Measuring the Engagement Level of TV Viewers

Javier Hernandez[1] Zicheng Liu[2] Geoff Hulten[2] Dave DeBarr[2] Kyle Krum[2] Zhengyou Zhang[2]

[1] Media Lab, Massachusetts Institute of Technology
Cambridge, United States of America
javierhr@media.mit.edu

[2] Microsoft Research, Microsoft
Redmond, United States of America
{zliu, ghulten, dave.debarr, kkrum, zhang}@microsoft.com

***Abstract*— This work studies the feasibility of using visual information to automatically measure the engagement level of TV viewers. Previous studies usually utilize expensive and invasive devices (e.g., eye trackers or physiological sensors) in controlled settings. Our work differs by only using an RGB video camera in a naturalistic setting, where viewers move freely and respond naturally and spontaneously. In particular, we recorded 47 people while watching a TV program and manually coded the engagement levels of each viewer. From each video, we extracted several features characterizing facial and head gestures, and used several aggregation methods over a short time window to capture the temporal dynamics of engagement. We report on classification results using the proposed features, and show improved performance over baseline methods that mostly rely on head-pose orientation.**

Keywords-component; engagement; attention; market research; facial expression analysis; face and head features

I. INTRODUCTION

The problem of automatically measuring the engagement level of people in a specific task is an area of great interest across a wide variety of fields. For instance, people in the automotive industry have been working on detecting driver inattention to prevent car accidents and improve overall car safety [10]**.** In online educational settings, educators want to monitor the interest levels of students to appropriately change their lessons and speed-up the learning process [11]. Similarly, in the field of market research there is a growing interested in finding objective measures that quantify the engagement level of the audience to better address market needs [17]. Automatic systems that understand and provide this type of information promise to dramatically change the way humans interact with computers.

Humans usually rely upon verbal and non-verbal behavior to estimate the level of engagement of people. In some situations, finding this set of behaviors may be relatively easy. In the context of car safety, for example, looking away from the road or closing the eyes for extended periods of time are good indicators of driver inattention. Unfortunately, in other real-life settings it is more difficult to define an appropriate set of behaviors for engagement, and it requires some context to make more accurate predictions. For instance, a professor may think that a student is deeply engaged if the student is leaning forward with his eyes wide open, but may discover that the students' mind is somewhere else when he is unable to answer simple questions. Similar cases may appear in the context of TV viewers. While not looking at the TV can be a good indicator of low engagement (top-left image of Fig. 1), it may also be the case that the viewer is commenting the program with another person (top-right). Similarly, while looking at the TV can be a good indicator of high engagement (bottom-left), it may also be the case that the viewer is thinking about something else (bottom-right). The behaviors associated with engagement may manifest differently in each situation and need to be interpreted within specific contexts.

Figure 1. Examples of viewers' responses. After analyzing the context in which each frame was taken, the engagements levels were coded as None (top-left), High (top-right), High (bottom-left) and None (bottom-right).

In this work we focus on the problem of engagement measurement of TV viewers in a naturalistic and ubiquitous setting such as a living room, where viewers are more likely to respond naturally and freely interact with real-life distracters (e.g., laptops, magazines, snacks or other people). In particular, we recorded 47 people watching TV programs with an RGB video camera mounted on the TV set. Then, we extracted several face and head features and used them to automatically predict the engagement level of viewers.

The paper is organized as follows. First, we review relevant work in the field of engagement measurement. Second, we describe the data collection and annotation process. Third, we

present several features and aggregation methods to capture facial and head changes over time. Fourth, we provide details about the experimental setting and evaluation methods. Finally, we present classification results and feature analysis.

II. Related Work

The problem of engagement measurement has been widely studied in the laboratory or semi-structured environments where variables that introduce noise are controlled or eliminated. Engagement and similar types of affective information are usually captured through two main sources of information: self-reports and autonomic information.

