Towards Sensing the Influence of Visual Narratives on Human Affect

Mihai Burzo
Mechanical and Energy Engineering
University of North Texas
Mihai.Burzo@unt.edu

Daniel McDuff
Media Lab
Massachusetts Institute of Technology
djmcduff@mit.edu

Rada Mihalcea
Computer Science and Engineering
University of North Texas
rada@cs.unt.edu

Louis-Philippe Morency
Institute for Creative Technologies
University of Southern California
morency@ict.usc.edu

Alexis Narvaez
Mechanical and Energy Engineering
University of North Texas
AlexisNarvaez@my.unt.edu

Verónica Pérez-Rosas
Computer Science and Engineering
University of North Texas
veronica.perezrosas@gmail.com

ABSTRACT

In this paper, we explore a multimodal approach to sensing affective state during exposure to visual narratives. Using four different modalities, consisting of visual facial behaviors, thermal imaging, heart rate measurements, and verbal descriptions, we show that we can effectively predict changes in human affect. Our experiments show that these modalities complement each other, and illustrate the role played by each of the four modalities in detecting human affect.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Discourse*

General Terms

Algorithms, Experimentation

Keywords

Multimodal signal processing, Multimodal sensing, Affective Behavior

1. INTRODUCTION

Narratives are a constant presence in our everyday lives, and can have significant influence on one's thoughts and actions. Narratives are often designed to explicitly appeal to the emotions of the reader or listener, and act as an "emotional prime" [11, 16]. Once an affective state has been induced, it can also lead to changes in cognition and action, in agreement with the large body of previous research on emotions [15].

A specific type of narrative that is becoming extremely popular with the Internet age is the visual narrative. With more than 10,000 new videos posted online every day, social websites such as YouTube and Facebook are an almost infinite source of visual narrative. People are posting videos to express their opinion and sentiment about different topics, products and events. To better understand how these online videos are influencing individuals and eventually the society at large, it is imperative that we develop automatic techniques to analyze human reactions to visual narrative.

In this paper, we propose a non-invasive multimodal approach to sense and interpret human reaction while watching online videos. This is a first important milestone toward a deeper understanding of visual narrative influence on human affective states. Our approach senses changes in human affect through four different modalities: visual facial behaviors, physiological measurements, thermal imaging, and verbal descriptions. The first three modalities are recorded live during the narrative interaction while the verbal descriptions are acquired during a post-study interview. We evaluate our multimodal approach on a new corpus of 70 narrative interactions. Figure 1 shows the overall flow of our approach.

The following section summarizes related work in visual, physiological, and linguistic analysis of human affective state. Section 3 presents our experimental methodology to create this new visual narrative corpus. Section 4 presents a detailed description of the multimodal features automatically extracted. Section 5 presents experimental results comparing the performance of our multimodal predictive models. Section 6 discusses our results and shows an analysis of the multimodal features and their effectiveness to predict human affective state. Section 7 presents our conclusions and future directions.

*The order of the authors is alphabetical, as all the authors have equally contributed to this work.

ICMI'12, October 22–26, 2012, Santa Monica, California, USA.

