# CONTINUOUS FACIAL EXPRESSION RECOGNITION FOR AFFECTIVE INTERACTION WITH VIRTUAL AVATAR

*Zhengkun Shang, Jyoti Joshi, Jesse Hoey*

David R. Cheriton School of Computer Science
University of Waterloo, Waterloo, ON
{z4shang, jyoti.joshi, jhoey}@uwaterloo.ca

## ABSTRACT

Affect-Sensitive Human-Computer Interaction is enjoying growing attention. Emotions are an essential part of interaction, whether it is between humans or human and machine. This paper analyses the interaction of a user with four different virtual avatars, each manifesting distinct emotional displays, based on the principles of Affect Control Theory. Facial expressions are represented as a vector in a 3D continuous space and different sets of static visual features are evaluated for facial expression recognition. A probabilistic framework is used to simulate the interaction between the user and the virtual avatar. The results demonstrate that the probabilistic framework enables the system to perceive user's and agent's feelings.

***Index Terms*—** Emotion, Facial Expression, Affect Control Theory, Human Computer Interaction, Affective Computing

## 1. INTRODUCTION

Affect (emotion) is an integral part of human perception and communication. Recently, integration of affect into Human-Computer Interaction (HCI) has gained much attention [1, 2]. It has been argued that systems capable of sensing and responding appropriately to users' affective feedback are likely to be perceived as more natural, persuasive, and trustworthy [3]. Facial expressions are the predominant non-verbal cues observed to infer an emotional state. Automatic facial expression recognition has been focused on categorising the emotions in predefined sets of universal categories. However, these categories do not cover the vast spread of emotions and the subtle expressions that humans display and perceive. On the other hand, representing emotions in a continuous multi-dimensional space enables encoding subtly different expressions and mixed emotions. Corneanu et al. [4] presents a survey on the latest trends in facial expression analysis.

In this paper, we first evaluate three different pseudo-static features for continuous facial expression recognition. Furthermore, we propose a technique that predicts a virtual avatar's preception of a user's affective state during an interaction based on the continuous facial expressions of the user. The paper demonstrates preliminary results and shows that this information can be leveraged to design emotionally aware human-computer interaction systems.

## 2. MODEL AND DATA

The proposed system is developed based on the social-psychological principles of Affect Control Theory (ACT) [5]. ACT proposes that people have fundamental (out-of-context) sentiments, which are representations of social objects, such as interactants' identities and behaviors or environmental settings in a 3D affective space Evaluation-Potency-Activity (EPA). These fundamental sentiments are *culturally shared*, meaning that there is implicit agreement amongst people of a similar culture about the affective connotations (meanings) of things. Fundamental sentiments are measured in large-scale human surveys and stored in dictionaries.[1] Transient impressions, which are also three-dimensional vectors in EPA space, result from the interaction in a social event which may cause deviation in the identity or behavior from their corresponding fundamental sentiments. The formation of transient impressions is also a culturally shared phenomenon. ACT proposes that people try to maintain consistency (alignment) in an interaction, and keep transient impressions close to fundamental sentiments. Emotions in ACT have a clear definition as the vector difference between fundamental sentiments and transient impressions. Emotions are signals sent from one interactant to another in order to help maintain the alignment.

In this work, we used BayesACT [6, 7, 8], a generalization of ACT. It keeps multiple hypothesis about both identities and behaviors as a probability distribution and is a sequential Bayesian model that estimates and updates distributions over fundamental sentiments, transient impressions and emotions over time from actions and observations. The distributional modeling of sentiment enables an artificial agent to provide more enriched interaction experience to users by learning their identities, taking into account their behaviours and emotions. The BayesACT engine is able to choose an action that minimizes deflection according to ACT principles.

The Semaine database [9] provides extensive annotated audio and visual recordings of a person interacting with an emotionally limited avatar, or sensitive artificial listener (SAL), to study natural social behavior in human interac-

[1] We use the Indiana 2002-2004 study in this work.

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

tion. Each video is a recording of a conversation between a user and a human actor. The avatar was asked to act in one of the four SAL avatars in each video: *"Poppy"* is happy and tries to make the user happy; *"Spike* acts angry; *"Obadiah"* is sad and depressed; and *"Prudence"* is sensible and even-tempered. The video recordings were transcribed and annotated frame by frame by 6 to 8 raters into five affective dimensions: Valence, Power, Activation, Anticipation, and Intensity. The first three dimensions constitute our EPA space in the studies. Since only a few videos of the avatar are annotated and the avatar is acting in different characters, we only use user's clips instead, which gives us 93 clips with 20 persons involved. The video is recorded at 49.979 frames per second and the emotions are annotated from -1 to 1.

