Estimating Audience Engagement to Predict Movie Ratings

Rajitha Navarathna, Peter Carr, Patrick Lucey, and Iain Matthews

Abstract—While watching movies, audience members exhibit both subtle and coarse gestures (e.g., smiles, head-pose change, fidgeting, stretching) which convey sentiment (i.e., engaged or disengaged) during feature length movies. Noticing these behaviors using computer vision systems is a very challenging problem—especially in a movie theatre environment. The environment is dark and contains views of people at different scales and viewpoints. Feature length movies typically run 80-120 minutes, and tracking people uninterrupted for this duration is still an unsolved problem. Facial expressions of audience members are subtle, short, and sparse; making it difficult to detect and recognize activities. Finally, annotating audience sentiment at the frame-level is prohibitively time consuming. To circumvent these issues, we use an infrared illuminated test-bed to obtain a visually uniform input of audiences watching feature length movies. We present a method which can automatically detect the change in behavior (key-gestures) using "key-frames", which can convey audience sentiment. As the number of key-frames are many orders of magnitudes lower than the number of frames, the annotation problem is reduced to assigning a sentiment label for each key-frame. Using these discovered key-gestures, we create a movie rating classifier from crowd-sourced ratings and demonstrate its predictive capability. Our dataset consists of over 50 hours of audience behavior collected across 237 subjects.

Index Terms—Audience, behaviour, engagement, movie, film

1 Introduction

Crowd sourced reviews, such as *Rotten Tomatoes*, capture the overall rating of a movie, but rarely contain detailed information about specific scenes or moments. Measuring viewer sentiment (i.e., engaged or disengaged) for long continuous time-series signals like movies is very useful for writers, directors, marketers and advertisers. The de-facto standard for measuring audience sentiment is via self-report [1]. Self-reporting is subjective and does not provide feedback with precise time stamps. Completing a report during the movie would require a person to consciously think about and document what they are watching and subjects may miss important parts of the movie. Although wearable sensors that gather physiological data (e.g., heart-rate, galvanic skin response [2], [3], [4], [5]) or continuous dial ratings [6] could be used, such approaches are invasive and unnatural, and may not be a good indicator of the actual rating. Vision-based approaches are ideal as they can be done unobtrusively and allow viewers to watch the stimuli uninhibited.

However, monitoring an audience in a movie theatre using computer vision is difficult. The environment is dark, and light spill from the screen causes drastic variations in illumination conditions. Moreover, the physical configuration makes it difficult to observe facial expressions on all audience members. To circumvent these issues, we created an infrared (IR) illuminated testbed. We screened feature-length movies, and collected over 50 hours of video footage across 237 subjects from 10 different movies. We captured audience footage at 15 fps from an IR camera, two IR illuminators and a IR band-pass filter to give a uniform visual signal (see Section 3 for more details).

Audiences responses can be quick and subtle (i.e., a smile at a joke or jumping at a scary moment). Manually identifying these events is prohibitive given the size of the dataset. As audience members are often stationary for long periods of time, the annotation of sentiment levels can be expedited by exploiting the significant redundancy in the input signal. After collecting the audience data, we identified *key frames* for each audience member and then mapped each key frame to a sentiment label. Recently Whitehill et al. [7] found that annotating engagement level for static images is more reliable than watching the video clip and continuously labeling engagement levels (in a classroom environment). They observed that labeling the video clips at normal viewing speed is difficult to execute in practice.

Using the labeled *key frames* and following a supervised learning approach, we present a framework which can automatically predict audience sentiment levels. Using the audience engagement levels, we predict movie ratings and demonstrate its predictive capability compared to self-reports and our previous work [8] (see Fig. 1). Our results show that audience sentiment levels are better for predicting movie ratings compared to self-reports. Overall, we show that the proposed pipeline can be used to predict movie ratings solely using audience behaviors, which is a

- R. Navarathna, P. Carr, and I. Matthews are with Disney Research, Pittsburgh, PA 15213.
E-mail: {rajitha.navarathna, peter.carr, iainm}@disneyresearch.com.
- P. Lucey is with STATS, Chicago, IL 60601. E-mail: plucey@stats.com.

Manuscript received 14 Dec. 2016; revised 22 May 2017; accepted 3 June 2017. Date of publication 3 July 2017; date of current version 7 Mar. 2019. (Corresponding author: Rajitha Navarathna.)
Recommended for acceptance by S. Zafeiriou, G. Zhao, I. Kotsia, M. Nicolau, and J. Cohn.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TAFFC.2017.2723011

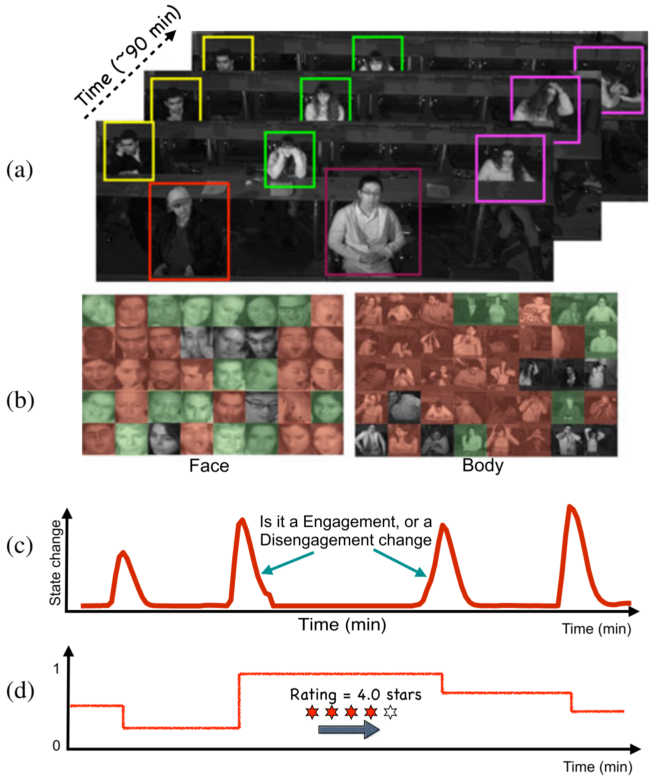

Fig. 1. (a) From a single video stream we capture the motion of all audience members during a feature-length movie. (b) Based on facial expressions and body motions, we then compactly represent each person's behavior as a series of *key gestures*, which are the distinct changes in face expressions (left) and body motions (right). Once the key gestures are determined, positive (green) and negative (red) sentiment is annotated for each key gesture. Using these sentiment levels, we predict audience movie ratings and demonstrate how visual input can be more informative than self-reports.

potential solution to the problems with current standard self-report measures [9].

2 Related Work

Having a large window of time to monitor human behavior introduces a broad gamut of additional gestures/activities associated with engagement or boredom—meaning that both coarse and fine behaviors in the face (e.g., smiles/yawning versus head-pose change) and body (e.g., fidgeting/doodling versus stretching) maybe indicators of engagement/disengagement [10]. In his book "In the Blink of an Eye" [11], Walter Murch speculates that the engagement of an audience can be measured through the synchronicity of motion.

Conventional methods of estimating the sentiment of an audience member during long-term experiences such as movies, stage drama-shows and TV shows are based on self-reports [1], [9]. Similarly, subjects could be instrumented with a myriad of wearable sensors, but such approaches are invasive and unnatural which may not be a good indicator of the actual rating.

2.1 Sentiment Detection from Physiological Changes

Use of physiological signals provides continuous affective states. These physiological changes can be used to measure the peripheral nervous system functions such as electrodermal activity, heart and blood circulation, skin conductivity, muscular activity, etc. Work from Picard et al. [12] and Healey et al. [13] show that certain affective states may be recognized using physiological data. Levenson et al. [14] used three physiological signals namely; heart rate, skin conductance, and finger temperature to measure autonomic nervous system (ANS) patterns for emotions anger, sadness, disgust and fear using subjects from America and the West Sumatra. In terms of user experience for musical events Vaitl et al. [15] found the ANS differentiation while subjects were listening to Wagner operas during the Bayreuth Festival. Electrodermal response and respiratory activity measurements were used to analyze emotional arousal. Also, Krumhansl et al. [16] recorded physiological data while listeners were hearing music and analyzed them to find out what relationship existed between the physiological measurements and the dynamic ratings of emotions (i.e., happy, sadness and fear). Authors found that emotion state happy is linked to the largest changes in respiration measure, changes in heart rate, blood pressure, and skin temperature associated for sadness and the rate of blood flow associated with the emotion state fear.