Verbal self-reports in the form of interviews or surveys are widely used in the field of market research to assess the level of recall and comprehension of certain stimuli such as advertisements (see Poels and Dewitte [17] for an extensive review). However, participants can also provide their emotional responses through more visual surveys such as the Self-Assessment Manikin [4] and FaceTrace™ [22]. A more subtle approach is the use of electronic dials that allow participants to provide moment-to-moment ratings. Although self-report methods provide first-hand information about what the participant experiences, they require cognitive effort to process and may divert their attention during the experiments. An alternative method is to utilize the reports of a human coder. This method is widely used in the field of facial expression analysis where certified Facial Action Unit coders examine videos and provide annotations [7]. Although this method avoids disrupting the participants, the annotation process is considerably slower and more expensive.

A typically less disruptive method captures the autonomic responses of the body. One of the most widely used autonomic signals to measure engagement is eye gaze behavior. For instance, Wedel and Pieters [21] studied the eye movements when looking at advertisements and found that the number of eye fixations, and not their duration, was a good indicator of attention and memory processing. However, eye trackers are expensive devices and cannot be easily applied in real-life settings without disrupting participants' behavior. Researchers have successfully used other biosensors such as heart rate [12] or galvanic skin response monitors [16]. However, these sensors can be easily influenced by external factors (e.g., movement, temperature) and require very controlled settings to clearly identify the stimulus eliciting the response. More advanced systems combine multiple modalities to clarify ambiguous cases and capture a wider range of responses. For instance, McDuff et al. [15] used a Kinect camera, a galvanic skin response sensor, and a video webcam to capture the engagement level of workers in an office environment, reaching 71% recognition accuracy. In a separate study, Kapoor and Picard [11] combined a pressure sensor on the chair and an Infra-Red camera to capture the interest level of students with an accuracy of 86%. Although combining multiple sensors typically increases recognition accuracy, it also increases instrumentation costs and limits its scalability to larger populations.

When exploring real-life scenarios, RGB video cameras are arguably one of the most ubiquitous and less invasive sensors because they are widely accessible to the public and can be used from a distance to measure rich psychophysiological information such as head gestures and facial expressions. In a recent attempt to capture information in less controlled settings, McDuff et al. [14] used the webcam of computer users to capture their responses while watching an advertisement. However, the short duration of the stimulus (around 3 minutes) severely limits the appearance of natural distracting behavior (e.g., speaking with someone else, reading a magazine).

III. Data Collection

This section provides details about the experimental setup and the process of engagement annotation.

A. Setting

In order to measure the level of engagement of TV viewers, we created a naturalistic environment that recreates a living room. As shown in Fig. 2, the experimental setting included a television (60 inches), an RGB video camera, a table and a couch. Additional chairs were placed in the background if there were many participants. The video camera was mounted on top of the TV and recorded viewer's responses with a resolution of 1280x720 pixels. The separation between the couch and the camera was 2.5 meters and the spatial distribution of the objects remained constant throughout all of the recordings.

Figure 2. Schematics of the experimental setup.

Data collection was performed over a series of sessions. During each session, a group of participants watched a TV show selected from a large online library. The only requirement for the program was to be more than 15 minutes and less than 1 hour. Throughout the experiment participants had access to drinks, snacks, magazines and a tablet computer with an Internet connection. Moreover, they were allowed to use personal electronic devices (e.g., cellphone, laptop) and to interact with other participants as they felt appropriate. The goal of this experimental setup was to elicit natural distracting behavior and to introduce challenges of real-life scenarios (e.g., face occlusions, body movement, lack of engagement).

We recorded the responses of 47 participants grouped in 14 different sessions. The maximum number of participants per session was 7, the minimum was 2, and the average was 3.36. The average duration of the sessions was 25 minutes including advertisements. Participants in each session knew one another from work or socially before the experiments. Four of the sessions recorded families who frequently watched TV together in their homes. People were recruited through e-mail and poster advertisements, and were selected to represent a wide range of demographics in terms of ethnicity, gender and age.

