AUTOMATIC INFERENCE OF MENTAL STATES FROM SPONTANEOUS FACIAL EXPRESSIONS

Yanjia Sun and Ali N. Akansu

New Jersey Institute of Technology
Department of Electrical & Computer Engineering
University Heights Newark, NJ 07102 USA
{Yanjia.Sun, Akansu}@NJIT.edu

ABSTRACT

Human face is a display of mental states that reflect the true feelings of a person. In this paper, we propose a framework for the video analysis of spontaneous facial expressions using an automatic facial emotion recognition system. Regional Hidden Markov Models (RHMMs) are created to describe the states of facial attributes for eyebrows, eyes, and mouth regions registered in a video sequence. The performance results reported in the paper show that the proposed technique outperforms the designated HMM for each emotion type [1, 2] tested with the Cohn-Kanade database for the person-independent case. More importantly, we used the proposed system to infer the mental states of a person based on spontaneous facial expressions. Merit of the proposed system is validated with human based evaluations.

Index Terms—Automatic facial emotion recognition, facial perception, Regional Hidden Markov Model, states of face regions, mental states

1. INTRODUCTION

Human facial expressions play significant role in conveying a person's mental states that reflect his internal cognitive states [3]. Methods to classify facial expressions are generally grouped as static and dynamic ones [4]. Static methods, taking advantage of Support Vector Machine, Neural Network, Bayesian Network, and others are based on the information obtained from a single video frame. Dynamic classifiers like Hidden Markov Model (HMM) utilize temporal information to analyze facial expressions, and are strongly suggested by the psychological experiments as reported in [5]. However, most of the prior research classify the entire image sequence into one of the expression types. In contrast, such a system should be able to classify every frame of a sequence highlighting temporal dynamics. Moreover, most of the earlier research employed emotion-specific HMMs for the entire frame to train and test a single emotion type [1, 2]. The information describing a facial emotion is mainly registered in the movements of facial regions [6]. High recognition rates of facial emotions have been achieved in existing systems based on the benchmarked databases containing posed facial emotions. However, the posed facial emotions are not proven yet to interpret the true feelings of humans [4]. They are generated by asking subjects to perform a series of exaggerated expressions that are quite different from spontaneous ones [7, 8]. One of the most interesting challenges in the area of Human-Computer Interaction (HCI) is how to make computers more human-like for intelligent user interfaces.

In this paper, a Regional Hidden Markov Model (RHMM) based facial emotion recognition system is proposed. It exploits the spatio-temporal dynamics of the facial video sequence by analyzing the movements (motions) of three regions eyebrows, eyes, and mouth in a frame. It is a marked departure from HMM based earlier studies that modeled emotion types for the entire image [1, 2]. The recent research in psychology field empirically revealed that people reliably infer others' preferences from spontaneous facial expressions [9]. To assess the plausibility of the proposed system, we compare its performance results with the human's inference of other people's mental states by analyzing the same datasets.

We propose an automatic facial emotion recognition system in Section 2. The experiments based on the Cohn-Kanade database are presented in Section 3. The experiments of inference to the human's mental states are detailed in Section 4. Finally, conclusions are drawn in Section 5.

2. SYSTEM DESCRIPTION

2.1. Feature extraction

We define typical states of face regions as shown in Table 1. 41 facial feature points are identified on each frame of video, as displayed in Fig. 1. They are comprised of 10 salient points on the eyebrows region, 12 points on the eyelids, 8 points on the mouth, 10 points on the corners of the lips, and one anchor feature point on the nose. The 2D coordinates of facial feature points in various face regions are extracted to form corresponding observation sequences for classification. The facial feature points are tracked using a constrained local model [10]. It utilizes a non-parametric method to represent the distribution of candidate locations

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

by using an optimization method called constrained mean-shifts. It outperforms the other methods for deformable model fitting [11, 12, 13].

Table 1. States of Three Face Regions.

Face Regions		Observable States
Eyebrows		*raise, fall, neutral*
Eyes		*open, close, neutral*
Mouth	Mouth	*open, close, neutral*
	lips corners	*up, down, pull, pucker, neutral*

2.2. Regional Hidden Markov Model

14 different RHMMs, labeled λ_1 to λ_{14}, are generated. Each represents one of the 14 observable states of the three facial regions. The observation sequences are formed using the coordinates of the facial feature points in a 2D representation. In Table 2, we tabulate six basic emotion types [14] and their states.

