressions features. However, we increase the features set by also considering the duration of speech segments in each cluster along with the number of speech segments in each cluster. We extracted the speech expressions features with different cluster sizes (10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60): this helped us find the best cluster size for engagement detection.

Experiment Three: we fused the previous two experiments' feature sets.

5.1. Classification Methods

The classification is performed using Linear Discrimination Analysis (LDA) in 10-fold cross-validation setting. This classifier is employed in MATLAB[3] using the statistics and machine

[3]http://uk.mathworks.com/products/matlab/



(a) *Distance between neighbouring clusters* (b) *Sample hits*

Figure 3: *Left Figure (a) indicates the distance between clusters (darker colour indicates more distance between clusters than lighter colours) and the right Figure (b) indicates the number of speech segments present in each cluster*

learning toolbox. LDA works by assuming that the feature sets of the classes to be discerned are drawn from different Gaussian distributions and adopting a pseudo-linear discriminant analysis (i.e. using the pseudo-inverse of the covariance matrix [27]).

6. Results and Discussion

The duration and number of speech segments in each cluster are used as a feature vector in detecting engagement, as defined earlier. The results are depicted in Table 3. All engagement levels are detected using the proposed feature vector above the blind guess baseline (50% A-weighted F-score (harmonic mean)). The results show that the visual and paralinguistic features (Vis+Para) extracted from video and subtitles provide better results than speech expressions (SE: clusters) features for 7 out of 14 user ratings. The fusion of speech expression and visual+para features (Fusion: clusters) improve results for 11 out of 14 user rating.

Table 3: *A-weighted F-score (harmonic mean). Where Vis+Para means high-level visual and paralinguistic features, SE: Clusters means Speech Expressions and the corresponding number of clusters and Fusion: Clusters mean Fusion of Vis+Para and SE along with the corresponding number of clusters.*

Rating	Vis+Para	SE: clusters	Fusion: clusters
Beautiful	55.28	60.58:15	61.70:30
Confusing	58.21	53.47:30	56.18:10
Courageous	60.33	58.08:20	61.25:10
Fascinating	52.65	54.02:55	58.04:45
Funny	70.96	61.61:35	71.85:10
Informative	59.08	61.24:40	64.09:40
Ingenious	57.42	56.67:30	57.35:55
Inspiring	54.85	53.05:55	56.13:60
Jaw-dropping	58.33	58.38:10	59.39:10
Longwinded	64.17	62.53:40	64.45:15
Obnoxious	48.87	52.14:45	54.25:45
OK	64.46	61.86:50	63.68:15
Persuasive	56.67	59.02:60	60.35:35
Unconvincing	56.4	56.86:20	58.11:10

In [22] we evaluated the relationship between high-level features (camera angles, pitch, laughter and applause) and user ratings. It only showed that high-level features are statistically different for user rating. This current study, however, not only analyses the relationship between speech expressions and user rating but it also proposes a system to detect the engagement with a novel combination of speech expressions and high-level features. The proposed system also generates feedback for the presenter or viewers in the form of video segments. The clustered video segments can potentially be used as training material for presenters to advise about using certain speech expressions for a particular type of engagement with viewers. For example in Table 2 it shows that clusters 3 and 7 have significant p-value with Longwinded. It may be used to guide a presenter to avoid speech expression of cluster 3 and utilisation of speech expression in cluster 7 to make sure that their presentations do not become Longwinded. Similarly, from a viewers perspective, it is a recommender system that can predict the engaging talks using multi-modal features. It can guide a potential viewer to avoid segments that have a higher number of speech segments in that cluster. From the Figure 3a, we can also see that the cluster 3 and 2 have a lesser distance between them than cluster number 3 and 7. Due to that lesser distance, the probability of sounding similar can be high for both clusters and the cluster number 2 instances may also be named as engaging one.

The clustering approach may also support in video summarization and segmentation e.g. summarising all the Inspiring parts of a video etc.

7. Conclusion and Future Work

The proposed approach demonstrates that characteristics of speech can be used to detect the engagement level of a talk. It is also a step forward towards generating feedback in the form of video chunks for presenters so that they will know the parts of videos which are engaging or not. However, the results are preliminary at this moment, and a detailed perception test with humans is needed for either all the clusters or the most significant clusters for positive and negative ratings to differentiate the engaging versus non-engaging speech expression. Possible future work following from this study include finding the relationship of gestures and facial expression with the engagement level of a talk. Another possible line of future work is to perform audio or video summarization based on the current study, and run perception tests to evaluate the outcomes.

8. Acknowledgements

This research is supported by Science Foundation Ireland (SFI) through the CNGL Programme (Grant 12/CE/I2267 and 13/RC/2106) in the ADAPT Centre (www.adaptcentre.ie) and is co-funded under the European Regional Development Fund at the School of Computer Science and Statistics, Trinity College Dublin, the University of Dublin, Ireland.