B. Annotation

The engagement level of viewers' responses was rated by an experienced human coder. In the context of TV viewers, we define the engagement level as the extent to which the viewer is focused on the experience being shown on the TV. These ratings were used as a ground truth to train and evaluate our recognition system. We defined four different states/levels:

- High: The viewer's main focus for an extended period is the content and he is eager to move past distracters to return to the content. For example, if several viewers are watching a football game and they are on the edge of their seat, they are highly engaged. Similarly, if they turn to each other and talk about the last play for a few seconds, then come back to the game, the engagement probably has remained High the whole time.
- Medium: The viewer's main focus is the content but he is also aware of other things in the environment and may check email, chat, etc.
- Low: The viewer's main focus is on something other than the content, but the viewer may check into the content periodically.
- None: The viewer's main focus is on something other than the content and he is not checking into the content. For instance, if a commercial comes on and he turns to his friend and chats about what they did last night, their engagement is probably Low or None.

Using the above guidelines, the human coder watched the entire video to understand the overall context, then went back and annotated state changes between the engagement states. A list of things that were considered (but not limited to) are: topic of conversation, visual focus of attention, expressive state, time scale of disengagement, and body pose. Although these factors are typically good indicators of engagement, everything had to be interpreted within a specific context. For instance, someone who is very relaxed and has no expression can still be highly engaged. Note that at the end of the annotation process, we obtained the engagement level of each viewer for every frame.

IV. FEATURE EXTRACTION

In the setting of meeting rooms, head pose is arguably one of the most reliable and widely used features to determine the focus of people's attention [1][2][5][20]. The underlying assumption is that people pay attention to whatever they are looking at. In this work, we utilize an off-the-shelf head pose estimation method [23] as a baseline feature to compare classification performance. This method detects 5 different states: undetected face, frontal face, right half, left half, full right and full left. Since different people may sit at different locations of the living room, we grouped the left and right sides in one single state (i.e., right/left half, full right/left). We hypothesize that only using head pose may be too limited to capture the engagement of TV viewers as they may multitask and can frequently change their head orientation.

A complementary and more complex approach consists of analyzing the facial expressions of people. This task usually involves finding several interesting points around the face (e.g., lips, eyebrows) and analyzing their change over time. Different researchers utilize different number of points and distributions around the face but, usually, the larger the number of points, the more information can be gathered, but the more computationally expensive it becomes as well. For instance, El Kaliouby and Robinson [8] tracked 24 facial points and found strong correlation of their movements with certain cognitive states (e.g., agreeing, confused, interested). Similarly, McDuff et al. [15] and Teixeira et al. [18] tracked 100 and 64 facial points, respectively, to recognize engagement in different settings. When analyzing social spaces such as living rooms, where many people may gather at the same time, less computational expensive algorithms are preferred. In this work we utilize an off-the-shelf system [13] that uses 5 facial points (see Fig. 3). Using these points, we extract the following types of features:

- Face distances and angles. Fig. 3 shows the 8 distances and the 12 angles computed between the different facial points. These features have been shown to capture facial expressions and activities such as smiling and talking.
- Head roll. We computed the rotation of the head by calculating the absolute angle between the two eyes and the horizontal plane. The goal of this feature is to capture certain gestures such as head tilts that may appear when someone is confused or bored.

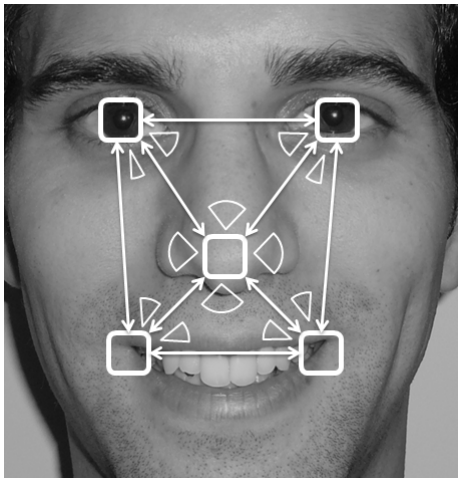

Figure 3. Facial points (squares), distances (lines), and angles (segments).

