International Journal on Smart Material and Mechatronics
IJSMM Vol. 4 No. 1 2017 DOI. 10.20342/IJSMM.4.1.269

p-ISSN: 2356-5314
e-ISSN: 2460-075X

Design and Implementation of HMM for 3D Emotion Recognition

Yu Lang
Division of Mechanical and Systems Engineering
Okayama University
Okayama, Japan
lang@usmm.sys.okayama-u.ac.jp

Maierdan Maimaitimin and Keigo Watanabe
Department of Intelligent Mechanical Systems
Okayama University
Okayama, Japan
merdan-m@usmm.sys.okayama-u.ac.jp
watanabe@sys.okayama-u.ac.jp

Received: 14 October 2016 / Revised: 29 November 2016 / Accepted: 29 January 2017 / Published online: 20 April 2017 © IJSMM2017

***Abstract*—Facial expression is one of the most useful information in human robot interaction. To improve the accuracy in 3-dimension based facial expression recognition, Hidden Markov Models (HMMs) are used to recognize the emotion from facial expressions in this study. In particular, facial expressions are measured by two parameters, which are given by previous work. The human emotions are defined as: anger, smile, normal, sadness, fear, and surprise. The referred parts in human face are selected based on the activeness during the facial expression. The activity and arousal values of each facial part are used as the observations for each hidden state in HMMs. Baum-Welch algorithm is used to train the hidden Markov model. As a result, six different emotions are very efficiently recognized through the trained HMMs.**

***Index Terms*— Hidden Markov model, emotion recognition, facial expression**

I. INTRODUCTION

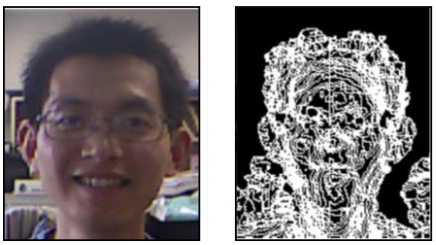
Facial expression is one of the most powerful, natural, and immediate means for human beings to communicate their emotions and purposes [1]. To understand human emotion is an essential issue for intelligent robots and systems, to obtain an effective communication with human in providing the desirable services.

In the past research of facial expression recognition, Ekman and Friesen [2] developed a Facial Action Coding System (FACS) to describe facial expressions. They separated the facial expression as an upper face and a lower face. Ying-li and her coworkers [1] used a neural network to make another feature-based automatic face analysis system due to Ekman's study. Furthermore, Hong-Wei Ng [5] used a deep Convolutional Neural Network on facial expression recognition. These studies have found that neural networks are effective in emotion expression recognition. However, most of above systems are used to predict a facial expression based on the single frame of image. Instead, emotions are more often shown as a time sequence of facial expressions. It is important to develop an emotion recognition model based on a time sequence database to increase the system accuracy.

This paper follows from the previous work [4], in which the Kinect sensor is used to capture human facial expression data. The activity and arousal [3] values of each facial part are given by [4], as the observation in Hidden Markov Models (HMMs). As the learning method, Baum-Welch algorithm is used as the training approach in HMM.

II. RELATED WORK

A. Surface Common Feature



RGB Image SFG Image

Fig. 1. Comparison of surface common feature image and RGB image

In machine learning processing, a logical, meaningful data is very important. A 3D point cloud data only has 3D spatial position information, which means the point cloud data do not have too much meaning for machine learning. In order to mine the valid data, a method named surface-common feature (SCF) map [3] is used to reveal the point cloud data. This method is based on the surface normal vectors. The SCF image is shown in Fig. 1. This method will be more practical than other approaches with inadequate illumination.

B. Autoencoder

An autoencoder is a wildly used machine learning approach. It can obtain the initial parameters of a neuron by inputting the feature itself. Suppose that there is an input sequence $\{x_i\}_{i=1}^{n}$, and map them with an encoder to a hidden representation $\{y_i\}_{i=1}^{m}$, e.g.

