Facial expression recognition with regional hidden Markov models

Y. Sun and A.N. Akansu

A regional hidden Markov model (RHMM) for automatic facial expression recognition in video sequences is proposed. Facial action units are described by RHMMs for the states of facial regions: eyebrows, eyes and mouth registered in a video. The tracked facial feature points in the spatial domain form observation sequences that drive the classification process. It is shown that the proposed technique outperforms other methods reported in the literature for the person-independent case tested with the extended Cohn-Kanade database.

Introduction: Classification is the most significant part of a facial expression recognition system. It provides classified features for recognition. Classification methods are generally grouped as static or dynamic ones [1]. Static methods, taking advantage of the support vector machine, neural network, Bayesian network etc., are based on the information obtained from the input image. Dynamic classifiers like the hidden Markov model (HMM) utilise temporal information to analyse facial expressions, and are strongly suggested by the psychological experiments carried out as reported in [2]. However, the researches employing the temporal information classify the entire image sequence into one of the expression types. For practical application, a facial expression system must be able to classify frames as they come into analysis. Moreover, most prior research reported in the literature applies the emotion-specific HMMs to train and test a single emotion type for the entire face image [3]. The information describing a facial emotion is mainly registered in the movements of facial regions [4].

In this Letter, a regional HMM (RHMM)-based facial expression recognition system is proposed. It emphasises the movements of three facial regions, rather than modelling emotion types for the entire face. The proposed RHMM method is also applicable to video sequences as expected. A set of experiments with the extended Cohn-Kanade (CK+) to validate the merit of the proposed method was performed [5]. A recognition rate of 90.9% was achieved for the person-independent case. The details of the system are given in the following Sections.

Feature extraction: To describe the motions (movements) of facial regions, 41 facial feature points were identified in each video frame, as shown in Fig. 1. They comprised 10 points on the eyebrow region, 12 points on the eyelids, eight points on the mouth, 10 points on the corners of the lips and one anchor feature point on the nose. The facial feature points are tracked using a constrained local model [6]. It utilises a non-parametric method to represent the distribution of candidate locations using an optimisation method called constrained mean-shifts. It outperforms the well-known methods for deformable model fitting.

Fig. 1 *Forty-one facial feature points on face image*

Regional hidden Markov model: To emphasise the importance of facial regions and their dynamic spatio-temporal information, 14 different RHMMs, labelled from λ_1 to λ_{14}, were generated, each representing one of the 14 action units (AUs) of the three facial regions, as in Table 1. AUs are the actions of individual muscles or a group of muscles on the face, which are utilised by facial action coding systems that describe facial expressions [7]. According to the tracked feature points in the corresponding facial regions, we code the prototypical facial emotions using various AUs, which is in line with the AU-coded emotion description given by Ekman in [7].

Table 1: Facial regions and corresponding AUs [7]

Facial regions	Action units
Eyebrows	1, 2, 4
Eyes	5, 7
Mouth	10, 12, 15, 16, 20, 23, 24, 25, 27

Formation of observation sequences: The observation sequences are formed using the coordinates of the facial feature points in a two-dimensional representation. In the proposed method, the images in the database are classified in a unique order to train RHMMs for different AUs of facial regions. For instance, AU 1 describes 'inner brow raiser'. To train RHMMs for AU 1 of the eyebrow region, we classify images following its states as inner brow raiser, inner brow lowerer and neutral. In the same way, we classify other AUs of the eyebrows, eyes and mouth regions as shown in Table 1. The advantage of this manipulation is to process various forms of videos, and it avoids defining specific forms of videos, such as starting from the neutral emotion to the peak and then fading back to the neutral emotion as described in [3]. It also brings additional benefits to the model selection step. We regard the left-to-right model as the optimal one for the training and recognition RHMM. It not only uses fewer parameters for training than the ergodic model, but can also model a natural sequential event always starting and ending in the two fixed states.

To reflect the motions of facial regions, the relative variations of feature points are calculated. The procedure to form observation sequences is the same for each AU of all regions. To illustrate this function, AU 1 of the eyebrow region is used in this example.