Recently Kim et al. [17] investigated the potential of physiological signals as reliable channels for emotion recognition using a musical induction (i.e., subjects are listening to music) which spontaneously leads subjects to real emotional states. They collected 360 samples (samples were between 3–5 min) from three subjects using the Procomp Infinity which is an eight-channel multi-modal Biofeedback system with 14-bit resolution and a fiber optic cable connection to the computer. They measured electromyogram (EMG), skin conductivity, electrocardiogram, and respiration. Finally, they used multi-class classification using an extended linear discriminant analysis to recognize musical emotional states of subjects.

2.2 Engagement Analysis in Computer Vision

A survey of recent work in automatically measuring a person's behavior using vision-based approaches is presented in [18]. Much of this work has centered on recognizing an individual's facial expression, with notable progress made in the areas of smile detection in consumer electronics [19], pain detection [20] and human-computer-interaction [21]. An emerging area of research over the last couple of years is the use of affective computing for marketing and advertising purposes. When users watch video clips or listens to music, they may experience certain feelings and emotions [17] that manifest through gestural and physiological cues, such as laughter.

Shan et al. [22] studied the relationship between music features and emotions from film music. In a recent study, Joho et al., [23] showed that facial expression is a good feature to predict personal highlights in media content. Hoque et al. [24] further showed that these facial behaviors vary from the laboratory setting to real-world. Teixerira et al. [25] demonstrated that joy (i.e., smiles) was the most reliable emotion that accurately reflects the user's sentiments when analyzing the engagement with commercials. McDuff et al. [26] utilized crowd-sourcing to collect responses from people watching commercials and used

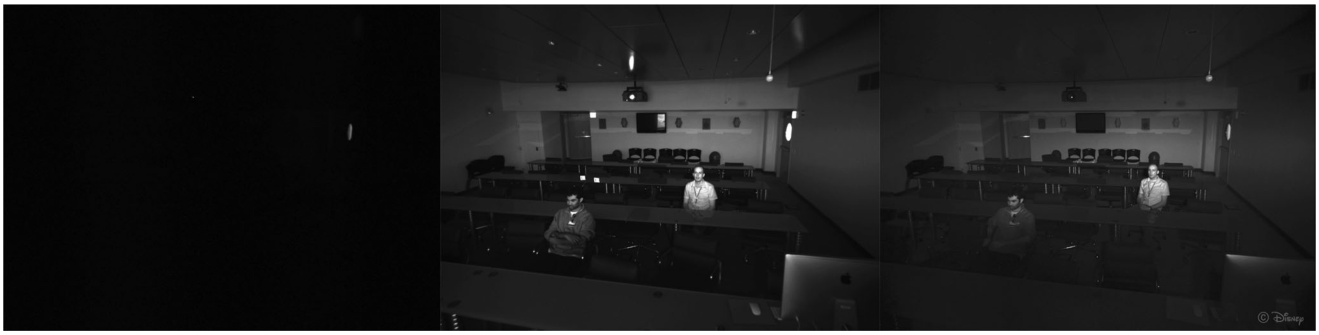

Fig. 2. (Left) Capturing video in a movie environment without IR illumination. (Middle) Example of the screening room with IR illuminators on—reflectance from the screen is problematic. (Right) We used an IR band-pass filter to remove the illumination reflected from the screen to obtain a uniform lighting environment.

smiles to gauge their reaction. They extended this work to predict the effectiveness of advertisements using smiles instead of "likes" [27]. Hernandez et al. [28] used a similar approach to measure the engagement of a single person watching a TV show. They mounted a camera on top of a TV set and recorded the responses of 47 participants. The Viola-Jones face detector [29] was used to locate the face, and conducted classification into four states of engagement based on facial movements. Recently, Whitehill et al. [7] used facial expression to understand student engagement in a classroom setting.

The above prior work was applied only to individuals and limited to stimuli of short duration. In this work we include simultaneous recordings of multiple individuals and continuous tracking of audience behaviour over long periods of time (e.g., up to 2 hours). Automatic long-term monitoring of human behavior is difficult: tracking people for this period of time is still an unsolved problem in computer vision. Additionally, being in a group environment introduces extra variability as behavior can be altered by other audience members as well as by the stimuli.

2.3 Key Poses for Long-Term Signals

Discovering key frames or key poses of human behaviors has been widely used in computer vision literature due to their ability to compactly represent the feature space both spatially and temporally. The simplest approach is to select key frames by randomly or uniformly sampling the video frames at predefined intervals [30]. Even though this method is very simple and fast, the major drawback is it neglects the content of the video (i.e., may not have selected the correct dictionary for optimal compression). Zhuang et al. [31] proposed an unsupervised clustering algorithm to extract key frames using color features. Li et al. [32] propose an algorithm based on color histograms to extract the key frames from face videos. The more advanced approaches use motion patterns to extract key frames [33], [34]. Liu et al. [33] used perceived motion energy to model motion patterns and extracted key frames using a threshold free approach. They defined key frames as the turning point of the motion acceleration and motion deceleration. In [31] authors computed the optical flow for each frame to measure the motion and selected key frames at the local minima of motion. However, calculating optical flow is computationally expensive, which can make large-scale analysis time consuming.

3 Dataset

Observing people watching movies is difficult because the environment is very dark, and light reflected from the screen creates fluctuating illumination (see Fig. 2). Wide aperture lenses and sensor sensitivity are two important features when working in low-light conditions. We instrumented a test-bed with an infra-red sensitive low-light camera (Allied Vision GX 1920 with a 2/3″ Sony ICX674 CCD sensor and a f/1.4 9 mm wide angle lens), two IR illuminators (Bosch UFLED95-8BD AEGIS illuminators with 850 nm wavelength and 95 degree wide beam pattern), and an IR band-pass filter to reduce reflections from the viewing screen (850 nm ± 5 nm). The IR camera is able to see in dark without effecting viewing conditions and the band-pass filter removes visible light reflected from the screen. The IR illuminators have 18 high efficiency surface mounted LED arrays and can spread around 50 m distance. The resulting captured images are 1,936 × 1,456 pixels captured at 15 frames per second. The schematic diagram of the IR illuminated testbed is given in Fig. 3.

3.1 Audience Footage

We selected movies from the *Animation, Comedy, Kids & Family* genre to screen. From those movies from 1998-2013, we selected a subset of ten movies (Table 1) with varying crowd sourced audience ratings from [35]. We chose three good movies (ratings greater than 80 percent), three average movies (ratings from 60 to 80 percent), and four bad movies (ratings below 60 percent).