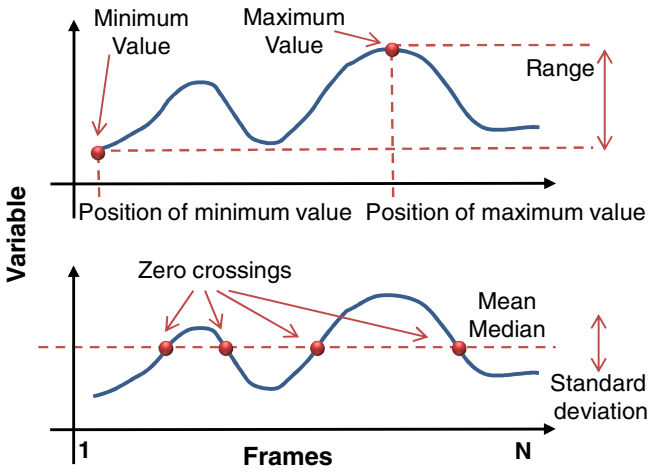

Figure 4. Graphical representation of temporal aggregation methods.

TABLE I. SUMMARY OF HEAD AND FACE FEATURES

Modules	Types of features	Aggregation methods
Head Pose Tracker	Head orientation	Amount of time at each state
Face Tracker	Face distances Face angles Head roll	Mean Median Standard deviation Minimum Maximum Position of maximum value Position of minimum value Range Number of zero crossings
Face Detector	Head size Head position	

Additionally, we utilize the output of the Viola and Jones face detector [19] to extract two more types of features: 1) head size (area of the detected head) and 2) head position (location of the head). While the first feature may be relevant to indicate when someone is leaning forward or backward, the second one is useful to capture head shaking or nodding. Since different viewers may have different head sizes, we normalized all of the distances with respect to the distance between the two eyes. In the case of face size and face position, the distance between the two eyes was averaged over a period of time. These types of features not only require low-computational cost but are also robust to many of the aspects that affect pixel intensity and color (e.g., illumination changes, skin tones).

All of the above features were computed in a frame by frame basis. In order to capture the temporal dynamics of face and head gestures, we aggregated the features with different functions over a short time window. For the head pose estimation, we created a histogram of state percentages, capturing the proportion of frames the viewer displayed each pose. For the rest of the features, the aggregation functions were as follows: mean, median, standard deviation (STD), minimum value, maximum value, position of the minimum value, position of the maximum value, range (difference between maximum and minimum values), and number of zeros crossings after mean subtraction (ZC). Fig. 4 shows the graphical interpretation of the aggregation methods for a specific variable (e.g., distance between mouth points) throughout N frames. Different aggregation functions can potentially capture different characteristics of the temporal dynamics. For instance, the position of the maximum value computed on the distance between the mouth points may indicate if the person is opening (i.e., maximum position at the end) or closing the mouth (i.e., position at the beginning). Another example: the number of zero crossings of the face position may represent the amount of movement of the person. TABLE I shows a summary of the different modules, types of features, and temporal aggregation methods.

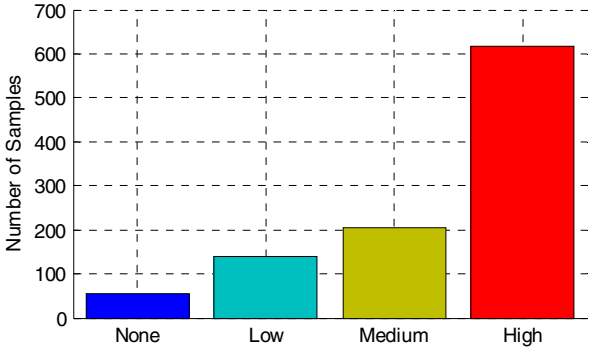

Figure 5. Distribution of engagement levels after 1-minute segmentation.

V. EXPERIMENTAL SETTING

This section provides details about the data preprocessing, classification method, and performance metric used in our experiments.

A. *Preprocessing*

In this work we want to predict the engagement level of 47 viewers while they watch TV. Since each participant displayed various engagement levels throughout a session, we divided each session into smaller non-overlapping clips (e.g., 1 minute) with the response of one viewer at a time. For instance, from a 10 minute session with 2 TV viewers, we extracted 20 1-minute long clips. The goal of the classification will be the prediction of the most common rating for a specific time window (i.e., mode of the engagement levels).

