Neurocomputing 149 (2015) 1604–1612

Contents lists available at ScienceDirect

Neurocomputing

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

Unsupervised adaptive sign language recognition based on hypothesis comparison guided cross validation and linguistic prior filtering

Yu Zhou [a], Xiaokang Yang [b], Yongzheng Zhang [a,*], Xiang Xu [a], Yipeng Wang [a], Xiujuan Chai [c], Weiyao Lin [b]

[a] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[b] School of Electronic, Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China
[c] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 24 September 2013
Received in revised form
20 May 2014
Accepted 12 August 2014
Communicated by Qingshan Liu
Available online 23 August 2014

Keywords:
Signer adaptation
Cross validation
Unsupervised learning
Sign language recognition

ABSTRACT

Signer adaptation is important for sign language recognition systems because a fixed system cannot perform well on all kinds of signers. In supervised signer adaptation, the labeled adaptation data must be collected explicitly. To skip the data collecting process in signer adaptation, we propose a novel unsupervised adaptation method, namely the hypothesis comparison guided cross validation method. The method not only addresses the problem of the overlap between the data set to be labeled and the data set for adaptation, but also employs an additional hypothesis comparison step to decrease the noise rate of the adaptation data set. We also utilize linguistic prior knowledge to down sample the adaptation data list to further decrease the noise rate. To evaluate the effectiveness of the proposed method, the CASIIE-SL-Database is formed, which is the first specialized data set for unsupervised signer adaptation to the best of our knowledge. Experimental results show that the proposed method can achieve relative word error rate reductions of 3.93% and 4.05% respectively compared with self-teaching method and cross validation method. Though the method is proposed for signer adaptation, it can also be applied to speaker adaptation and writer adaptation directly.

1. Introduction

Sign language recognition (SLR)[1] is an important and challenging task in pattern recognition [2] field. SLR not only benefits the communication between the hearing-impaired people and the hearing people, but also is a good test bed for more general gesture recognition and human–computer interaction research. Since the 1990s, many research works on SLR have been reported [3–11], especially for the signer dependent (SD) case. Though the SD SLR systems have made remarkable advances, the performance is poor when the test signer is different from the training signer. The degradation arises from the large diversity of different signers' signing styles. Collecting data from large number of signers to train a signer independent (SI) model set is an alternative to solve this problem [12–15]. However, the SI model set is not comparable to the SD one in that the SI model set is fixed. The SI model set can have acceptable performance for most signers, but can not have outstanding performance as the SD model set does. Adaptation techniques [16] in speech recognition [17] and handwriting recognition [18] provide an alternative solution to the problem. Adaptation techniques utilize some labeled and/or unlabeled data from a new person to tailor the parameters of the original model set so that the tailored model set can better model the new person. For signer adaptation, the tailored model set is called signer adapted (SA) model set. If the labels of the adaptation data are available, the adaptation is called supervised adaptation; if the labels are unavailable, the adaptation is called unsupervised adaptation. The classic adaptation algorithms include eigenvoice (EV) [19], maximum likelihood linear regression (MLLR) [20] and maximum a posteriori (MAP) [21]. EV is fit for the case that the SI model set is trained by relatively large number of training subjects. MLLR is fit for rapid adaptation with small amount of adaptation data. MAP is fit for incremental adaptation with large amount of adaptation data. MAP is superior to other algorithms in that the SA model set generated by MAP can be comparable with the SD model set in performance if the amount of adaptation data is large enough.

To the best of our knowledge, several signer adaptation works have been reported in the literature. Ong et al. [22,23] used a two-step classifier to recognize sign language words. The first step was the channel-level classification, and at the second step they used

* Corresponding author.
E-mail address: zhangyongzheng@iie.ac.cn (Y. Zhang).

[1] An earlier version of this paper was presented at the IEEE International Conference on Multimedia & Expo 2011 [1].

http://dx.doi.org/10.1016/j.neucom.2014.08.032
0925-2312/

a Bayesian network framework to combine the channel classification results. Since the signer combined model set gave low recognition rate on the data from the test signer, they implemented an adaptation scheme to obtain a model set that were close to the SD one. MAP was adopted as the adaptation algorithm. Other than taking the signer combined model set as the initial model set, they first selected a model set that gave the best recognition accuracy for the adaptation data from the training subjects' model sets. By this way more accurate prior distribution for the parameters could be estimated. On the vocabulary of 20 gestures the recognition rate for the test signer was 88.5%, which was a 75.7% reduction in error rate from the one of the signer combined mode set. Agris et al. [24,25] combined MLLR and MAP for signer adaptation. On the MLLR step, they modified MLLR with dedicated one-hand transformation. With 80 and 160 labeled sign samples, they achieved 78.6% and 94.6% accuracy respectively on a vocabulary of 153 signs. In their succeeding work [26], they proposed the "EV + MLLP + MAP" approach. The model set that outputted by EV was taken as the initial model set for MLLR, and the resulted model set by MLLR was taken as the input model set for MAP. By this way they combined the benefits of these three algorithms, hence yielded rapid adaptation speed and slow performance saturation. Experiments showed that the "EV + MLLR + MAP" method was superior to "MLLR + MAP" method. To reduce the number of samples in the adaptation data set and preserve the recognition rate to an acceptable level at the same time, Wang et al. [27] presented a supervised adaptive method based on data generating. They analyzed the structure of the Chinese sign language words and realized that the samples of the whole vocabulary can be generated by the samples of a subset. Their idea was that different words shared some similar segments in part, and the data pooling among different word segments can be utilized. About 2/3 reduction of the size of the adaptation data set was achieved, and acceptable recognition accuracy was obtained at the same time. Oya et al. [28] proposed a multi-class classification strategy for Fisher scores and applied it to sign language recognition. In their score space selection step, they found that the subsets selected from the validation data set did not perform well on the test data set. Therefore they proposed a signer adaptation method during the score space selection phase. By applying the adaptation method on the multi-class classification strategy they achieved an average recognition rate of 68.35%.