We recruited participants to be part of an audience test screening ranging in size of 5-10 people (mean 8 people) for a session. This work was approved by our Institutional Review Board, and participants were compensated for their time. We screened the movies from 6:00 pm–8:30 pm and for each screening, ensured participants had not seen the movie previously and had normal or corrected-to-normal vision and hearing. We held three sessions for each movie (total 30 sessions) and each subject could only watch one movie. Our audience footage consists of 237 audience members (125 male and 112 female). The participants ranged in age from 18 to 70 and 63.3 percent were from 18-24 age group, 20.7 percent were from age 25-30 age group, 6.8 percent were from age 31-39 age group, 5.1 percent were from age 40-59 age group and 4.1 percent were from over 60 age group. The majority of the audience members were Caucasian (49.4 percent) with the

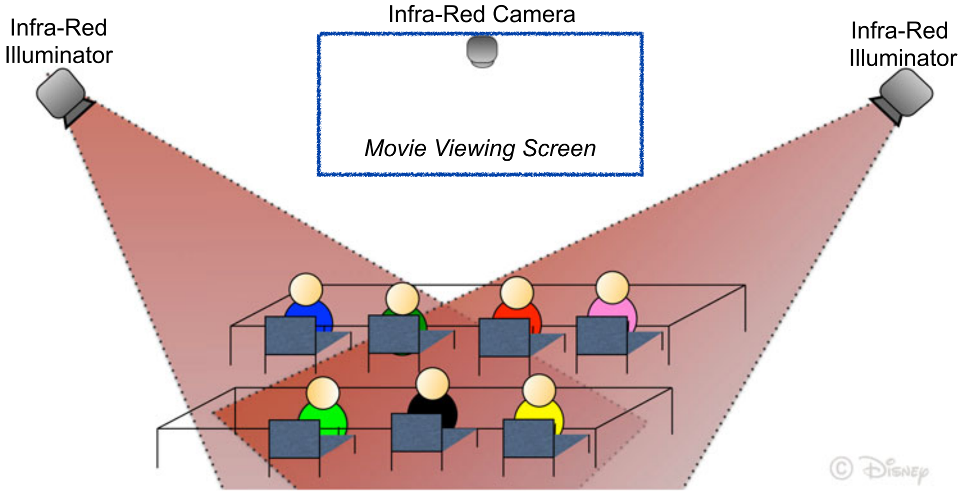

Fig. 3. A schematic of the audience testbed. We capture audience footage at 15 fps from an infrared camera, two IR illuminators and an IR band-pass filter to give a uniform visual signal.

reminder Asian (27.0 percent), African American (11.8 percent) and Hispanic (11.8 percent). At the completion of each session, every participant completed a survey asking about their overall rating of the movie, as well as their age, gender, movie genre preference, and expectation/recommendation of the movie.

We used the same approach as [35] to generate aggregate ratings from our individual survey responses: each audience member rated the movie on a scale of 1–5 stars, and the aggregate rating for the movie is the fraction of user ratings of 3.5 stars or higher. The overall correlation of our audience responses to *Rotten Tomatoes* users is 0.917 with $p \leq 0.001$. For the majority of movies our audience ratings were compatible with the *Rotten Tomatoes* users except movies M-8 and M-10. Therefore, the ratings from our sample audience are representatives of the general population. The detailed comparison of movie ratings using the self-report method from our audience members to the *RottenTomato* users for each movie is given in Fig. 4.

4 Discovering Audience Member Key-Frames

Talking to another person, checking phones/watches, smiling, large body pose changes such as stretching arms and changing sitting pose, eating and drinking, and falling asleep are key audience behaviors widely found when people are watching movies. The benefit of analyzing an audience environment is its low variance in behaviors compared to "in-the-wild" conditions due to the many constraints that exists. Such constraints are: i) every person shares the same input stimuli at the same time, ii) people tend to be stationary and are sitting, and iii) due to proxemics each persons tends to limit or maintain their personal space between each person. As audiences tend to adhere to this principal, we pre-defined an area of the camera's image for each seat. From videos cropped from these seat areas, we discover the *key frames*.

TABLE 1
An Inventory Showing the Number of Audience Members, Attributes, and the Rotten Tomatoes Rating per Movie

Movie	Sessions	Viewers	Duration [min]	Rating [%]
M-01	3	25	103	87
M-02	3	25	81	53
M-03	3	25	96	72
M-04	3	27	101	89
M-05	3	24	96	87
M-06	3	22	83	47
M-07	3	25	87	35
M-08	3	23	93	76
M-09	3	22	86	43
M-10	3	19	88	62

4.1 Per-Member Key Frame Dictionary

We follow an online learning approach to discover the audience key frame dictionary, as it has the ability to dynamically adapt to the incoming frames. Audience members do not move substantially during the movie—they tend to stay within the confines of their seat to maintain space between other audience members. Following the methodology of [8], we use the first frame to define a volume that the person will occupy.

Once we define a volume for each audience member, we calculate the similarity using a template matching approach [36]. We used the first frame of each audience member as the first template and calculate the similarity with the incoming frames in the video. For a given audience member i we assume our initial key frame is the first frame $I_{i,1}$ and update our key frame dictionary for audience member i to $d_i = \{I_{i,1}\}$. Then we calculate the similarity score $\alpha_{i,1}$ between frame $I_{i,1}$ and $I_{i,2}$. If the value $\alpha_{i,1}$ is less than predefine threshold β then we define the frame $I_{i,2}$ as another key frame and update template dictionary to $d_i = \{I_{i,1}, I_{i,2}\}$. We continue this process for each full-length movie for each audience member, which results in a key per-movie frame dictionary

$$\mathbf{D}_a = [d_1, d_2, \ldots, d_n], \tag{1}$$

where n is number of audience members watching the movie. Once we identify these key frames, it allows us to quickly annotate the interesting behaviors. As the number of key-frames are many orders of magnitudes lower than

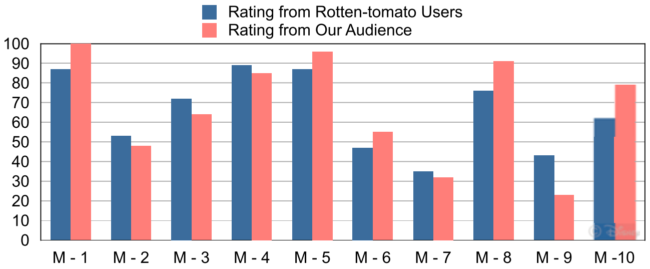

Fig. 4. A bar chart comparing the ratings of our recruited audience to the crowd-sourced ratings from rottentomatoes.com. M-1, M-4 and M-5 are high rated movies, M-3, M-8 and M-10 are well rated movies, and M-2, M-6, M-7 and M-9 are low rated movies from rottentomatoes.com.

the number of frames,[1] the annotation problem is reduced to assigning a label for each key frame from a small dictionary of activities which allows us to estimate audience sentiment. We use the similarity threshold value $\beta = 0.7$ as it shows a reasonable number of key frames and also has the ability to discover both subtle and coarse face (e.g., smiles, disgust, eye-closure versus head-pose change) and body (e.g., fidgeting versus stretching) behaviors. An example of the relationship between the threshold β and number of audience key frames is shown in Fig. 5. We discovered the interesting audience key frames across all the 30 movie sessions in the collected audience dataset. The break down of the proposed approach is given in Algorithm 1.

Algorithm 1. Discovering Audience Key Frames

```
Data: Audience video or frames $V = \{I_1, I_2, \ldots, I_n\}$
Result: Dictionary D with key frames
$\beta \leftarrow$ Threshold;
index $\leftarrow$ 1;
$D_{index} \leftarrow I_1$;
Initialize q $\leftarrow 0$;
for $i = 2$ to $n$ do
  nD $\leftarrow$ length($D$);
  for $j = 1$ to $nD$ do
    Template $\leftarrow D_j$;
    Image $\leftarrow I_i$;
    Compute $\alpha$ between $I_i$ and $D_j$ ;
    Add $\alpha$ to q;
  end
  a $\leftarrow$ maximum of q;
  if $a < \beta$ then
    index $\leftarrow$ index+1;
    $D_{index} \leftarrow I_i$ ;
  end
  empty q;
end
return D;
```

5 Extracting Audience Features

Whitehill [7] found that humans rely on head pose, and elementary facial actions like brow raise, eye closure, and upper lip raise to make judgments of engagement. In audience domains, visual components such as visibility of face, whether an audience member is looking at the screen, large body motions, smiling, yawning and sleeping may be indicative of engagement levels. In this study, we extract visual features for body motion and face expression, and investigate their predictive power.

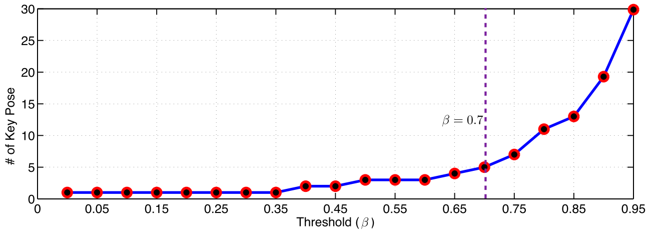

Fig. 5. Variation in the number of audience key frames using different thresholds β.