$$y = s(\omega x + b) \tag{1}$$

p-ISSN: 2356-5314
e-ISSN: 2460-075X

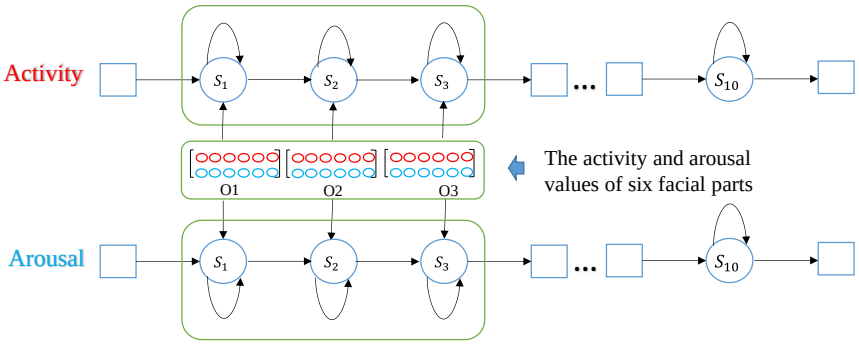


Fig. 2. Structure of HMMs

where s is a nonlinear activation function (such as sigmoid function).

After that, y is mapped onto the reconstruction z of the same shape as x :

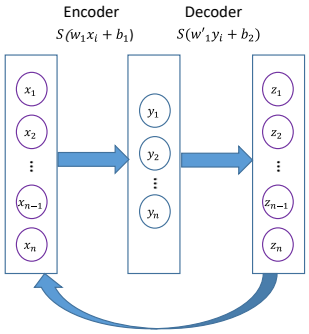


Fig. 3. Concept of autoencoder

$$z = s(\omega' y + b) \tag{2}$$

Autoencoder is also trained to minimize the reconstruction errors (such as squared errors).

$$L_H(x,z) = -\sum_{k=1}^{d}[x_k \log z_k + (1-x_k)\log(1-z_k)] \tag{3}$$

In this research, we used an autoencoder as a pre-training method for the deep neural network.

C. Convolutional Neural Network

As a famous model on deep learning, a convolutional neural network (CNN) has a complicated structure, which will have at least three different kinds of layers:

a) Convolutional layer: The convolutional layer is the core building block of a CNN which consists of a grid of neurons.

b) Pooling layer: The pooling layer is the another important concept of the CNN, which is a form of nonlinear down-sampling.

c) Fully-connected layer: The fully-connected layer always appears after several convolutional and pooling layers, and the high-level reasoning in the neural network is performed via fully connected layers.

For instance, there is a $W \times W$ pixels image, using an $H \times H$ pixels filter to convolute it:

$$u_{ij} = \sum_{p=0}^{H-1}\sum_{q=0}^{H-1} x_{i+p,j+p} h_{pq} \tag{4}$$

We extract the activity and arousal parameters by a convolution neural network from surface common feature maps. Then, it uses the two parameters as observations of Hidden Markov Models.

III. Current Methodology

A. Data Obtaining

The processing is calculated as follows:

In previous work, the Kinect sensor was used to capture the point cloud data of facial expressions. Common feature maps are extracted from each facial expression. Seven referred parts in human face are selected based on the activeness during the facial expression, such as left eye, right eye, nose, upper lip, under lip and cheek. Filters of each part are pre-trained through an auto-encoder.

Then, the CNN is used to obtain their activity and arousal parameters from feature maps. The activity and arousal parameters are defined as observations of HMMs.

Considering the resolution of the data, the fixed distance D between the camera and the human face is set to 90 cm. Facial

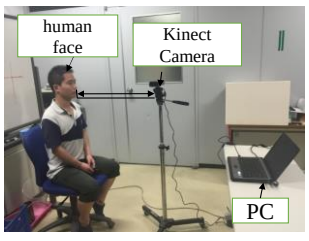


Fig. 4. Experimental environment

parts are segmented, such as nose, left eye, right eye, left cheek, right cheek, left mouth corner, and right mouth corner, as in total 7 parts, so that the capture range is defined as 32 × 26 pixel.

p-ISSN: 2356-5314
e-ISSN: 2460-075X

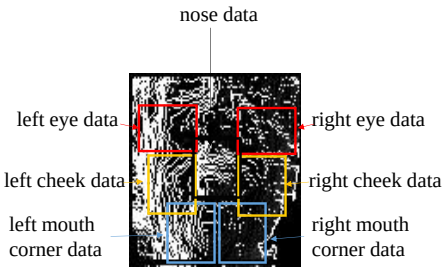


Fig. 5. Segmented facial data

Six different facial expressions {Anger, Happy, Fear, Sadness, Normal, Surprise} are captured for 23 persons, in which each facial part has 230 frames for each emotion, and results in 6 × 6 × 230 groups in total. Those facial expression data are fed into the CNN to obtain emotional activity and arousal values. The captured data are segmented as Fig. 5.