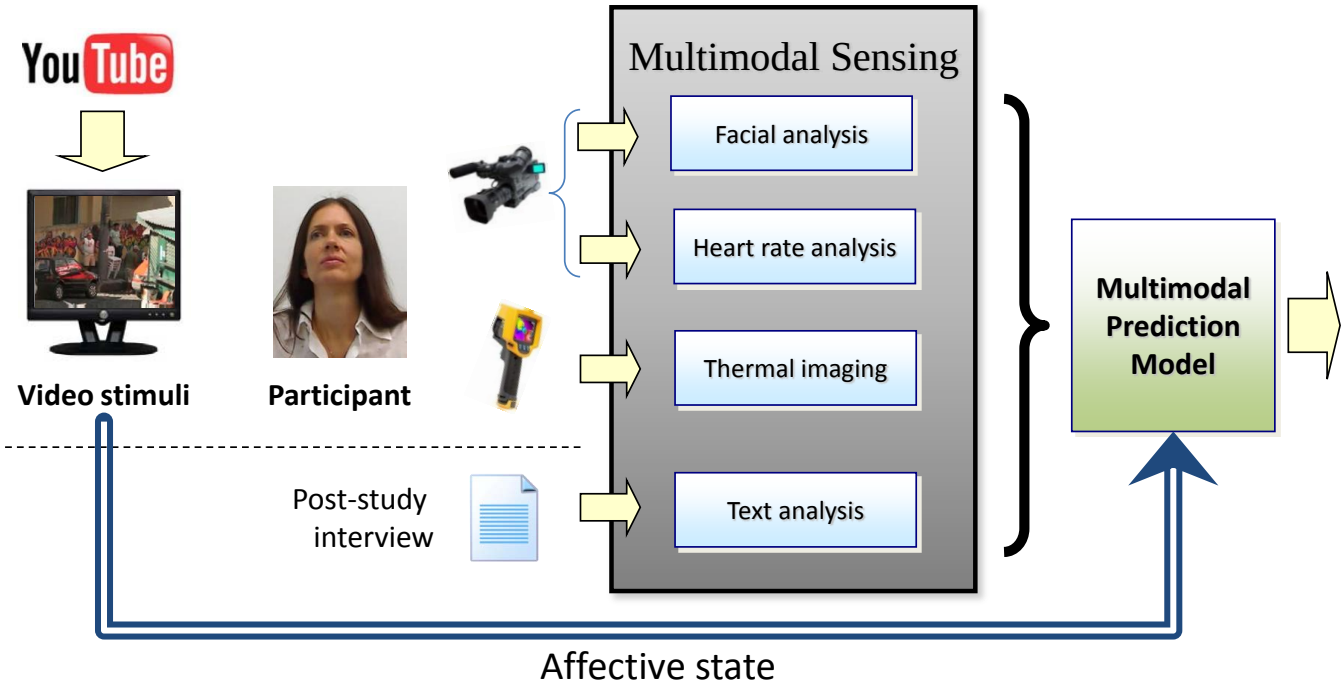

Figure 1: Overview of our multimodal approach for sensing affective state.

2. RELATED WORK

Influence models have been widely used to detect and predict human behavior, including game theory [5], statistical Bayesian models [26], and Markov models [18]. In our work, we build upon our own previous experience and related work in several research areas, including natural language processing, computer vision, and computational physiology.

2.1 Affect Recognition in Text

There has been growing interest in the recognition of affect in text, focusing on the automatic identification of private states, such as opinions, emotions, sentiments, evaluations, beliefs, and speculations. Most of the work in this area has addressed the analysis of the affective property (positive, negative, or neutral), where affective properties are associated with words [43, 39], phrases [45], sentences [33, 35], or entire documents [27, 19].

Finer-grained affect recognition has also been addressed, where emotions (e.g., anger, disgust, fear, joy, sadness, surprise) are automatically identified in text [37, 2, 35, 20, 36].

2.2 Physiological Sensing

Previously proposed solutions for non-contact measurement of vital signs, such as heart rate (HR) and respiratory rate (RR), include laser Doppler [40], microwave Doppler radar [13], variations in transmitted or reflected light [1], and thermal imaging [12, 10]. In particular, skin temperature detected from thermal images was found to be correlated with changes in affect [17], deceptive behavior [29], and stress [31].

2.3 Visual Sensing

There have been numerous approaches to track and analyse facial feature points from single images or image sequences [28, 48]. Approaches that can be considered model based are Active Shape Models [8], Active Appearance Models [7], 3D Morphable Models [3], and Constrained Local Models [9]. There exist also many techniques for estimating head pose [23, 22]. Research work in this area has also shown that visual features can be effectively used for dimensional and categorical affect recognition [48, 14, 46, 24, 25].

2.4 Multimodal Analysis of Affect

Also related to our work is the research done on multimodal emotion analysis [4, 34, 49]. For instance, a novel algorithm is defined in [47], based on a combination of audio-visual features for emotion recognition. The features used by these novel algorithms are usually basic and low level like tracking points for collecting visual data. An engineering approach is then applied to this large set of data points, in order to extract the ones that would be useful for the actual analysis. It is also worth noting the recent work on multimodal sentiment analysis, where textual, visual, and audio sensing is combined in order to determine the polarity orientation of statements expressed in YouTube videos [21]. Finally, somehow related is also the research reported in [32], where speech and text have been analyzed jointly for the purpose of opinion identification.

