Predicting Movie Ratings from Audience Behaviors

Rajitha Navarathna[1], Patrick Lucey[1], Peter Carr[1], Elizabeth Carter[1,2], Sridha Sridharan[3], Iain Matthews[1]
[1]Disney Research, Pittsburgh, USA
[2]Carnegie Mellon University, USA, [3]Queensland University of Technolgy, Australia
{rajitha.navarathna,patrick.lucey,peter.carr,iainm}@disneyresearch.com
lizcarter@cmu.edu, s.sridharan@qut.edu.au

Abstract

We propose a method of representing audience behavior through facial and body motions from a single video stream, and use these motions to predict the rating for feature-length movies. This is a very challenging problem as: 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) so tracking people uninterrupted for this length of time is an unsolved problem; and iii) expressions and motions of audience members are subtle, short and sparse making labeling of activities unreliable. To circumvent these issues, we use an infra-red illuminated test-bed to obtain a visually uniform input. We then utilize motion-history features which capture the subtle movements of a person within a pre-defined volume, and then form a group representation of the audience by a histogram of pair-wise correlations over small time windows. Using this group representation, we learn a movie rating classifier from crowd-sourced ratings collected by rottentomatoes.com *and show our prediction capability on audiences from 30 movies across 250 subjects ($>$ 50 hours).*

1. Introduction

Having the ability to objectively measure group experience would be of major benefit within the educational, marketing, advertising and behavioral science domains. However, due to the complexities of the observed environments and task, the de-facto standard of measuring audience or group experience is still via self-report [4]. As self-report measures are subjective, labor intensive, and do not provide feedback at precise time-stamps; an automated and objective measure is desirable. In an attempt to provide an objective measure, Madan et al. [19] utilized a wearable device which measured audio, head movement and galvanic skin responses of a group interacting. Eagle and Pentland [7] developed a system using a PDA which required continuous user input. While both are interesting approaches, our goal is to implement a less invasive solution.

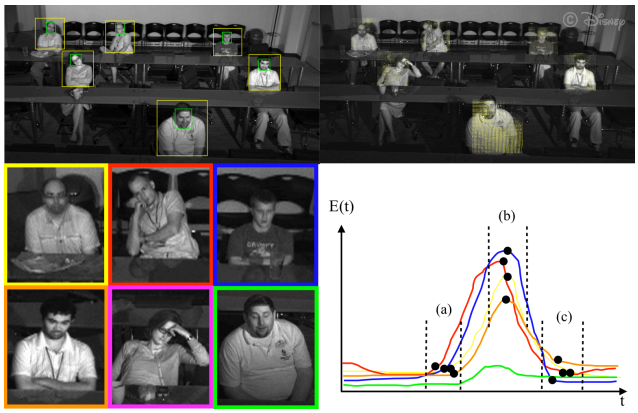

Figure 1. In our infra-red illuminated screening room, we use both face (top left) and body motion features (top right) to profile each audience member (bottom left) and find the synchrony or coherence of motion to analyze, summarize and predict audience ratings to movies (bottom right - each curve color corresponds to an audience member).

For measuring reactions to consumer products, almost all ratings are via self-report (i.e., "likes" or a Likert-type scale [5]). Given enough crowd-sourced ratings (100k's), useful measures can be obtained which can be used to predict other products that a person maybe interested in based on their previous behavior. Such *recommendation* systems are often based on matrix factorization approaches. *Pandora*[1] (songs), *Netflix*[2] (movies/tv-shows) and *Amazon*[3] (products) are popular examples for content-based and collaborative filtering approaches [15].

For movies, *Rotten Tomatoes* [1] have both critic and crowd-sourced audience ratings. Such information is only useful at a coarse level as it captures the overall global reaction to the stimuli and does not contain any specific local "interest" information. For long continuous time-series signals like movies, knowing which parts the audience (or sub-groups of the audience) like and do not like would be very beneficial to writers/directors/marketers/advertisers. Achieving this through self-report is subjective and diffi-

[1] pandora.com [2] netflix.com [3] amazon.com

cult, as it would require a person to consciously think and document about what they are watching (most likely causing subject to miss important parts of the movie). 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.

In this paper, we use a single camera as our input sensor and use face and body motion features to predict and summarize audience ratings of full-length movies (see Figure 1). Our work is motivated by the noted film editor Walter Murch who speculates in his book "In the Blink of an Eye" [22], that the engagement of an audience can be gauged through the synchrony of audience motion. Apart from the very dark environment, monitoring an audience from a single vantage point for a full-length feature film is a challenging problem because: i) it spans a very long time period (typically movies normally range from 80-150 minutes) which is an enormous amount of video data to process; ii) people are at different vantage points and resolutions; iii) we required frame-based measurements to measure synchrony; and iv) getting ground-truth labels of activity is subjective and time-consuming.