B. Hidden Markov Model

In this paper, the face expression based emotion recognition task is formulated as a sequence recognition problem.

The HMMs are set as two Markov chains: one uses activity values as observations and the other uses arousal values. The whole architecture of the HMMs is showed in Fig. 2.

The necessary parameters in HMMs are shown as follows:

$$\lambda = (A,\ B,\ \pi) \tag{5}$$

- A: State transition probability matrix

$$A = [a_{ij}]_{N \times N} \tag{6}$$

where a_{ij} is the transition probability from the state q_i to q_j,

- B: Observation probability distribution, and
- π: Initial state distribution.

We treat the whole dataset as a concatenated timing sequence of frames. The frame sequence is represented as:

$$O_i = \{O_{i1}, O_{i2}, \ldots, O_{iT}\},\ \mathrm{i} \in \{1,2,\ldots,12\} \tag{7}$$

where O_i corresponds to the feature of the ith frame. The length of the sequence is T = 10.

Define $Y = \{y_1, y_2, y_3, \ldots, y_6\}$ as the label of emotions, where y_i is the ith data's label.

The emotional activity and arousal values of facial expression are treated as being generated sequentially from a Markov process that trainsits between states $S = \{s_1, s_2, s_3, \ldots, s_6\}$.

C. Training Approach

The Baum-Welch algorithm is used to infer unknown parameters of HMMs. The training procedure is shown as follows.

1) In the first place, the emotional activity values of each facial part are combined to one value by Bayesian network. The emotional arousal values are managed with the same way.

anger	Activity		Arousal	
Facial parts	Plus	Minus	Plus	Minus
Left eye	0.2	0.8	0.3	0.7
Right eye	0.2	0.8	0.3	0.7
Left cheek	0.5	0.5	0.6	0.4
Right cheek	0.5	0.5	0.6	0.4
Left mouth corner	0.3	0.7	0.3	0.7
Right mouth corner	0.3	0.7	0.3	0.7

Fear	Activity		Arousal	
Facial parts	Plus	Minus	Plus	Minus
Left eye	0.3	0.7	0.6	0.4
Right eye	0.3	0.7	0.6	0.4
Left cheek	0.5	0.5	0.5	0.5
Right cheek	0.5	0.5	0.5	0.5
Left mouth corner	0.4	0.6	0.4	0.6
Right mouth corner	0.4	0.6	0.4	0.6

normal	Activity		Arousal	
Facial parts	Plus	Minus	Plus	Minus
Left eye	0.5	0.5	0.5	0.5
Right eye	0.5	0.5	0.5	0.5
Left cheek	0.4	0.6	0.3	0.7
Right cheek	0.4	0.6	0.3	0.7
Left mouth corner	0.4	0.6	0.5	0.5
Right mouth corner	0.4	0.6	0.5	0.5

Sadness	Activity		Arousal	
Facial parts	Plus	Minus	Plus	Minus
Left eye	0.3	0.7	0.3	0.7
Right eye	0.3	0.7	0.3	0.7
Left cheek	0.5	0.5	0.5	0.5
Right cheek	0.5	0.5	0.5	0.5
Left mouth corner	0.4	0.6	0.4	0.6
Right mouth corner	0.4	0.6	0.4	0.6

Surprise	Activity		Arousal	
Facial parts	Plus	Minus	Plus	Minus
Left eye	0.6	0.4	0.7	0.3
Right eye	0.6	0.4	0.7	0.3
Left cheek	0.5	0.5	0.5	0.5
Right cheek	0.5	0.5	0.5	0.5
Left mouth corner	0.6	0.4	0.7	0.3
Right mouth corner	0.6	0.4	0.7	0.3

happy	Activity		Arousal	
Facial parts	Plus	Minus	Plus	Minus
Left eye	0.8	0.2	0.7	0.3
Right eye	0.8	0.2	0.7	0.3
Left cheek	0.5	0.5	0.6	0.4
Right cheek	0.5	0.5	0.6	0.4
Left mouth corner	0.6	0.4	0.7	0.3
Right mouth corner	0.6	0.4	0.7	0.3

Fig. 6. The probability to combine emotional values.

2) Then, 10 sequential emotional activity values and 10 sequential emotional arousal values are fed into two Markov chain as observations.

