Machine Vision and Applications (2001) 12: 243–258

Machine Vision and Applications

A model-based hand gesture recognition system

Chung-Lin Huang, Sheng-Hung Jeng

Electrical Engineering Department, National Tsing Hua University, Hsin Chu, Taiwan, ROC; e-mail: clhuang@ee.nthu.edu.tw

Received: 19 May 1999 / Accepted: 4 September 2000

Abstract. This paper introduces a model-based hand gesture recognition system, which consists of three phases: feature extraction, training, and recognition. In the feature extraction phase, a hybrid technique combines the spatial (edge) and the temporal (motion) information of each frame to extract the feature images. Then, in the training phase, we use the principal component analysis (PCA) to characterize spatial shape variations and the hidden Markov models (HMM) to describe the temporal shape variations. A modified Hausdorff distance measurement is also applied to measure the similarity between the feature images and the pre-stored PCA models. The similarity measures are referred to as the possible observations for each frame. Finally, in recognition phase, with the pre-trained PCA models and HMM, we can generate the observation patterns from the input sequences, and then apply the Viterbi algorithm to identify the gesture. In the experiments, we prove that our method can recognize 18 different continuous gestures effectively.

Key words: Hand gesture recognition – Principal component analysis (PCA) – Hidden Markov model (HMM) – Hausdorff distance measurement – Viterbi algorithm

1 Introduction

Hand gesture is normally used in our daily life to communicate with one another. Children know how to make gesture communication before they can talk. Clearly, gesture recognition has become one of the most interesting research topics in human-computer interface. Most of the recent works [1] related to hand gesture interface techniques have been categorized as: glove-based methods and vision-based methods. The vision-based methods, based on the computer vision techniques, have been proposed for locating objects and recognizing gestures. The gloved-based gesture recognition methods require expensive wired “Dataglove” equipment [2]. Gesture recognition research has many applications such as window-based user interface [3] and video coding [4].

The model-based static gesture recognition approach proposed by Davis and Shah [5], uses a finite-state machine to model four qualitatively distinct phases of a generic gesture. Hand shapes are described by a list of vectors and then matched with the stored vector models. A dynamic gesture recognition system for American sign language (ASL) interpretation has been developed by Charayaphan et al. [6]. They propose a method to detect the direction of hand motion by tracking the hand location, and use adaptive clustering of stop location, simple shape of the trajectory, and matching of the hand shape at the stop position to analyze 31 ASL signs.

A more reliable method called the space-time gesture recognition method developed by Darrell et al. [7] represents gestures by using sets of view models. It recognizes the gestures by matching the view models to stored gesture patterns using dynamic time warping. Cui and Weng [8] propose a learning-based hand sign recognition framework by using the multiclass, multivariate discriminant analysis system to select the most discriminating feature (MDF), and then applying a space partition tree to reduce time complexity. Hunter et al. [9] explore posture estimation based on the 2D projective hand silhouettes for vision-based gesture recognition. They use Zernike moments and normalization to separate the rough posture estimate from specific translation, rotation, and scaling.

The most difficult part of gesture identification is to classify the posture against complex backgrounds. Triesch et al. [10] employ elastic graph matching for the classification of hand postures in gray-level images. Heap et al. [11] construct a 3D deformable point distribution model of the human hand. Then, they use this model to track an unmarked human model with six degrees of freedom. Another simplified method (by Lee and Kunii [12]) assumes that the positions of fingertips in the human hand, relative to the palm, is almost always sufficient to differentiate the gestures. They propose the skeleton-based model consisting of 27 bones and 19 links, each link has different degrees of freedom.

However, gesture recognition is more generally treated as a time variation problem, therefore, more and more com-

