HYPOTHESIS COMPARISON GUIDED CROSS VALIDATION FOR UNSUPERVISED SIGNER ADAPTATION

Yu Zhou\textsuperscript{1}, Xiaokang Yang\textsuperscript{1}, Weiyao Lin\textsuperscript{1}, Yi Xu\textsuperscript{1}, Long Xu\textsuperscript{2}

\textsuperscript{1}Institute of Image Communication and Information Processing, Shanghai Jiao Tong University
\textsuperscript{2}Department of Computer Science, City University of Hong Kong
\{eeyzhou, xkyang, wylin, xuyi\}@sjtu.edu.cn, lxu222@cityu.edu.hk

ABSTRACT
Signer adaptation is important to sign language recognition systems in that a one-size-fits-all model set can not perform well on all kinds of signers. Supervised signer adaptation must utilize the labeled adaptation data that are collected explicitly. To skip the data collecting process in signer adaptation, we propose an unsupervised adaptation method called hypothesis comparison guided cross validation (HC-CV) algorithm. The algorithm not only addresses the problem of 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. Experimental results show that the HC-CV adaptation algorithm is superior to the CV adaptation algorithm and the conventional self-teaching algorithm. Though the algorithm is proposed for signer adaptation, it can also be applied to speaker adaptation and writer adaptation straightforwardly.

Index Terms—Unsupervised signer adaptation, cross validation, sign language recognition, maximum a posteriori

1. INTRODUCTION
Since the 1990s, many research works on sign language recognition (SLR) have been reported \cite{1}, especially for the signer dependent (SD) case. Though the SD systems have been made remarkable advances, the performance is poor when the test signer is different with the training signer. The degradation arises from the large diversity of different signers' signing styles. Adaptation techniques supply an alternative solution to the problem, and several signer adaptation works have been done. Ong et al. \cite{2} applied supervised maximum a posteriori (MAP, \cite{3}) to adapt their system and yielded 88.5% accuracy on a vocabulary of 20 gestures. von Agris et al. \cite{4} combined maximum likelihood linear regression (MLLR, \cite{5}) and MAP for signer adaptation. With 80 and 160 labeled signs, they achieved 78.6% and 94.6% accuracy respectively on a vocabulary of 153 signs. In their latest work \cite{6}, they proposed the eigenvoice + MLLP + MAP approach. Wang et al. \cite{7} presented a supervised adaptive method based on data generating, in which they reduced the size of adaptation data set from 350 to 136 with an acceptable recognition rate.

In summary, previous signer adaptation works mainly focus on the supervised signer adaptation, in which an explicit enrollment session for labeled adaptation data collecting is required. Since the enrollment session is usually boring to users, unsupervised signer adaptation is of great significance. However, to the best of our knowledge, only von Agris addressed the unsupervised signer adaptation problem in \cite{4}, and achieved 72.9% recognition rate with the ideal confidence measure.

In this paper, we propose a novel unsupervised signer adaptation algorithm called hypothesis comparison guided cross validation (HC-CV) adaptation. The organization of this paper is as follows. The self-teaching and the CV adaptation algorithms are reviewed in Section 2, and the HC-CV algorithm is proposed in Section 3. Experimental setup and results are described in Section 4. Finally the conclusion is given and the future works are presented.

2. THE SELF-TEACHING AND CROSS VALIDATION ADAPTATION ALGORITHMS
A straightforward method for unsupervised adaptation is the self-teaching algorithm. Self-teaching includes two steps: in the decoding step, the models to be adapted are used to generate the hypothesis for unlabeled data; in the adaptation step, a supervised adaptation algorithm takes the model set and the data with the hypothesis as inputs and outputs the adapted models. Though self-teaching can improve the performance to some extent, it suffers from the error reinforcement problem. A wrongly labeled sample is used to tailor the corresponding model parameters, and the deviated model will label the same sample more wrongly. The error is reinforced iteration by iteration.
To solve the error reinforcement problem, Shinozaki et al. [8] proposed the CV adaptation algorithm. The CV adaptation is based on the idea of cross validation. By this way the data overlap between the decoding step and the adaptation step can be avoided. First, $K$ CV model sets are generated by copying the initial model set. Afterwards, the unlabeled data are randomly and evenly divided to $K$ unlabeled (U) subsets with each corresponding to a CV model set. Each CV model set decodes its U subset to get a hypothesis, and the U subset and the hypothesis form the “labeled” (L) subset. Later on, we copy all L subsets other than its own one to construct the adaptation (A) data set for each CV model, and adapt each CV model set with its A data set. If the number of maximum iteration does not arrive, the $K$ adapted model sets are copied as the CV model sets for the next iteration, and the labeling and the adaptation processes are repeated. Finally the test data set is decoded with all CV model sets, and the final results are obtained by combining all the CV results.