To the best of our knowledge, our paper is the first to integrate these four modalities (visual, thermal, physiologic and linguistic) to predict affective response to visual narrative.

3. METHODOLOGY

Our goal is to identify cues that are indicative of people's reactions when exposed to visual narratives. Specifically, in this first set of experiments, we target the identification

of affective states (as compared to neutral states), and furthermore we also aim to classify the valence of the affect experienced by a person (positive or negative).

We first selected four video stimuli from the YouTube website, all of them in English, two of them with negative content (one showing the effects of a tsunami, 153s; and one showing an accident, 145s), and two of them with positive content (one about the Coca-Cola happiness truck in Brazil, 152s; and one about an amusing incident during a wedding, 143s). These videos were selected based on popularity and the positive/negative comments written on YouTube. Several videos were considered as candidates, and the four final videos are the ones for which two of the authors of the paper agreed on their affective properties.

The participants consisted of fourteen subjects, three women and eleven men, all of them within the age range of 25-45, and all of them living in the United States although coming from different backgrounds (Asian, African-American, Caucasian, Hispanic). Each participant was first recorded during a "neutral" state while simply looking at the recording station and/or the lab were the recording took place. The four video stimuli were then played in random order, and the participant was recorded while watching the videos.

The recordings were done with two cameras. A regular Logitec Web camera, with a resolution of 980x720 and a frame rate of 15 frames per second, and a FLIR ThermoVision A40 thermal camera with a resolution of 340x240 and a frame rate of 60 frames per second.

After each recording, the participants were asked to write 2-4 sentences reflecting how they felt about the video they just saw.

Figures 2 and 3 show sample frames from the visual and the thermal camera recordings. Table 1 shows sample textual statements provided by the participants.

It is important to note that in these experiments we use the prior classification of the videos (as either positive or negative) as our approximation of the induced affective state. We thus assume that a positive video induces a positive affective state, while a negative video induces a negative affective state. In the future, we plan to use the PANAS-X survey [42] to capture the actual affective state experienced by a person during exposure to the visual narratives.

Figure 2: Sample snapshots taken by the Web camera showing participants exposed to positive and negative video narratives.

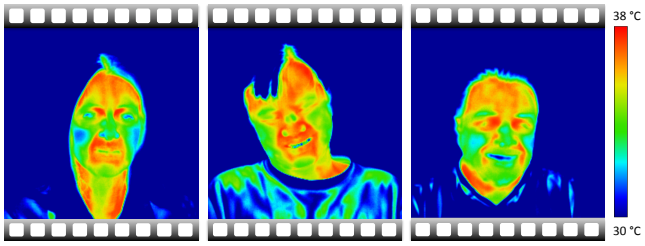

Figure 3: Sample snapshots taken by the thermal camera.

4. SENSING AFFECTIVE RESPONSE TO VISUAL NARRATIVES

We are sensing changes in human affect through four different modalities: (1) facial expressions obtained by processing the visual recording of the participants; physiological features, including (2) thermal features obtained from the thermal recording, and (3) heart rate measurements obtained from the visual recording; and (4) linguistic descriptions of the participants' state after exposure to the stimuli.

We describe below each of these modalities in detail, along with the features obtained from each of them.

4.1 Facial Expressions

The visual features are automatically extracted from the video sequences. Since only one person is present in each video clip and they are all the time facing the camera, current technology for facial tracking can efficiently be applied to our dataset. We use a commercial software called OKAO Vision that detects at each frame the face, it extracts the facial features, and extrapolates some basic facial expressions as well as eye gaze direction. The main facial expression being recognized is smile. This is a well-established technology that can be found in many digital cameras. For each frame, the vision software returns a smile intensity (0-100) and the gaze direction, using both horizontal and vertical angles expressed in degrees. The sampling rate is the same as the video framerate: 15Hz.

An important aspect when generating visual features is the quality of the video, and correspondingly the quality of the visual processing that can be automatically performed on the video. OKAO provides a confidence level for each processed frame in the range 0-1000. We discounted all the frames with a confidence level below 700.