Correspondence to: C.-L. Huang

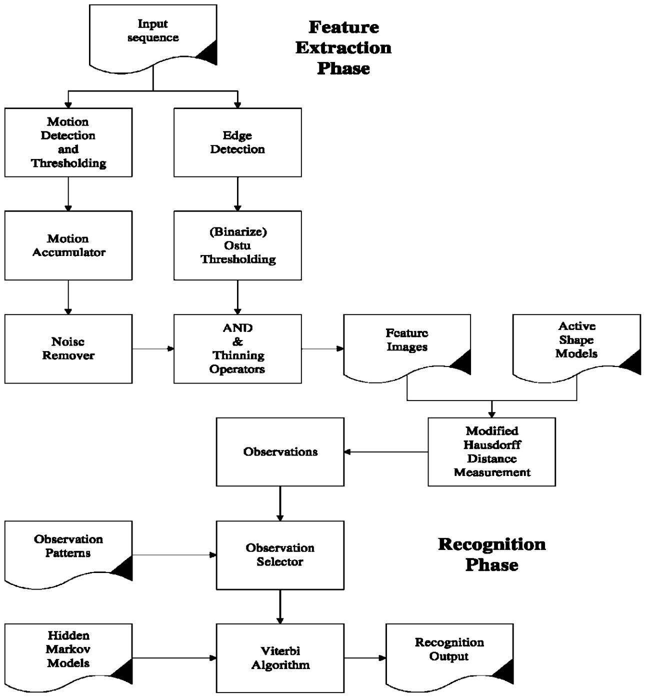

Fig. 1. The flow diagram of hand gesture recognition system

puter vision researchers have become aware of using hidden Markov models (HMMs) to model the image sequence of gestures. Starner et al. [13] use HMM to recognize a full sentence and demonstrate the feasibility of recognizing a series of complicated series of gesture. Bobick et al. [14] present a state-based method for representation and recognition of gesture from a continuous stream of sensor data. The variability and repeatability evidence in a training set of a given gesture is classified by states.

The major difficulties of the complex articulated-objects analysis are the appearance of large variation of 2D hand shapes, the view point sensitive for 2D hand shapes and motion trajectories, the transition between the meaningful gestures, and the interference of complex background. The gesture image sequence is basically composed of spatial and temporal variation signals, so we need to apply the principal component analysis (PCA) and HMM to model the spatial and the temporal shape variation of the gestures. Figure 1 shows the flow diagram of our model-based hand tracking and recognition. We use the Hausdorff distance measurement to measure the differences of the input gestures and pre-stored PCA shape models. The differences are then converted to observations for HMM training (in the training phase) or for state sequence evaluation (in the recognition phase) by using the Viterbi algorithm based on the pre-trained HMM. The most likely state transition sequence is associated with the gesture to be recognized.

The hand gesture recognition system can be described in three phases: the training phase, the feature extraction phase, and the recognition phase. We represent each gesture by a sequence of states. The state transition indicates the spatial temporal variability of the gesture, which is invariant to the speed of motion. In the feature extraction phase, we develop a new method, which combines the edge and motion information to extract the shape of the moving object. Then, in training phase, we develop a simple training algorithm for the PCA models and HMM which characterize the spatial and the temporal variation of gestures. Finally, in the recognition phase, we apply pre-stored PCA models to observe the input gesture, and use the Hausdorff distance measure to find the similarity between the extracted features and the pre-stored PCA models. In the experiments, we show that our gesture recognition system is insensitive to motion speed and trajectory direction, and it can precisely recognize 18 different gestures in complex background.

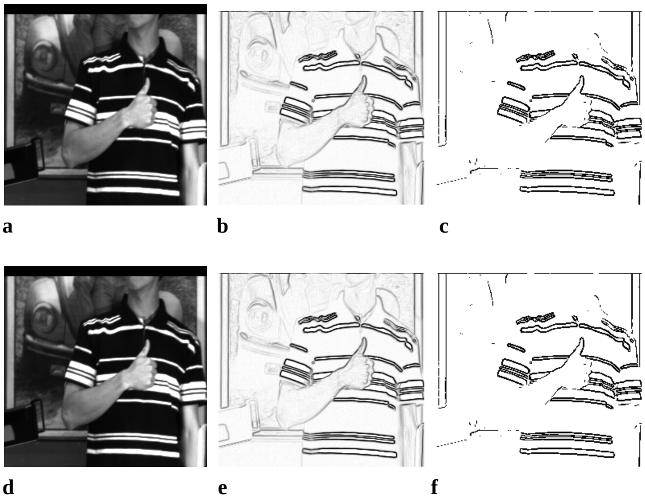

Fig. 2a–f. Edge information. **a** and **d** are the original images. **b** and **e** are the gradient strength images. **c** and **f** are results after the Ostu thresholding

2 Feature extraction phase

Here, we assume that the moving objects in complex background are somehow identifiable by their edge boundaries. Usually, the edge information is too noisy to be applicable for computer vision system, and most of the edge information is redundant. Here, we assume that the background is complex but stationary, the moving hand is the only moving object in the scene. Using the frame difference, we can partially capture the motion information. By accumulating the motion information of the moving objects in several consecutive frames, we may localize the moving pixels more accurately.

First, we apply the Sobel operators [17] and Ostu thresholding method [18] to extract the edges in the scene (see Fig. 2). Second, we find the motion information by using the motion accumulator and the noise remover. Finally, we apply the AND operation on the edges and the accumulated motion pixels to acquire the real moving edges.

2.1 Motion accumulator

To find the edge information from the object movement, we assume that the gesture in the sequence is non-stationary. In the spatial-temporal space, the motion detector may capture all the possible moving objects by examining the local gray-level changes. Let F_i be the ith frame of the sequence and