5.1 Body Motion

In terms of recognizing individual and specific actions, there is a plethora of research that has solely focused on this domain, with excellent progress being made [37]. Efros et al. [38] used optical flow features to recognize actions from ballet, soccer and tennis. More recently, Rodriguez et al. [39] used similar features to analyze crowds. However, we are not interested in the specific actions of one person but instead the synchronicity of actions (i.e., is everyone doing the same thing at the same time?). The screening room environment introduced a natural spacing of audience members so each person could watch the movie unoccluded and in comfort, resulting in each person occupying a minimum uninterrupted 3D volume. We experimented with an aggregated real-time approach to represent the spatio-temporal motion that recursively integrates frame differences into a motion history image [40]. We found that this representation is equally reliable in audience environment to optical flow [41], but with substantially less computational burden. This is done by layering the threshold differences between consecutive frames one over the other. This represents *how much* motion is present in the image as opposed to *magnitude and direction* of the motion. A motion history image is calculated as

$$H_\gamma(x,y,t) = \begin{cases} \gamma & \text{if } D(x,y,t) = 1, \\ \max(0, H_\gamma(x,y,t-1) - 1) & \text{otherwise}, \end{cases} \quad (2)$$

where $D(x,y,t)$ is a binary image sequence indicating regions of motion at pixel (x,y) at time t, and the parameter γ is the temporal duration of the motion history images. We then calculate the normalized local 3D energy for person q as $\mathbf{e}(q,t) = \frac{1}{N_q}\sum H_\gamma(x,y,t)$ where N_q is the size of the predefined volume.

5.2 Face Attributes

Faces provide useful information such as gaze angle (e.g., is the person looking at the screen) and expression (smiling, yawning and sleeping). These attributes are strong cues for estimating engagement level (i.e., the person is engaged or disengaged with the movie). To extract these attributes from each audience face, the first task is to register the location of the face of a given audience member.

[1] Due to the long length of input stimuli (approximately 1-2 hours per movie), it is highly impractical and unscalable to get higher level of annotation in every frame for each audience and it would be expected that the reliability of annotation would greatly diminish due to the high level of subjectivity.

5.2.1 Face Detection

Despite an audience member remaining relatively stationary whilst watching a movie, continuous face detection/tracking is challenging because there are considerable appearance changes due to out-of-plane head motion or self-occlusion (e.g., hands on the face). While face tracking is a mature area of research, most of the previous work has only looked at videos of small periods of time (i.e., up to one minute). The intuitive method of registering each audience member would be to use an off-the-shelf face-detector/template update approaches [42], [43], [44] on each frame and then track each detection. This approach works well in ideal conditions but not so well in our test-bed because we are capturing faces from a different viewpoint (i.e., camera is looking down on the audience), we are operating in the infrared spectrum, and the resolution of faces can be small [8].

Recently, King et al. [45] proposed a method known as Max-Margin Object Detection (MMOD), which optimizes over all sub-windows to detect objects in images. This approach learns a Histogram of Oriented Gradients (HoG) [46] template on training images using structural support vector machines which enables it to train on all the sub-windows in every training image (efficiently finding the 'hard negatives' automatically). This approach works well for a fixed environment as the detector is discriminative. As we are operating in a fixed IR spectrum, we use such an MMOD implementation to train models and detect faces. The face detector was trained on labeled faces in our audience dataset. In particular, we trained a HOG face detector using about 800 images. We use the DLib C++ Library[2] to train an environment specific face model and then to detect audience faces the entire dataset. We found this implementation runs at 4-6 frames per seconds. On a validation set of 10,000 images, the precision and recall values are 99.5 percent and 94.2 percent. The missed faces were manually cropped for analysis and all face images were normalized into 48×32 image patches.

5.2.2 Gaze Angle

Head pose of an audience member (i.e., frontal/near frontal, looking away from the screen and looking down), while watching the movie provides useful information since the head pose usually indicates the focus of attention. To create a focus of attention, calculating the head pose is crucial since it usually coincides with the gaze direction [47].

A plethora of work has been conducted to estimate the head pose estimation based on appearance-based and model-based approaches. Appearance-based methods [48], [49], [50] concentrate on face detection and consider the pose estimation problem as a classification problem using pre-defined head orientation classes. These approaches are quite efficient in terms of computation time, but do not estimate all three rotation angles (i.e., roll, pitch and yaw). Model-based methods use a geometric model of the face to estimate the head pose. Stiefelhagen et al. [51] and Gee et al. [52] extract a set of facial features and map the features onto the 3D model using perspective projection, while Dornaika et al. [53] apply an active appearance model and use the contours and features of the face to estimate the head poses. In this work, we use such an approach to calculate the rotation matrix of a given audience member.

TABLE 2
Description of the Features

Feature	Description
$\mathbf{f}_{angle}$	Visual focus for movie screen
$\mathbf{f}_{hog}$	Facial expressions
$\mathbf{x}_{face}$	Face features: visual focus + facial expressions
$\mathbf{b}_{w}$	Body motion feature
$\mathbf{x}_{facebody}$	Combination of face and body features

For each detected face, we use the DLib 68 landmark shape model to generate face landmark locations [54]. Then we associate the 68 fitted 2D face landmark locations to a 3D face mesh from Face Warehouse [55] to calculate the 3D rotation matrix $\mathbf{R}$ to estimate the roll, pitch and yaw of the audience member ($\mathbf{f}_{angle}$).

5.2.3 Visual Expressions

Humans use facial expressions such as smiling/laughter, yawning and sleeping as very strong cues to understand if an audience member is in engaged or disengaged. In terms of understanding these behaviors automatically, concatenating of filter responses before learning a classifier has found particular success in facial expression recognition [56] compared to learning those classifiers with appearance intensities. We designed such an approach by calculating HOG features [46] from the given face image. Our visual expression feature $\mathbf{f}_{hog}$ consists of representing the input face image via HOG descriptor using 9 orientation bins with overlapping regions with block size of 2×2, and cell size of 8×8.

A summary of the features is given in Table 2.

6 Estimating Audience Engagement

The visual appearance of an audience member may give an indication of how *engaged* or *disengaged* they are during various segments of the movie. The problem of defining or learning affective states such as disengagement and engagement is difficult [57] compared to facial expression recognition such as happy, sad, angry, or surprised [58]. Devising an adequately clear definition of labeling procedure is important for the reliability and validity of the training labels. Recently Whitehill et al. [7] found that viewing static images and labeling engagement levels is a more reliable solution than watching the video clip and continuously labeling engagement levels in a classroom student engagement setting. Motivated by this approach, we organized a team of three annotators to label the key frames. These annotators viewed each key frame and labeled each frame as either engaged or disengaged. Annotators were instructed to annotate key frames according to *how engaged does the subject appear to be* rather than *predict what they were actually thinking*. Specific instructions to the annotators were:

- *Engaged.* The audience member's main focus is on the screen. Visual components such as facial expressions (*smiles/laughter*) and *leaning forward* can be identified as key attributes. Additionally, audience members that look relaxed and have no expression can still be highly engaged.