9. References

[1] S. Attfield, B. Piwowarski, and G. Kazai, "Towards a science of user engagement (Position Paper)," in *WSDM Workshop on User Modelling for Web Applications*, Hong Kong, 2011.

[2] M. J. Albers and M. B. Mazur, *Content and complexity: information design in technical communication*. Routledge, 2014.

[3] H. L. O'Brien and E. G. Toms, "What is user engagement? a conceptual framework for defining user engagement with technology," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 6, pp. 938–955, 2008.

[4] H. L. OBrien and E. G. Toms, "Examining the generalizability of the user engagement scale (ues) in exploratory search," *Information Processing & Management*, vol. 49, no. 5, pp. 1092–1107, 2013.

[5] F. Dobrian, A. Awan, D. Joseph, A. Ganjam, J. Zhan, V. Sel, I. Stoica, and H. Zhang, "Understanding the Impact of Video Quality on User Engagement," in *Communications of the ACM*, 2013, pp. 91–99.

[6] P. J. Guo, J. Kim, and R. Rubin, "How Video Production Affects Student Engagement : An Empirical Study of MOOC Videos," in *L@S 2014 - Proceedings of the 1st ACM Conference on Learning at Scale*, 2014, pp. 41–50.

[7] S. Benini, P. Migliorati, and R. Leonardi, "Statistical Skimming of Feature Films," *International Journal of Digital Multimedia Broadcasting*, vol. 2010, pp. 1–11, 2010. [Online]. Available: http://www.hindawi.com/journals/ijdmb/2010/709161/

[8] M. Haesen, J. Meskens, K. Luyten, K. Coninx, J. H. Becker, T. Tuytelaars, G.-J. Poulisse, P. T. Pham, and M.-F. Moens, "Finding a needle in a haystack: an interactive video archive explorer for professional video searchers," *Multimedia Tools and Applications*, vol. 63, no. 2, pp. 331–356, May 2011. [Online]. Available: http://link.springer.com/10.1007/s11042-011-0809-y

[9] F. A. Salim, "From artifact to content source: Using multimodality in video to support personalized recomposition," in *User Modeling, Adaptation and Personalization*. UMAP, 2015.

[10] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," Idiap, Tech. Rep., 2013.

[11] S. Wernicke, "Lies, damned lies and statistics (about tedtalks)," http://go.ted.com/bDrm, 2010.

[12] S. Tan, J. Bu, X. Qin, C. Chen, and D. Cai, "Cross domain recommendation based on multi-type media fusion," *Neurocomputing*, vol. 127, pp. 124–134, Mar. 2014. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0925231213009260

[13] D. Brezeale and D. J. Cook, "Learning video preferences using visual features and closed captions," *IEEE Multimedia*, vol. 16, no. 3, pp. 39–47, 2009.

[14] A. Anwar, G. I. Salama, and M. B. Abdelhalim, "Video Classification And Retrieval Using Arabic Closed Caption," in *ICIT 2013 The 6th International Conference on Information Technology VIDEO*, 2013.

[15] F. Chen, C. De Vleeschouwer, and A. Cavallaro, "Resource allocation for personalized video summarization," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 455–469, 2014.

[16] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.

[17] A. Dong and H. Li, "ONTOLOGY-DRIVEN ANNOTATION AND ACCESS OF PRESENTATION VIDEO DATA." *Estudios de Economía Aplicada*, 2008.

[18] F. Haider, L. Cerrato, N. Campbell, and S. Luz, "Presentation quality assessment using acoustic information and hand movements," in *Proceeding of 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.

[19] K. Curtis, G. J. Jones, and N. Campbell, "Speaker impact on audience comprehension for academic presentations," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 129–136.

[20] X. Ochoa, M. Worsley, K. Chiluiza, and S. Luz, "Mla'14: Third multimodal learning analytics workshop and grand challenges," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 531–532.

[21] F. A. Salim, F. Haider, O. Conlan, S. Luz, and N. Campbell, "Analyzing multimodality of video for user engagement assessment," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 287–290.

[22] F. Haider, F. A. Salim, S. Luz, O. Conlan, and N. Campbell, "High level visual and paralinguistic features extraction and their correlation with user engagement," in *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2015, pp. 326–331.

[23] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[24] F. Eyben, M. Wllmer, and B. Schuller, "Openear–introducing the munich open-source emotion and affect recognition toolkit," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, Sept 2009, pp. 1–6.

[25] R. Lienhart, A. Kuranov, and V. Pisarevsky, "Empirical analysis of detection cascades of boosted classifiers for rapid object detection," in *Proceedings of the 25th DAGM Pattern Recognition Symposium*, 2003, pp. 297–304.

[26] G. Bradski, "The OpenCV Library," *Dr. Dobbs Journal of Software Tools*, 2000.

[27] S. Raudys and R. P. W. Duin, "Expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix," *Pattern Recognition Letters*, vol. 19, no. 5-6, pp. 385–392, Apr. 1998.