## 3. EMOTION PREDICTION

### 3.1. Implementation

We first experimented with three commonly used pseudo-static visual feature descriptors extracted from raw face images and compared their accuracy in predicting a person's Evaluation, Potency, and Activity (EPA) scores. These descriptors are Action Unit (AU) [10], Histogram of Oriented Gradient (HOG) [11], and Felzenszwalb's HOG (FHOG) [12]. The three feature descriptors characterize human faces from different aspects:

- AU encodes muscular activity that produces appearance changes, commonly used to analyze facial expressions. The AU activation information was used as a high level feature descriptor.

- A typical low-level feature descriptor, HOG splits an image into a number of non-overlapping cells. For each cell, it computes a histogram of gradients, discretized into 9 orientation bins and normalized with the total energy of the four $2\times2$ blocks containing this cell. The parameters mentioned here are chosen manually. This descriptor is commonly used for localizing face positions and is paired with an SVM classifier.

- FHOG is a variant of HOG that reduces the HOG feature space using principal component analysis, which makes it possible to use fewer parameters in its models to speed up detections and learning processes.

For each feature descriptor, we employed the same workflow illustrated in Fig. 1 for training and testing models. At first, we randomly split 93 video clips into a training set and a test set. The training set has 75 clips and the test set has 18 clips for tuning and evaluating the model. For each image, we localized and aligned the face in the image, then extracted the feature using one of the three descriptors. We leveraged OpenFace [13], an open-source framework, to localize face positions and extract features from aligned facial images. We used LibSVM [14] for training and testing models.

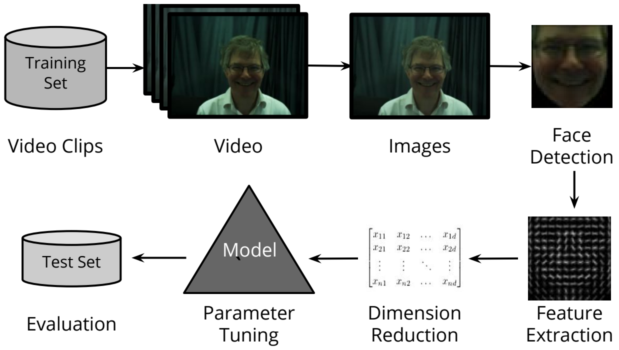


**Fig. 1**. An overview of training and continuous prediction.

**Table 1**. Number of dimensions used for each feature descriptor before and after dimension reduction.

| Feature Descriptor | Before PCA | After PCA |
|---|---|---|
| **AU** | 17 | 7 |
| **HOG** | 5184 | 49 |
| **FHOG** | 4464 | 153 |

We initially employed all the 93 annotated videos from the Semaine database for model training and this translated to well over one million images. However, we soon found this large amount of data exceeding the memory capacity, during the training of the models. We therefore sampled one frame every 0.4 second from the videos, resulting in 50,601 images for training and 12,315 for testing. To further optimize the time required for training, we used principal component analysis (PCA) to reduce the feature space. The first $k$ eigenvectors that contain at least 80% variance in the training set were selected. Table 1 shows the dimension number after feature extraction and after applying PCA.

For each feature descriptor, three models that correspond to evaluation, potency, and activity were trained. We used support vector regression (SVR) for training all models. The radial basis function (RBF) kernel was used and 5-cross validation was performed towards the training set to tune parameters $C$, $\gamma$, and $\epsilon$.

To evaluate a model's prediction performance, we used the root-mean-square error (RMSE) as our metric. We chose the parameters with the least RMSE from cross validation to train the models. The fitted models were evaluated with the test set. Table 2 shows the RMSEs for each feature descriptor. There are four conditions: AU without PCA, AU with PCA, HOG with PCA, and FHOG with PCA. Because the original AU descriptor has only 17 dimensions, applying PCA was not necessary. However, we still applied PCA to AU for fair