For each video in our dataset, we define three series of summary features, which were used in previous work on affect analysis [21]:

- **Smile duration:** Given all the frames in a video, how many frames are identified as "smile." In our experiments, we use three different variants of this feature with different intensity thresholds: 25, 50, 75.
- **Look-away duration:** Given all the frames in a video,

Positive	Negative
I like it when people have fun even during the more "official" moments of their life. This seems like a joyful couple, I hope they will have many years together. The priest was a bit disheartened at first (or so it seemed), but then he managed to joke himself about the whole situation. I am sure this made for an even more memorable event for all those present.	This is so distressing! It may have to do with the fact that I have a small daughter myself, or with the innocence of the child who suffered in this video, but I found it extremely sad. It is so unfair how we always depend on other people's decisions - like in this video, one's rush brought an end to somebody else's life. It almost makes you think that life before the technology age was better and safer.
Nice to see people feel happy at an unexpected moment during their regular daily routine. I felt it was a very creative way of making people smile. Feels relaxing when you watch happy stuff.	Interesting to get to view natural disasters and its effects on structures and other things around it. Its sad to think of how people who live in places which are hit by natural calamities suffer and recover from something like that. The video was interesting.

Table 1: Sample textual statements made by the participants after watching the videos.

in how many frames is the speaker looking at the camera. The horizontal and vertical angular thresholds were experimentally set to 10 degrees.

- **Eye gaze direction:** Given all the frames in a video, in how many frames is the speaker looking to the left with an angle higher than N degrees, where N can be 10, 20, or 30. Variants of these features were also created for eye gaze to the right, eye gaze up, and eye gaze down.

All the visual features are normalized by the total number of frames in the video. Thus, if the person is smiling half the time, then the smile feature will be equal to 0.5 (or 50%).

4.2 Thermal Analysis of the Face

Another major sensorial input is the skin temperature, which relies on thermal features drawn from images captured with the thermal camera.

Using the thermal recordings, we infer several features that reflect the temperature of the face and the overall temperature (including face, neck, and shoulders). Starting with a map of pixel-level temperature measurements obtained from each frame in the recording, we collect the following features:

- **Face temperature features:** Using the temperature measured on the face, we calculate: average temperature; overall minimum and overall maximum temperature; average of the minimum and maximum temperature observed in each frame; standard deviation; standard deviation for the minimum and maximum temperature observed in each frame; difference between the minimum and the maximum temperature.

- **Overall temperature features:** We also calculate the same features listed above, but using the entire frame as input, which includes the face, neck, and shoulders of the subject (as well as a static white background, which did not contribute to changes in temperature).

4.3 Heart Rate

In addition to the physiological signals captured through thermal imaging, we also measure the heart rate using the recording made with the video camera. In order to calculate the heart rate from the video sequences we used an adapted version of the method presented by Poh et al. [30]. We performed analysis on a sliding window of 30s length (15fps x 30s = 450 frames) with a 1s (15 frame) time increment.

For each video frame within a particular time window the face was segmented using the Open Computer Vision library (OpenCV) face detector [41]. The facial region of interest (ROI) was determined as the full height and central 60% of the width of the face box as determined by OpenCV. The ROI was separated into the three RGB (red, green and blue) channels and a spatial average of the resulting image components calculated. This was performed for every frame to yield the raw time varying signals r(t), g(t) and b(t). The raw traces were detrended using a procedure based on a smoothness priors approach [38] with a smoothing parameter λ=10. The channels were each normalized by subtracting their mean and then dividing by their standard deviation. The normalized signals were decomposed into the three independent source signals using the JADE implementation of Independent Component Analysis (ICA) [6].

Each of the source signals were band-pass filtered with normalized low and high 3dB frequency cut-offs of 0.08 and 0.2 respectively. The power spectrum of each of the resulting source signals was then calculated. The source channel containing the BVP was selected as the signal with the greatest frequency peak within the frequency range of interest: 0.08 to 2. For the first window the frequency with the maximum power within the range was selected as the heart rate frequency. Artifacts due to motion of the subject, ambient lighting changes or other noise can be problematic. Therefore for the remaining windows historical estimations of the pulse frequency were performed to reject artifacts by fixing a threshold for maximum change in pulse rate between successive measurements (taken 1s apart). If the difference between the current pulse estimation and the previously computed value exceeded the threshold (threshold = 10% of previously computed value) the algorithm rejected it. The frequency range was searched again for the second highest peak, this was repeated until the conditions were satisfied. If no frequency peaks met the criteria the previous pulse frequency estimation was maintained.

