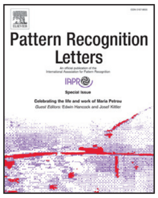
Pattern Recognition Letters 78 (2016) 28–35

Contents lists available at ScienceDirect

Pattern Recognition Letters

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



Continuous sign language recognition using level building based on fast hidden Markov model[☆]

Wenwen Yang, Jinxu Tao[*], Zhongfu Ye

National Engineering Laboratory for Speech and Language Information and Processing, Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China

ARTICLE INFO

Article history:
Received 13 August 2015
Available online 16 April 2016

Keywords:
Continuous sign language recognition
Level Building
Fast hidden Markov model
Grammar constraint
Sign length constrain

ABSTRACT

Sign sequence segmentation and sign recognition are two main problems in continuous sign language recognition (CSLR) system. In recent years, dynamic time warping based Level Building (LB-DTW) algorithm has successfully dealt with both two challenges simultaneously. However, there still exists two crucial problems in LB-DTW: low recognition performance due to bad similarity function and offline due to high computation. In this paper, we use hidden Markov model (HMM) to calculate the similarity between the sign model and testing sequence, and a fast algorithm for computing the likelihood of HMM is proposed to reduce the computation complexity. Furthermore, grammar constraint and sign length constraint are employed to improve the recognition rate and a coarse segmentation method is proposed to provide the maximal level number. In experiments with a KINECT dataset of Chinese sign language containing 100 sentences composed of 5 signs each, the proposed method shows superior recognition performance and lower computation compared to other existing techniques.

1. Introduction

In recent years, there has been increasing interest in developing automatic sign language recognition (SLR) systems to enhance communication between normal hearing and deaf people. Generally, this kind of systems most focus on the manual aspect of signs and recognize hand configurations including hand shape, position, orientation and movement. The systematic change of these hand configurations produces massive different signs, which are well-defined in Chinese sign language dictionaries. Usually, there are three levels of sign language recognition: finger spelling (alphabets), isolated words, and continuous sign language (sentences), while main researches focus on the latter two. In the isolate sign language recognition (ISLR), a hand gesture is a sequence with fixed starting/ending points as the sign boundary, while there is no such explicit sign boundary in CSLR. In the CSLR, a sentence sequence consists of several true-sign sequences and non-sign sequences, also called movement epenthesis (ME), which connect the end location of the previous sign to the start location of the next sign. The task of ISR is to label sign with right sign label, while the main tasks of the CSLR are: (1) splitting sentence sequence into true-sign sequences and movement epenthesis sequences; (2) labeling each true-sign sequence with right sign label. To handle these crucial problems of ISLR and CSLR, many statistical methods or machine learning methods are proposed and employed.

Towards the ISLR, HMM [9,14,22], conditional random field (CRF) [10] and dynamic time warping (DTW) [18] are mainstream methods. Otherwise, convolutional neural network (CNN) [23] and deep neural network (DNN) [20] have been applied in SLR. Starner and Pentland [2] utilize HMM training sign model with feature vector consisting of each hand's x and y position, eccentricity of the bounding ellipse and angle of axis of least inertia in American sign language (ASL) recognition system. For large-scale ASL applications, Vogler and Metaxas [3] first break down sign sequences into phonemes with Movement–Hold model, then utilize parallel HMM to model movement and hand shape features for each phonemes respectively, and combine the probability of each channel as final output in recognition stage. Product-HMM, a variant of multistream HMM, is employed for fusing movement and shape information in the Greek Sign Language and achieves higher performance than parallel HMM [7]. Inspired by Movement–Hold model, Theodorakis et al. [8] propose a phonetic modeling framework for sign language recognition based on dynamic–static (D–S) subunits. Firstly, signs are segmented into dynamic or static segments, which are clustered to construct D–S subunits. Then parallel HMM are utilized to model these D–S subunits. Further experiments on Boston University ASL, Greek SL lemmas and ASL Large Vocabulary Dictionary show the effectiveness of their method. Sminchisescu et al. [5] use CRF to recognize human motion, which

[☆] This paper has been recommended for acceptance by Maria De Marsico.

[*] Corresponding author. Tel.: +86 551 63601329, +8615656910022.
E-mail address: tjingx@ustc.edu.cn, yww2013@mail.ustc.edu.cn (J. Tao).

http://dx.doi.org/10.1016/j.patrec.2016.03.030
0167-8655/© 2016 Elsevier B.V. All rights reserved.

outperforms HMM. For incorporating hidden structures of gesture sequences, Wang et al. [6] propose hidden state conditional random field (HCRF), a discriminative hidden-state approach for the recognition of gestures, which outperforms CRF. Lichtenauer et al. [17] utilize statistical DTW only for time warping, and a combined statistical classifier is employed to model signs. Pigou et al. [23] utilize two CNNs to extract features, one for hand features and one for upper body features. And artificial neural network (ANN) is employed as a classifier of CNN. Wu and Shao [20] utilize a deep dynamic neural networks (DDNN) for gesture segmentation and recognition in the ChaLearn Looking at People 2014 challenge.