Fig. 5 shows the distribution of engagement levels for the 1-minute clips. As can be seen, the engagement levels are skewed toward High. We believe this is because we allowed participants to choose their preferred programs, which is more representative of real-life scenarios. Moreover, we observed that viewers of the same session showed correlated responses, which is to be expected because they were watching the same TV program.

B. *Classification*

In this study, we address the problem as a binary classification, in which None and Low levels of engagement were grouped into one class (186 samples), and Medium and High levels of engagement were grouped into a different class (806 samples). Future work will focus on collecting more data for the None and Low levels so we can appropriately address the problem as 4-class classification problem.

To perform classification, we used the publicly available LIBSVM library [6] that provides an efficient implementation of Support Vector Machines (SVMs) [3]. We used 5-fold-cross

TABLE II. CLASSIFICATION SCORE (%) FOR EACH ONE OF THE FEATURES AND AGGREGATION METHODS

	Mean	Median	STD	Maximum	Minimum	Max. Pos	Min. Pos	Range	ZC
Face distances	78.76	**79.30**	63.50	43.89	68.81	44.28	45.78	45.42	62.52
Face angles	**73.57**	72.86	73.54	69.72	68.07	43.76	44.46	69.50	51.23
Head roll	44.54	43.82	42.90	44.36	44.04	43.70	43.60	44.67	**49.03**
Head size	43.53	42.95	56.56	63.97	53.38	44.17	43.74	61.89	**69.91**
Head position	43.53	43.06	59.10	52.66	47.34	43.99	43.07	55.14	**66.59**
Average	56.79	56.40	59.12	54.92	56.33	43.98	44.13	55.32	**59.86**

validation for testing and 10-fold-cross-validation for training. That is, we divided the dataset into 5 different groups and used four of them as a training set and the remaining one as the testing set. This process was then repeated until performance was obtained for all the groups. In the training phase, we divided the training set into 10 different groups and followed the same iterative process to gather performance for different misclassification costs of SVMs. Once the process was completed, we used the whole training set and the best cost to obtain the final classifier model, which then was used in the testing set. In order to ensure that our algorithm would generalize to other participants, video samples of the same participant always belonged to the same fold (i.e., they were never used for training and testing at the same time). We utilized a Linear Kernel for each type of feature and combined them with equal weights. The misclassification costs of SVMs that were used during the cross validation were: $\log_2 C \in \{5, 6, ..., 18\}$.

In order to find the most discriminative subset of features we used the Forward Feature Selection (FFS) method [9]. Starting with the most discriminative feature, this method iteratively adds the feature that yields the highest performance. The method stops either if there are no additional features or the performance was better with a smaller subset of features.

C. Performance

Traditionally, accuracy has been the most common metric to report classification performance. However, in more realistic datasets where class distributions may be unbalanced, a classifier that always predicts the most common class may deceptively yield high accuracy. A metric that addresses this problem is the F1 score, which is the harmonic mean between precision and recall. However, this metric provides performance on the detection of one class, which usually is the most relevant but least common.

Both of these metrics capture the performance of a classifier for a specific configuration (e.g., one threshold on the confidence values). However, when analyzing the discriminative power of a classifier, Receiver Operating Characteristic (ROC) and/or Precision/Recall (PR) curves are preferred. These metrics capture the performance of the classifier for different configurations of the classifier (e.g., different thresholds), and their area under the curve (AUC) is traditionally used as a representative score, ranging from 1 (maximum score) to 0 (minimum score). In this paper, we want to capture the performance of both classes (lower and higher engagement levels) so we defined our score as the average between the ROC AUC and the AUCs of two PR curves: one assuming the class of interest is lower engagement levels, and the other one assuming the class of interest is higher engagement levels. This score was used to find the best parameter of the algorithm and will be used in the following section to report classification performance. For some of the results, we will also include traditional metrics for comparison. Note that a classifier that always predicts the most likely class will obtain a performance of 0, since none of the curves can be computed (i.e., confidence values are always one). In order to provide a baseline method, we used a classifier trained with only the head pose feature, which is the most common feature when measuring attention from a distance.