Using the extracted heart rate vectors, we calculate the following features:

- **Average, minimum, and maximum heart rate:** We calculate statistics over the entire vector, reflecting the average heart rate, as well as the minimum and the maximum heart rate.

- **Total heart rate changes:** Given the heart rate vector for a video, we determine the total absolute changes in heart rate over time, normalized with the duration of the video.

4.4 Verbal Descriptions

After each recording, the participants were asked to make a verbal statement describing how they felt about the video they just saw. We use a bag-of-words representation of these textual statements to derive unigram counts, which are then used as input features. First, we build a vocabulary consisting of all the words, including stopwords, occurring in the transcriptions of the training set. We then remove those words that have a frequency below three (value determined empirically on a small development set). The remaining words represent the unigram features, which are then associated with a value corresponding to the frequency of the unigram inside each transcription. These simple weighted unigram features have been successfully used in the past to build sentiment classifiers on text [27, 19]. The remaining words represent the unigram features, which are then associated with a value corresponding to the frequency of the unigram inside each line.

We also derive and use coarse textual features, by using mappings between words and semantic classes. Specifically, we use the OpinionFinder lexicon [44] to derive coarse textual features. The OpinionFinder lexicon was compiled from manually developed resources augmented with entries learned from corpora. It contains 6,856 unique entries, which are labeled as strong or weak clues of subjectivity, and also annotated for their polarity. In our classification, we use the *positive* and *negative* classes, and disregard the strength annotations. We thus derive two main types of linguistic features:

- **Unigrams:** For each of the unigrams selected as part of the vocabulary, we create a vector of features reflecting the frequency of the unigram.

- **Affective lexicon classes:** For each of the positive and negative affective classes obtained from the OpinionFinder lexicon, we infer a feature indicating the number of words in the verbal description belonging to that class.

5. PREDICTION OF AFFECTIVE RESPONSE

Our goal is to explore the changes in human affect during exposure to visual narratives. We formulate the task as a prediction problem, and use the features obtained from the four modalities described above in order to infer changes in affect and to classify them as either positive, negative, or neutral.

We run three main experiments. First, we run an experiment where we try to predict whether a person is being exposed to a stimulus that induces an affective state (either positive or negative), or is in a neutral state. Second, we also experiment with a three-way classification, where we differentiate in a single classifier between positive affect, negative affect, and a neutral state. Finally, we also run an experiment where we try to predict the valence of the affective state experienced by a person, and classify it as either positive or negative.

In all the experiments, we use the Ada Boost classifier with decision stumps as the classification algorithm,[2] and we run ten-fold cross validation experiments, meaning that we repeatedly train on a subset of the data and test on a separate subset. Thus, we do not use the entire data set for training, and the validation is independent.

5.1 Affective State versus Neutral State

We first build classifiers that attempt to determine if a person is in a positively or negatively valenced affective state, regardless of the valence of the affect she experiences. Since no linguistic descriptions have been collected from the participants during the neutral state, these classifiers are build using the non-verbal features, namely the facial expression features, thermal features, and heart rate. For these classifiers, we use all 70 videos that we collected, which include 56 videos recorded when the subjects were exposed to positive or negative stimuli, and 14 videos recorded when the subjects were in a neutral state.

Table 2 shows the results obtained with individual classifiers based on one modality at a time, and a combined classifier that makes use of all the non-verbal modalities. The figures in the table represent the percentage of times the classifiers have correctly identified the correct state (i.e., the percentage of times the classifiers have correctly labeled the state as either affective or neutral). The baseline for these classifiers is 80%, which corresponds to selecting by default an affective state (the majority class in this dataset).

Modality	Accuracy
Baseline	80.00%
Facial expressions	81.42%
Thermal features	90.00%
Heart rate features	88.57%
All modalities	**92.85%**

Table 2: Automatic classification performances to differentiate between an affective and a neutral state for three different non-verbal models: facial expressions, thermal features, and heart rate features. The integration of the three models provides the best results.

5.2 Three-way Classification: Positive, Negative, or Neutral

The second set of classifiers is concerned with the identification of the presence of affect, as well as the valence of the affect (positive or negative). As in the previous set of experiments, since no linguistic descriptions are available for the neutral state, we only use the features obtained from the non-verbal sensing.

Table 3 shows the results obtained by this three-way classification, using one non-verbal modality at a time, and all three modalities combined. Note that these classifiers use all 70 videos, and have a baseline of 40%, which corresponds to selecting by default a positive (or negative) affective state.