To counter these issues, we calculate the motion-history features of each audience member within a 3D volume to capture his/her face and body movements. We then propose an entropy of pair-wise correlations measure to gauge the collective behavior of the audience. We show that our approach outperforms human-annotated labels which do not pick up on these fine details. Using the audience ratings from *rottentomatoes.com*, we then use this feature to predict the movie rating solely from audience behaviors. Additionally, we use change-point detection to temporally cluster and summarize audience behaviors into a series of interest segments.

2. Related Work

A survey of recent work in automatically measuring a person's behavior using vision-based approaches is presented in [33]. 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 [32], pain detection [17] and human-computer-interaction [29]. An emerging area of research over the last couple of years is the use of affective computing for marketing and advertising purposes. When a user watches video clips or listens to music, they may experience certain feelings and emotions [14] which manifest through gestural and physiological cues such as laughter. These emotional responses to multimedia content have been studied in the research community [25]. Shan et al., [25] studied the relationship between music features and emotions from film music. In a recent study, Joho et al., [13] showed that facial expression is a good feature to predict personal highlights in media content. Hoque et al., [11] further showed that these facial behaviors vary from the laboratory setting to real-world. Teixerira et al.,l [28] 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., [21] utilized crowdsourcing to collect responses from people watching commercials and used smiles to gauge their reaction. They extended this work to predict the effectiveness of advertisements using smiles instead of "likes" [20]. Finally, Hernandez and colleagues [10] 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, using the Viola-Jones face detector [30] to locate the face, and detected in which of four states of engagement the viewer was based on facial movements.

This prior work was applied only to individuals and limited to stimuli of short duration (i.e., $10-60$ seconds), with the exception of [10]. We expand this research to include simultaneous recording of multiple individuals and continuous tracking 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 vision (see Section 4). Additionally, being in a group environment introduces extra variability as behavior can be altered by other audience members as well as by the stimuli.

3. Experimental Setup

3.1. InfraRed Illuminated Testbed

Observing people watching visual stimuli from a screen is difficult because: 1) the environment is very dark, and 2) the reflected lights from the visual stimuli causes a non-uniform illumination environment. Wide aperture lenses and sensor sensitivity are two important features to consider when selecting a good camera to capture the objects in low-light conditions. We instrumented a test-bed with an infra-red (IR) sensitive low-light camera (Allied Vision GX 1920 with a 2/3" Sony ICX674 CCD sensor and a f/1.4 9mm wide angle lens), two IR illuminators (Bosch

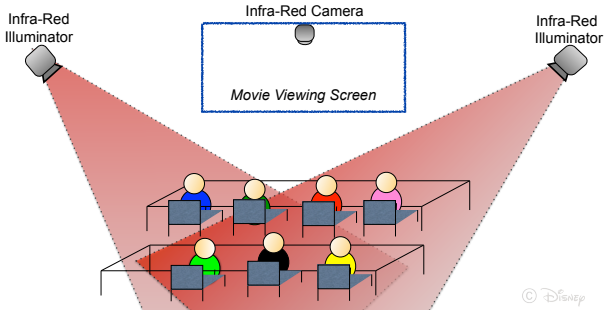

Figure 2. A schematic of the audience-test bed used in this work.

Figure 3. (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) To remove the reflected illumination from the screen we used a band-pass filter to obtain a uniform lighting environment.

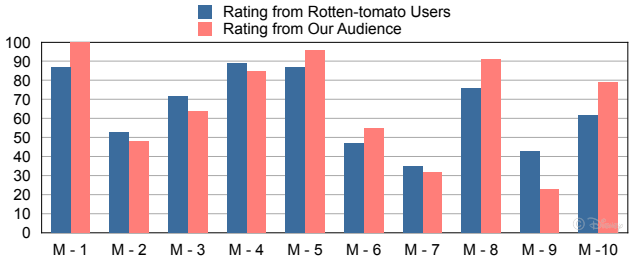

Figure 4. A bar chart comparing the ratings of the audience compared to the crowd-sourced ratings from rotten-tomatoes.com.