[2] http://dlib.net/

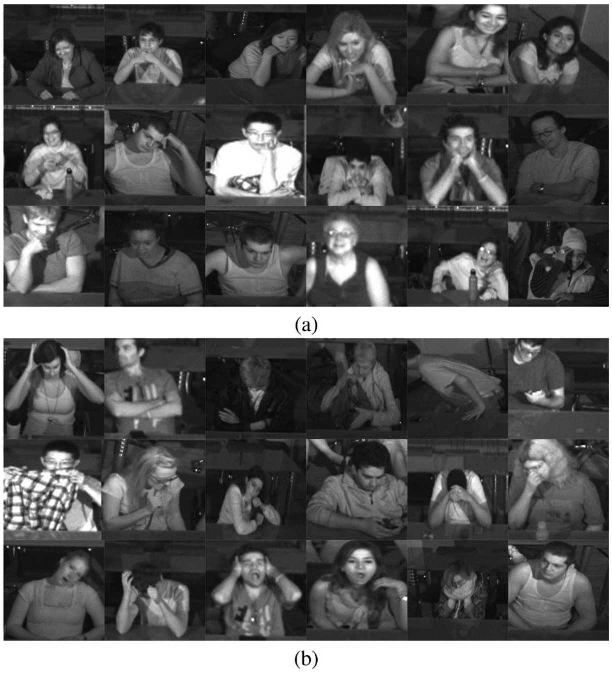

Fig. 6. Sample of audience behavior for (a) engaged and (c) disengaged.

- *Disengaged.* The audience member's focus is not on the movie screen. Visual components such as: looking away from the screen, looking at his/her phone/watch, eating/drinking, sleeping, large-body motions/doodling, yawning can be identified as disengaged attributes.

Visual examples are given in Fig. 6.

6.1 Discovery of Engagement Levels Using Human Labels

Across our audience data set we discovered 10,787 key frames. The discovered key frames are many orders of magnitudes lower than the number of frames. The discovered key frames were consistent with visual components such as visibility of face, view of the face (i.e., frontal or near frontal, looking away, looking down, occluded), larger body motions, smiling, yawning, sleeping (or eye closure) and eating/drinking.

The key frames were shuffled both in time and across subjects. Human observers labeled these key frames for the appearance of engagement, as described in Section 6. Among these three annotators, Fleiss' Kappa value was 0.59, and the average pairwise observed agreement was 0.812 (with an expected agreement of 0.542). All annotators fully agreed for 7,741 key frames (i.e., ≈ 72 percent of the data). Among these fully agreed key frames there are 5,403 engaged key frames and 2,338 disengaged key frames.

7 Learning

We focus our study of automatic engagement classification based on the features in Section 5. Our focus is not only to understand how predictive each feature type is for engagement classification but also to asses these state-of-the-art computer vision architectures for a novel application. We propose an automatic engagement classification pipeline using (a) face features only and (b) motion features only and (c) a combination of face and motion features. The propose pipeline is given in Fig. 7.

We conducted our experiments using the key frames that all annotators were in agreement over. We divided these key frames into 5 different groups. We conducted 5-fold cross-validation using these groups. We trained a binary random forest classifier. The experiments were conducted using the feature types shown in Table 2. The main interest to perform individual experiments using each attribute is to discover how well each attribute can distinguish engagement levels.

7.1 Model Complexity

To understand the model complexity of the random forest for the task of engagement classification, initially we extracted different features as shown in Table 2 namely: a) motion, b) gaze c) face HOG, d) (gaze + HOG) and e) motion + gaze + HOG. Then we conduct binary engagement classification experiments to see the model complexity using HOG features for given number of trees (see Fig. 8a). We also extended the experiments using the other four features and the variation of error with the number of trees is given in Fig. 8b. The performance of the test error after 100 trees has no significant difference. We also tested the variation of error with different tree depths. In all cases, we set the number of trees to 100. The variation of error with the depth of the trees is shown in Fig. 9.

8 Estimating Audience Member Engagement

In this section we describe the engagement classification accuracy using the proposed pipeline: (a) face features only, (b) motion features only, and (c) combination of face and motion features.

8.1 Only Body Motion Features

We calculated the motion history images as descried in Section 5.1. We calculate the motion magnitudes for each audience member and normalized and scaled according to

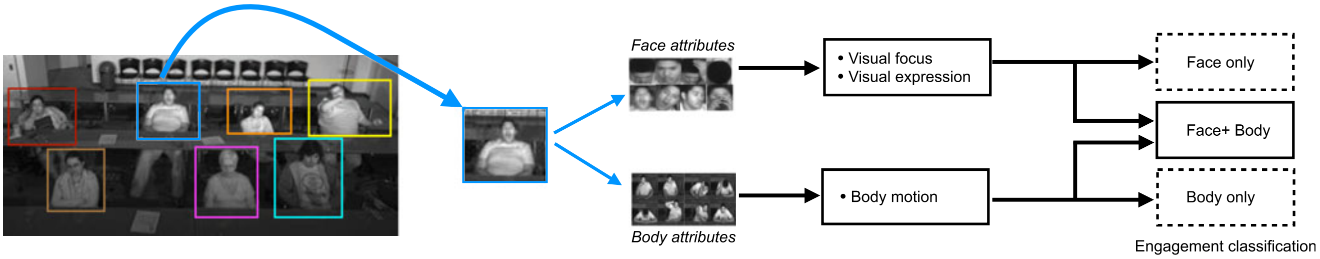

Fig. 7. The proposed engagement pipeline consists of first capturing both face and body key frames. Face attributes consist of (i) visual focus (looking at the screen) or expressions such as smiling, yawning and sleeping. Body attributes consist of stretching or fidgeting. The final sentiment score is estimated using face only, body only and a combination of face and body features.

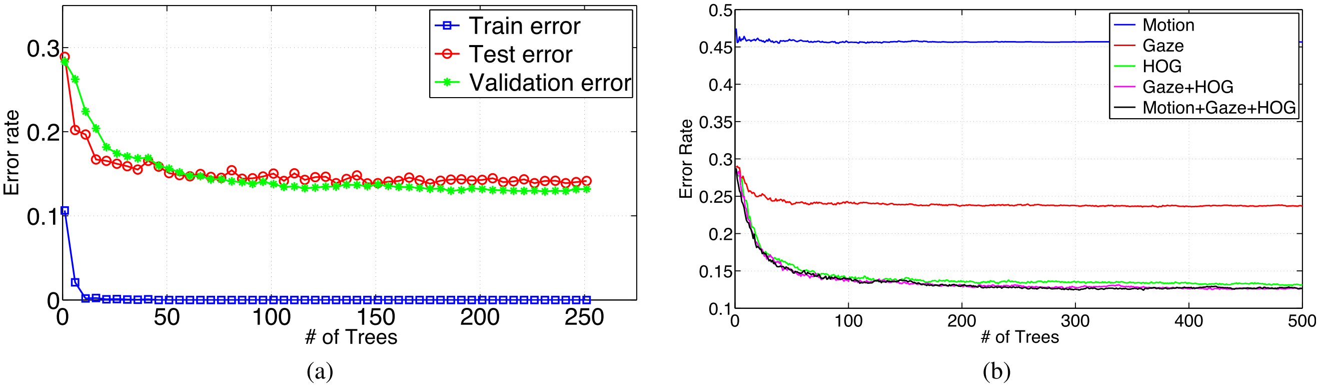

Fig. 8. (a) Model complexity using the HOG features. (b) The variation of test error with the number of trees.

the image window size. We obtain motion feature $\mathbf{b}_w$ for each key frame and conducted 5 fold-cross validation engagement classification experiments. The average accuracy using the motion feature is 0.5866 ± 0.05. The average accuracy with $\pm 2\sigma$ is given in Table 3.

8.2 Only Facial Features

8.2.1 Pose Angle: $\mathbf{f}_{angle}$

Feature $\mathbf{f}_{angle}$ consist of yaw, roll and pitch angle from each face in key-frames. The average engagement classification accuracy is 0.7973±0.03.

8.2.2 Visual Appearance: $\mathbf{f}_{hog}$

In terms of understanding facial expression individually, the most common approach is a concatenation of filter responses, before learning a classifier. Instead of maximizing the likelihood for recognizing facial expressions such as smiling, yawning and sleeping; we maximize the likelihood for engagement. As described in the Section 5.2, we extracted HOG features $\mathbf{f}_{hog}$ (756 in dimensions) and used engagement labels to learn a classifier. The accuracy was 0.8412, which is higher than both body motion and gaze angle.