1. Calculate the relative distance d_i, $i = 1, \ldots, N_{\text{eyebrows}}$ between each feature point in the eyebrow region and the anchor point on the nose for each frame of a video in the database. N_{eyebrows} is the number of facial feature points in the eyebrow region.
2. Then, form

$$\begin{aligned} S_{\text{eyebrows},n} &= \{d_1, \ldots, d_{N_{\text{eyebrows}}}\} \\ S^{\text{AU1}}_{\text{eyebrows},m} &= \{d_1, \ldots, d_{N_{\text{eyebrows}}}\} \\ m &= 1, \ldots, N^{\text{AU1}}_{\text{frame}} \end{aligned} \quad n = 1, \ldots, N_{\text{frame}}, \tag{1}$$

where $S_{\text{eyebrows},n}$ represents a set of relative distances in the eyebrow region of the nth frame in the database N_{frame}. $S^{\text{AU1}}_{\text{eyebrows},m}$ represents another set of relative distances for AU 1 of the eyebrow region in the mth frame of $N^{\text{AU1}}_{\text{frame}}$. $N^{\text{AU1}}_{\text{frame}}$ is the number of frames showing AU 1.
3. Project $S^{\text{AU1}}_{\text{eyebrows},m}$ onto $S_{\text{eyebrows},n}$ to measure the distance ε_j, $j = 1, \ldots, N_{\text{frame}}$. In some cases, ε_j may be small, even zero if $S^{\text{AU1}}_{\text{eyebrows},m}$ is projected onto the same state of the eyebrow region.
4. The feature vector for the mth frame of $N^{\text{AU1}}_{\text{frame}}$ is formed as

$$\boldsymbol{V}^{\text{AU1}}_{\text{eyebrows},m} = [\varepsilon_1, \ldots, \varepsilon_n, \ldots, \varepsilon_{N_{\text{frame}}}] \tag{2}$$

5. Group the vectors of AU 1 for the eyebrow region to obtain the training matrix as

$$\boldsymbol{Tr}^{\text{AU1}}_{\text{eyebrows}} = \left[V^{\text{AU1}}_{\text{eyebrows},1}, \ldots, V^{\text{AU1}}_{\text{eyebrows},m}, \ldots, V_{\text{eyebrows},N^{\text{AU1}}_{\text{frame}}}\right]^{\text{T}} \tag{3}$$

Experimental results: The performance of the proposed automatic facial expression recognition system was tested for the widely used extended Cohn-Kanade database (CK+) in the literature [5] to recognise the seven universal facial emotions including neutral. We conducted the person-independent experiment to classify facial features using the proposed RHMM framework for the AUs of facial regions, in which the leave-one-subject-out cross-validation method is used. To emphasise the temporal variations of facial regions, we consider the expression onset and peak for all sequences instead of the peak expression in the last frame [8].

In the training step, the Baum-Welch algorithm [9] is used to re-estimate the initial parameters of each RHMM. During testing, each

frame of the test videos goes through the same process to form the observation sequences O_i corresponding to different facial regions. Here, i is one of the three facial regions. The probability of the observation sequence for the related RHMM $P(O_i|\lambda_j)$ is calculated using the forward–backward algorithm [9], where j is the AU of a facial region. A prototypical emotion is expressed by combining AUs of different facial regions. Therefore we calculate the probability of an emotion type P_k by summing the probabilities of AUs of different facial regions as

$$P_k = \sum P(O_i|\lambda_j) \tag{4}$$

where k is one of the seven emotion types. A video frame is recognised with the facial emotion type that yields the highest measured probability, $\max(P_k)$.

Fig. 2 displays the recognition rates for emotion types against frames (time) in various test sequences. The performance of the proposed method is better than the prior work. The experimental results are listed in Table 2. The system achieved the highest accuracy of 100% in the emotion types happiness and surprise. The average accuracy is 90.9%, which is significantly higher than the baseline results reported in [5]. For comparison, we also conducted experiments using the traditional emotion-specific HMM for the entire face image [3]. Its average accuracy is 85.8%. The recognition rates reported in [8, 10] are 86 and 85.84%, respectively.

