Study of Extracting Multi-modal features for Recognizing Sign Language Vocabularies that Combine Facial Expressions and Hand Gestures

Luo Dan† Jun Ohya†

1. Introduction

The use of gesture as a natural interface plays an utmost important role for achieving intelligent Human Computer Interaction (HCI). In this paper, we describe a two-stage method for extracting multimodal features, including facial expression, hand motion and hand shape features which are extracted from image frames. The first stage uses Modified Census Transform (MCT) based detector to propose face and hand position using Adaptive GMM. A second stage we combine the DCT based facial feature and hog hand shape feature with hand location temporally, which are dimensionally reduced.

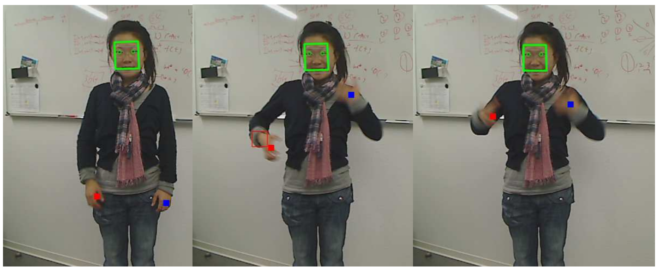
2. System Overview

Human gestures include different components of visual actions such as motion of hands, face, and torso, to convey meaning. So far, in the field of gesture recognition, most previous work has focused on classifying a couple simple hand gestures [1]. 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. In this paper, we present a multimodal-based gesture recognition framework, which combines different groups of features, facial expression, hand shape and motion which are extracted from the image sequences acquired by a single web camera. The system refer 12 classes of human gestures with facial expression including neutral (e.g. a sign "feel"), negative (e.g. "angry") and positive (e.g. "excited") meanings from American Sign Languages.

In our approach, we build on ideas from the previous work [3] and extend them to extract sufficient multimodal features. Our aim is to implement an integrated system which extracts different modalities of features (hand motion, and facial expression features) and combination strategies. Our method for building the multimodal features consists of two main modules: hand feature extraction, and face feature extraction. We combined these features as multimodal features using weighted sum method. The next two sections elaborate on the multimodal features extraction and experiment, respectively.

2. Multimodal Features

The MCT-based face detector is used to localize the position of face in each frame and locate the eyes within the detected face region. The detected eye positions suffice to normalize the face localization. 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 build the skin color database and non-skin color database for hand segmentation using adaptive GMMs (Gaussian Mixture model). Here we obtain the face blob and hand blobs to build multimodal features [1]. Fig.1 shows some frames of a signer performing the sign gesture,"excited". Face is detected and overlaid with bounding box normalized by eye location. The red and blue points indicate the centroids of the segmented left and right hand blobs.

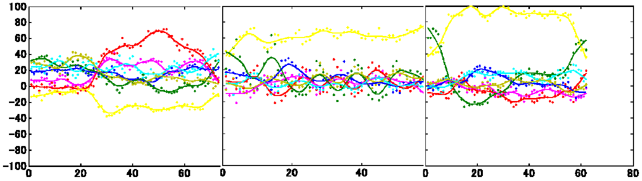


frame #2 frame #134 frame #198

Fig.1 Detection and segmentation result:"excited".

3.1 Facial Features

We compute the facial feature vector according to the method which has proven to provide a robust representation of the facial appearance in real-world applications. In short, the aligned face is divided into non-overlapping blocks of 8×8 pixels resulting in 64 blocks. On each of these blocks, the 2-dimensional discrete cosine transform (DCT) is applied and the resulting DCT coefficients are ordered by zig-zag scanning (i.e. $C_{0,0}$, $C_{1,0}$, $C_{0,1}$, $C_{0,2}$, $C_{1,1}$, $C_{2,0}$, …). From the ordered coefficients, the first is discarded for illumination normalization. The following 5 coefficients from all blocks, respectively, are concatenated to form the facial appearance feature vector (5 × 64=320 dimensional). See the proposed method for details [2].



angry_p1 excite_p2 happy_p1

Fig.2 Examples of facial expression trajectories.

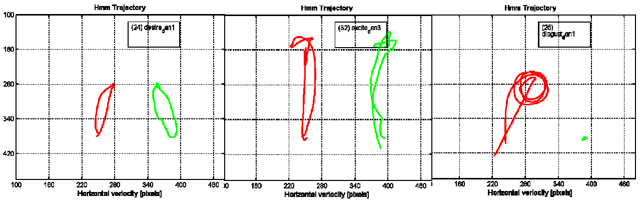
Since the signer may have expression in different intensity, and some signs may not correspond to specific expression class, we extract a low dimensional representation for facial expression. The face feature vector is projected onto an "expression sub-space" using PLS (partial least squares). The "expression sub-space" is learned on a subset of the FEED database [4] and CK+ [5]. We select face images in different expression intensities of seven different expressions. After PLS, we transform a face feature vector into a 6 dimensional vector in the "expression sub-space". Similar facial expression should have low distance in this sub-space. Similar to the hand trajectory, we represent facial

† Waseda University, Tokyo, Japan

expression with “expression trajectory” in the “expression subspace” over a video sequence. Fig.2 shows facial expression trajectories of example gestures from two people p1 and p2. We could see the yellow line indicates the energy of “smiling” during the sign gesture. Obviously, “angry” has fewer smiles than “excite” and “happy” had more smiles than “excite” through the gesture sequences. Note that the curves are very noisy because of 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” during classifying.