8.2.3 Face Feature: $\mathbf{x}_{face}$

Finally, we combined the pose angle feature $\mathbf{f}_{angle}$ and HoG features from the face $\mathbf{f}_{hog}$ to obtain the final face feature vector such that $\mathbf{x}_{face} = [\,\mathbf{f}_{angle} \quad \mathbf{f}_{hog}\,]$ with a dimension of 759. We achieved best accuracy with the combination of these two attributes with a average of 0.8532 over the 5 folds.

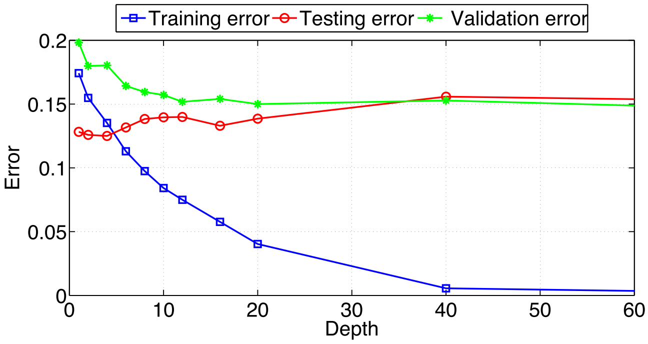

Fig. 9. Model complexity with the variation of tree depth.

8.3 Body Motion and Facial Features

Finally we combined the face features $\mathbf{x}_{face}$ and motion feature $\mathbf{b}_w$ to obtain the final face & gesture vector $\mathbf{x}_{facebody} = [\,\mathbf{x}_{face} \quad \mathbf{b}_w\,]$ which is 760 dimensions. Across all the 5-folds, we found that adding face features improves the engagement classification accuracy. Overall average engagement accuracy is given in Table 3 and confusion matrices are shown in Fig. 10.

As shown in Fig. 10, the combination of face and body features improved the diagonal values compared to all the other feature representations. The overall accuracy for this configuration with $\mathbf{x}_{facebody}$ is 0.8582±0.07. We observed that adding motion features combined with the face features helped to distinguish more disengaged behaviors. The results suggest that the proposed framework can be used to distinguish extreme sentiment states very well in an audience environment. The finding may be applicable to other domains such as educational, behavioral science and entertainment.

8.4 Significance of Features

As shown in Table 3 the engagement classification performance of the features clearly shows (motion + HOG + gaze) > (HOG + gaze) > HOG > gaze > motion. To see whether the performance is statistically significant, we tested the classifiers using McNemar's test [59].

McNemar's test values are given in Table 4. The cut off for the χ^2 test value at 99 percent is 6.635 (i.e., p value is $p < 0.01$). From our analysis, we observe that HOG features have significant improvement over motion only, and gaze only. Combining gaze + HoG is not significant, but the combination of all three features (HOG + Gaze + Motion) is.

9 Predicting Movie Ratings

Finally, we investigated the effect of our engagement analysis framework to the task of movie prediction. To gauge

TABLE 3
Engagement Classification Accuracy

Feature Type		Average Accuracy (Test)
Body	$\mathbf{b}_w$	0.5886 ± 0.05
	$\mathbf{f}_{angle}$	0.7914 ± 0.03
Face	$\mathbf{f}_{Hog}$	0.8412 ± 0.04
	$\mathbf{x}_{face}$	0.8532 ± 0.04
Face + Body	$\mathbf{x}_{facebody}$	0.8582 ± 0.07

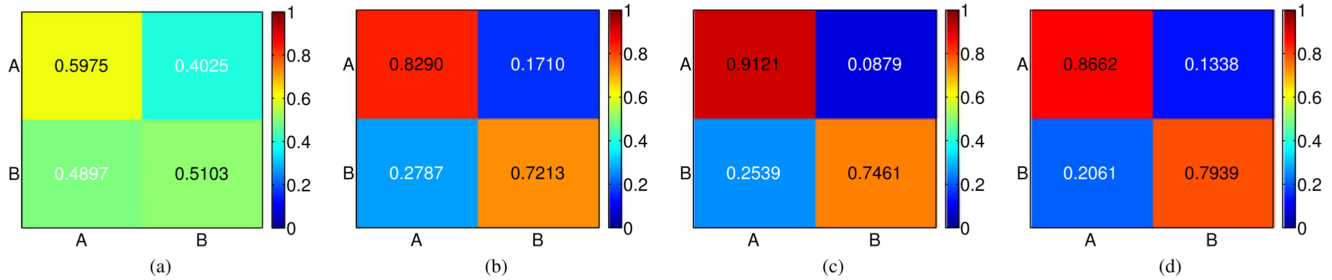

Fig. 10. Confusion matrices using body and face features: (a) $\mathbf{b}_w$, (b) $\mathbf{f}_{angle}$, (c) $\mathbf{x}_{face}$ and (d) $\mathbf{x}_{facebody}$. The terms: $\mathbf{A}$ engage and $\mathbf{B}$ disengage classes.

how much the general public likes a particular movie, *rottentomatoes.com* has an interactive feature that allows people to submit a rating. Over time the number of ratings aggregate (100k's) and based on these crowd-sourced ratings, they generate an average audience measure. In *rottentomatoes.com*, a movie with an average audience measure of 60 percent or higher is deemed a good movie and below 60 percent denotes a bad movie.

9.1 Audience Feature Representation

9.1.1 Body Motion

In previous work, we used the synchronicity of body motion to predict movie ratings. Specifically, we compared body motion features $\mathbf{f}_{i,t}$ and $\mathbf{f}_{j,t}$ of two audience members concatenated over a temporal window t, and calculated the pairwise similarity

$$s_{ij,t} = \exp\left(\frac{- \parallel \mathbf{f}_{i,t} - \mathbf{f}_{j,t} \parallel^2}{2\sigma^2}\right), \tag{3}$$

where $\sigma = 0.5$. We then exhaustively calculated all of the pairwise correlations between audience members to produce the collection $\mathbf{S}_t$ of similarity scores $s_{ij,t}$. When everyone is doing something at the same time (e.g., laughing/smiling) the cohesion is high; similarly, when everyone is doing nothing, the audience cohesion is still high. Given the collection of similarity scores for a given temporal window t, we generate a probability distribution $p(s_{ij,t})$, allowing us to gauge the synchronicity of the audience reaction for the given time window t using entropy [60]

$$X_t^{\text{body}} = \sum_{ij} p(s_{ij,t}) \log p(s_{ij,t}). \tag{4}$$

9.1.2 Engagement Level

Alternatively, we use the combination of face and body features to predict an engagement score for each audience member $e_{i,t}$ over time.[3] For a given time window t, we compute the average audience engagement

$$X_t^{\text{engaged}} = \frac{1}{N}\sum_{n=1}^{N} e_{i,n}. \tag{5}$$

9.2 Temporal Aggregation

The instantaneous state X_t of the audience is measured from multiple frames of data collected over the time window t (using either body motion or predictions of engagement). Because movies have different temporal durations, we sample feature data from the first, middle and last 30 minutes of the movie to create a fixed length feature vector regardless of the duration of a movie.

In order to determine the optimal temporal window size over which to compute audience state using either entropy (body motion) or average (engagement level), we selected different window sizes from 10s-150s. For each window size, we created corresponding feature vectors $\mathbf{X}$ for the first, middle and last 30 minutes of the movie by concatenating the feature representation X_t for each temporal window. We validate our framework using leave-one-out cross-validation experiments leaving out an entire movie. The parameters for support vector regression were chosen using a cross-validation method as described in [61]. The average mean squared validation error with different window sizes is given in Fig. 11. A 30s window generally produces the lowest prediction error for every segment of the movie.

9.3 Audience Ratings Prediction

As movies are variable length, we extracted features for (a) audience body motion and (b) audience engagement levels methods in Section 9.1 using first, middle and last 30 minutes of the movies to create a fixed length feature representation. We used these features and *Rotten Tomatoes* ratings (see Table 1) to predict the audience rating for an unseen movie. The full comparison of prediction error for all the movies from (a) audience body motion and (b) audience engagement levels is given in Table 5. In addition to the proposed method, we also report the error from the audience survey responses. The average prediction error from (i) audience body motion, (ii) engagement levels and (iii) survey responses are 19.8, 12.6 and 14.2 respectively. The results imply that our engagement pipeline can

TABLE 4
McNemar's Test Values

Classifiers	McNemar's test χ^2
Motion versus HOG	334.89
Gaze versus HOG	54.23
(Gaze + HOG) versus HOG	5.32
(Motion + Gaze + HOG) versus HOG	14.49

[3] For simplicity, we assume an audience member has the same engagement score from one key frame to the next key frame.

