Facial Analysis Aided Human Gesture Recognition for Human Computer Interaction

Dan Luo[*a,b], Hua Gao[b], Hazım Kemal Ekenel[b], Jun Ohya[a]
[a]Waseda University, 1-3-10 Nishi-Waseda Shinjuku-ku, Tokyo, Japan;
[b]Karlsruhe Institute of Technology, Adenauerring 2, Karlsruhe, Germany

Abstract

Human gesture recognition systems are natural to use for achieving intelligent Human Computer Interaction (HCI). These systems should memorize the specific user and enable the user to gesture naturally without wearing special devices. Extracting different components of visual actions from human gestures, such as shape and motion of hand, facial expression and torso, is the key tasks in gesture recognition. So far, in the field of gesture recognition, most of the previous work have focused only on hand motion features and required the user to wear special devices. In this paper, we present an appearance-based multimodal gesture recognition framework, which combines the different modalities of features such as face identity, facial expression and hand motions which have been extracted from the image frames captured directly by a web camera. We refer 12 classes of human gestures with facial expression including neutral, negative (e.g. "angry") and positive (e.g. "excited") meanings from American Sign Language. A condensation-based algorithm is adopted for classification. We collected a data set with three recording sessions and conducted experiments with different combination techniques. Experimental results showed that the performance of hand gesture recognition is improved by adding facial analysis.

1 Introduction

The recognition of natural human gestures from video sequences is very challenging problem with diverse applications in human computer interaction (HCI), especially in robotics. It is desirable that humans can use the natural gestures to give commands without operating any device such as a remote controller, or data glove. Due to different components of human gestures performed by signer using the hand, face, and torso, vision-based gesture recognition approaches need efficient detection and feature extraction. Many of the existing approaches use only hand feature to recognize gesture. Facial expressions also play a very important role in human gestures. Many manual gestures are ambiguous in isolation, and need to be accompanied by facial expressions in order to convey a specific sign. Moreover, facial expressions represent a continuous stream of supplementary information in gesture communication, offering clarity and sensitivity to the viewer.

In vision-based gesture recognition, capturing, tracking and segmentation problems occur, and it is hard to build a robust recognition framework. Most of the current systems use specialized hardware [1], or work on the simpler case of hand gesture recognition with small vocabularies [2]. For dealing with the hand motion, previous vision-based methods have demonstrated some successes using Hidden Markov Models (HMMs) [3, 4, 5] and conditional random fields (CRFs) [6]. Black and Jepson [9] extend the Condensation algorithm proposed by Isard and Blake proposed [7, 8] to recognize gestures and facial expressions in which human motions are modeled as temporal trajectories from image sequences without hand drawn templates. Our aim is to build an appearance-based multimodal gesture recognition framework, which combines the different modalities of features such as face identity, facial expression and hand motions which have been extracted from the image frames captured directly by a web camera. Condensation-based algorithm is applied for multimodal feature without training models such as models based on HMM.

In the following, Section 2 outlines the proposed multimodal construction and combination method. Section 3 explains the Condensation-based algorithm for recognizing multimodal features. Experimental results and discussions are presented in Section 4. Section 5 concludes this paper.

2 Multimodal Features

An integrated approach to human gesture recognition is required that combines the various visual cues available using specialized, complementary techniques, aiming to extract sufficient aggregate information for robust recognition. Our strategy to implement such an integrated system relies on extracting different modalities of features (hand motion features and facial expression features) and combination strategies.