Table 2. Library of Emotion Types and States of Face Regions.

Emotion Type	States of Face Regions
Anger	Eyebrows fall, eyes close, mouth close, and lips corners pucker
Disgust	Eyebrows fall, eyes close, mouth close and lips corners pull
Sadness	Eyebrows raise, eyes close, mouth close and lips corners down
Happiness	Eyebrows neutral, eyes neutral, mouth open, and lips corners up
Surprise	Eyebrows raise, eyes open, mouth open, and lips corners neutral
Fear	Eyebrows raise, eyes open, mouth open, and lips corners pull

2.3. Training RHMMs and recognition

We select the left-to-right model for training and recognition RHMMs. Gaussian type is chosen to represent the observation probability distribution. RHMMs are trained using the Baum-Welch algorithm [15]. In the recognition step, the probability of the observation sequence for the related RHMM $P(O_i \mid \lambda_j)$ is calculated using the forward-backward algorithm [15], where the observation sequence O_i corresponds to the i^{th} facial region and j is the state of a face region. A prototype emotion is expressed by combining states of the all facial regions. Therefore, we calculate the probability of an emotion type P_k by summing the probabilities of the states of different facial regions according to Table 2 as follows

$$P_k = \sum P(O_i \mid \lambda_j) \quad (1)$$

where k is one of the six basic emotion types. The emotion type for a facial video frame is decided based on the highest measured probability $\max(P_k)$.

3. SYSTEM PERFORMANCE

The widely referenced Cohn-Kanade database [16] is used to recognize the six basic facial emotions listed in Table 2. We conduct two sets of experiments for the person-independent case and compare our results with prior art. 180 video sequences and 30 different subjects from the Cohn-Kanade database are used such that one sequence per subject per universal emotion type is included. 24 out of the 30 subjects are utilized for training, with the remaining 6 subjects used for performance tests. Two sets of experiments are run using 5-fold cross validation. One is to classify facial features using the proposed RHMM framework for the states of face regions. The other uses the emotion-specific HMM [1, 2].

Fig. 1 displays the recognition rates for emotion types as a function of frames (time) in a video sequence. The recognition performances for the proposed RHMM and emotion-specific HMM systems are tabulated in Tables 3 and 4, respectively. The former performs 2.84% better than the latter tested on the same database. The performance is also better than earlier work reported in the literature for the person-independent case. The recognition rates given in [17] and [18] are 86% and 85.84%, respectively.

Table 3. Recognition performance (in %) of the proposed method (average recognition rate is 86.67%).

	A	D	F	H	N	SA	SU
A	**76.67**	10	3.33	0	0	10	0
D	3.33	**83.33**	0	3.33	0	10	0
F	0	0	**86.66**	3.33	0	3.33	6.67
H	0	0	3.33	**96.67**	0	0	0
N	0	0	0	0	**100**	0	0
SA	3.33	16.67	10	0	6.67	**63.33**	0
SU	0	0	0	0	0	0	**100**

Table 4. Recognition performance (in %) of emotion-specific HMM per emotion type [1, 2] (average recognition rate is 84.28%).

	A	D	F	H	N	SA	SU
A	**73.33**	10	3.33	0	0	13.34	0
D	6.67	**80**	0	3.33	0	13.34	0
F	0	0	**83.33**	6.67	0	3.33	6.67
H	0	0	3.33	**93.33**	0	0	3.33
N	0	0	0	0	**100**	0	0
SA	3.33	23.33	6.67	0	6.67	**60**	0
SU	0	0	0	0	0	0	**100**

