J-018

Study of Human Gesture Recognition by Integrating Face and Hand Motion Features

Luo Dan†,‡ Hazım Kemal Ekenel‡ Jun Ohya†

1. Introduction

In order to understand natural human sign gestures for Human Computer Interaction (HCI), it is necessary and important to recognize the multimodal nature of the visual cues such as hand motion, facial motion, body pose, etc., which is known to be a very challenging task. Our strategy to implement an integrated system that aims at extracting sufficient information for recognizing human gestures selected from sign language relies on three modules. The first module uses active appearance models for detailed face tracking, allowing the quantification of facial expressions such as mouth and eye aperture and eyebrow rise. The second module is dedicated to hand motion understanding using color and trajectories. Finally, the third module combines the information coming from the first two modules to provide robust human gesture recognition.

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 the hand gestures. In this paper, we present an appearance-based multimodal gesture recognition framework, which combines different groups of features such as head motion, facial expression and hand motion which have been extracted from the images captured directly by a 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. Active Appearance Model [1] is used for detailed face tracking, allowing the quantification of facial expressions such as mouth and eye aperture and eyebrow rise. For hand motion understanding, we use color and trajectories that are described in Section 3. The system then can combine the features in two different levels. At the feature level, an early feature combination can be performed by concatenating features extracted from face and hands, and employing statistical models to choose the most discriminate elements from the combined feature set. At the decision level, weighted decisions from single modalities can be fused in as late stage.

† Waseda University, Tokyo, Japan

‡ Karlsruhe Institute of Technology, Karlsruhe, Germany

3. Face and Hand Motion

3.1 Facial Features

Facial expressions and head motion play an important role in human gestures. In the initial work [3], some hand gestures are ambiguous in isolation, and need to be accompanied by appropriate facial expressions in order to convey a specific message. For face and hand feature extraction, facial parameters such as eye and mouth apertures can be inferred from the configuration of a set of relevant facial features such as the positions of fiducially points on eyelids and lips. Active Appearance Model (AAM) is a statistical generative model. Shape and texture variations of the human face as well as the correlations between them are learned from a set of example face images, on which corresponding "landmark" points have to be marked priori. Here, Active Appearance Model is used to track such facial features for the face tracking system [2]. Fitting the AAM to an input image is done by finding the values of the parameters that minimize the difference between the synthesized model image and the input image using a gradient descent-based approach.

A shape in AAM is defined as a set of normalized 2D facial landmarks. An instance of the linear shape model can be represented as $s = s_0 + \sum_{i=1}^{n} p_i s_i$, where s_0 is the mean shape, s_i is the i^{th} shape basis, and $P = [p_1, p_2, ..., p_n]$ are the shape parameters. The texture model is defined inside the mean shape, which explains the variations in texture caused by changes in illumination, identity, and expression, etc. It represents an instance appearance as $A = A_0 + \sum_{i=1}^{n} \lambda_i A_i$, where A_0 is the mean appearance, A_i is the i^{th} appearance basis, and $\lambda = [\lambda_1, \lambda_2, ..., \lambda_n]$ are the appearance parameters.

3.1 LDA-based Feature

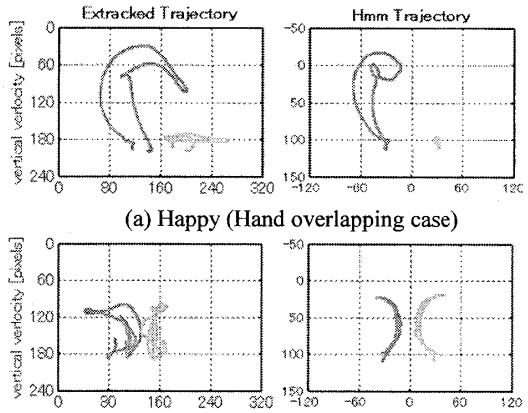
The complete framework of the face tracker is composed of an offline part where the face model is built that contains all the facial appearance variation information as well as preprocessed data for the step of fitting, and using this model to track the facial features. Since the fitting method is a local search, AAM is initialized by the face detector [4]. Eye and mouth apertures shown here are quantified by the normalized area of the contours delimited by eye and mouth point features respectively.

(a) Satisfied (b) Happy (c) Neutral

Fig.1. Samples of Tracked AMM Shape and Normalized Texture for Different Facial Expression (Upper row: AAM Shape; Lower row: Normalized Texture [Appearance])

3.2. Hand Motion Features

In each frame of the video sequence, we segment the image using the color database so that the body blob, face blob and hand blobs are obtained. Based on positional information on these blobs, we construct HFLC (Human-Following Local Coordinate) system, which follows the human body in the video sequence [3].



(a) Happy (Hand overlapping case)

(b) Satisfied

Fig.2. Samples of Extracted Hand Trajectories (Upper row: hand trajectories of "happy"; Lower row: hand trajectories of "satisfied")

By obtaining the hand's trajectory with respect to the HFLC system, the effect of the camera motion is suppressed from the obtained trajectory. A hand trajectory motion model (HTMM) database stores hand gestures to be recognized, where the database is constructed using the trajectories of the hands. Based on the constructed HFLC system, the temporal modeling of a gesture is important since human gestures are dynamic processes. Psychological studies show that a hand gesture consists of three phases. These phases are: Preparation, Nucleus, and Retraction. Every hand gesture in the experiment consists of these three strokes. To construct models for the gestures, each gesture was performed approximately half a dozen times and the trajectories were manually aligned and the mean trajectories were computed. A standard deviation from the mean trajectories was also computed for each curve. Some examples of the trajectory models for each gesture are shown in Fig. 2. Fig. 2(a) shows that even the face-hand overlapping case can be accurately extracted by HFLC.

4. Feature Combination

We get the facial parameter (shape parameter and normalized texture parameter), which is described in Section 3. Gesture feature is built by face feature acquired from shape parameter and texture parameter and hand trajectories. For face feature, linear discriminate analysis (LDA) is used for feature dimensional reduction and to select discriminative features. Two different combination strategies can be employed to fuse the information coming from face and hands. The first one is at feature level by combining the feature vectors extracted from face and hands. A statistical method can be used afterwards to select the most discriminative features for classification. The second one is at decision level by combining the classification scores of each modality.

5. Conclusion

We proposed an integrated framework that aims at extracting multimodal information for recognizing human gestures selected from 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.

Acknowledgment

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.

Reference

[1] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models." In Proc. of the Eur. Conf. on Comp. Vis., pages 484–496, 1998.

[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 and Jun Ohya, "Hand-gesture extraction and recognition from the video sequence acquired by a dynamic camera using condensation algorithm." Proc. SPIE 7252, 72520S, 2009.

[4] P. Viola and M. Jones. 2001. Rapid Object Detection using a Boosted Cascade of Simple Features. In Proc. IEEE CVPR 2001.