**Table 2**. EPA prediction with different feature descriptors

| Dimension | Feature Descriptor | Training Set |  | Test Set |  |
|---|---|---|---|---|---|
|  |  | **RMSE** | **COR** | **RMSE** | **COR** |
| Evaluation | AU | 0.251 | 0.127 | **0.272** | **0.235** |
|  | AU+PCA | 0.252 | 0.081 | 0.273 | 0.205 |
|  | HOG+PCA | 0.251 | 0.082 | 0.274 | 0.213 |
|  | FHOG+PCA | 0.251 | 0.161 | 0.277 | 0.121 |
| Potency | AU | 0.225 | 0.176 | **0.233** | 0.189 |
|  | AU+PCA | 0.227 | 0.024 | 0.234 | 0.081 |
|  | HOG+PCA | 0.228 | 0.040 | 0.235 | 0.122 |
|  | FHOG+PCA | 0.229 | 0.087 | **0.233** | **0.196** |
| Activity | AU | 0.220 | 0.303 | 0.221 | 0.352 |
|  | AU+PCA | 0.231 | 0.181 | 0.229 | 0.339 |
|  | HOG+PCA | 0.219 | 0.336 | 0.216 | 0.414 |
|  | FHOG+PCA | 0.217 | 0.350 | **0.214** | **0.444** |
| (Average) | AU | 0.232 | 0.202 | 0.242 | **0.259** |
|  | AU+PCA | 0.237 | 0.095 | 0.245 | 0.208 |
|  | HOG+PCA | 0.233 | 0.153 | 0.242 | 0.250 |
|  | FHOG+PCA | 0.232 | 0.199 | **0.241** | 0.254 |

comparisons. We observed a worse performance after applying PCA to AU, understandably due to 20% loss of variance. It is to be understood that AUs are high-level feature representations, hence, already in a lower dimension. FHOG had the best averaged accuracy among all four conditions. Moreover, activity received the best accuracy among the three EPA dimensions, perhaps because it is easier to detect from facial expression than evaluation and potency.

## 4. BAYESACT SIMULATIONS

With facial features transformed into the EPA space that characterizes the users' sentiments, we demonstrate the feasibility of further building artificial systems that track users' real-time emotions through BayesACT simulations. The BayesACT simulations allow the avatar to learn a model of the user's emotional state based on the ACT principles. In this work, we study how and to what extent avatars with distinct emotion characteristics affect the user's emotional states. Our analysis provides preliminary evidence for BayesACT as a model for integrating emotional intelligence within artificial agents.

BayesACT simulates the interactions between a user and an avatar from the avatar's perspective. In the simulation, the user and the avatar take turns, and provide an EPA action that denotes the emotional content of their current behaviour (e.g. 'talking to' someone is good and powerful, while 'yelling at' someone is bad and powerful). In each turn, in addition to providing an action, the user also supplies the current emotional state. The avatar perceives the user's emotional signal, and uses it to update its estimate of transient impressions and fundamental sentiments. In this experiment, we supplied the emotional signals extracted from the conversations in the video as the input to BayesACT and analyzed the learned users' sentiment change for four avatars with different emotional characteristics.

We set the user's identity to 'student' (EPA: $1.5, 0.3, 0.8$), since all participants in the experiment were undergraduate or graduate students. In each turn, the user or the avatar performed the action 'talk to' (EPA: $1.5, 1.3, 0.9$) to the other party. When it was the user's turn, the current (sampled every 5 seconds) EPA values from the facial expression database were supplied as the emotional signal.[2] The raw EPA values ranged $[-1, 1]$ were transformed using a tangent function $2.77 * tan(EPA)$ to range $[-4.3, 4.3]$, as required by the BayesACT engine. After each turn, the avatar generates two groups of EPA values: the user's emotion and the transient impression of the user's identity. In addition, the avatar converts these sentiments into labels. A label is an adjective chosen from the sentiment dictionary that has the maximum cosine similarity from the user's current emotion EPA. When aggregating the conversation simulations, the mean values of EPA were calculated as a measurement of tendency. Since the lengths of the videos varied, we aligned our data to use only the first 200 seconds of each conversation. Therefore, each simulation has at most 40 rounds.

### 4.1. Emotion Change during Simulations

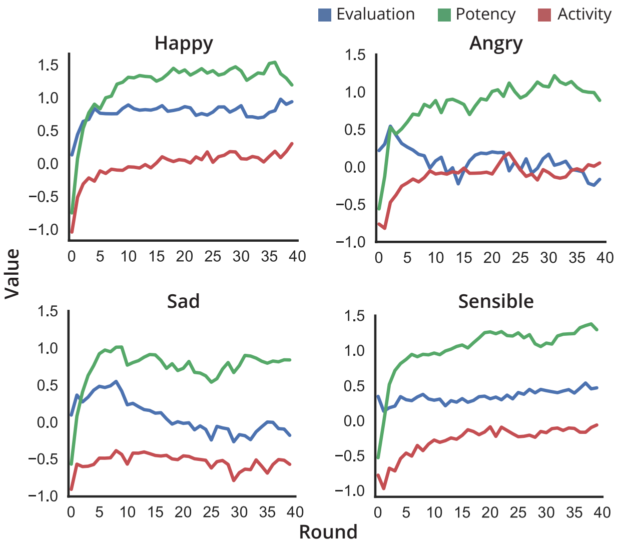


**Fig. 2**. Avatar's posterior estimate of the user's emotions.

The experiment results are illustrated in Fig. 2 and Fig. 3, showing the avatars posterior estimate of the user's emotions and transient impressions of the user's identity, respectively. The figures illustrate the averaged values at each time stamp for Evaluation, Potency, and Activity across all 18 test videos. The user's emotion changed significantly with different avatars. The users were positive, confident (powerful) and calm when talking to a happy avatar, since EPA of the happy avatar's behaviors was similar to how a student should