Movie No	No Sess	No People	Time (min)	Budget ($ mill)	Box Off. ($ mill)	Rating (%)
1	3	25	103	200	1063	87
2	3	25	81	150	315	53
3	3	25	96	150	310	72
4	3	27	101	165	471	89
5	3	24	96	175	731	87
6	3	22	83	105	172	47
7	3	25	87	30	16	35
8	3	23	93	185	555	76
9	3	22	86	47	38	43
10	3	19	88	95	877	62

Table 1. An inventory showing the number of audience members, attributes and the rotten-tomoatoes.com rating per movie.

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 (850nm $\pm$ 5nm). The resulting images are 1936×1456 pixels captured at 15 frames per second. The schematic diagram of the infra-red illuminated test-bed and effects of those instruments are shown in Figure 2 and in Figure 3 respectively.

3.2. Audience Footage

We used www.rottentomatoes.com [1] to select movies from the genre "Animation, Comedy, Kids & Family". Out of a total of 62 movies (year 1998 – 2013) in that genre, we selected a subset of ten movies (refer to Table 1) with varying crowd-sourced audience ratings again

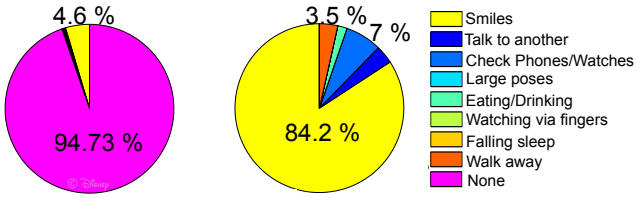

Figure 5. (Left) An example of the he distribution of labeled activities for an entire movie - about 95% of the time audience members do nothing. (Right) The distribution of activities when audience members are active.

from [1]. To do this, we chose three good movies (ratings greater than 80%), three average movies (ratings from 60% to 80%), and four bad movies (ratings below 60%).

For this study we sought subjects (age 18-70) to be apart of an audience ranging in size of 5-10 people (mean 8 people). This work was approved by an Institutional Review Board, and participants were compensated for their time. We screened the movies at the same time (6.00pm) and for each screening, only participants who had not seen the movie previously, and had normal or corrected-to-normal vision and hearing were used. We had three sessions for each movie (total 30 sessions) and each subject could only participant once. At the completion of each session, every participant completed a survey asking about their overall rating of the movie (similar to a self-report), age, gender, movie genre preference, and expectation/recommendation of the movie. A comparison of movie ratings using the self-report method from our audience (mean 67.3) to the *rotten-tomatoes.com* users (mean 65.1) for each movie is given in Figure 4. As shown in Figure 4, our audience had a reasonably good compatibility compared to the crowd-sourced measure.

To get a sense of how many different actions and activities a person normally performs while viewing a movie, we selected a subset of sessions for human annotation. As we were interested in both facial expressions and body movements, we manually annotated the following gestures at the frame-level. A description of these actions and activities are

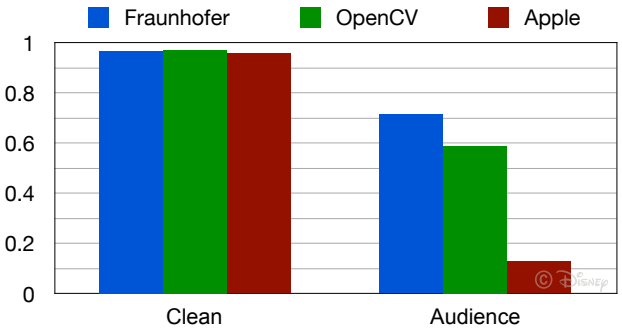

Figure 6. Face detection performance in clean and audience environment. Off -the -shelf face detectors performs poorly in audience environment mainly, due to low lighting (we used IR camera to capture footage) and different view point.

given below:

Smiles/laughter: Using FACS [9], we annotated smiles and laughter. The onset of smiles/laughter were labelled as the onset of AU12 and the offset was labelled at the end of that occurrence.

Body movements: We annotated the following common actions: talking to another person, raising arm, moving hand to head/table, moving within chair, eating/drinking, watching through fingers, using laptop/iPad, checking phone/watch.

In terms of activity, approximately 90% of the time no activity was observed, as can be seen in Figure 5. This could be due to: i) people not moving at all, ii) intensity or duration of activity being so low or short that it does not warrant labelling, iii) the activity not fitting into the pre-set activities vocabulary. It can be argued that ii) and iii) are due to problems with annotations, but as a result of the long length of input stimuli (approximately 1-2 hours per movie), it is highly impractical and unscalable to get this level of annotation[4]. Even if it is possible to get the level of annotation it would be expected that the reliability of annotation would greatly diminish due to the high level of subjectivity. Motivated by this analysis, we require a solution that captures both facial and body movements. In terms of automatic analysis, this can be circumvented as the continuous flow features of each person can be used to temporally segment potentially interesting behaviors.