VI. RESULTS

In this section we report results on the automatic recognition of engagement levels. The analysis is divided as follows. First, we evaluate the predictive value of each of the feature types and temporal aggregation methods. Second, we provide results for the best combination of features, and compare performance with standard metrics. Finally, we explore the use of different time windows. For the first two parts, we will use a 1-minute segmentation window.

A. Feature Types and Aggregation Methods

Using the experimental setting described in the previous section, we computed the performance obtained for each of the features. When only using head pose information, the overall classification score was 62.37%. TABLE II shows the scores for each of the other features and aggregation methods. As can be seen, distances and angles between facial points outperformed head pose with 79.30% and 73.57%, respectively. This finding supports our hypothesis that facial expressions provide relevant information to detect the engagement levels. Interestingly, head position and head size also yielded better performance than head pose. A careful review of the predictions indicated that head size and head position successfully captured the cases where the viewers were deeply engaged and did not move, or were distracted and moved around (e.g., to speak with others, look at distracters). Among the feature aggregation methods, both mean and median performed at the same level and were the best methods for the facial distances and angles. Number of zero crossing was the most useful method for head features (i.e., roll, size, and position) and, on average, both ZC and standard deviation yielded the best performance across features.

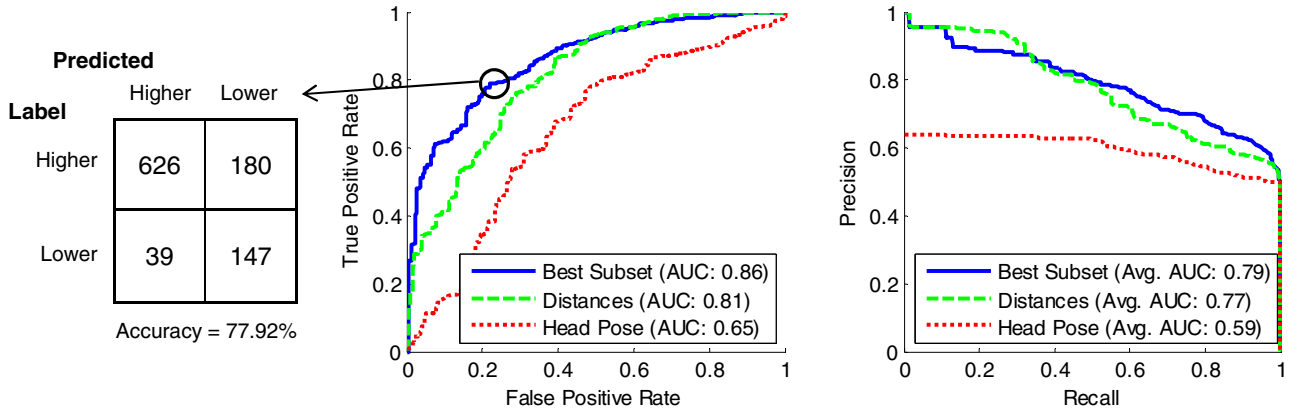

Figure 6. Confusion matrix (left), ROC curves (center) and average of Precision/Recall curves (right) for head pose (baseline method), distances between facial points (best single feature), and best combination of features.

B. *Best Subset of Features*

In this experiment, we wanted to explore the performance when combining different features.

After running the FFS method for all possible combinations of feature types and aggregation methods, the best subset of features was: distances between facial points with mean aggregation, ZC of head size, and range of head roll. This combination achieved a score of 82.54%, which is 3.24% higher than the best feature. Note that even though head roll showed poor performance when used independently, it provided additional discriminative information when used in combination with other features. On the other hand, angles between facial points was not selected as part of the best subset of features, probably indicating that the information encoded by this feature is similar to that encoded by distances between facial points. Similarly, combining head pose with facial distances did not yield higher performance, which is expected as different head orientations affected the distances between facial points.

Fig. 6 shows the confusion matrix, ROCs and the average of Precision/Recall curves for the head pose, the best single feature, and the best combination of features. Using these metrics, it becomes easier to analyze the performance of the models for different configurations. For instance, when using the best combination of features, we can correctly detect 79% of the lower engagement responses (i.e., False Positive Rate of 21%) while correctly detecting 77.67% of the higher engagement responses. The overall accuracy for this specific configuration is 77.92%.

