Pattern Recognition 43 (2010) 3059–3072

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Hand gesture recognition based on dynamic Bayesian network framework ☆

Heung-Il Suk [a], Bong-Kee Sin [b,*], Seong-Whan Lee [a]

[a] Department of Computer Science and Engineering, Korea University, Anam-dong, Seongbuk-ku, Seoul 136-713, Republic of Korea
[b] Department of Computer Engineering, Pukyong National University, Daeyon-dong 599-1, Nam-ku, Busan 608-737, Republic of Korea

ARTICLE INFO

Article history:
Received 21 June 2009
Received in revised form
9 February 2010
Accepted 22 March 2010

Keywords:
Hand gestures recognition
Dynamic Bayesian network
Coupled hidden Markov model
Continuous gesture spotting

ABSTRACT

In this paper, we propose a new method for recognizing hand gestures in a continuous video stream using a *dynamic Bayesian network* or DBN model. The proposed method of DBN-based inference is preceded by steps of skin extraction and modelling, and motion tracking. Then we develop a gesture model for one- or two-hand gestures. They are used to define a cyclic gesture network for modeling continuous gesture stream. We have also developed a DP-based real-time decoding algorithm for continuous gesture recognition. In our experiments with 10 isolated gestures, we obtained a recognition rate upwards of 99.59% with cross validation. In the case of recognizing continuous stream of gestures, it recorded 84% with the precision of 80.77% for the spotted gestures. The proposed DBN-based hand gesture model and the design of a gesture network model are believed to have a strong potential for successful applications to other related problems such as sign language recognition although it is a bit more complicated requiring analysis of hand shapes.

1. Introduction

Since Johansson's work on human motion perception and analysis [1], many researchers in computer vision have tried to analyze and understand human motion in video. Aggarwal and Cai reviewed literatures related to human motion analysis. In the paper, they divided human motion analysis into three areas, i.e. body structure analysis, tracking, and recognition, and addressed the relationships among these areas [2]. In this paper, we focus on the recognition of human hand motions occurring in a video sequence. Pavlovic et al. [3] surveyed problems and issues in visual hands gestures. To date a large body of literatures focuses on isolated hand gesture recognition [4–10], whereas only a small number of works dealt with detecting and recognizing hand gestures from video frames [11–14].

A hand gesture can be described by a locus of hand motion recorded in a sequence of signal frames. To model these signals *hidden Markov models* (HMMs) have been widely accepted as the choice of the models with applications to video analysis problems, such as recognizing tennis motions [15], identifying humans by their gaits [16,17], browsing PowerPoint™ slides using hand commands [11] and so on. Brand et al. suggested a coupled HMM that combines two HMMs with causal, possibly asymmetric link to recognize three T'ai Chi gestures [4].

Recently, there has been an increasing interest in a more general class of probabilistic models, called the *dynamic Bayesian network* (DBN), which includes HMMs and Kalman filters as special cases. DBN is a generalized version of the Bayesian network (BN) with an extension to temporal dimension. Du et al. defined five classes of interactions that could happen between two persons and developed a DBN-based model which took local features such as contour, moment, height and global features such as velocity, orientation, distance as observations [18]. Park et al. employed a DBN to analyze the change of the poses of body parts and recognized the interaction between two persons [19]. Avilés-Arriaga et al. extracted the region and the center of a hand as input features and used a näive DBN to recognize 10 one-hand gestures [20].

Early on, Pavlovic proposed the use of DBN for gesture recognition that can be seen as a combination of an HMM and a dynamic linear system [5]. Wilson also presented modeling techniques to adapt gesture models using the DBN scheme [6]. Yang et al. used time-delayed neural network to analyze feature vectors from hand trajectories [7]. With 40 different isolated signs they achieved a recognition rate up to 96.21%. Nefina et al. compared several different methods of audio-visual speech recognition and suggested the use of coupled HMMs and factorial HMMs by showing that coupled HMMs outperformed all the other models in the performance of recognition [21]. The coupled HMM will be compared with the proposed model in our experiments.

☆ A preliminary partial version of the paper has been presented in the 19th IAPR International Conference on Pattern Recognition, Tampa, USA, December 2008, and the 8th IEEE International Conference on Automatic Face and Gesture Recognition, Amsterdam, The Netherlands, September 2008.

* Corresponding author. Tel.: +82 51 629 6256; fax: +82 51 629 6210.
E-mail addresses: hisuk@image.korea.ac.kr (H.-I. Suk), bkshin@pknu.ac.kr (B.-K. Sin), swlee@image.korea.ac.kr (S.-W. Lee).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.03.016

These previous works considered recognizing isolated gestures rather than spotting gestures in a continuous stream of motion. León et al., on the other hand, used a sliding window of 10 frames to represent the local trajectory with 10 observation nodes in a BN [13]. They showed that, even though some of the observations were missing, the method could still distinguish similar gestures such as "Good-bye" from "Move-Right." Shi et al. considered to segment and recognize human activities from a continuous action stream and presented a semi-Markov model [22]. Voglar and Metaxas proposed a framework for recognition of an American sign language based on an HMM. In their experiment, they extracted the signer's arm and hand motion information using three video cameras and an electromagnetic tracking system. The method achieved a recognition rate of 94.5% in isolated single signs and 84.5% in whole sentences [12]. Recently, Yang et al. proposed a threshold model which extends conventional conditional random fields model to dealing with the task of spotting and recognizing American signs in a set of vocabulary [23].