4. Extracting Audience Features

To extract features from each audience member, we first register the image region that he or she occupies over the course of the movie, and then extract motion features. The following section describes each method.

[4] Note that this process was very time consume (annotation time was $>$ 90 hours per session)

4.1. Registering Audience Members

Despite a person remaining relatively stationary whilst watching a movie, continuous tracking is a 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). In contrast, our problem represents a paradigm shift in this area called long-term face tracking. To illustrate the issues in this method, we provide the following example. First, the intuitive method of registering each audience member would be to use a "off-the-shelf face-detector on each frame and then track each detection. As can be seen in Figure 6, 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 infra-red spectrum, and the resolution of faces can be small. An example of the "off-the-shelf" face detector performance is given in Figure 7(a). Alternatively, we could use a template update method, where we register an initial face and then update the template at every frame [3, 18]. This works reasonably well, but it tends to drift over long periods of time (Figure 7(b)). New methods that use a dictionary of templates have worked reasonably well, especially those of the l_1 variety [12]. However, as shown in Figure 7(c), they perform worse when there is considerable change in appearance or pose - i.e., when a key frame is not in the dictionary. A solution to this is to have prior knowledge of the key frames in the dictionary, but this is not ideal as it requires manual intervention (Figure 7(d)).

But this begs the question: *do we actually need to track each person?* As the person does not move substantially during the movie - they are basically restricted within the confines of their volume to maintain space between other audience members - a more reliable solution is to pre-define a volume that the person occupies throughout the movie. In this work, we implemented such an approach by using the first frame to define a volume that the person would occupy. Across the 250 subjects, we found that this method worked very well, even in cases where the person left to go to the bathroom, as our feature extraction was robust to this issue. In this work, we implemented such an approach and it was much more reliable than the tracking approach, which constantly failed.

4.2. Motion Features

In terms of recognizing individual and specific actions, there is a plethora of research which has solely focussed on this domain, with excellent progress being made [2]. Efros et. al., [8] used optical flow features to recognize actions from ballet, soccer and tennis. More recently, Rodriguez et. al., [23] used similar features to analyze crowds.

Figure 7. Examples of various face detectors/trackers: (a) Fraunhofer face detector fails due to the low light conditions, viewpoint (camera is looking down on the audience) and resolution of the faces are small (b) Template tracking method fails when there is pose/appearance change, (c) l_1 tracker breaks when the key frames are not in the dictionary, and (d) Modified l_1 tracker works reasonably well but requires key frames to be found manually.

However, we are not interested in the specific actions of one person but instead the synchrony of actions (i.e., is everyone doing something at the same time?). The screening room environment introduced a natural spacing of audience members so each person could watch the movie with unoccluded and in comfort, resulting in each person occupying a minimum uninterrupted 3D volume. We examined features based on optical flow [16] and motion history images [6].

Optical Flow Features: To measure the synchronous body movement of an individual, we developed an energy-based flow-profile measure [27]. Having N audience members, we initialize a local 3D volume for each person in the horizontal and vertical directions x and y over the time t as $Q = f(x, y, t)$. We generate a flow-profile of each person contained within their 3D temporal volume (which was defined manually by a human) using optical flow components V_x and V_y respectively. In this work, we used the following optical flow formulation:

$$I_x V_x + I_y V_y + I_t = 0 \tag{1}$$

where V_x and V_y are the optical flow components in x and y directions and I_x, I_y and I_t are the image derivatives at point (x, y) at time t. Using these flows, we calculate the normalized local 3D energy for person q as,

$$E_{q,t} = \frac{1}{a_q}\sqrt{V_{q,x,t}^2 + V_{q,y,t}^2} \tag{2}$$

where the a_q is the area defined for an individual to move over the time.