C. *Time Windows*

For this part of the analysis we wanted to explore whether different window sizes have an impact on the classification performance. In practice, the distribution of the engagement levels for different window sizes was very similar to that observed when using a 1-minute window (see Fig. 5).

Fig. 7 shows the performance achieved when using FFS on different time windows. As can be seen, windows above 30 seconds yielded 7% higher performance than some of the smaller windows. This finding is consistent with the hypothesis that engagement needs to be analyzed within a specific temporal context. We also observed that different types of features performed differently on different time scales. For instance, while the zero crossings of head size improved performance with larger windows, head pose improved performance with smaller windows. Therefore, the best combinations of features and aggregation methods across windows was slightly different.

In real-life settings, different window sizes may be appropriate for different applications. For instance, smaller windows may be more useful in safety scenarios where immediate predictions are critical, and engagement levels tend to fluctuate faster. On the other hand, larger windows can be more useful in market research scenarios where accurate predictions are more important, and engagement levels are likely to be measured over the duration of a stimulus (e.g., commercial advertisements or long television programs).

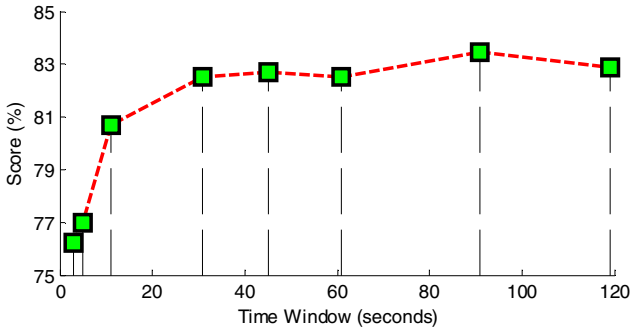

Figure 7. Performance achieved using Forward Feature Selection on different time windows (ranging from 3 seconds to 2 minutes).

VII. CONCLUSIONS

In this work, we study the problem of engagement measurement of TV viewers from an RGB video camera. In particular, we collected facial responses of 47 participants while watching a TV show, used a human coder to manually annotate perceived engagement levels, and built an automated system to predict the engagement levels from head and facial gestures. We were able to successfully discriminate between

lower engagement levels and higher engagement with a score of 79.30%, when using one single feature, and a score of 82.54% when using the most discriminative subset of features. These results significantly outperformed predictions based on head pose (62.37%), which could be considered as the baseline method in the literature.

Among all of the features, the distances and angles between facial points were the most discriminative features for engagement recognition. Moreover, head size and head position successfully captured some of the relevant temporal dynamics of engagement such as the lack of movement. Among the temporal aggregation methods, we found that mean, number of zeros crossings, and standard deviation were some of the most useful methods. We also studied the classification results for different time windows (from 3 seconds to 2 minutes) and observed that 30 seconds seemed sufficient to reach top performance, which is 7% higher than smaller windows. However, different features and aggregation methods may work better at different time granularities. Future work will focus on finding the optimal combination of features from different time windows. Additionally, temporal models such as Hidden Markov Models or Conditional Random Fields could be used to further encode the temporal dependencies.

In this study, we chose to use 5 facial points to perform our analysis. Although using 5 points significantly reduced the computational cost of our approach, some additional points (especially around the eyebrows) could help characterize other relevant expressions such as frowning. It would also be worth exploring other features such as eye openness and eye blinking rate. However, obtaining these metrics reliably from a long distance can be a challenge. Finally, although this work only considered the use of an RGB video camera, we acknowledge the value of adding other sensors. For instance, a microphone could be used to detect when participants are talking or laughing, or a 3D camera could be used to capture fine grain resolution of the body position. However, new sensors may introduce new challenges and may not be easily accessible in real-life settings.

This work has shown that automatic engagement recognition is possible in a naturalistic environment, with low-computational cost algorithms, and non-invasive and accessible sensors. Although we studied engagement in the context of TV viewers, similar systems can be applied in a wide variety of fields such as automobile safety, education and gaming. We look forward to the near future where such systems will be used in our daily life to increase our safety, assist our decision-making, and enhance our interactions with computers.