[2] We present the simluations using the ground truth as input to more clearly show the emotional prediction mechanism at work.

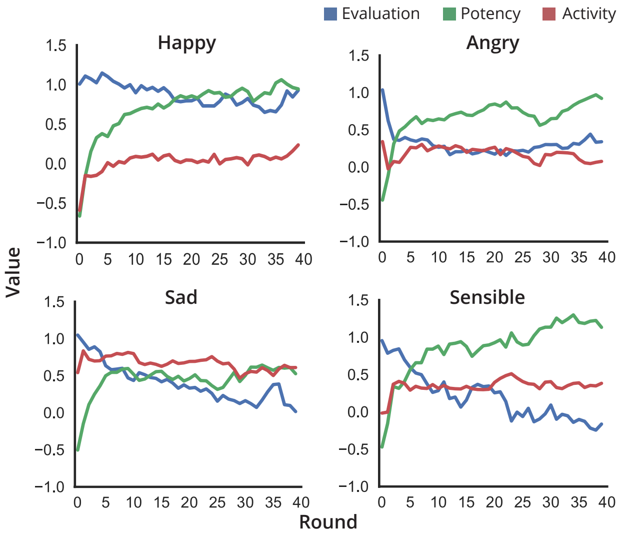
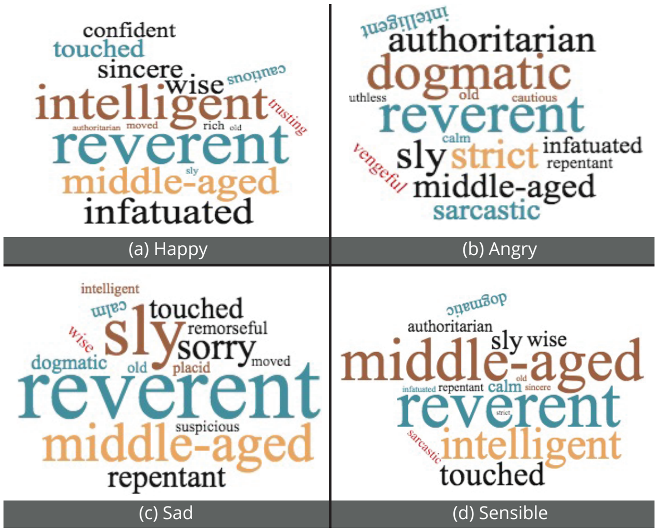


**Fig. 3**. Avatar's posterior estimate of the user's transient impressions.

behave. Therefore, the deflection was small and the conversation went smoothly. However, when the avatar was angry, the avatar behaved unpleasantly (more powerful and active), which made the users feel more negative. When the avatar was sad, the users maintained were less positive and less active. Moreover, when the avatar changed to be sensible, the users felt less positive compared with talking to a happy avatar. User's were evaluated as having more positive identities when interacting with the happy avatar, more powerful when interacting with a sensible avatar, and less powerful when interacting with a sensible avatar.



(a) Happy

(b) Angry

(c) Sad

(d) Sensible

**Fig. 4**. Four word clouds created by the top 15 words that describe the user's feeling when talking to Poppy (Happy), Spike (Angry), Obadiah (Sad), and Prudence (Sensible).

## 4.2. Sentiment Labels

To better visualize how different avatars affect the user's emotion, we aggregated the sentiment labels generated from the BayesACT simulations, which approximates users' emotional states. The aggregated labels were used to create four word clouds as shown in Fig. 4.[3] Each word cloud contains the top 15 words that describe the user's emotions from the avatar's perspective, for different avatars. The size of each word is dictated by its frequency of appearance.