5.3 Positive State versus Negative State

Finally, we also build classifiers that try to determine the valence of the affect experienced by a person, by differentiating between a positive affect and a negative affect. Here,

[2]We use the implementation available in the Weka package www.cs.waikato.ac.nz/ml/weka/

Modality	Accuracy
Baseline	40.00%
Facial expressions	51.42%
Thermal features	54.28%
Heart rate features	**57.14%**
All modalities	55.71%

Table 3: Automatic classification performances for three-way classifiers that differentiate between a positive state, a negative state, and a neutral state for three different models: facial expressions, thermal features, and heart rate features.

we build individual classifiers for all verbal and non-verbal modalities, and also a combined classifier that includes all four modalities. For these experiments, we only use the videos recorded during exposure to positive or negative stimuli. We thus use features extracted from 56 videos, including 28 positive videos and 28 negative videos.

Table 4 shows the results obtained with individual classifiers based on one modality at a time, and a combined classifier that makes use of all the modalities. The baseline for these classifiers is 50%, which corresponds to selecting by default a positive (or negative) affective state.

Modality	Accuracy
Baseline	50.00%
Facial expressions	61.02%
Thermal features	50.00%
Heart rate features	53.57%
Linguistic features	67.51%
All modalities	**73.21%**

Table 4: Automatic classification performances to differentiate between a positive and a negative affective state for four different models: facial expressions, thermal features, heart rate features, and linguistic features. The integration of the four models provides the best results.

6. DISCUSSION

These initial experiments reveal interesting findings about the influence of visual narratives on affect, and how that can be captured through four different modalities.

The first experiment on differentiating between an affective and a neutral state clearly shows that physiological modalities, such as heart rate and changes in skin temperature, are more effective than a visual modality that relies primarily on facial expressions. This is probably explained by the fact that a state of excitement, which typically corresponds to an affective state, induces changes in our physiological functions that are very well captured by thermal recordings and measurements of heart rate. On the other side, it appears that the facial expressions made during a neutral state are not very different from those that are made in an affective state, as a neutral state may also include occasional smiles or look aways that can confuse a classifier that relies primarily on visual clues. The joint use of all non-verbal modalities provides the best result, representing a relative error rate reduction of 64% compared to the baseline.

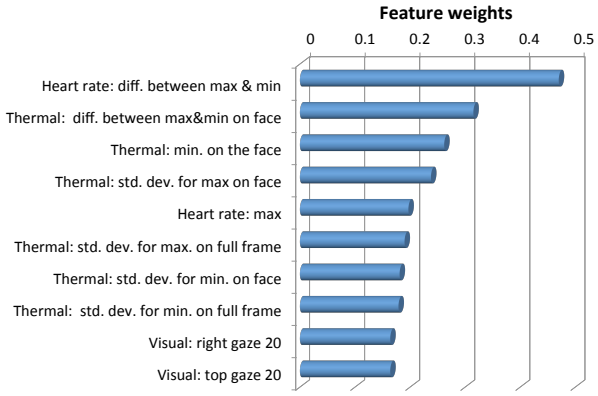

Figure 4: Top ten non-verbal features selected by the affective state classifier.

The second experiment shows a similar trend, where the physiological measures provide the best results. An analysis of the confusion matrix produced by these classifiers shows that once again these measures are doing very well in separating an affective state from a neutral state. Unlike the previous experiments, the visual features are also effective when compared to the baseline, which can be explained by their ability to differentiate positive and negative excitement (as also shown by the third experiment).

Finally, the last experiment shows the effectiveness of the facial expressions to distinguish between different types of affect (positive and negative). The thermal features and the heart rate features are significantly less effective here. This is probably due to the fact that the excitement experienced in an affective state induces changes in heart rate and skin temperature regardless of the valence of the affect (e.g., our heart beats faster when we are sad and it also beats faster when we are happy). In this experiment, we were also able to add the linguistic modality, which is the most effective overall. This suggests that when present, verbal communication can be a very useful clue for the prediction of affective state. The combination of all four modalities provides the best results, with a relative error rate reduction of 46% compared to the baseline.