3. THE HYPOTHESIS COMPARISON GUIDED CROSS VALIDATION ADAPTATION ALGORITHM

The CV adaptation algorithm is superior to the conventional self-teaching method, but the “labeled” data sets are directly copied to the adaptation data set, which may lead to high noise rate in the adaptation data set. In [9][10], they show that if there are $N$ data samples, which satisfies:

$$N \geq 2 \cdot \ln \left(2 \cdot \frac{H}{\delta} \cdot \left[\frac{e^2 \cdot (1 - 2\eta)^2}{\eta} \right] \right)^{-1}$$  \hspace{1cm} (1)

where $\varepsilon$ is the upper bound of classification error rate, $\eta$ is the upper bound of noise rate, $\overline{H}$ is the number of hypothesis, $\delta$ is the confidence, then a specific hypothesis $H_i$ that minimizes disagreement with $\delta$ will be PAC learnable:

$$\Pr \left[ d(H_i, H_i) \geq \varepsilon \right] \leq \delta$$  \hspace{1cm} (2)

If a constant $c$ that makes (1) hold equality is selected, then:

$$\varepsilon^2 \cdot N \cdot (1 - 2\eta)^2 = c$$  \hspace{1cm} (3)

which indicates that if the noise rate ($\eta$) decreases and (or) the number of the training set ($N$) increases, the classifier’s classification error rate ($\varepsilon$) will be decreased, which means a better generalization ability.

To decrease the noise rate of the adaptation dataset we propose the HC-CV adaptation algorithm. The HC-CV

Fig. 1. Illustration of HC-CV adaptation algorithm. We select 5-fold CV adaptation as the example. For clear illustration we copy the No. 5 items at the most right to the most left, and the copied items are all dashed. “M” represents “the model set”, “U” represents “the unlabeled dataset”, “L” represents “the labeled dataset”, “SL” represents “the same labeled dataset”, “Fr” represents “from”, “A” represents “the adaptation dataset.”
adaptation employs a hypothesis-comparing step to select more reliable “labeled” data as the adaptation data. The HC-CV adaptation is illustrated in Fig. 1, and the procedure is as follows:

1. Generate K CV model sets by copying the initial model set. Randomly and evenly divide the unlabeled data to K unlabeled (U) subsets with each corresponding to a CV model set.
2. Each CV model set decodes its own and a neighboring U subset to get hypothesis, then the two U subsets and the hypothesis form two “labeled” (L) subsets.
3. Compare two L subsets from the same U subset, and the data that have the same decoding results by two CV model sets are copied to form the “same labeled” (SL) subset.
4. Copy the SL subset to the adaptation (A) subsets whose corresponding CV models do not decode the SL subset’s corresponding U sets;
5. Adapt each CV model set with its A data set.
6. If the number of maximum iteration does not arrive, go to step 7, or else go to step 8.
7. Copy the K adapted model sets as the K CV model sets for the next iteration, go to step 2.
8. Recognize the test data set with all CV model sets, and combine them by probability normalization to get the final results.

At the first iteration, all the CV model sets are same. After several iterations, the CV model sets become diverse because their adaptation data sets are different from each other. The final recognition results are obtained by combining all model sets’ decoding hypothesis. Let \( O \) denotes a sequence to be decoded, then \( N \) probabilities \( \{ P(W_i|O) | i = 1,2,…,N \} \) can be obtained for each CV model set. For each set of probabilities, we normalize them by:

\[
P'(W_i|O) = \frac{P(W_i|O) - MinP}{MaxP - MinP}\]

where MaxP and MinP are the maximal and the minimal probabilities respectively. Then, we add all the normalized probabilities corresponding to the same word as the final probabilities. The word with the largest normalized probability is the final recognition result. Because the adaptation data sets of the CV model sets in HC-CV algorithm are more diverse than those in CV algorithm, the combination can achieve more effective results.