ACKNOWLEDGMENT

We would like to thank Michael Conrad for his contribution and support to this project.

REFERENCES

[1] R. Akker, D. Hofs, H. Hondorp, H. Akker, J. Zwiers, and A. Nijholt, "Supporting engagement and floor control in hybrid meeting," in Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions, 2009, pp. 276-290.

[2] S. O. Ba and J. Odobez, "Head pose tracking and focus of attention recognition algorithms in meeting rooms," in Proc. of Classification of Events, Activities and Relationships, 2006, pp. 345-357.

[3] B. E. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in Proc. of the Fifth Annual Workshop on Computational Learning Theory, 1992, pp. 144-152.

[4] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," Journal of Behavior Therapy and Experimental Psychiatry, 25 (1), 1994, pp. 49-59.

[5] L. Dong, H. Di, L. Tao, G. Xu, and P. Oliver, "Visual focus of attention recognition in the ambient kitchen," in Proc. of the Asian Conference on Computer Vision, vol. 3, 2009, pp. 548-559.

[6] C. C. Chang and C. J. Lin, "LIBSVM: a library for Support Vector Machines," 2001, software: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[7] P. Ekman and W. Friesen, "Facial Action Coding System: a technique for the measurement of facial movement," Palo Alto, CA: Consulting Psychologists Press, 1978.

[8] R. El Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in Proc. of Real-time Vision for Human Computer Interaction, 2005, pp. 181-200.

[9] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, 3, 2003, pp. 1157-82.

[10] J. Jo, S. J. Lee, H. G. Jung, K. R. Park, and J. Kim, "Vision-based method for detecting driver drowsiness and distraction in driver monitoring system," in Optical Engineering, vol. 50 (12), 2011, pp. 127202/1-24.

[11] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in Proc. of ACM International Conference on Multimedia, 2005, pp. 677-682.

[12] A. Lang, "Involuntary attention and physiological arousal evoked by structural features and mild emotion in TV commercials," in Communication Research, vol. 17 (3), 1990, pp. 275-299.

[13] L. Liang, R. Xiao, F. Wen, and J. Sun, "Face alignment via component based discriminative search," in Proc. of European Conference on Computer Vision, vol. 2, 2008, pp. 72-85.

[14] D. McDuff, R. El Kaliouby, and R. W. Picard, "Crowdsourcing facial responses to online videos," IEEE Transactions on Affective Computing, vol. 3 (4), pp. 456-468.

[15] D. McDuff, A. Karlson, A. Kapoor, A. Roseway, and M. Czerwinski, "AffectAura: an intelligent system for emotional memory," in Proc. of Computer Human Interaction, 2012, pp. 849-858.

[16] J. Peacock, S. Purvis, and R. L. Hazlett, "Which broadcast medium better drives engagement? Measuring the powers of radio and television with electromyography and skin-conductance measurements," Journal of Advertising Research, vol. 51 (4), 2011.

[17] K. Poels and S. Dewitte, "How to capture the heart? Reviewing 20 years of emotion measurement in advertising," Journal of Advertising Research, vol. 46 (1), 2006.

[18] T. Teixeira, M. Wedel, and R. Pieters, "Emotion induced engagement in Internet video advertisements," Journal of Marketing Research, vol. 49 (2), 2012, pp. 144-159.

[19] P. Viola and M. J. Jones, "Robust real-time face detection," International Journal of Computer Vision, vol. 57 (2), 2004, pp. 137-154.

[20] M. Voit and R. Stiefelhagen, "Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios," in Proc. of Multimodal Interfaces, 2008, pp. 173-180.

[21] M. Wedel and R. Pieters, "Eye fixations on advertisements and memory for brands: a model and findings," Marketing Sciences, vol. 19 (4), 2000, pp. 297-312.

[22] O. Wood, "Using faces; measuring emotional engagement for early stage creative," in European Society for Opinion and Marketing Research, 2007.

[23] Z. Zhang, Y. Hu, M. Liu, and T. Huang, "Head pose estimation in seminar room using multi view face detectors," in Proc. of Classification of Events, Activities and Relationships, 2007, pp. 299-304.