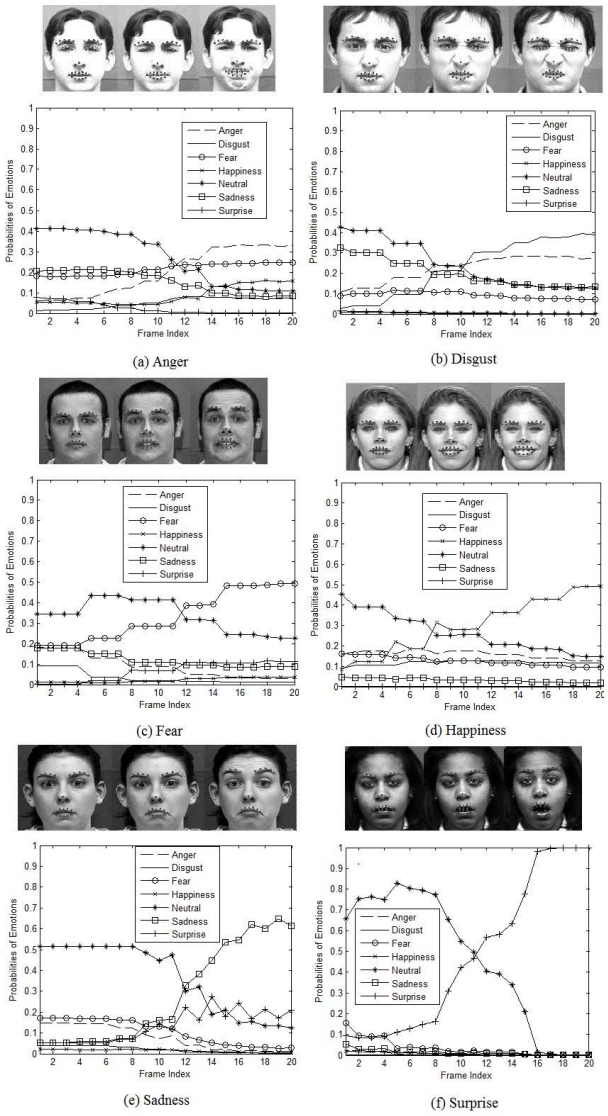

Fig. 1. Recognition rates (probabilities) of emotion types as a function of frame index (time) in a video sequence.

4. INFERENCE TO MENTAL STATES

To assess the plausibility of the proposed system to infer mental states, we compare its performance results with the human's by analyzing the same dataset. Affect [19] refers to the emotion elicited by stimuli and objectively reflects a person's mental state. It has been generally classified in the field of psychology using positive and negative dimensions [20]. Based on the basic facial emotion types, we classify Happiness as the positive affect and Anger, Disgust, Fear, and Sadness as the negative affect.

This experiment consists of two steps. In the first step, 24 professional actors (mean age = 29.167, SD = 9.907, 11 females) are selected as target participants to express five prototype emotion types elicited by stimuli from real-life activities, such as conversations and events. They are Anger, Disgust, Fear, Happiness, and Sadness. Since Surprise can be revealed as a result of positive or negative affect, it is not used for measuring affect in this experiment. Their frontal-view facial emotions are extracted from video clips of the TV series in which the actors perform. In this novel paradigm, subjects express their facial emotions triggered by real environment or events without bias. A total of 120 video clips with 640x352 pixel resolution are extracted, each of which starts and ends with a single emotion type per person. The auditory information is filtered and the motions of the mouth due to speaking or chewing are isolated.

In the subsequent step, both the human perceivers and the proposed system separately judge target participants' affects by watching these video clips. 15 people (mean age = 24.933, SD = 5.483, 1 female) served as perceivers. Both the perceivers and the system are assigned to watch 120 videos clips. After watching a video clip, the perceivers evaluate the type of affect (positive or negative) and the degree of the affect using a 21-point scale ranging from (-10 = extremely negative, +10 = extremely positive) [9]. Analogously, the system provides its affect rating in the range between -1 to +1. To calculate the similarity between ratings of the system and the perceivers, the perceivers' ratings are also scaled between -1 to +1. Both the perceivers and the system do not know the emotion type (affect) labeled on the video clips. The resulting advantage is that they naturally give the final ratings in the absence of prior knowledge.

4.1. Evaluation of video clips

We asked all perceivers to independently evaluate all video clips after watching them. Each video clip is classified as an explicit video if the number of perceivers inferring the same facial affect as that labeled on the video clip is over 50% of all perceivers. Otherwise, it is classified as an ambiguous video clip. A total of 13 of the 120 videos were classified as ambiguous videos. The remaining 107 explicit video clips were utilized for the experiments. Some samples are shown in Fig. 2.

Fig. 2. Samples of spontaneous facial expressions.