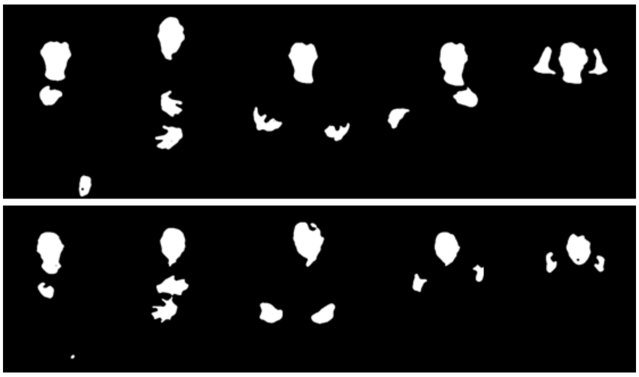
3.2. Hand Features

We use the centroids of the left and right hand blobs to generate hand motion trajectories over the whole video sequence shown in Fig.3 by three gesture samples. The generated spatial hand motion trajectories are normalized using the distance between face and hands in the first frame of a video, which normalize the scale variation of the trajectories from different signers and recordings.



“desire” “excite” “disgust”

Fig.3 Examples of hand motion trajectories.



“anger” “appreciate” “desire” “excite” “surprise”

Fig.4 Segmentation mask results

In each frame of the video sequence, we segment the image with the color database built during face detection progress so that face blob and hand blobs are obtained showed in Fig.4 with mask. Here we normalize and rescale the detection region using the radius of face region to 128×256. The main features for characterizing our blob images are normalized bin values from a Pyramid of Histogram of Oriented Gradients (PHOG) [6] with 8 bins/level, and 2 sublevels, resulting in a feature vector size of 8 $\times(1+2^2+4^2)$ = 168. Obtaining a Histogram of Oriented Gradients (HOG) involves calculating a gradient direction and magnitude for every pixel in the segmented hand region and the edge of the face region binning these gradients by their direction with a weight based on their magnitude. We next add to our feature vector a histogram of intensities, binning intensity values along with our original oriented gradient values. This quick addition effectively doubles the length of our feature vector to 336 dimensions. Using a simple Support Vector Machine (SVM) with a linear classifier, and using 70-30 cross validation to train and test hand shape features, we find that we are able to quickly achieve a precision of 86.1% on a training set of size 36 with test set 144.

4. Experiment

The database contains 180 video clips of 12 sign gesture vocabularies with facial expression performed 3 to 7 times by 3 signers. Each video clip has a spatial resolution of 640×480 pixels with 25fps from frontal view. The data-set is split into two independent data-sets: a training set and a testing data-set for evaluation. The training set contains one recording session per person, i.e. 12×3=36 video clips. The rest of the clips are used for test. Global Multimodal feature is built by facial expressional features, hand trajectories and hand shape features. In the experiment, we first conduct experiments on recognizing sign gestures using single modality based on condensation algorithm [3]. Using hand trajectory only, 85.4% of the video clips in the testing set can be correctly recognized. By combining the classification scores of facial feature and hand shape modality with scores of weighted sum rule, the accuracy is up to 93.4%.

5. Conclusion

We proposed an integrated framework that aims at extracting multimodal information for recognizing human gestures selected from American Sign Language. AAM is used to extract facial motion feature to capture face expression and pose information. Linear discriminate analysis is used to select most discriminative facial features for gesture recognition. Hand trajectories are obtained with respect to the HFLC system. The system can exploit both feature level and decision level fusion strategies.

Reference

[1] M.B. Holte, T.B. Moeslund, P. Fihl, “View-invariant gesture recognition using 3D optical flow and harmonic motion context, ” Computer Vision and Image Understanding, Volume 114, Issue 12, Pages 1353-1361, December 2010.

[2] Hua Gao, Hazım Kemal Ekenel, and Rainer Stiefelhagen, “Pose Normalization for Local Appearance-Based Face Recognition.” 3rd Int.l Conference on Biometrics (ICB 2009), LNCS 5558, pp. 32–41, 2009.

[3] Dan Luo, Hua Gao, Hazim Kemal Ekenel and Jun Ohya, "Appearance-based human gesture recognition using multimodal features for human computer interaction", Proc. SPIE 7865, 786509 (2011).

[4] F. Wallhoff, “Facial Expressions and Emotion Database,” http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html, Technische Universit at Munchen, 2006.

[5] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I., “The Extended Cohn-Kande Dataset (CK+): A complete facial expression dataset for action unit and emotion-specified expression,” the Third IEEE Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB), 2010.

[6] A. Bosch, A. Zisserman, “Pyramid Histogram of Oriented Gradients (PHOG)”. University of Oxford Visual Geometry Group, http://www.robots.ox.ac.uk/vgg/research.