4. EXPERIMENTAL RESULTS

Since validating the CV and HC-CV methods needs large repetitions of samples for each gesture, and such dataset is not available to the best of our knowledge, we form our own dataset. We use a camera to collect the gesture data. Because we focus on the adaptation module, the vocabulary and the feature extraction are simplified. The vocabulary is composed of 10 gestures, each of which is a number with 3 connected digits. The digit is signed by Chinese spelling way. We extract eight features from these gestures: the area, the circumference, the length of two axes of the ellipse to fit the gesture region and their derivatives. Experimental data set consists of 1200 samples over 10 gestures and 5 subjects. Among the 5 subjects, 4 of them are selected as the training subjects, and each of them signs 5 times of each gesture. Since collecting data is time consuming and tedious, only the fifth subject is taken as the test subject. The test subject signs 100 times of each gesture. The samples of 20 groups of the test subject are taken as the test data set, and the samples of the other 80 groups are taken as the adaptation data set. Till now, HMM is still the most popular model for SLR in real application, so HMM is taken as the statistical model. Each gesture is modeled by a 3-state Bakis HMM, and the observation probability distribution is unimodal multivariate Gaussian. Because the mean vectors are the most important for adaptation [2][4], only the mean vectors are adapted in our work. HTK [11] is used to build the models and to adapt the models’ mean vectors.

At the adaptation step, the MAP [3] algorithm is adopted. For CV adaptation and HC-CV adaptation, the problem of bias adaptation may appear. As shown in Figure 1, the adaptation data sets for CV model sets may consist of unbalanced adaptation data for all classes. For example, for some classes the number of adaptation data may be very large, and for other classes the number may be very small. Therefore some models are weakly adapted, which will affect the next decoding step. We use MAP/VFS [12] to smooth the model set.

Fig. 2 shows the comparison of HC-CV adaptation result with the self-teaching (ST) result and the CV adaptation result. The horizontal axis represents the number of iterations, and the vertical axis represents the recognition rate or the labeling rate. The recognition rate of the SI models is 68%, and the labeling rate is 60%. If all the labels of the adaptation data are available, the models with the supervised adaptation can achieve 93.5% recognition rate. From figure 2.a we can see that both the CV adaptation and HC-CV adaptation obtained better recognition results than self-teaching adaptation after several times of iteration. This proves that the separation of decoding data set and adaptation data set is effective. At the first iteration, because the CV model sets in HC-CV are all the same copies of the initial model set, then the two decoding results for the same unlabeled sample are equal. Therefore all data are sent to the adaptation data set. But the number of adaptation data for HC-CV is less than the corresponding one of CV, so the recognition rate of HC-CV is a little less than the rate of CV. Starting from the second iteration, as can be seen from figure 2.b that the noise rate of the HC-CV adaptation data set is decreased by comparing the decoding results of two CV models, so the recognition rate is higher than that of CV adaptation. At the 5th iteration, the recognition rate of the HC-HC adapted models to the test data set can arrive at 88.5%. Considering that in the supervised case the
What is worth mentioning is that in real application collecting the unlabeled data is feasible, and the number of data can be very large. Therefore the number of fold of CV can be large too. For a specific adaptation sub set, we can use 3 or more model sets to decode it, and the noise rate can be decreased further by comparing multiple labeling results.

5. CONCLUSION

To address the problem of unsupervised adaptation in SLR, we propose the HC-CV adaptation algorithm. The novel algorithm not only separates the dataset to be labeled and the dataset for adaptation but also utilizes a comparing procedure to filter out some wrongly labeled samples to decrease the noise rate for the adaptation dataset. Experimental results show the effectiveness of the algorithm. Though the improvement has been achieved, there are still some problems to be addressed in our future works. For example, the performance is uncontrollable after some times of iteration; the recognition rate can not be improved much if the initial recognition rate of the SI model set is low. How to decide the number of iteration and the relationship between the SI recognition rate and the performance of adaptation are the main focus of our future work. The algorithm is independent of the SLR or gesture recognition, so it can be directly applied to speaker adaptation or writer adaptation.

6. REFERENCES