4.2. Inference of automatic system and human perceivers

Inference of targets' mental states by the system is based on facial emotion recognition. The system recognizes emotion type of a video clip v by calculating the largest probability $P_v^{Emotion}$ among Anger, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise expressed as

$$P_v^{Emotion} = \max(P_i), \quad v = 1,...,N_{video},\ i = 1,..,7 \qquad (2)$$

where N_{video} is the total number of videos. P_i is the probability of an emotion type, that is obtained by summing the overall probabilities p_{ij} of this emotion type in each frame of a video clip. p_{ij} is normalized to one. Hence, P_i is calculated as follows

$$P_i = \sum_{j=1}^{N_{frame}} p_{ij},\ \ subject\ to\ \sum_{i=1}^{7} P_{ij} = 1 \qquad (3)$$

where N_{frame} is the number of fames of a test video.

The ratings of both the positive affect P_{pos} and the negative affect P_{neg} for a video are separately calculated by the system as

$$P_v^{Affect} = \max(P_{pos},\ P_{neg}),\ v = 1,...,N_{video},$$
$$\text{subject to} \begin{cases} P_{pos} = \dfrac{P_{Happiness}}{P_{Happiness} + \max(P_{Anger}, P_{Disgust}, P_{Fear}, P_{Sadness})} \\ P_{neg} = \dfrac{\max(P_{Anger}, P_{Disgust}, P_{Fear}, P_{Sadness})}{P_{Happiness} + \max(P_{Anger}, P_{Disgust}, P_{Fear}, P_{Sadness})} \end{cases} \qquad (4)$$

where the larger affect rating P_v^{Affect} of the two values is selected as the recognized affect of the corresponding video clip.

Affect recognition accuracy of the system is measured by comparing its results with the actual labels of test video clips. The emotion recognition rate of the system is 76.64% and the affect recognition rate reaches to a higher value of 85.98%. This is mostly due to the fact that some emotion types are misrecognized although they may be affiliated with the same affect type. The perceivers' inferences are given in Table 5. $P_i,\ i = 1,...,15$ represents the i^{th} perceiver. The perceivers' average accuracy for affect recognition is 90.47%.

Table 5. Accuracy (in %) of human perceivers' inferences

P1	P2	P3	P4	P5	P6	P7	P8
86.92	96.26	96.26	87.85	92.52	94.39	81.31	82.24
P9	**P10**	**P11**	**P12**	**P13**	**P14**	**P15**	
92.52	92.52	91.59	79.44	94.39	94.39	94.39	

4.3. Similarity of the system and the human perceiver

Similarity scores are calculated by the percentages of affect recognitions by the system the same as the perceivers' in all test videos. For this purpose, we use the one-sample t-test that assesses whether the sample mean is statistically different from the population mean [21]. The similarity score for positive affect (mean = 97.39%, SD = 5.14, t(14) =35.68, p<0.001) is higher than that for negative affect (mean = 76.90%, SD = 6.84, t(14) = 15.24, p<0.001), where p is the probability of obtaining a test statistic close to the one that was observed, given that the null hypothesis is true. It is in line with the fact that the system is more sensitive to Happiness emotion. We also calculate the similarity score for the overall affects (mean = 81.31%, SD = 5.22, t(14) = 23.24, p<0.001). The results of one-sample t-test show that similarity scores of inferences between the system and perceivers are significantly above 50% for both affects. The proposed system is able to perform as the human does to infer people's mental states from their spontaneous facial expressions estimated from a video sequence.

To check the reliability of our results, we examine whether some targets' spontaneous facial expressions lead to predict similarity of inference between the system and perceivers. For each emotion type, we calculate the correlation between (a) recognition rates of the system for each target, and (b) similarity scores for each target. Fisher's z transform is used to evaluate for sample correlations in the range of -1 and +1 to a near Gaussian population [9, 22]. The correlations for each emotion type are close to zero (r_{anger} = -0.23, $r_{disgust}$ = -0.15, r_{fear} = 0.02, $r_{happiness}$ = -0.18, $r_{sadness}$ = 0.08, mean = -0.10, SD = 0.16, t(4) = 1.38, p>0.1). The results show that the correlations for the emotion types are not significantly higher or lower than zero. It explains that the human-like ability of the system to infer people's mental states is quite independent of the recognized spontaneous emotion types.