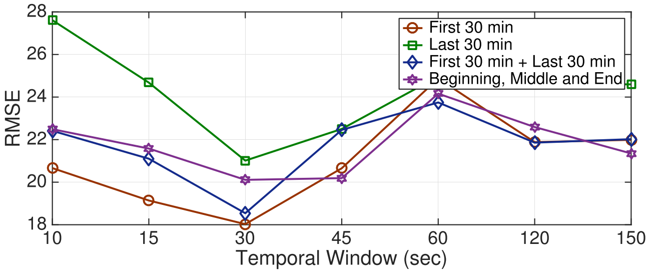

Fig. 11. Variation of average RMSE with respect to different temporal window sizes using the interesting segments of the movie.

TABLE 5
The Prediction Error of our Automatic Audience Rating Approaches Compared to the Crowd-Sourced Ones from *rottentomatoes.com*

Movie	Rotten Tomatoes		Prediction Error	
	Rating	Survey Model	Motion Model	Engagement Model
M-01	87	5.0	26.0	**4.5**
M-02	53	16.0	**12.2**	13.1
M-03	72	6.0	16.0	**1.5**
M-04	89	**8.0**	26.4	17.2
M-05	87	6.0	20.7	**5.6**
M-06	47	29.0	14.8	**5.6**
M-07	35	31.0	30.0	**20.2**
M-08	76	**10.0**	11.7	22.0
M-09	43	**12.0**	32.5	17.4
M-10	62	19.0	**7.8**	20.0
Average	-	14.2	19.8	12.6

generate very good predictions of overall audience enjoyment of a movie—even outperforming predictions based on audience responses to exit surveys.

10 Conclusions

In this paper, we present a framework to estimate the engagement of audience members and to predict movie ratings, based on face expressions and body motions. The problem is challenging because: i) the movie viewing environment is dark and contains views of people at different scales and viewpoints, ii) the duration of feature-length movies is long (80-120 mins) and tracking people uninterrupted for this length of time is difficult, iii) expressions and motions of audience members are subtle, short and sparse making labeling of activities unreliable, and iv) annotating the sentiment at the frame-level is prohibitive. To circumvent these issues, we use an infrared illuminated testbed to obtain a visually uniform input video. Due to the enormous amount of video data to process, we first discovered the key frames within a predefined image region for each audience member. We extracted face and body features for each key frame and learned classifiers to estimate whether a key frame represents an engaged or disengaged behavior.

In addition to the proposed audience engagement level architecture, we also proposed an automatic approach to predict movie ratings solely using audience behaviors. We showed that audience sentiment levels can be more predictive of overall movie rating than self-report measurements. We tested the utility of our approach using 30 movie sessions across more than 200 subjects ($<$ 50 hours of video data).

References

[1] R. Bales, *Social Inteaction System: Theory and Measurement*. Piscataway, NJ, USA: Transaction Publishers, 1999.

[2] M. Chaouachi, P. Chalfoun, I. Jraidi, and C. Frasson, "Affect and menatl engagement: Towards adaptability for intelligent systems," in *Proc. 23rd Int. Conf. FLAIRS*, 2010, pp. 355–360.

[3] B. Goldberg, R. Sottilare, K. Brawner, and H. Holden, "Predicting learner engagement during well-defined and ill-defined computer based intercultural interactions," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2011, pp. 538–547.

[4] A. Pope, E. Bogart, and D. Bartolome, "Biocybernetic system evalutes indices of operator engagement in automated task," *Biol. Psychology*, vol. 40, no. 1–2, pp. 187–195, 1995.

[5] S. Makeig, J. Westerfield, J. Townsend, T. Jung, E. Courchesne, and T. Sejnowski, "Functionally independent components of early event-related potentials in a visual spatial attention task," *Philosoph. Trans. Roy. Soc.: Biol. Sci.*, vol. 354, no. 1387, pp. 1135–1144, 1999.

[6] R. Anna Marie and R. Levenson, "Continuous measurement of emotion: The affect rating dial," *Handbook of Emotion Elicitation and Assessment*, New York, NY, USA: Oxford University Press, 2007.

[7] J. Whitehill, Z. Serpell, Y. Lin, A. Foster, and J. Movellan, "The faces of engagement: Automatic recognition of student engagement from facial expressions," *IEEE Trans. Affect. Comput.*, vol. 5, no. 1, pp. 86–98, Jan.-Mar. 2014.

[8] R. Navarathna, P. Lucey, P. Carr, E. Carter, S. Sridharan, and I. Matthews, "Predicting movie ratings from audience behaviors," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2014, pp. 1058–1065.

[9] N. Schwarz and F. Strack, "Reports of subjective well-being: Judgmental processes and their methodological implications," *Well-Being: Foundations Hedonic Psychology*, 1999.

[10] C. Varner and G. Dickinson, "The lecture, an analysis and review of research," *Adult Edu. Quart.*, vol. 17, pp. 85–100, 1967.

[11] W. Murch, *In the Blink of an Eye: A Perspective on Film Editing*. West Hollywood, CA, USA: Silman-James Press, 2001.

[12] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, pp. 1175–1191, Oct. 2001. [Online]. Available: http://dx.doi.org/10.1109/34.954607

[13] J. Healey and P. W. Picard, "Digital Processing of Affective Signals," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Proc.*, 1998, pp. 3749–3752.

[14] R. Levenson, P. Ekman, K. Heider, and W. Friesen, "Emotion and autonomic nervous system activity in the minangkabau of west sumatra," in *J. Personality Soc. Psychology*, vol. 62, no. 6, pp. 972–988, 1992.

[15] D. Vaitl, W. Vehrs, and S. Sternagel, "Prompts-leitmotif-emotion: Play it again, Richard Wagner," *The Structure of Emotion: Psychophysiological, Cognitive, and Clinical Aspects*, Oxford, United Kingdom: Hogrefe & Huber, 1993.

[16] C. Krumhansl, "An exploratory study of musical emotions and psychophysiology," *Can. J. Experimental Psychology*, vol. 51, no. 4, pp. 336–353, 1997.

[17] J. Kim and E. Andre, "Emotion recognition based on physiological changes in music listening," in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2067–2083, Dec. 2008.

[18] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[19] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan, "Towards practical smile detection," in *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 31, no. 11, pp. 2106–2111, Nov. 2009.

[20] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The UNBC-McMaster shoulder pain expression archive database," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognition Workshops*, 2011, pp. 57–64.

[21] A. Vinciarelli, M. Pantic, and H. Bourland, "Social signal processing: Survey of an emerging domain," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1743–1759, 2009.

[22] M. Shan, F. Kuo, M. Chiang, and Y. Lee, "Emotion-based music recommendation by affinity discovery from film music," *An Int. J. Expert Syst. Appl.*, vol. 36, no. 4, pp. 7666–7674, 2009.

[23] H. Joho, J. Staiano, N. Sebe, and J. Jose, "Looking at the viewer: Analysing facial activity to detect personal highlights of multimedia contents," in *Multimedia Tools Appl.*, vol. 51, no. 2, pp. 505–523, 2011.

[24] M. Hoque and R. Picard, "Acted versus natural frustration and delight: Many people smile in natural frustration," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognition Workshops*, 2011, pp. 354–359.

[25] T. Teixerira, M. Wedel, and R. Pieters, "Emotion-induced engagement in internet video advertisements," *J. Marketing Res.*, vol. 49, no. 2, pp. 144–159, Apr. 2012.