2.1 Hand Feature

Following the face detection step, skin color is trained for extracting hand region. Face blob and hand blobs are found from each frame of the video sequence by using skin color segmentation based on skin and non-skin color Gaussian Mixture Model. Based on the color segmentation results, we use the centroids of the left and right hand blobs to generate hand motion trajectories over the whole video sequence. There is three cases of blobs overlapping: 1) Face and one hand blob overlapping, 2) Two hands overlapping, 3) Face and two hands overlapping. For each frame, the three biggest blobs are considered as face and hand blobs. Once one of the three overlapping cases occurs, these three blobs will connect to each other and k-means algorithm is used to find the centroid of the overlapped blobs. During such an overlapping case,

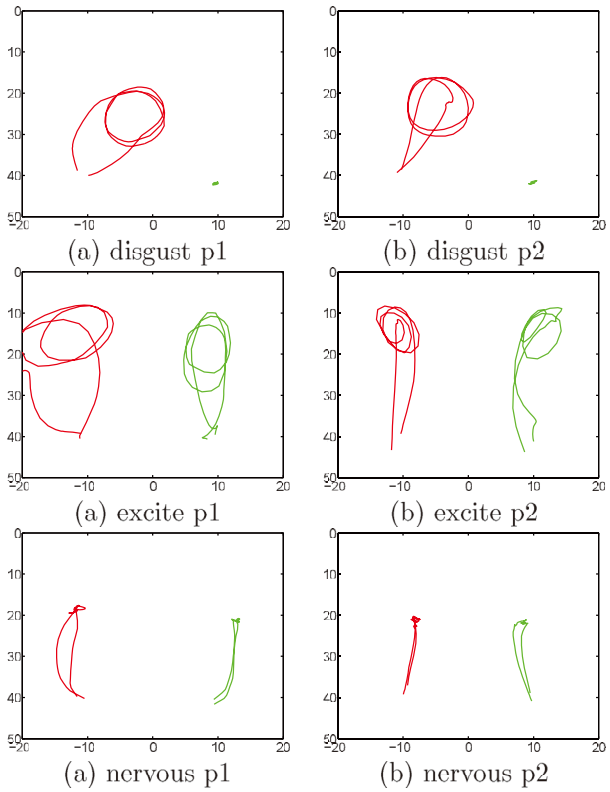

Figure 1. Samples of hand motion trajectories

the three blobs' trajectories cannot be obtained separately from each other, but we found that a low-pass filter based operation can obtain smoother and better prediction for the trajectories for the overlapped blobs during the overlapping period. Fig. 1 shows the sample hand trajectories from different signers.

2.2 Facial Analysis for Gesture Recognition

We investigate two alternative facial analysis to aid hand gesture recognition. Intuitively, facial expression has strong correlation with hand gesture. The other possibility is face identity, which can be use to do person specific gesture recognition, in which different styles of hand gesture by various signer is eliminated.

2.2.1 Feature Extraction

For both expression recognition and face identification we use the local appearance based face representation [10]. An modified census transform (MCT)-based face detector [13] is used to localize the position of face in each frame. The MCT face detector is a cascade of boosted MCT features. It is faster and more robust compared to the state-of-the-art face detectors such as the Viola-Jones method [14]. We also use an MCT-based eye detector to localize the center of the eyes, which are used for face alignment. A rigid transformation is applied so that the eyes are located in a fixed position in the aligned face image. From the aligned image, we compute a feature vector according to the method in [10] which has proven to provide a robust representation of the facial appearance in real-world applications. In short, the aligned face of 64×64 pixels is divided into non-overlapping blocks of 8×8 pixels resulting in 64 blocks. On each of these blocks, the two-dimensional discrete cosine transform (DCT) is applied and the resulting DCT coefficients are ordered by zig-zag scanning (i.e. $c_{0,0}, c_{1,0}, c_{0,1}, c_{0,2}, c_{1,1}, c_{2,0}, \ldots$). From the ordered coefficients, the first is discarded for illumination normalization. The following 5 coefficients from all blocks, respectively, are normalized and concatenated to form the facial appearance feature vector ($5 \times 64 = 320$ dimensional). See [10] for details.