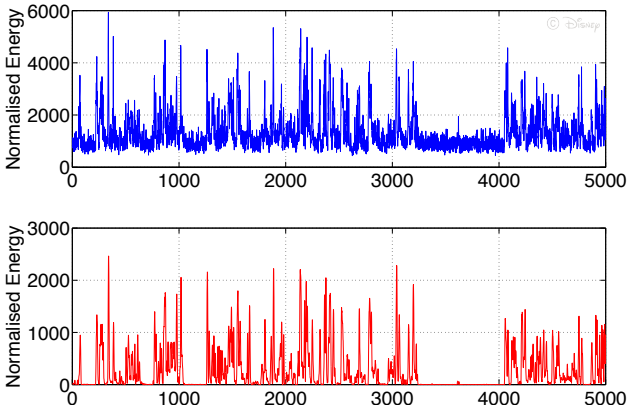

Figure 8. (Top) An example of the magnitude of the optical flow of an audience member, (bottom) compared to the magnitude of the motion-history features which had over 85% correlation.

Motion History Images: Using optical flow is very computationally expensive to compute[5], which limits the usefulness of this approach for this work. In order to overcome the computation time from the optical flow method, we used an aggregated real-time approach to represent the spatio-temporal motion that recursively integrates into a single motion history images [6]. This is done by layering the threshold differences between consecutive frames one over the other. This represents *how* motion in the image is moving opposed to *where*, which is our interest. These motion history images can be calculated as follows:

$$H_\gamma(x, y, t) = \begin{cases} \gamma & \text{if } D(x, y, t) = 0 \\ \max(0, H_\gamma(x, y, t-1) - 1) & \text{otherwise} \end{cases} \tag{3}$$

where, $D(x, y, t)$ is a binary image sequence indicating regions of motion at pixel (x, y) in time t and parameter γ is the temporal duration of the motion history images. Then, we calculate the normalized local 3D energy for person q as, $E_{q,t} = \frac{1}{a_q}\sum H_\gamma(x, y, t)$.

The normalized energy from optical flow and motion history can be vectorized over the duration of the movie time T as $\mathbf{e}_q = [E_{q,1}, E_{q,2}, \cdots, E_{q,T}]$. Finally, we define a aggregate normalized measure of overall audience engagement over the movie time T as $\mathbf{e}_{\text{movie}} = \frac{1}{N}\sum_{q=1}^{N} \mathbf{e}_q$.

Comparison: To see how reliable each feature was, we analyzed the correlation between flow features (i.e optical flow features and motion history images) for a one movie. An example of an individual flow-field for an audience member using these features is given in Figure 8. We observed 85% of average cross-correlation between motion history features and optical flow magnitudes.

[5] Calculating the optical flow of an audience for a 2-hour feature length movie took more than 2-3 days on a high-performance computing cluster which is not tractable for our application

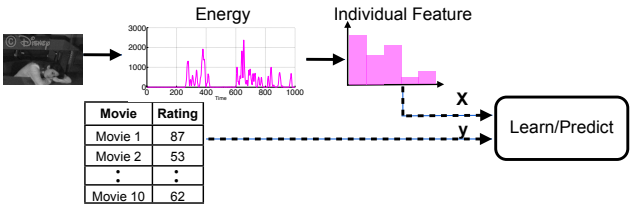

Figure 9. Individual Representation: We first break the motion-history time-series into chunks across a small-window of time and then form a histogram based on the mean energy for each chunk. This gives us a feature representation for each movie, and we learn a classifier by using crowd-sourced ratings from *rottentomatoes.com*.

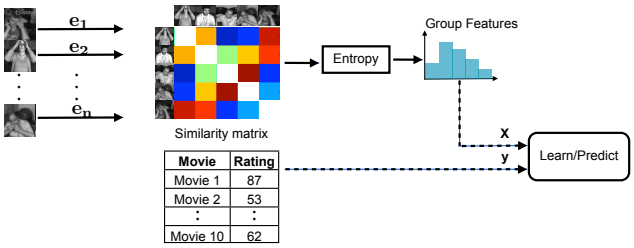

Figure 10. Group Representation: To capture the group interaction we calculate the pair-wise correlations and then the entropy for each time-chuck. The final representation is the histogram of entropy values across the movie.

5. Predicting Movie Ratings

To gauge how much the general public likes a particular movie, *rottentomatoes.com* has an interactive feature which allows people to go online and give a rating. Over time the number of ratings aggregate (100k's) and based on these crowd-sourced ratings, they generate an "audience measure". Based on these scores, an average audience measure is obtained, with a movie rating 75% or higher been deemed a good movie, a movie rating between 50-75% being ok and below 50% denotes a bad movie.

Achieving this using self-report is difficult as it would require a person to consciously think and document what they are watching and subjects may miss important parts of the movie, due to distractions. Similarly, subjects could be instrumented with a myriad of wearable sensors, but such approaches are invasive and unnatural and therefore may not result in good indicators of the actual rating. Alternatively, we derive the following representations of individual audience members as well as the entire group solely on the audience reaction to predict movie ratings.