In summary, previous signer adaptation works mainly focus on the supervised signer adaptation. In supervised adaptation, an explicit enrollment session for labeled adaptation data collecting is required. The enrollment session is usually tedious to users. At the same time, unlabeled data can be collected more easily and the data collecting does not need the user's intervention. For example, the data that are produced when users are manipulating the system can be automatically stored to the hard disks, and the stored data can be used as the unlabeled adaptation data to adapt the original model set. Therefore, unsupervised signer adaptation is more important and useful to SLR than the supervised one. In [1], we propose a novel unsupervised signer adaptation method called the hypothesis comparison guided cross validation (HCCV) adaptation method, which is a preliminary version of this paper. Compared with [1], our contributions in this paper are summarized as follows. First, we utilize the linguistic prior knowledge to filter the adaptation data list, with which the noise rate of the adaptation data set can be decreased further. Secondly, inspired by [29], we introduce a global model set which is served as the final model set to recognise the test data set. As a result, we do not need to combine the recognition results of multi model sets. Finally, we create a sign language database named CASIIE-SL-Database, which is the first database that is specifically designed for the unsupervised signer adaptation research to the best of our knowledge. In the future, we expect this database may be used as a potential standardized benchmark for unsupervised signer adaptation, and can boom the unsupervised signer adaptation research.

The organization of this paper is as follows. In Section 2 we review the three classic adaptation algorithms, and MAP is finally selected as the algorithm for our work. The next section describes the proposed hypothesis comparison guided cross validation adaptation method. CASIIE-SL-Database, the experimental setup and results are described in Section 4. Finally the conclusion and potential future works are given in the last section.

2. Selection of the adaptation algorithm

There are three classical adaptation algorithms, namely EV, MLLR and MAP.

EV is based on the dimensionality reduction techniques. First, N SD model sets should be trained with the data from N signers. A supervector for each model set can be constructed by concatenating the Gaussian mean vectors of all the models in the set, and totally N supervectors can be generated. By applying the dimensionality reduction techniques (for example, principle component analysis) on the N supervectors, we can yield M principal supervectors, which are the basis vectors and are called eigenvoices.[2] The supervector of the new signer's model set can be represented by weighing the M eigenvoices. The parameters that need to be estimated in EV are the M weights only, where M is much smaller than the number of the model set's parameters. As a result, the adaptation speed with EV is rapid and only a small amount of adaptation data are needed. However, EV has two limitations for signer adaptation. First, in signer adaptation EV needs relatively large numbers of signers to collect data in order to train the SD model sets, which is much more difficult than that in speaker adaptation. Moreover, EV saturates fast, so even large amounts of adaptation data are available, the recognition rate cannot be increased to a very high level as in the SD case.

MLLR supposes the adapted mean vectors can be transformed with a set of transformation matrices from the initial mean vectors:

$$\tilde{m} = Wm + b \tag{1}$$

where $\tilde{m}$ is the adapted mean vector, m is the initial mean vector, W is the transformation matrix, and b is the bias vector. MLLR utilizes a regression class tree to share adaptation data among different mean vectors. The regression class tree is a binary tree, and each leaf node is a base class and corresponds to a cluster of the mean vectors that are similar. The regression class tree can be constructed by linguistics or by data-driven manners. The mean vectors in the same node share the same transformation matrix. The regression class tree makes MLLR flexible. If the amount of adaptation data is large, and all leaf nodes have enough data, a transformation matrix can be estimated for each leaf node. If the amount of the adaptation data is small, and only the nodes close to the root have enough data, some leaf nodes may share the same transformation matrix. The estimation of the transformation matrices is based on the maximum likelihood criterion. MLLR can also rapidly adapt the model set by a small amount of data, but it still falls into the problem of fast saturation and the performance is not comparable to that of the SD case.

MAP, which is also denoted Bayesian adaptation, involves the use of prior knowledge about the model parameter distribution. The informative priors that are generally used are from parameters of the SI model set. If the conjugate priors are used, MAP results in

[2] EV is named eigen***voice*** because it is originated from the speech recognition field.