H-007

# Visual-based Human Gesture Recognition Using Multi-modal Features

Luo Dan† Jun Ohya†

## 1. Introduction

The use of human gesture as a natural interface plays an utmost important role for achieving intelligent Human Computer Interaction (HCI). This paper presents a visual-based gesture recognition framework, which combines different groups of features: hand shape, hand motion and facial expression features. We employ two fusion strategies in the feature level and decision level to the combined multi-modal features. An adaptive Conditional Random Field (CRF) and condensation-based algorithm is adopted for classification. Experimental results show that facial analysis and hand shape information improved hand gesture recognition and decision level fusion performs better than feature level fusion.

## 2. Multimodal Features

In our approach, we build on ideas from the previous work [1] and extend them to extract sufficient multimodal features. Our aim is to implement an integrated system which extracts different modalities of features and combination strategies.

To locate face and hand position, the MCT-based face detector is used to 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). Fig.1 shows some frames of a signer performing the sign gesture,"excited".

frame #2 frame #134 frame #198

Fig.1 Face detection and hand 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 $8\times8$ 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 $5\times64=320$ dimensional facial appearance feature vector. See the proposed method for details [1].

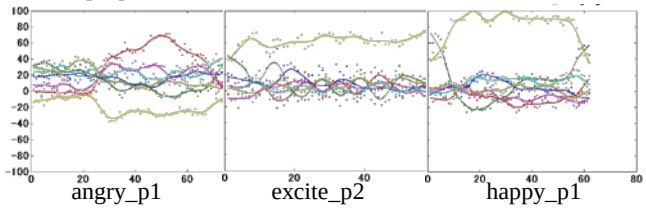



Fig.2 Examples of facial expression trajectories.

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 provided by Wallhoff and CK+ database provided by Cohn and Kanede. 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 expression with "expression trajectory" in the "expression sub-space" 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 "smiling" energy 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.

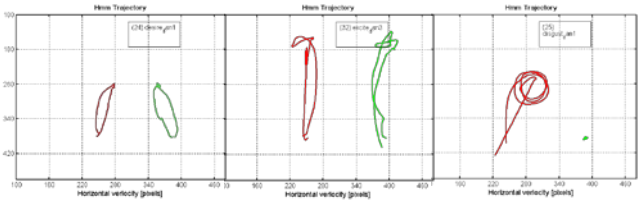
### 3.2. Hand Features

**Hand Motion 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.

**Hand Shape Features**: 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 show in Fig.4 with mask. Here we normalize and rescale the detection region using the radius of face region to $128\times256$. The main features for characterizing our blob images as Fig. 4 up show up are normalized bin values from a Pyramid of Histogram

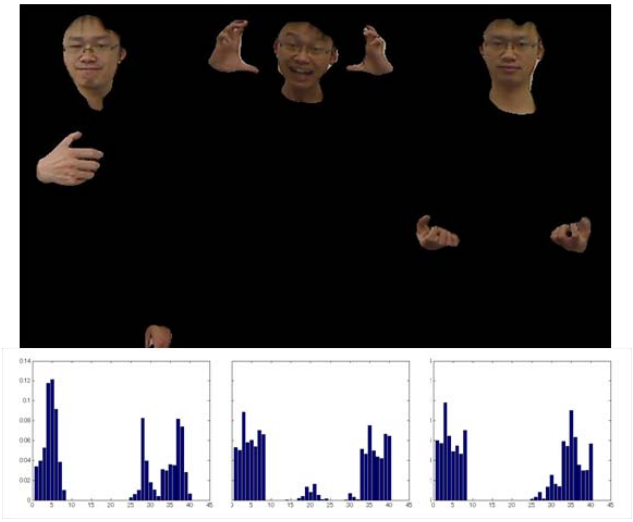
† Waseda University, Tokyo, Japan

of Oriented Gradients (PHOG) [3] with 8 bins/level, and 2 sublevels, resulting in a feature vector size of $8\times(1+2^2+4^2)=168$, then all the histograms in Fig. 4 down can be concatenated to form a raw shape feature vector, which is normalized to generate the shape descriptor Ds. An appearance likelihood map and the shape descriptor are computed from it. Our method for estimating appearance likelihoods is explained in Section 4. The distance between two shape descriptors is computed using the Euclidean distance metric.



“desire” “excite” “disgust”

Fig.3 Examples of hand motion trajectories.



“anger” “surprise” “desire”

Fig.4 Segmentation mask results

## 3. Discriminative Models

**Conditional Random Field**: The goal is to learn a classifier $p(y\,|\,\mathrm{X})$ that predicts a gesture label $y\in\mathrm{Y}$ given a temporal sequence of input images $\mathrm{X}=\{\mathrm{X}_1,\ldots,\mathrm{X}_T\}$. For each image $\mathrm{X}_T$, we extract hand motion features $\phi(\mathrm{X}_t^2)\in\Re^{N2}$, hand shape features $\phi(\mathrm{X}_t^3)\in\Re^{N3}$ and facial features $\phi(\mathrm{X}_t^2)\in\Re^{N2}$; each $\mathrm{X}_T$ is presented as a multi-signal feature-vector

$$\phi(\mathrm{X}_t)=(\phi(\mathrm{X}_t^1)\phi(\mathrm{X}_t^2)\phi(\mathrm{X}_t^3))^T.$$

Our ladent-dynamic CRF model was trained using the objective function described in [4] Section 3.1 with class labels (but not hidden stated). For testing, given a new test sequence, we estimate the most probable label sequence that maximizes our conditional model.

**Condensation algorithm**

The Condensation algorithm (Conditional Density Propagation over time) makes use of random sampling in order to model arbitrarily complex probability density functions. Each sample consists of a state and a weight proportional to the probability that the state is predicted by the input data. See detail in [1].

## 4. Experiment

The system refers 12 classes of human gestures with facial expression selected from American Sign Languages. The database contains 144 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. In the decision level weighted sum rule are used to each feature’s likelihood. The rest of the clips are used for test. The hand motion only result is 85.2%, Table 1 show the combined multi-modal recognition result respectively.

Table 1 Multi-modal Recognition Result (CRF/Condensation)

| Method | Joint Face | Joint Hand Shape | Joint Multi-modal |
|---|---|---|---|
| Decision Level | 89.8%<br>87.0% | 92.7%<br>91.7% | 94.4%<br>92.6% |
| Feature Level | 88.0%<br>86.1% | 91.7%<br>90.7% | 92.6%<br>92.6% |

## 5. Conclusion

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. We proposed an integrated framework that aims at extracting multimodal information for recognizing human gestures selected from American Sign Language. The system can exploit both feature level and decision level fusion strategies. Experimental results show that facial analysis and hand shape information improved hand gesture recognition from 85.2% to 94.4% using CRF based classifier and decision level fusion performs better than feature level fusion.

### Acknowledge

This paper is a part of the outcome of research performed under a Waseda University Grant for Special Research Projects (Project number: 2012A-890).

### Reference

[1] Luo Dan, Hazim Kemal Ekenel, and Ohya Jun, “Human Gesture Analysis Using Multimodal Features”, IEEE International Conference on Multimedia and Expo Workshops (ICMEW2012), pp. 471-476, (2012.07).

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

[4] Morency, L.; Quattoni, A.; Darrell, T., "Latent-Dynamic Discriminative Models for Continuous Gesture Recognition," IEEE Conference on Computer Vision and Pattern Recognition 2007 (CVPR '07), pp.1-8, 17-22 June 2007.