In the word clouds, the word 'reverent' appeared the most number of times in all conditions, understandably representing the mental state of a student interacting with a person who leads the conversation with more or less greater extents of power. Other than 'reverent', the user showed different emotions towards different types of avatars. From the avatar's perspective, the users felt 'infatuated', 'intelligent', 'wise', 'touched', and 'confident' when talking to a happy person. They felt 'strict', 'dogmatic', and 'authoritarian', interacting with an angry avatar. When talking to a sad avatar, the users felt 'sorry', 'touched', 'remorseful', 'repentant', and 'middle-aged'. When talking to a sensible avatar, the users felt 'middle-aged', 'sly', 'intelligent', and 'touched'. The word cloud provides us with an impression of how the avatar processes users' emotions given their facial expressions in the BayesACT simulations.

## 5. CONCLUSION

The paper addressed an interesting and challenging requirement of HCI, i.e. integration of emotional intelligence to add empathy to the system. Facial expressions were recognised in a 3D continuous domain. Based on the principles of ACT, BayesACT was used to simulate interactions between a user and distinct behaviour-styled avatars. The result demonstrates affective information of users and avatars can be perceived and thus can be used to build affective HCI systems. In the simulations, we assumed the identity of user to be student and action as 'talk to' in all the cases. However, this might not always hold true. Despite these assumptions we see some distinctive words generated in the word cloud based on the behaviour of different avatars. We presented BayesACT simulation results using the database labels as input, but clearly it would be preferable to use the automated facial expression recognition. Our analysis shows that improvements will be needed in 3D continuous facial expression recognition in order to make this feasible. Once these improvements are made, we plan to test this technique in assistive systems for persons with dementia, e.g. by modeling the identity of the user and avatar as 'patient' and 'assistant' [15, 16].

**Acknowledgments:** We thank the support of AGE-WELL NCE Inc., the Canadian Consortium on Neurodegeneration and Aging, and of the Alzheimers Association (grant number ETAC-14-321494).

[3] Generated with `http://www.wordclouds.com/`.

## 6. REFERENCES

[1] Christian Peter and Russell Beale, Eds., *Affect and Emotion in Human-Computer Interaction: From Theory to Applications*, Springer-Verlag, Berlin, Heidelberg, 2008.

[2] Rachel Kirby, Jodi Forlizzi, and Reid Simmons, “Affective social robots,” *Robotics and Autonomous Systems*, vol. 58, no. 3, pp. 322–332, 2010.

[3] Maja Pantic and Leon JM Rothkrantz, “Toward an affect-sensitive multimodal human-computer interaction,” *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, 2003.

[4] Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F Cohn, and Sergio Escalera Guerrero, “Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications,” *IEEE TPAMI*, vol. 38, no. 8, pp. 1548–1568, 2016.

[5] David R Heise, *Expressive order: Confirming sentiments in social actions*, Springer Science & Business Media, 2007.

[6] Jesse Hoey, Tobias Schroder, and Areej Alhothali, “Bayesian affect control theory,” in *ACII, 2013 Humaine Association Conference on*. IEEE, 2013, pp. 166–172.

[7] Jesse Hoey, Tobias Schröder, and Areej Alhothali, “Affect control processes: Intelligent affective interaction using a partially observable Markov decision process,” *Artificial Intelligence*, vol. 230, pp. 134–172, January 2016.

[8] Tobias Schröder, Jesse Hoey, and Kimberly B. Rogers, “Modeling dynamic identities and uncertainty in social interactions: Bayesian affect control theory,” *American Sociological Review*, vol. 81, no. 4, 2016.

[9] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, “The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent,” *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, Jan. 2012.

[10] Y. I. Tian, T. Kanade, and J. F. Cohn, “Recognizing action units for facial expression analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97–115, Feb. 2001.

[11] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *CVPR’05)*, June 2005, vol. 1, pp. 886–893 vol. 1.

[12] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan, “Object detection with discriminatively trained part-based models,” *IEEE TPAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[13] T. Baltruaitis, P. Robinson, and L. P. Morency, “OpenFace: An open source facial behavior analysis toolkit,” in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2016, pp. 1–10.

[14] Chih-Chung Chang and Chih-Jen Lin, “LIBSVM: A Library for Support Vector Machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[15] Luyuan Lin, Stephen Czarnuch, Aarti Malhotra, Lifei Yu, Tobias Schröder, and Jesse Hoey, “Affectively aligned cognitive assistance using bayesian affect control theory,” in *Proc. of International Workconference on Ambient Assisted Living (IWAAL)*, Belfast, UK, December 2014, pp. 279–287, Springer.

[16] Alexandra König, Linda E. Francis, Jyoti Joshi, Julie M. Robillard, and Jesse Hoey, “Qualitative study of affective identities in dementia patients for the design of cognitive assistive technologies,” *Journal of Rehabilitation and Assistive Technologies Engineering*, vol. 4, 2017.