Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 40 (2007) 3012–3026
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Simultaneous gesture segmentation and recognition based on forward spotting accumulative HMMs

Daehwan Kim, Jinyoung Song, Daijin Kim*

Department of Computer Science and Engineering, Pohang University of Science and Technology, San 31, Hyoja-Dong, Nam-Gu, Pohang 790 784, Republic of Korea

Received 12 May 2006; received in revised form 2 February 2007; accepted 12 February 2007

Abstract

Existing gesture segmentations use the backward spotting scheme that first detects the end point, then traces back to the start point and sends the extracted gesture segment to the hidden Markov model (HMM) for gesture recognition. This makes an inevitable time delay between the gesture segmentation and recognition and is not appropriate for continuous gesture recognition. To solve this problem, we propose a forward spotting scheme that executes gesture segmentation and recognition simultaneously. The start and end points of gestures are determined by zero crossing from negative to positive (or from positive to negative) of a competitive differential observation probability that is defined by the difference of observation probability between the maximal gesture and the non-gesture. We also propose the sliding window and accumulative HMMs. The former is used to alleviate the effect of incomplete feature extraction on the observation probability and the latter improves the gesture recognition rate greatly by accepting all accumulated gesture segments between the start and end points and deciding the gesture type by a majority vote of all intermediate recognition results. We use the predetermined association mapping to determine the 3D articulation data, which reduces the feature extraction time greatly. We apply the proposed simultaneous gesture segmentation and recognition method to recognize the upper-body gestures for controlling the curtains and lights in a smart home environment. Experimental results show that the proposed method has a good recognition rate of 95.42% for continuously changing gestures.

Keywords: Gesture segmentation; Gesture recognition; Hidden Markov model; Forward spotting; Accumulative HMM; Competitive differential observation probability; Association mapping; Smart home control

1. Introduction

A gesture is a movement that we make with a part of our body, face and hands as an expression of meaning or intention. Gestures are classified into two forms according to the intention: natural and artificial gestures. The natural gesture is meaningless and uncertain, and it has cultural and local diversity. However, the artificial gesture can express more detailed and various meanings using predefined motions. We focus on upper-body artificial gestures in this work.

Many existing studies [1–3] have applied gesture recognition for human–computer interaction (HCI). Usually, these approaches used an hidden Markov model (HMM) and manually segmented image sequences for gesture recognition, so they are difficult to apply to continuous gesture recognition [4].

One main concern of gesture recognition is how to segment some meaningful gestures from a continuous sequence of motions. In other words, we need to spot the start point and the end point of a gesture pattern. This is considered a highly difficult process because gestures have segmentation ambiguities [5] and spatio–temporal variability [6]. The first property is caused by the fact that we do not know exactly when a gesture starts and ends in a continuous sequence of motions. The second property is caused by the fact that the same gesture varies in shape, duration, and trajectory, even for the same person.

To alleviate this problem, many researchers have used the HMM because it can model the spatial and temporal characteristics of gestures effectively. Wilpon et al. [7] used an HMM

* Corresponding author. Tel.: +82 54 279 2249; fax: +82 54 279 2299.
E-mail addresses: msoul98@postech.ac.kr (D. Kim), jysong@postech.ac.kr (J. Song), dkim@postech.ac.kr (D. Kim).

0031-3203/$30.00
doi:10.1016/j.patcog.2007.02.010

for keyword spotting and proposed a garbage or filler model to represent the extraneous speech. Lee and Kim [8] proposed an HMM-based threshold model that computed the likelihood threshold of an input gesture pattern and could spot the start and end points by comparing the threshold model with the predefined gesture models. Deng and Tsui [9] proposed an evaluation method based on an HMM for gesture patterns that accumulated the evaluation scores along the input gesture pattern. Kang et al. [10] proposed a novel gesture spotting method that combined gesture spotting with gesture recognition. It recognized the meaningful movements while concurrently separating unintentional movements from a given image sequence.