2.2.2 Facial Expression Analysis

The 320-dimensional facial appearance feature vector is projected onto an "expression sub-space" using linear discriminant analysis (LDA). The "expression sub-space" is learned on a subset of the FEED database [12]. We select face images from FEED in seven different expression classes. Sample images of different facial expression are displayed in Fig. 2. After LDA, we transform a face feature vector into a six dimensional vector in the "expression sub-space". Similar facial expression should have low distance in this LDA sub-space. Similar to the hand trajectory, we represent facial expression with "expression trajectory" in the "expression sub-space" over a video sequence. Fig. 3 shows some examples of the trajectories. Note that the curves are very noisy because the noise in face alignment. We smooth the curves with a low pass filter. The similarity of facial expression is calculated by matching the "expression trajectory" using the condensation-based curve matching algorithm.

2.2.3 Face Identity Analysis

The face identity analysis is implemented as open set face recognition problem. If the face of a signer is identified, we only use the gesture models of corresponding signer. If the signer can not be identified, all trained gesture models will be used for gesture recognition.

Open set face recognition is different from the traditional face identification in that it also involves rejection of impostors in addition to identify accepted genuine members that are enrolled in the database. We formulate the open-set face recognition as a multiple verification problem as proposed in [11]. Given a claimed identity, the result of an identity verification is whether the claimed identity is accepted or rejected. Given a number of positive and negative samples it is possible to train a classifier that models the distribution of faces for both cases. Based on this idea, we trained an identity verifier for every one of the n known subjects in the gallery using support vector machines (SVM) classifiers. Once a new probe is presented to the system, it is checked against all classifiers; if all of them reject, the person is reported as unknown; if one accepts, the person is accepted with that identity; if more than a single verifier accepts, the identity with the highest score wins. Scores are linearly proportional to the distance to the hyperplane of the corresponding SVM classifier.

Since a person's identity does not change within a face track if there is no track switching error, we can enforce temporal consistency. In order to make it possible to revise a preliminary decision later on, instead of relying on a single classification result for every frame an n-best list is used. N-best lists store the first n

Figure 2. FEEDTUM data expression

highest ranked results. We choose $n = 3$ in this work. For each hypothesis a cumulated score is stored that develops over time.

3 Condensation-based Gesture Recognition

The Condensation algorithm (Conditional Density Propagation over time) makes use of random sampling in order to model arbitrarily complex probability density functions. That is, rather than attempting to fit a specific equation to observed data, it uses N weighted samples to approximate the curve described by the data. Each sample consists of a state and a weight proportional to the probability that the state is predicted by the input data. As the number of samples increases, the precision with which the samples model the observed pdf increases [9].

We increase the sample set to apply condensation to recognize more complex human gesture models. Specifically, a state at time t is described as a parameter vector: $s_t = (\mu, \phi^i, \alpha^i, \rho^i)$ where:

μ is the integer index of the predictive model,
ϕ^i indicates the current position in the model,
α^i refers to an amplitude scaling factor,
ρ^i is a scale factor in the time dimension,
where, $i \in \{l, \gamma, f\}$.

The models could contain data about the motion trajectory of both the left hand and the right hand or face feature vectors; by allowing three sets of parameters, the motion trajectory of left, right hands and facial feature trajectories to be scaled and shifted separately from each other. In summary, there are seven parameters that describe each state for hand motion trajectories, four parameters for facial feature trajectories, and ten parameters for both hand motion and facial feature trajectories.

The sample set is initialized with N samples uniformly distributed among the possible states. In the prediction step, each parameter of a randomly sampled S_t is used to determine S_{t+1} based on the model parameter (Φ_t) of that particular S_t. The weight of S_{t+1} is calculated based on how well the observed trajectory matches the models trajectory (parameterized according to the parameter vector in S_{t+1} over a historical time window w. At a given time t, we can determine the most likely gesture being observed by the system by finding the model with the most cumulative weight. A series of gestures can be recognized by introducing transition probabilities between models and choosing a threshold for the probability of a state at time which μ is classified as recognized. With this algorithm in place, all that remains is actually classifying the video