Towards the CSLR, mainstream methods are based on HMM, CRF and DTW. Lee and Kim [4] propose a threshold-model with HMM for ME which calculates the adaptive likelihood threshold of an input pattern. The result show that the proposed method can successfully extract trained gestures from continuous hand motion with a 93.14% reliability, when testing sentence consists of ten hand gestures. Fang and Gao [11] propose a simple recurrent networks/HMMs (SRNs/HMMs) for signer-independent CSLR, where SRN is used as soft segmentation of continuous sign language. The system obtains a 85% accuracy in recognizing 100 sentences from seven signers on a vocabulary of 208 signs. Fang et al. [12] utilize a transition movement model (TMM) to handle ME in large-vocabulary CSLR. Testing on 1500 sentences composed of 5113 Chinese signs yields an average accuracy of 91.9%. However, the data is acquired by the data Glove. Kelly et al. [13] also propose a parallel HMM threshold model to handle ME based on the threshold HMM (T-HMM). Yang et al. [15] adopt an enhanced Level Building algorithm to simultaneously segment and match signs to the testing continuous sentence. With the trigram grammar constraint, the system obtains 83% recognition rate in sentence level. Based on Microsoft depth camera KINECT, Zafrulla et al. [16] construct an American sign language recognition system for deaf children education games, where HMM is employed to depict each sign. Kong and Ranganath [19] propose a segment-based probabilistic approach to robustly recognize continuous sign language sentence. Firstly, the sentences are segmented into sign or ME sub-segments by utilizing Bayesian network fusing the outputs of CRF and support vector machine (SVM). Then a sign sub-segments are merged and recognized by a two-layer CRF classifier. Making tests on the data from 8 signers, the system obtains a recall rate of 95.7% and a precision of 96.6% for unseen samples from seen signers, and a recall rate of 86.6% and a precision of 89.9% for unseen signers. This approach achieves 0.8162 score in the gesture spotting challenge. Koller et al. [21] aim at building a real-life continuous sign language recognition system. HMM-based visual models are employed to complete the recognition task together with the class language model and the constrained maximum likelihood linear regression (CMLLR) is utilized to deal with signer-dependency.

In this paper, we only consider the manual part of Chinese sign language signs in our work and work on the problems of Yang's method: enhanced Level Building [15]. In Yang's work, DTW is utilized to calculate the distance between the sign model and the candidate sign sequence at each level, and then search a global optimal matching distance, accordingly yielding the segments and recognition result of the sentence through the backtracking path. So, the distance function is crucial in the whole process. As known, DTW is not the better model in isolated words recognition compared to HMM, which results in low sentence recognition rate. What is worse, the system runs very slowly due to high computation caused by massive search times at each level and calculating DTW during each search. In order to overcome these two problems, HMM is employed to calculate the similarity between sign model and test sign sequence in our paper, since HMM characterizes the sign better than DTW. Then the sign length constraint and grammar constraint are embedded into Level Building recognition process to enhance the recognition performance. Furthermore, to reduce the computation of HMM-based Level Building (LB-HMM), we propose a fast algorithm (called Fast-HMM) to calculate the likelihood of the HMM approximately. In Fast-HMM, given a test sentence sequence with M frames, we firstly calculate the optimal decoding probabilities of M sequences via Viterbi algorithm [1] where the ith decoded sequence is from 1st frame to$i$th frame. Thus, we can get M decoded probabilities for each sign model and this process executes only one time before Level Building. In Level Building, through several basic mathematic operation using M probability values, we can obtain the similarity between the sign model and the arbitrary candidate sign sequence.

The main contributions of this paper are concluded as follows:

1. HMM is embedded into the Level Building algorithm, which will improve the recognition rate at sentence level.
2. Grammar and sign length constraints are employed to reduce substitution, insertion and deletion errors.
3. A fast calculation algorithm is proposed to approximately compute the likelihood of HMM, which reduces the computation of LB-HMM.
4. Coarse segmentation method is employed to obtain the number of levels adaptively for each sentence, not the fixed.

In the following parts of the paper, Section 2 describes the Level Building algorithm based on HMM. We will present fast algorithm for calculating the likelihood of HMM and the coarse segmentation method to decide the number of levels in Section 3. Section 4 presents the experiment results of our method compared with other methods and our conclusion is in Section 5.

2. The Level Building algorithm based on HMM

We declare our notations referring to Yang et al. [15]:

(1) $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{N_\lambda - 1}, \lambda_{N_\lambda})$: the sign model set where true sign models are from λ_1 to $\lambda_{N_\lambda - 1}$ and λ_{N_λ} is a non-sign model, with N_λ as the number of sign model.
(2) T: a M frame sequence composed of several signs.
(3) $\boldsymbol{e}_L = (e_0, e_1, \ldots, e_l, \ldots, e_L)$: a sign boundaries sequence of a query sentence, where e_l is a frame number on which $l$thsign ends and e_0 represents 0th frame.
(4) $\boldsymbol{S}_L = (S_1, \ldots, S_l, \ldots, S_L)$: a sign label sequence where S_l represents one sign model in $\boldsymbol{\lambda}$.
(5) $L_{\max}$: the maximal number of signs in a test sentence, also considered as the level number in Level Building algorithm.
(6) $T(i:j)$: a subsequence of T from frame i to frame j, which is considered as a candidate sign segment in the searching process.
(7) $prob(\lambda_i, T(j:m))$: the probability or likelihood of the subsequence $T(j:m)$ generate by sign model λ_i.
(8) $ll(\lambda_i, T(j:m))$: log of $prob(\lambda_i, T(j:m))$.
(9) $P(\boldsymbol{S}_L, T)$, $\tilde{P}(\boldsymbol{S}_L, T)$: the probability and log probability of the sentence T labeled as $\boldsymbol{S}_L$.

2.1. The Level Building algorithm based on HMM

In this paper, HMM is employed to train the sign model. $\lambda_i = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$ is often employed to indicate the probabilistic parameter of HMM. Here, $\boldsymbol{\pi}$ denotes the vector of the initial probability π_i that hidden state i as starting state. $\mathbf{A}$ stands for the matrix of state transition probabilities a_{ij} that a transition from state i to j. And $\mathbf{B}$ represents the matrix of the observation probability $b_j(\mathbf{O}_t)$ that observation $\mathbf{O}_t$ emitted at time t in state j.

For a test sequence, we often use likelihood to measure the similarity between the test sequence and sign model, where the