5.1. Individual Representation

To represent the individual behavior, we used individual motion features $\mathbf{e_i}$ using motion history images. Given an audience energy signal smoothed over 6 seconds, we generate histogram distribution $\mathbf{X} = p(\mathbf{e_i})$, which allows us to represent a measure of each audience behavior during the movie. Given this representation and known movie ratings $\mathbf{y}$ from [1], we learn a regression model to predict the movie ratings solely on the individual audience reaction as shown in Figure 9.

5.2. Group Representation

5.2.1 Joint Representation

We develop an objective measure using the facial expressions and body motion of audience members to gauge the synchrony of behavior. In order to represent the group, we initially used the joint distribution of the audience. We used an aggregate normalized measure of overall audience engagement over 30-second temporal segments, $\mathbf{e}_{\text{movie}}$. Once we derive the aggregate measure, we generate the joint distribution/histogram $\mathbf{X} = p(\mathbf{e}_{\text{movie}})$ for all audience members (similar to Figure 9) that is used for prediction.

5.2.2 Mid-Level Representation

We utilized an entropy of pair-wise similarity between each audience member at the local-level (i.e. pair-wise comparison) as well as the global-level (i.e. compared to the whole group). In this regard, we first compare the small feature segment between two audience members, $\mathbf{e_1}$ and $\mathbf{e_2}$, and calculate the pair-wise similarity by using,

$$C_{\mathbf{e_1 e_2}} = \exp\left(\frac{- \parallel \mathbf{e_1} - \mathbf{e_2} \parallel^2}{2\sigma^2}\right) \tag{4}$$

where σ is an adjustable parameter for each similarity matrix. We then exhaustively calculated all of the pairwise correlations between audience members, yielding a similarity matrix. 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 that the similarity matrix of piece-wise correlations can be represented by $\mathbf{S}$, we can generate a probability distribution of $\mathbf{S}$ for that time segment $p(\mathbf{S})$, allowing us to gain a measure of audience disorder via entropy [26]

$$H(S) = -\sum_{i=0}^{N-1} p(i) \log p(i) \tag{5}$$

A high value of entropy means that there is great disorder (i.e., random behavior), while a low value of entropy means that there is cohesion or predictability of behavior. Finally, we generate a probability distribution $\mathbf{X} = p(H(S))$ to gain a measure of synchrony of audience for predicting movie ratings. The system is shown in Figure 10.

5.3. Performance Evaluation

Once we extract the features from individual and group representation, we used those features to learn audience behaviors from a library of movies (See Table 1) and use these features to predict the audience rating for an unseen movie. We analysed this prediction using different predictors such as linear, logistic and support vector regression (SVR). There was not a big discrepancy between these methods, and we present the results for SVR. Given the feature representation $\mathbf{X}$ and known movie ratings $\mathbf{y}$ from [1], we learn $\mathbf{w}$ for SVR by minimizing the following objective function,

$$\begin{aligned} \arg\min_{\mathbf{w}} \quad & \frac{1}{2} \parallel \mathbf{w} \parallel^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \\ \text{subject to} \quad & y_i - \mathbf{w}^T \mathbf{x}_i - b \leq \epsilon + \xi_i \\ & \mathbf{w}^T \mathbf{x}_i + b - y_i \leq \epsilon + \xi_i^* \\ & \xi_i \geq 0, \xi_i^* \geq 0 \end{aligned}$$

where $C > 0$ is a parameter to control the amount of the influence and ξ_i, ξ_i^* are slack variables.

We validate our framework using a leave-one-out cross validation strategy (leaving out entire an movie). The parameters for SVR were chosen using a cross-validation method as described in [31] with a polynomial kernel. For a quantitative assessment, we compute the root mean squared error (RMSE) between the predicted rating value $\hat{y}_i$ and the audience rating y_i such that:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (\hat{y}_t - y_t)}{n}} \tag{6}$$

For the mid-level group representation, we tested different timing window segments (i.e 30, 45, 60, 120 seconds) to obtain pair-wise entropy values and different σ values. We observed that 30 second window segments with $\sigma = 0.5$ gave the best prediction values.

The experimental results for leaving out an entire movie in terms of average RMSE are shown in Table 2. As shown in Table 2, audience behavior (i.e., synchrony/coherency of audience motion) for a group is more robust than for each individual. Overall, our framework showed that we can predict movie ratings solely using audience behaviors, a potential solution to the problems with current standard self-report measures. Using the mid-level group representation and SVR, we show our average movie prediction (i.e., average from all the 3 sessions for a movie) results for a movie in Figure 11. As can be seen from this result, we get a reasonable approximation to the *rottentomatoes.com* crowdsourced ratings.