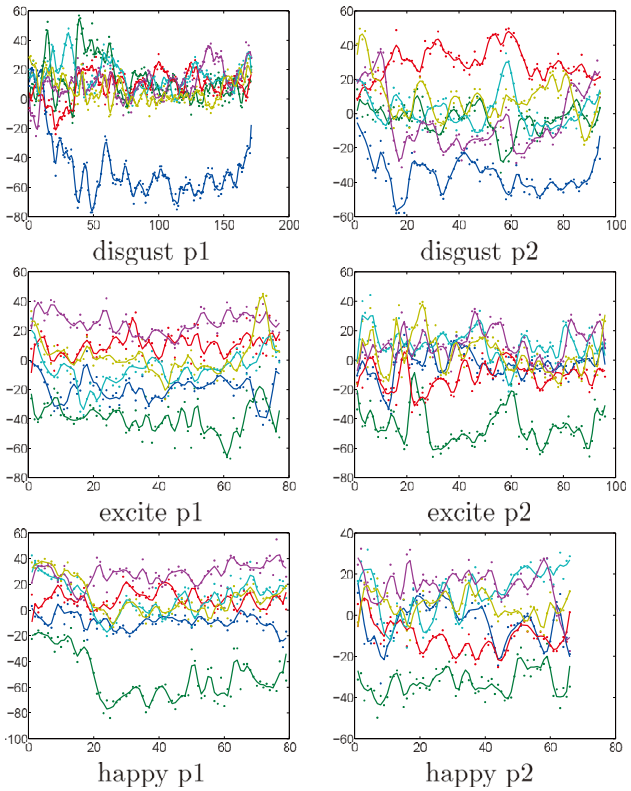

Figure 3. Samples of facial expression trajectories

sequence as one of the existing signs. Since the whole idea of Condensation is that the most likely hypothesis will dominate by the end, we chose to use the criterion of which model was deemed most likely at the end of the video sequence to determine the class of the entire video sequence. Determining the probability assigned to each model is a simple matter of summing the weights of each sample in the sample set at a given moment whose state refers to the model in question.

4 Experiments and Results

The database we used for the experiments contains 180 action clips for 12 sign gestures with facial expression and each of them is performed 3 to 7 times by 3 actors. Each video clip has a spatial resolution of 640×480 pixels, a frame-rate of $25fps$ and it is captured by a web camera Logitech Webcam Pro 9000 from front view. The classes of the gestures are: (1)anger, (2)apologize, (3)appreciate, (4)desire, (5)disgust, (6)excite, (7)fear, (8)happy, (9)nervous, (10)sad, (11)soso and (12)surprise. The data-set is split into a training set and testing set. The training set contains one recording session per person, i.e. 36 video clips. The rest of the clips are used for testing.

4.1 IIand Gesture Recognition with Facial Expression Analysis

Using the hand trajectory only, 85.4% of the video clips in the testing set can be correctly recognized. Some gestures can be easily confused with each other due their high similarity, such as "appreciate" and "fear". However, the corresponding expression of the confused gesture can be different enough, so that after

Table 1. Classification results

Modality	Recognition rate
Hand gesture	85.4%
Facial expression(FE)	45.0%
Hand + FE	89.5%
Hand + ID	92.6%
Hand + FE + ID	92.6%

combining the decision from the two modalities, some classification error can be resolved. We observed that the recognition rate of facial expression on this dataset is low (45%), which is kind of expected because some gestures do not correspond any obvious facial expression. Nevertheless, if we combine the scores of hand gesture recognition and facial expression recognition with weighted sum rule, the recognition rate is improved to 89.5% as listed in Table 1.