3) Calculate the temporary variables according to Bayes' theorem. Given the observed sequence Y and the parameters θ, the probability of being in state i at time t is obtained by

$$\gamma_i(t) = P(X_t = i|Y,\theta) = \frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^{N}\alpha_j(t)\beta_j(t)} \tag{8}$$

where α_i represents a message from state i-1 to state i and β_i represents a message from i+1 to i. Given the observed sequence Y and the parameters θ, the probability of being in states i and j at times t and t+1 respectively is

$$\xi_{ij}(t)=P(X_t=i,X_{t+1}=j|Y,\theta)=\frac{\alpha_i(t)a_{ij}\beta_j(t+1)b_j(y_{t+1})}{\sum_{k=1}^{N}\alpha_k(T)}$$

(9)

4) Update Initial State Probability: It is an expected frequency spent in state *i* at time t_1,

$$\pi_i=\gamma_i(1) \tag{10}$$

5) Update Transition Matrix: It is the expected number of transitions from state *i* to state *j*, divided by the expected total number of transitions away from state *i*, which is given by

$$a_{ij}=\frac{\sum_{t=1}^{T-1}\xi_{ij}(t)}{\sum_{t=1}^{T-1}\gamma_i(t)} \tag{11}$$

6) Update Emission Matrix: It is the expected number of times that the output observations have been equal to v_k are divided by the expected total number of times in state *i*,

$$b_i(v_k)=\frac{\sum_{t=1}^{T}1_{y_t=v_k}\gamma_i(t)}{\sum_{t=1}^{T}\gamma_i(t)} \tag{12}$$

where $v_k$is an observing symbol and

$$1_{y_t=v_k}=\begin{cases}1 & \text{if } y_t=v_k\\ 0 & \text{otherwise}\end{cases} \tag{13}$$

IV. Result

The trained HMMs reached a convergence region, it confirmed that the training is successful. The trained HMMs parameters are shown in Eqs. 14 and 15,

$$A=\begin{pmatrix}0.4 & 0 & 0.24 & 0 & 0 & 0.35\\ 0.15 & 0.3 & 0 & 0 & 0 & 0.53\\ 0.66 & 0 & 0 & 0 & 0 & 0.34\\ 0 & 0.64 & 0 & 0.36 & 0 & 0\\ 0 & 0.63 & 0 & 0 & 0.37 & 0\\ 0.37 & 0 & 0.12 & 0 & 0 & 0.51\end{pmatrix} \tag{14}$$

where *A* is a transition matrix and *B* is the emission matrix,

$$B=\begin{pmatrix}1 & 0 & 1 & 0 & 0 & 0.99\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 1 & 0 & 0\end{pmatrix} \tag{15}$$

V. Conclusion

In this paper, according to recognizing six different emotions, an emotion recognition system is designed based on Hidden Markov Models. Using Baum-Welch algorithm as the training method. As a result, the system can recognize the six emotions successfully, as found as {'Normal,' 'Happy,' 'Happy,' 'Happy,' 'Happy,' 'Happy,' 'Happy,' 'Happy,' 'Happy,' 'Happy'} in the sample of smile.

In future work, we will optimize the probability of observation which is used to combine six emotional values to one observation. And we have to complete database by increasing the amount of data.

References

[1] Yl. Tian, T. Kanade and JF. Cohn, "Recognizing action units for facial expression analysis," Pattern Analysis And Machine Intelligence, vol. 23, no. 2, pp.97–115, 2001.

[2] P. Ekman and W. V. Friesen, "The facial action coding system: A technique for the measurement of facial movement," San Francisco, USA: Consulting Psychologists Press, 1978.

[3] M. Maierdan, "Development of an emotion recognition system based on human behaviors," Okayama University, Master Thesis, 2013

[4] M. Maierdan, K. Watanabe and S. Maeyama, "Surface-common-feature descriptor of point cloud data for deep learning," in Proc. of International Conference on Mechatronics and Automation, 2016, pp. 525–529.

[5] HW. Ng, VD. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in Proc. of International Conference on Multimodal Interaction, ACM, 2015.

[6] Q. Guo, D. Tu, J. Lei, and G. Li, "Hybrid CNN-HMM model for street view house number recognition," in Proc. of Asian Conference on Computer Vision, 2014, pp. 303–315.

[7] M. Bartlett, P. Viola, T. Sejnowski, B. Golomb, J. Hager, P. Ekman, and J. Larsen. "Classifying facial action," in Advances in Neural Information Procession Systems, D. Touretzky, M. Mozer and M. Hasselmo, Eds. Cambridge: MIT Press, 1996, pp. 823–829.