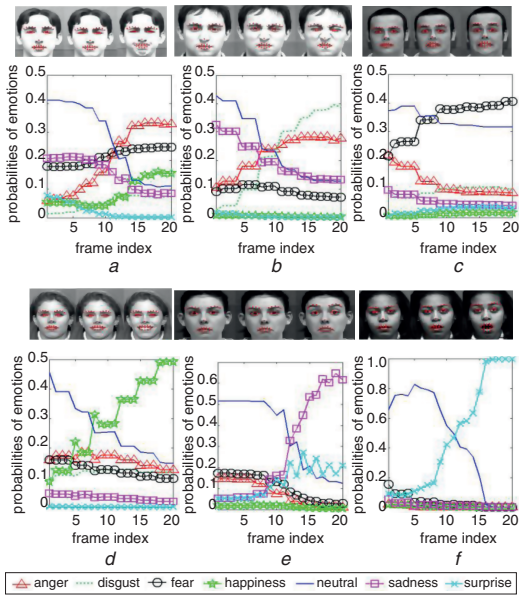

Fig. 2 *Probabilities of emotion types against frame index*

a Anger
b Disgust
c Fear
d Happiness
e Sadness
f Surprise

Table 2: Recognition performance (%) of proposed method

	A	D	F	H	N	SA	SU
A	**85.6**	10	0	0	0	4.4	0
D	8.8	**86.6**	0	3.4	0	1.2	0
F	0	0	**90.1**	0	0	0	9.9
H	0	0	0	**100**	0	0	0
N	0	0	0	0	**97.9**	2.1	0
SA	3.3	6.7	10	0	3.3	**76.7**	0
SU	0	0	0	0	0	0	**100**

Table 3: Fusion of facial regions: upper region and lower region

Facial regions	Action units
Upper region	1 + 4, 4 + 7, 1 + 4 + 5 + 7, 1 + 2 + 4 + 5, 1 + 2 + 5, 1 + 4 + 7
Lower region	10, 12, 15, 16, 20, 23, 24, 25, 27

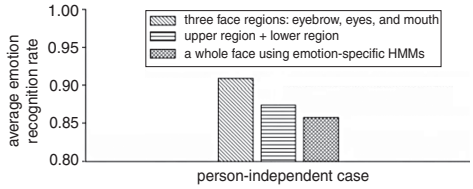

Fig. 3 *Performance of fusion of facial regions*

We performed additional experiments to determine the importance of facial regions for facial expression recognition. We combined eyebrows and eyes as the upper region in order to do this. Therefore the observation sequences for the upper region are formed by the proposed method, utilising the facial features of eyebrows and eyes. The corresponding combined AUs are shown in Table 3 [7]. The observation sequences for the lower region have no changes in this manner. The experimental results displayed in Fig. 3 show that the performance of the proposed method is superior to the other methods also combining facial regions. It is observed from the results that the recognition rate lowered with the fusion of facial regions.

Conclusion: In this reported work, we employed RHMMs to model the spatiotemporal dynamics of facial regions rather than using a single HMM to represent the entire face. Experimental results showed that the proposed method outperforms other techniques reported in the literature for the person-independent case.

17 February 2014
doi: 10.1049/el.2014.0441
One or more of the Figures in this Letter are available in colour online.

Y. Sun and A.N. Akansu (*Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ, USA*)

E-mail: yanjia.sun@njit.edu

References

1 Zeng, Z., Pantic, M., Roisman, G., and Huang, T.S.: 'A survey of affect recognition methods: audio, visual, and spontaneous expressions', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009, **31**, (1), pp. 39–58
2 Ambadar, Z., Schooler, J., and Cohn, J.: 'Deciphering the enigmatic face: the importance of facial dynamics to interpreting subtle facial expressions', *Psychological Sci.*, 2005, **16**, (5), pp. 403–410
3 Sun, Y., Reale, M., and Yin, L.: 'Recognizing partial facial action units based on 3D dynamic range data for facial expression recognition'. 8th IEEE Int. Conf. Automatic Face and Gesture Recognition, Amsterdam, The Netherlands, September, 2008, pp. 17–19
4 Padgett, C., and Cottrell, G.: 'Identifying emotion in static face images'. Proc. 2nd Joint Symp. Neural Computation, San Diego, CA, USA, June 1995, pp. 91–101
5 Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I.: 'The extended Cohn-Kanade dataset (ck +): a complete dataset for action unit and emotion-specified expression'. IEEE Conf. Computer Vision and Pattern Recognition Workshops, San Francisco, CA, USA, June 2010, pp. 94–101
6 Saragih, J., Lucey, S., and Cohn, J.: 'Deformable model fitting by regularized landmark mean-shifts', *Int. J. Comput. Vis.*, 2010, **91**, (2), pp. 200–215
7 Ekman, P.: 'Emotion in the human face' (Cambridge University Press, Cambridge, UK, 1982)
8 Jeni, L.A., Girard, J.M., Cohn, J.F., and De La Torre, F.: 'Continuous AU intensity estimation using localized, sparse facial feature space'. 2nd Int. Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space, Shanghai, China, April 2013, pp. 1–7
9 Rabiner, L.R.: 'A tutorial on hidden Markov models and selected applications in speech recognition', *Proc. IEEE*, 1989, **77**, (2), pp. 257–286
10 Jain, S., Hu, C., and Aggarwal, J.K.: 'Facial expression recognition with temporal modeling of shapes'. IEEE Int. Conf. Computer Vision Workshops, Barcelona, Spain, November 2011, pp. 1642–1649