4.2 Hand Gesture Recognition with Face Identity Analysis

We also combine hand gesture recognition with face identification system. As the faces of the signers are mainly frontal in the recorded dataset and the illumination condition did not change in different recording sessions, there is no challenge to identify the signers. In [11], the open set recognition system achieved a very high recognition rate with progressive-scoring in a real-world scenario with various conditions. In this study, all frames in the test set are correctly identified in controlled condition. With the identity information, we can do person specific gesture recognition, where we achieved 92.6% correct recognition rate. Unfortunately, there is no further improvement when we add facial expression analysis in the person specific gesture recognition.

5 Conclusions

This paper presents two facial analysis methods to improve hand gesture recognition. The analysis on facial expression helps to distinguish ambiguous hand gestures. The analysis on signer identity enables person specific hand gesture recognition, which makes the problem easier without variations in gesture style. Experiment results show that both facial analysis improve hand gesture recognition. In particular, utilizing signer identity improve the recognition rate from 85.4% to 92.6%.

Acknowledgments

This study is partially funded by InterACT program between Waseda University and Karlsruhe Institute of Technology (KIT) and by the "Concept for the Future" of KIT within the framework of the German Excellence Initiative.

References

[1] G. Yao, H. Yao, X. Liu, and F. Jiang, "Real time large vocabulary continuous sign language recognition based on op/viterbi algorithm," in Intl. Conf. Pattern Recognition, Hong Kong, Aug. 2006, vol. 3, pp. 312-315.

[2] S. B. Wang, A. Quattoni, Louis-Philippe Morency, David Demirdjian, and Trevor Darrell, "Hidden conditional random fields for gesture recognition", in Computer Vision Pattern Recognition, New York, USA, June 2006, vol. 2,pp. 1521-1527.

[3] W. Gao, G. Fang, D. Zhao, and Y. Chen, "Transition Movement Models for Large Vocabulary Continuous Sign Language Recognition", Proc. Sixth IEEE Int'l Conf. Automatic Face and Gesture Recognition, pp. 553-558, May 2004.

[4] T. Starner, J. Weaver, and A. Pentland, "Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 12, pp. 1371-1375, Dec. 1998.

[5] C. Vogler and D. Metaxas, "A Framework for Recognizing the Simultaneous Aspects of American Sign Language", Computer Vision and Image Understanding, vol. 81, no. 3, pp. 358-384, 2001.

[6] R.D. Yang and S. Sarkar, "Detecting Coarticulation in Sign Language Using Conditional Random Fields", Proc. 18th Int'l Conf. Pattern Recognition, pp. 108-112, Aug. 2006.

[7] M. Isard and A. Blake, "Contour tracking by sotchastic propagation of conditional density", In 11th European Conference on Computer Vision (ECCV'96), Cambridge, UK, pp. 343-356, 1996.

[8] M. Isard and A. Blake, "A mixed-state Condensation tracker with automatic model switching", In International Conference on Computer Vision (ICCV'98), Mumbai, India, pp. 107-112, Jan, 1998.

[9] M. J. Black and A. D. Jepson, "Recognizing temporal trajectories using the condensation algorithm", In Int. Conf. Automatic Face and Gesture Recognition, Nara, Japan, pp. 16-21, Apr., 1998.

[10] H.K. Ekenel and R. Stiefelhagen, "Analysis of Local Appearance-based Face Recognition: Effects of Feature Selection and Feature Normalization", In Proc. of CVPR Biometrics Workshop, New York, USA, June 2006.

[11] H.K. Ekenel, L. Toth and R. Stiefelhagen, "Open-Set Face Recognition-based Visitor Interface System", 7^{th} International Conference on Computer Vision Systems, Liege, Belgium, October 2009.

[12] F. Wallhoff, "Facial Expressions and Emotion Database", http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html, Technische Universität München, 2006.

[13] B. Fröba, A. Ernst, "Face detection with the modified census transform", In Proc. of 6^{th} IEEE International Conference on Automatic Face and Gesture Recognition, pp. 91-96, 2004.

[14] Paul Viola, Michael J. Jones, "Robust Real-Time Face Detection", International Journal of Cumputer Vision 57(2),vol.57, pp.137-154, 2004.