To gain further understanding into the role played by the non-verbal features in predicting the changes in affective state, we compare the information gain assigned by the learning algorithm to each of these features. Figure 4 shows the top ten non-verbal features selected by the classification algorithm, obtained during the experiment that differentiates between affective and neutral states.

In line with the observations above, the physiological measures are the most useful when making a distinction between affective and neutral states. Among these, the difference between the maximum and minimum heart rate and the difference between the maximum and minimum temperature on the face seem to be the most useful features, followed by

several other physiological features that reflect the minimum temperature, the standard deviation for the minimum and the maximum temperature, the maximum heart beat, and others. Among the visual features, the gaze to the right and gaze to the top appear to be the features that matter most for this classification.

For the distinction between positive and negative affect, the only significant features identified by the classifier are two smile features (number of smiles with intensity of 50 and 75 respectively), followed by the gaze to the right and gaze to the top features. None of the physiological features appear to be strong indicators of the valence of the affective state.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we explored a non-invasive multimodal approach to sense changes in human affect when exposed to emotionally loaded videos. We experimented with four different modalities: visual facial behaviors, physiological measurements - including thermal imaging and heart rate measurements, and verbal descriptions. The first three modalities are non-verbal and are recorded during the narrative interaction, while the fourth modality is acquired in a post-study interview. Our experiments show that these modalities complement each other in detecting human affect. Specifically, our initial findings seem to suggest that physiological measures are most effective in identifying the presence of an affective state (as compared to a neutral state), whereas facial behaviors and verbal descriptions are most effective at differentiating between positive and negative states.

To our knowledge, this is the first attempt to integrate these four modalities (visual, thermal, heart rate, and linguistic) to predict human affective response to visual narratives. In future work, we plan to perform a more fine-grained temporal analysis of the multimodal features as they align with the content of the visual stimuli, to reach a better understanding of when and why do changes in affect take place, and how they can be effectively sensed and predicted.

Acknowledgments

We are grateful to Prof. Bill Buckles from the University of North Texas for allowing us to use the thermal camera. This material is based in part upon work supported by National Science Foundation awards #0917170 and #0917321. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

8. REFERENCES

[1] J. Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement*, 28:R1, 2007.

[2] C. Alm, D. Roth, and R. Sproat. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, Canada, 2005.

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH '99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[4] C. Busso and S. Narayanan. Interrelation between speech and facial gestures in emotional utterances: a single subject study. *IEEE Transactions on Audio, Speech and Language Processing*, 15(8):2331–2347, November 2007.

[5] C. Camerer. *Behavioral Game Theory: Predicting Human Behavior in Strategic Situations*. Princeton University Press, 2004.

[6] J. Cardoso. High-order contrasts for independent component analysis. *Neural computation*, 11(1):157–192, 1999.

[7] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):681–685, Jun 2001.

[8] T. F. Cootes and C. J. Taylor. Active shape models - 'smart snakes'. In *Proceedings of the British Machine Vision Conference*, 1992.

[9] D. Cristinacce and T. Cootes. Feature detection and tracking with constrained local models. In *Brittish Machine Vision Conference*, 2006.

[10] J. Fei and I. Pavlidis. Thermistor at a distance: Unobtrusive measurement of breathing. *Biomedical Engineering, IEEE Transactions on*, 57(4):988–998, 2010.

[11] G. Fine. *The Storied Group: Social Movements as "Bundles of Narratives"*. SUNY Press, 2002.

[12] M. Garbey, N. Sun, A. Merla, and I. Pavlidis. Contact-free measurement of cardiac pulse based on the analysis of thermal imagery. *Biomedical Engineering, IEEE Transactions on*, 54(8):1418–1426, 2007.

[13] E. Greneker. Radar sensing of heartbeat and respiration at a distance with applications of the technology. In *Radar 97 (Conf. Publ. No. 449)*, pages 150–154. IET, 1997.

[14] H. Gunes and M. Pantic. Automatic, dimensional and continuous emotion recognition. *Int'l Journal of Synthetic Emotion*, 1(1):68–99, 2010.

[15] C. Izard. Emotion theory. *Annual Review of Psychology*, 60(1), 2009.

[16] A. Kane. *Finding Emotion in Social Movement Processes: Irish Land Movement Metaphors and Narratives*. University of Chicago Press, 2001.