Representation	Average RMSE
Individual	21.2
Joint	13.4
Mid-Level	12.7

Table 2. Average movie prediction error in terms of RMSE using SVR.

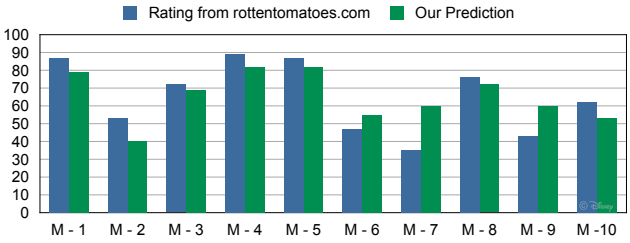

Figure 11. Average results of our automatic approach compared to the crowd-sourced ones from *rottentomatoes.com*.

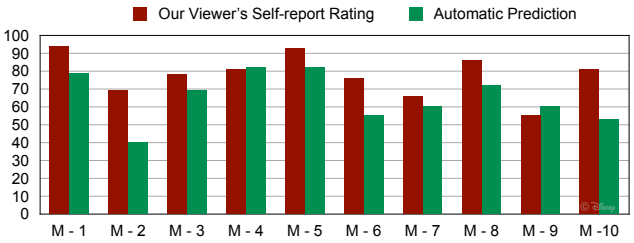

Figure 12. Average results of our automatic audience rating measure compared to the viewer's self-report measure.

Finally, we also compared our automatic prediction from audience behavior to their (our viewers) self-report audience rating, as shown in Figure 12. The average RMSE value compared to the viewer's self-report rating is 16.95. In this environment, the result makes sense: self-report is subjective and difficult as it would require an audience to consciously think about what they were watching. In addition, it does not contain feedback at precise timestamps [24].

5.4. Temporal Window Analysis

During the movie, audience members tend to move and react. In this work, we are interested in the synchrony of audience behavior (i.e., is everyone doing a particular thing at the same time?). We looked at what is the optimal timing window in which the audience behaves in an interesting way? To do this analysis, we selected different window sizes from 10sec – 5min. For these different windows sizes, we generated group representations and predicted movie ratings. The average RMSE with different window sizes is given in Figure 13. We observed that we can capture interesting audience behaviors using 30-second increments.

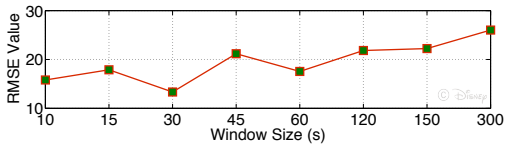

Figure 13. Variation of average RMSE with respect to different temporal window sizes.

(a) (b)

Figure 14. An example of movie summarization for a: (a) good movie and (b) bad movie. The green boxes show examples of similar activities while red boxes illustrates random activities.

6. Movie Summarization

As feature-length movies are very long in duration, often it is beneficial for a domain expert to quickly skim through the behaviors of an audience. Finally, to summarize the reaction of the audience to a movie signal $\mathbf{e_q}$ (smoothed over a 6-second window), we chunk the movie into 1-minute windows, and we find for each window the strongest audience change-point (i.e., zero-crossing values in audience signal $\mathbf{e_q}$). Using that as our index, we use a 1-second window centered at that change-point to summarize the audience behavior over that minute. We piece this together to form a summarization of the audience behavior, allowing someone to assess a 90-minute movie over the course of 90 seconds. Qualitatively, we found that we could find engaging and disengaging segments during the movie using this approach. Visual examples are given in Figure 14.

7. Summary

We proposed an automatic method of measuring, summarizing and predicting audience behavior using face and body motions from a single video stream. Due to the complexity and difficulty of this task, no one has previously looked at this problem. To do this: (i) we introduce an IR based test-bed as the movie viewing environment is dark and contains views of many people at different scales and viewpoints, and we use more than $>$ 50 hours of audience data; (ii) we then utilize motion-history features that can pick up on the subtle movements of a person within a predefined volume; (iii) we propose a method to learn individual and group behaviors; and (iv) we use these representations to learn our movie rating classifier from crowdsourced ratings collected by *rottentomatoes.com* and show our prediction capability on audiences from 30 movies and 250 viewers. We showed that we can give a reasonable approximation solely from audience behavior to the *rottentomatoes.com* crowd-sourced ratings.