In this paper, we propose a dynamic Bayesian network model for hand gesture recognition that can be used to control media players or slide presentation. Unlike previous systems the proposed model accepts both one and two hand gestures. Given a video sequence, it captures the hand motion trajectories and relations to the face. They are converted to time series signals, and analyzed by gesture models. In experiments with 10 isolated gestures, the proposed model achieved a recognition rate of 99.59% with cross validation. In addition, a more practical problem of continuous gesture recognition is addressed based on a cyclic spotting network connecting gesture DBNs. To simultaneously recognize gestures and detect the start and end points of embedded gestures in a sequence of motion signals we developed a Viterbi-like dynamic programming method. A test on long videos showed 84% in recall and 80.77% in precision.

In the rest of the paper, we will define 10 hand gestures and describe the methods of detecting and tracking hands, and describe features in Section 2. The proposed hand gesture recognition model and the inference and learning algorithm are explained in Section 3. Section 4 presents a circular network model for continuous gesture motion spotting and recognition for practical applications. The experimental results are presented and analyzed in Section 5. Finally Section 6 concludes the paper.

A preliminary partial version of this paper appeared in [24,25] with limited scope of isolated gesture classification. The main contribution of the current work compared to the previously published conference papers is that it analyzes the results of isolated gesture recognition by decoding the hidden states in DBNs. We then further extend the DBN-based hand gesture model to deal with continuous gestures stream by designing a gesture network model and developing a Viterbi-like dynamic programming method for more practical applications. The proposed gesture network model can detect the start and end points of the embedded meaningful gestures in a video stream. We also demonstrate many experimental results both on isolated and continuous gestures recognition.

2. Hands tracking and feature extraction

Successful dynamic hand gesture recognition requires accurate location of hands and face in space-time. The result of this step influences greatly on the performance of the target system.

2.1. Hand gesture classes

For potential applications to controlling media players or slide presentation, we define 10 different hands gestures including five two-hand gestures and five one-hand gestures as shown in Fig. 1. Each black dot in the figure represents the starting position of the hand and each directed curve the motion trajectory of the hands. In the case of one hand gestures the remaining hand not participating in the gesture may or may not appear in video frames.

2.2. Tracking

Hand detection and tracking in video, though simple, is by no means an easy task due to noise, uncertainty, and the variation in illumination conditions [26].

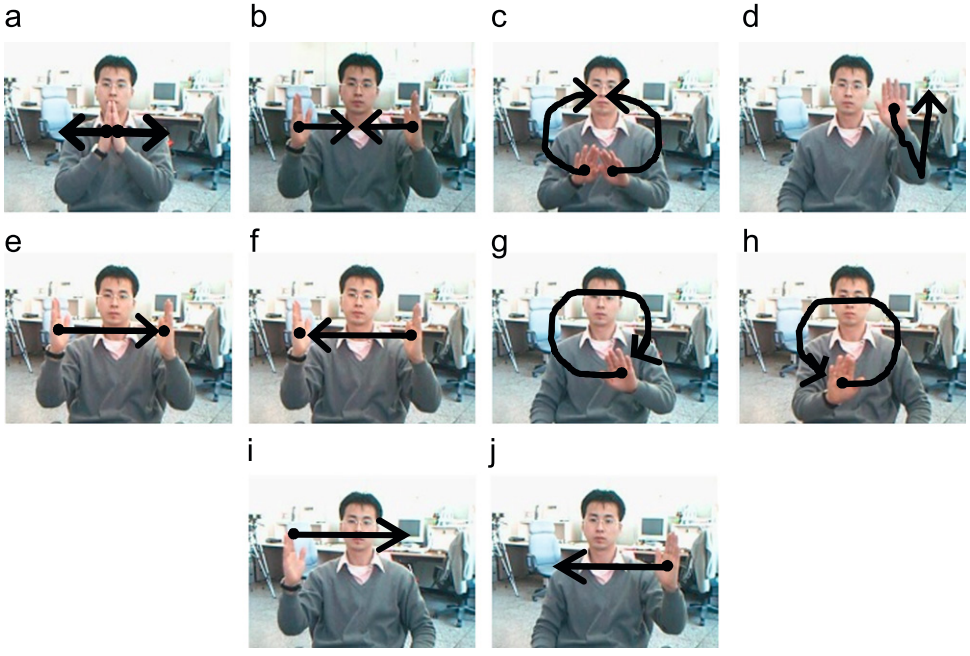

Fig. 1. Ten hands gestures: (a) open a file (OP), (b) close a file (CL), (c) play (PL), (d) pause (PA), (e) move to the last frame (ML), (f) move to the first frame (MF), (g) 10 seconds forward (TF), (h) 10 seconds backward (TB), (i) fast forward (FF), and (j) fast rewind (FR).