[17] M. Khan, R. Ward, and M. Ingleby. Classifying pretended and evoked facial expressions of positive and negative affective states using infrared measurement of skin temperature. *ACM Transactions on Applied Perception*, 6(1), 2009.

[18] A. Liu and D. Salvucci. Modeling and prediction of human driver behavior. In *Proceedings of the International Conference on Human-Computer Interaction*, 2001.

[19] A. Maas, R. Daly, P. Pham, D. Huang, A. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the Association for Computational Linguistics (ACL 2011)*, Portland, OR, 2011.

[20] G. Mishne. Experiments with mood classification in

blog posts. In *Proceedings of the 1st Workshop on Stylistic Analysis Of Text For Information Access (Style 2005)*, Brazile, 2005.

[21] L. Morency, R. Mihalcea, and P. Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the International Conference on Multimodal Computing*, Alicante, Spain, 2011.

[22] L. Morency, J. Whitehill, and J. Movellan. Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. In *Automatic Face and Gesture Recognition*, pages 1–8, 2008.

[23] E. Murphy-Chutorian and M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:607–626, 2009.

[24] M. Nicolaou, H. Gunes, and M. Pantic. Audio-visual classification and fusion of spontaneous affective data in likelihood space. In *ICPR*, 2010.

[25] M. Nicolaou, H. Gunes, and M. Pantic. Output-associative rvm regression for dimensional and continuous emotion prediction. In *IEEE FG'11*, 2011.

[26] N. Oliver, B. Rosario, and A. Pentland. A Bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 2000.

[27] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*, Barcelona, Spain, July 2004.

[28] M. Pantic and M. S. Bartlett. Machine analysis of facial expressions. In K. Delac and M. Grgic, editors, *Face Recognition*, pages 377–416. I-Tech Education and Publishing, Vienna, Austria, July 2007.

[29] I. Pavlidis and J. Levine. Thermal image analysis for polygraph testing. *IEEE Engineering in Medicine and Biology Magazine*, 21(6), 2002.

[30] M. Poh, D. McDuff, and R. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optical Society of America*, 2010.

[31] C. Puri, L. Olson, I. Pavlidis, J. Levine, , and J. Starren. Stresscam: non-contact measurement of users' emotional states through thermal imaging. In *Proceedings of the 2005 ACM Conference on Human Factors in Computing Systems (CHI)*, 2005.

[32] S. Raaijmakers, K. Truong, and T. Wilson. Multimodal subjectivity analysis of multiparty conversation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 466–474, Honolulu, Hawaii, 2008.

[33] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, pages 105–112, 2003.

[34] N. Sebe, I. Cohen, T. Gevers, and T. Huang. Emotion recognition based on joint visual and audio cues. In *ICPR*, 2006.

[35] C. Strapparava and R. Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on the Semantic Evaluations (SemEval 2007)*, Prague, Czech Republic, 2007.

[36] C. Strapparava and R. Mihalcea. Learning to identify emotions in text. In *Proceedings of the ACM Conference on Applied Computing ACM-SAC 2008*, Fortaleza, Brazile, 2008.

[37] C. Strapparava and A. Valitutti. Wordnet-affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, 2004.

[38] M. Tarvainen, P. Ranta-Aho, and P. Karjalainen. An advanced detrending method with application to hrv analysis. *Biomedical Engineering, IEEE Transactions on*, 49(2):172–175, 2002.

[39] P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 417–424, Philadelphia, 2002.

[40] S. Ulyanov and V. Tuchin. Pulse-wave monitoring by means of focused laser beams scattered by skin surface and membranes. In *Proceedings of SPIE*, 1993.

[41] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.

[42] D. Watson and L. Clark. *THE PANAS-X Manual for the Positive and Negative Affect Schedule - Expanded Form*. University of Iowa.

[43] J. Wiebe. Learning subjective adjectives from corpora. In *Proceedings of the American Association for Artificial Intelligence (AAAI 2000)*, pages 735–740, Austin, Texas, 2000.

[44] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210, 2005.

[45] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, Canada, 2005.

[46] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie. Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *INTERSPEECH*. ISCA, 2008.

[47] M. Wollmer, B. Schuller, F. Eyben, and G. Rigoll. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. *IEEE Journal of Selected Topics in Signal Processing*, 4(5), October 2010.

[48] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[49] Z. Zhihong, M. P. G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *PAMI*, 31(1), 2009.