5. CONCLUSIONS

We proposed an automatic facial emotion recognition system that employs RHMMs for states of face regions instead of using a traditional HMM to represent a single emotion type for the entire face. The results showed that the proposed method outperforms the traditional HMM method for the person-independent case. The performance was also superior to the state-of-the-art in the literature. More importantly, the results verify that the proposed system has the capability to infer mental states reported similar with human-to-human interactions. To the best of our knowledge, this is the first attempt to explore the psychological plausibility of the automatic system to analyze spontaneous facial expressions registered in video sequences. The authors are currently studying the implementation of such a system for various applications.

6. ACKNOWLEDGEMENT

The authors would like to thank Prof. A. Todorov and Prof. D.N. Osherson of Princeton University for their help to design human based evaluations in order to validate the performance of the proposed system.

7. REFERENCES

[1] Z. Liu and S. Wang, "Emotion recognition using Hidden Markov Models from facial temperature sequence," *Affective Computing and Intelligent Interaction, Springer-Verlag, Berlin Heidelberg,* pp. 240-247, 2011.

[2] Y. Sun, M. Reale, and L. Yin, "Recognizing partial facial action units based on 3D dynamic range data for facial expression recognition," *8th IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 17-19, 2008.

[3] R. El Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expression and head gestures," *in Proc. IEEE International Conf. Computer Vision and Pattern Recognition*, vol. 3, pp. 154, 2004.

[4] Z. Zeng , M. Pantic, G. I. Roisman and T. S. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39-58, 2009.

[5] Z. Ambadar, J. Schooler, J. Cohn, "Deciphering the enigmatic face: The importance of facial dynamics to interpreting subtle facial expressions," *Psychological Science*, vol. 16, no. 5, pp. 403-410, 2005.

[6] C. Padgett, G. Cottrell, "Identifying emotion in static face images," *Proc. of the 2nd Joint Symposium on Neural Computation*, pp. 91-101, 1995.

[7] M. Valstar, M. Pantic, Z. Ambadar, and J.F. Cohn, "Spontaneous versus posed facial behavior: automatic analysis of brow actions," *in Proc. Eight Int'l Conf. Multimodal Interfaces*, pp. 162-170, 2008.

[8] M.F. Valstar, H. Gunes, and M. Pantic, "How to distinguish posed from spontaneous smiles using geometric Features," *in Proc. Ninth ACM Int'l Conf. Multimodal Interfaces*, pp. 38-45, 2007.

[9] M.S. North, A. Todorov, and D.N. Osherson, "Inferring the preferences of others from spontaneous, low-emotional facial expressions," *Journal of Experimental Social Psychology*, vol. 46, no.6, pp. 1109-1113, 2010.

[10] J. Saragih, S. Lucey, and J. Cohn, "Deformable model fitting by regularized landmark mean-shifts," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200-215, 2011.

[11] T. Cootes, G. Edwards and C. Taylor, "Active appearance models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681-685, 2001.

[12] Y. Wang, S. Lucey, and J. Cohn, "Enforcing convexity for improved alignment with constrained local models," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2008.

[13] L. Gu and T. Kanade, "A generative shape regularization model for robust face alignment," *European Conference on Computer Vision*, pp. 413-426, 2008.

[14] P. Ekman and W.V. Friesen, *Emotion in the Human Face. Englewood Cliffs*, NJ: Prentice-Hall, 1975.

[15] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of IEEE*, vol. 77, no. 2, pp. 257-286, 1989.

[16] T. Kanade, J.F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," *in Proc. Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 46-53, 2000.

[17] L.A. Jeni, J.M. Girard, J.F. Cohn, and F. De La Torre, "Continuous AU intensity estimation using localized, sparse facial feature space," *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space*, pp. 1-7, 2013.

[18] S. Jain, C. Hu, J.K. Aggarwal, "Facial expression recognition with temporal modeling of shapes*," IEEE Int. Conf. on Computer Vision Workshops, Barcelona, Spain*, pp. 1642-1649, 2011.

[19] P.T. Trzepacz and R.W. Baker, *The psychiatric mental status examination*, Oxford University Press, 1993.

[20] D. Watson, L. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the PANAS scales," *Journal of Personality and Social Psychology*, pp.1063-1070, 1988.

[21] R.J. Freund, *Regression Analysis: Statistical Modeling of a Response Variable*, Academic Press, 1998.

[22] R.A. Fisher, "On the 'Probable Error' of a coefficient of correlation deduced from a small sample," *Metron*, vol. 1, pp. 3-32, 1921.