[26] D. McDuff, R. Kaliouby, and R. Picard, "Crowdsourcing facial responses to online videos," in *IEEE Trans. Affect. Comput.*, vol. 3, no. 4, pp. 456–468, Oct.–Dec. 2012.

[27] D. McDuff, R. Kaliouby, D. Demirdjian, and R. Picard, "Predicting online media effectiveness based on smile responses gathered over the internet," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognition Workshops*, pp. 1–7, 2013.

[28] J. Hernandez, L. Zicheng, G. Hulten, D. DeBarr, K. Krum, and Z. Zhang, "Measuring the engagement level of TV viewers," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognition Workshops*, 2013, pp. 1–7.

[29] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition*, pp. I-511–I-518, 2001.

[30] R. Hammound and R. Mohr, "A probabilistic framework of selecting effective key frames from video browsing and indexing," in *Proc. Int. Workshop Real-Time Image Sequence Anal.*, 2000, pp. 79–88.

[31] Y. Zhuang, Y. Rui, T. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proc. Int. Conf. Image Process.*, vol. 1, pp. 866–870, Oct. 1998.

[32] X. Li and T. Xu, "Face video key-frame extraction algorithm based on color histogram," *Proc. Int. Conf. Comput. Sci. Inf. Technol.*, 2011.

[33] T. Liu, H.-J. Zhang, and F. Qi, "A novel video key-frame-extraction algorithm based on perceived motion energy model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 10, pp. 1006–1013, Oct. 2003.

[34] W. Wolf, "Key frame selection by motion analysis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2, pp. 1228–1231, May 1996.

[35] [Online.] Available: http://www.rottentomatoes.com/

[36] J. P. Lewis, "Fast normalized cross-correlation," Tech. Rep., Industrial Light & Magic, 1995.

[37] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Comput. Surveys*, vol. 43, no. 3, Apr. 2011, Art. no. 16.

[38] A. Efros, C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, 2003, pp. 726–733.

[39] M. Rodriguez, J. Sivic, I. Laptev, and J. Audibert, "Data-driven crown analysis in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1235–1242.

[40] J. Davis and A. Bobick, "The representation and recognition of action using temporal templates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition*, 1997, pp. 928–934.

[41] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Int. Joint Conf. Artificial Intell.*, 1981, pp. 674–679.

[42] S. Baker and I. Matthews, "Lucan-Kanade 20 years on: A unifying framework: Part 1: The quantity approximated, the warp update rule, and the gradient descent approximation," *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 221–255, Feb. 2004.

[43] S. Lucey, R. Navarathna, A. Ashraf, and S. Sridharan, "Fourier Lucas-Kanade algorithm," *IEEE Tran. Pattern Anal. Mach.*, vol. 35, no. 6, pp. 1383–1396, Jun. 2013.

[44] S. Kalantari, R. Navarathna, D. Dean, and S. Sridharan, "Visual front-end wars: Violoa-Jones face detector versus Fourier Lucas-Kanade," in *Proc. Int. Conf. Auditory Vis. Speech Process.*, 2013, pp. 163–168.

[45] D. E. King, "Max-margin object detection," *CoRR*, 2015.

[46] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. Conf. Comput. Vis. Pattern Recognition*, 2005, pp. 886–893.

[47] M. Bennewitz, F. Faber, D. Joho, S. Schreiber, and S. Behnke, "Towards a humanoid museum guide robot that interacts with multiple persons," *In Proc. IEEE/RSJ Int. Conf. Humanoid Robots*, 2005, pp. 418–423.

[48] J. Meynet, T. Arsan, J. Mota, and J. Thiran, "Fast multiview tracking with pose estimation," Comput. Eng. Dept., Kadir Has Univ., Istanbul 34230, Turkey, Technical Rep. TR-ITS.2007.01, 2007.

[49] R. Rae and H. Ritter, "Recognition of human head orientation based on artificial neural nets," *IEEE Trans. Neural Netw.*, vol. 9, no. 2, pp. 257–268, Mar. 1998.

[50] R. Stiefelhagen, "Estimating head pose with neural networks–results on the pointing04 icpr workshop evaluation data," in *Proc. Pointing '04 ICPR Workshop Int. Conf. Pattern Recognition*, 2004.

[51] J. Stiefelhagen, R. Yang, and A. Waibel, "A model-based gaze tracking system," in *Proc. IEEE Int. Joint Symposia Int. Syst.*, 1996, pp. 304–310.

[52] A. Gee and R. Cipolla, "Determine the gaze of faces in images," *Image Vis. Comput.*, vol. 12, no. 10, pp. 639–647, 1994.

[53] F. Dornaika and J. Ahlberg, "Fast and reliable active appearance model search for 3D face tracking," *IEEE Trans. Syst., Man Cybern., Part B*, vol. 34, no. 4, pp. 1838–1853, Aug. 2004.

[54] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition*, 2014, pp. 1867–1874.

[55] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3D facial expression database for visual computing," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 3, pp. 413–425, Mar. 2014. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2013.249

[56] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: Machine learning and application to spontaneous behavior," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition*, Jun. 2005, vol. 2, pp. 568–573.

[57] K. Porayska-Pomsta, M. Mavrikis, S. D'Mello, C. Conati, and R. Baker, "Knowledge elicitation methods for affect modemodel in education," *Int. J. Artificial Int. Edu.*, pp. 107–140, 2013.

[58] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition*, Jun. 2010, pp. 94–101.

[59] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, pp. 153–157, 1947.

[60] C. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, pp. 379–423, 1948.

[61] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Tech. Rep., Taipei, 2003.

Rajitha Navarathna received the bachelor of computer engineering (with First Class Hons.) degree from the University of Peradeniya, Sri Lanka, in 2008. He received the PhD degree at the Queensland University of Technology in the SAIVT Laboratory, Australia. He is a postdoctoral associate at Disney Research in Pittsburgh, where his main focus is on "Video Analytics", where he analyze an enormous amount of video for analytical purposes. Generally, he is most interested in creating machine vision systems which can sense and understand the world. Before joining Disney Research.

Peter Carr received the bachelor's of applied science (engineering physics) from Queen's University, Kingston, Canada, the master's degree in physics from the Centre for Vision Research, York University, Toronto, Canada, and the PhD degree from the Australian National University, in 2010, under the supervision of Prof. Richard Hartley. He is a research scientist at Disney Research, Pittsburgh. His research interests lie at the intersection of computer vision, machine learning and robotics. He joined Disney Research in 2010 as a postdoctoral researcher.

Patrick Lucey received the bachelor of electrical engineering (with first class Hons.) degree from the University of Southern Queensland, Australia, in 2003 and the PhD degree from the Queensland University of Technology, Brisbane, Australia, in 2008, within the Speech, Audio, Image and Video Technology (SAIVT) Laboratory. He is the director of Data Science at STATS, where he conduct research into Group Behavior and Sport Analytics. His research centers on representing, learning and predicting both cooperative and adversarial groups using spatiotemporal data-with application to continuous sports and audience domains. In his previous position, he was a associate research scientist at Disney Research Pittsburgh. He also conducted research on using facial expressions to aid in the diagnosis of medical conditions (such as Pain, Depression and Facial Paralysis) with Prof. Jeff Cohn.

Iain Matthews received the BEng degree in electronic engineering and the PhD degree in computer vision from the University of East Anglia. He is a research scientist at Oculus Research working on social virtual reality. His research interests include computer vision and facial tracking, modeling, and animation. He then joined Carnegie Mellon University, first as a post-doctoral fellow then as faculty in the Robotics Institute. In 2006, he spent two years at Weta Digital creating the facial motion capture system for the movies Avatar and Tintin, and was awarded a Scientific and Engineering Award (technical Oscar) for this work, in 2017. He joined the newly formed Disney Research Pittsburgh, in 2008 to lead the computer vision group. In 2013 he became the associate director of Disney Research Pittsburgh. He holds an adjunct faculty appointment in the Robotics Institute at Carnegie Mellon University and an Honorary professor position at the University of East Anglia. He has published more than 100 academic papers and has a dozen awarded patents.