Most existing methods use the backward spotting scheme, first performing the gesture segmentation and then performing the recognition. First, they usually detect the end point of gesture by comparing the observation probability of the gesture model and the non-gesture model. Second, they trace back through an optimal path via the Viterbi algorithm [11] to find the start point of the gesture. Third, they send the extracted gesture segment to the HMM for gesture recognition. Thus, there is an unavoidable time delay between the gesture segmentation and the gesture recognition. This time delay is not appropriate for on-line continuous gesture recognition.

To solve this problem, we propose a forward spotting scheme that performs gesture segmentation and recognition at the same time. The forward scheme computes a competitive differential observation probability (CDOP) that is defined by the difference of observation probability between the gesture and the non-gesture, and detects the zero crossing points. The start (or end) points correspond to the zero crossing points from negative to positive (or positive to negative). From the start point, we obtain the posture type of the input frame using the predetermined association mapping between 2D shape and 3D articulation data and apply it to the HMM. Then, the HMM determines the gesture of each input frame until the end point.

We also propose a sliding window and accumulative HMM that can alleviate the problem of spatio–temporal variabilities. The sliding window technique computes the observation probability of gesture or non-gesture using a number of continuing observations within the sliding window. This reduces the undesirable effect of an abrupt change of observations within a short interval that can be caused by erroneous and incomplete feature extraction. The accumulative HMM decides the final gesture type by a majority vote of all recognition results that are obtained between the start and end point. This improves the classification performance of gesture recognition greatly.

To recognize the gesture, we need to extract the features from the input image. In general, two kinds of features, the 2D shape data and 3D articulation data are widely used for gesture recognition. Bobick and Davis [12] used a 2D view-based approach to represent and recognize human movements. The temporal templates containing the motion energy image (MEI) and the motion history image (MHI) were used to represent human movements. The Hu moments of the temporal templates were used to recognize the movements. Dong et al. [13] presented a method for human gesture recognition based on quadratic curves, where trajectory information of the center points of skin color and 2D foreground silhouettes were used to represent the movements and six invariants from the fitted quadratic curve was used to recognize them. However, these techniques can be used for the gesture recognition within a very limited view because the obtained 2D features were very dependent on the viewing angle between the human and the cameras.

To overcome these limitations, many researchers have tried to extract 3D articulation data. Agarwal and Triggs [14] recovered 3D body poses by the direct nonlinear regression of joint angles against shape descriptor vectors from the monocular silhouettes. Shakhnarovich et al. [15] introduced a new example-based algorithm for fast parameter estimation with parameter-sensitive-hashing (PSH) that could estimate the accumulated 3D human body poses. Sigal and Black [16] proposed a general process to infer the 3D poses from the silhouettes using the hierarchical Bayesian inference framework. Sminchisescu et al. [17,18] presented a mixture density propagation framework for the 3D pose recovery.

In this paper, we use the 3D articulation data as the input for the proposed gesture recognition method. Since we are using multiple cameras to capture the movements, it is necessary to estimate the 3D articulation data from the captured 2D images. There are two ways to do this: (1) the direct computation from corresponding multiple 2D images and (2) the indirect estimation by fitting multiple 2D images to the 3D articulation model. Both methods are not appropriate for real-time gesture recognition because they require a large amount of computation time.

In this paper, we propose an association mapping technique that correlates the 2D shape data to the 3D articulation data. It is built by the following manner. First, we prepare a large number of training samples with 2D shape data correlated to 3D articulation data. Second, we quantify the 2D shape and the 3D articulation data using the self-organizing map (SOM) for each. Third, we find an association mapping between the discrete 2D shape data and the discrete 3D articulation data using the learning technique by examples. This predetermined association mapping reduces the computation time for obtaining the 3D articulation data from the captured images greatly.

The rest of this paper is organized as follows. Section 2 presents the theoretical backgrounds of the HMM and SOM. Section 3 describes simultaneous gesture segmentation and recognition. Section 4 describes and discusses experimental results. Finally, Section 5 presents our conclusions.

2. Theoretical background

2.1. Hidden Markov models

An HMM is a statistical modeling tool which is applicable to analyzing time-series with spatial and temporal variability [8,19,20]. It is a graphical model that can be viewed as a dynamic mixture model whose mixture components are treated as states. It has been applied in classification and modeling problems such as speech or gesture recognition. Fig. 1 illustrates a simple HMM structure. Some parameters for defining