References

[1] http://www.rottentomatoes.com/. 1, 3, 6, 7

[2] J. Aggarwal and M. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 2011. 4

[3] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework: Part 1: The quantity approximated, the warp update rule, and the gradient descent approximation. *International Journal of Computer Vision*, 56(3):221–255, February 2004. 4

[4] R. Bales. Social inteaction system: Theory and measurement. *New Brunswick, NJ:Transaction Publishers*, 1999. 1

[5] D. Betram. *Likert Scales*. Topic Report, The Faculty of Mathematics University of Belgrad, 2009. 1

[6] J. Davis and A. Bobick. The representation and recognition of action using temporal templates. *in CVPR*, pages 928–934, 1997. 5

[7] N. Eagle and A. Pentland. Social network computing. *Ubicomp 2003: Ubiquitous Computing, Springer-Verlag Lecture Notes in Computer Science*, pages 289–296, 2003. 1

[8] A. Efros, C. Berg, G. Mori, and J. Malik. Recognizing Action at a Distance. In *ICCV*, 2003. 4

[9] P. Ekman and W. Friesen. Manual for the facial action coding system. *Consulting Psychologists Press*, 1977. 4

[10] J. Hernandez, L. Zicheng, G. Hulten, D. DeBarr, K. Krum, and Zhang.Z. Measuring the engagement level of tv viewers. *In FG*, 2013. 2

[11] M. Hoque and R. Picard. Acted vs natural frustration and delight: many people smile in natural frustration. *In FG*, 2011. 2

[12] H. Ji, H. Ling, Y. Wu, and C. Bao. Real time robust l1 tracker using accelerated proximal gradient approach. *In CVPR*, 2012. 4

[13] H. Joho, J. Staiano, N. Sebe, and J. Jose. Looking at the viewer: analysing facial activity to detect personal highlights of multime- dia contents. *In Multimedia Tools and Applications*, 2011. 2

[14] J. Kim and E. Andre. Emotion recognition based on physiolog- ical changes in music listening. *In TPAMI*, pages 2067–2083, 2008. 2

[15] Y. Koren, R. Bell, and C. Volinsky. Matrix Factorization Techniques for Recommender Systems. *IEEE Computer Society*, 2009. 1

[16] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *In Proceeding of the International Joint Conference on Artifical Intelligence*, 1981. 5

[17] 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 FG*, 2011. 2

[18] S. Lucey, R. Navarathna, A. Ashraf, and S. Sridharan. Fourier lucas-kanade algorithm. *IEEE Transactions on PAMI*, 2013. 4

[19] A. Madan, R. Caneel, and A. Pentland. Groupmedia: Distributed multimodal interfaces. *International Conference on Multimodal Interfaces*, 2004. 1

[20] D. McDuff, R. Kaliouby, D. Demirdjian, and R. Picard. Predicting Online Media Effectiveness Based on Smile Responses Gathered Over the Internet. In *FG*, 2013. 2

[21] D. McDuff, R. Kaliouby, and R. Picard. Crowdsourcing facial responses to online videos. *In IEEE TOAC*, 2012. 2

[22] W. Murch. *In the Blink of an Eye: A Perspective on Film Editing*. Silman-James Press, 2001. 2

[23] M. Rodriguez, J. Sivic, I. Laptev, and J. Audibert. Data-Driven Crowd Analysis in Videos. In *ICCV*, 2011. 4

[24] N. Schwarz and F. Strack. Reports of subjective well-being:judgmental processes and their methodological implications. *Well-being: The foundations of hedonic psychology*, 1999. 7

[25] M. Shan, F. Kuo, M. Chiang, and Y. Lee. Emotion-based music recommendation by affinity discovery from film music. *An International Journal Expert Systems with Applications*, 2009. 2

[26] C. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 1948. 6

[27] D. Sun, S. Roth, and M. Black. Secrets of optical flow estimation and their principles. *In CVPR*, pages 2432 –2439, 2010. 5

[28] T. Teixerira, M. Wedel, and R. Pieters. Emotion-induced engagement in internet video advertisements. *Journal of Marketing Research*, 2011. 2

[29] A. Vinciarelli, M. Pantic, and H. Bourland. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 2009. 2

[30] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001. 2

[31] C. wei Hsu, C. chung Chang, and C. jen Lin. A practical guide to support vector classification. 2010. 7

[32] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. Towards practical smile detection. *In TPAMI*, 2009. 2

[33] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *In TPAMI*, 2009. 2