Pattern Recognition 46 (2013) 2313–2322

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

A discriminative linear regression approach to adaptation of multi-prototype based classifiers and its applications for Chinese OCR

Jun Du [a,*], Qiang Huo [b]

[a] National Engineering Laboratory for Speech and Language Information Processing (NEL-SLIP), University of Science and Technology of China, No. 96, JinZhai Road, Hefei, Anhui, PR China
[b] Microsoft Research Asia, 13/F, Building 2, No. 5 Danling Street, Haidian District, Beijing, PR China

ARTICLE INFO

Article history:
Received 18 April 2012
Received in revised form
9 January 2013
Accepted 14 January 2013
Available online 23 January 2013

Keywords:
Discriminative linear regression
Sample separation margin
Minimum classification error
Rprop
Adaptation
OCR

ABSTRACT

This paper presents a new discriminative linear regression approach to adaptation of a discriminatively trained prototype-based classifier for Chinese OCR. A so-called sample separation margin based minimum classification error criterion is used in both classifier training and adaptation, while an Rprop algorithm is used for optimizing the objective function. Formulations for both model-space and feature-space adaptation are presented. The effectiveness of the proposed approach is confirmed by a series of experiments for adaptation of font styles and low-quality text, respectively.

1. Introduction

With the fast development of mobile internet, OCR-based applications are becoming increasingly more popular (e.g., [1–4]). However, the most off-the-shelf OCR engines were trained on scanned documents, and they may not work well for new application scenarios where the properties of the captured character images are significantly different from the ones in the training data set (e.g., [5]). One of the solutions to address this problem is to adapt a pre-trained classifier to deal with the new scenario by using the document to be recognized itself via an *unsupervised adaptation* strategy, or by using a small amount of adaptation data collected in the target scenario via a *supervised adaptation* strategy.

Using adaptation to improve the OCR accuracy has been a research topic for several decades. Some ingenious ideas specific to OCR have been tried. For example, Nagy et al. [6,7] demonstrated that a character classifier trained on many typefaces can be adapted effectively to text in a single unknown typeface by using a *self-adaptation* strategy. Hong [8] showed how to use an adaptation strategy that alternates between applying "visual constraints" and "linguistic constraints" to reduce errors for recognizing books printed in a single typeface. Authors in [9–13] investigated a family of style-conscious algorithms to improve the recognition accuracy on documents which contain only a few typefaces and limited variations in image qualities and other variabilities. More recently, Xiu and Baird [14] demonstrated how to use a self-adaptation technique to improve the whole-book recognition, where a so-called iconic model and a linguistic model are mutually leveraged and adapted.

In the past several decades, many effective adaptation techniques have also been invented for speaker adaptation in the area of speech recognition (see an overview paper [15] and the references therein) and for writer adaptation in the area of handwriting recognition (see a recent work in [16,17] and the references therein). Some of them have been applied to OCR adaptation. For example, in [18], a continuous density hidden Markov model (CDHMM) based English character recognizer was adapted to deal with unseen fonts with two popular adaptation algorithms, namely maximum-likelihood linear regression (MLLR) [19,20] and maximum *a posteriori* (MAP) estimation [21], which were developed originally in speech recognition area for speaker adaptation. Apparently, the style transfer mapping (STM) technique proposed in [16,17] for writer adaptation can also be used for OCR adaptation.

In this paper, we study the adaptation techniques for Chinese OCR. One of the state-of-the-art techniques to build a Chinese OCR engine is to use a discriminatively trained prototype-based classifier as reported in [22]. In spite of the large vocabulary of Chinese characters, such a classifier can be made both compact

* Corresponding author. Tel.: +86 551 63607863; fax: +86 551 63607863.
E-mail addresses: jundu@ustc.edu.cn (J. Du), qianghuo@microsoft.com (Q. Huo).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.01.021

(e.g., [22,23]) and efficient in the recognition stage (e.g., [24]). A high recognition accuracy can be achieved by using Gabor features, LDA (linear discriminant analysis) or MCE (minimum classification error) based discriminative feature extraction, and MCE-based classifier parameter training [22]. Recently, a so-called sample separation margin (SSM) based MCE training approach was proposed in [25] for training prototype-based classifiers, which performs better than the MCE training approach in [22]. In [23], the SSM-MCE training approach has been used to construct a state-of-the-art compact prototype-based handwritten Chinese character recognizer where a batch-mode Quickprop algorithm [26] is used for optimizing the SSM-MCE objective function. In [27], the SSM-MCE formulation has been extended to training pattern classifiers with quadratic discriminant functions (QDF), including the "modified quadratic discriminant function (MQDF)" [28] popular in the areas of OCR and handwriting recognition for East Asian languages. In this study, we have built our baseline classifier for Chinese OCR by using the techniques described in [22,25,23] with a minor difference: we used an Rprop algorithm (e.g., [29,30]) to optimize the SSM-MCE objective function because the setting of control parameters is much easier than the Quickprop algorithm used in [23].

The main contribution of this paper is to propose a new SSM-MCE linear regression (LR) approach to adaptation of an SSM-MCE trained prototype-based classifier and demonstrate its effectiveness for Chinese OCR as an illustrative example. Formulations for both model-space and feature-space adaptation are presented. In terms of general concept, our work is related to the MCE-LR approach reported in [31] for speaker adaptation of CDHMM-based speech recognizer, where a traditional MCE objective function is used. Our work is also relevant to the STM work on writer adaptation for handwritten Chinese character recognition reported in [16,17], where a similar MCE training approach as in [22] is used to train a prototype-based classifier, but a least regularized weighted squared error approach is used to estimate a global feature transform (a.k.a. style transfer mapping (STM)) for writer adaptation. The adaptation capability of the STM approach is similar to our feature-space adaptation approach, but is inferior to our model-space adaptation approach because multiple transforms can be used for model adaptation. Even for feature-space approach, our experimental results show that our approach performs significantly better than the original STM approach in [16] for both supervised and unsupervised adaptation of font styles and low-quality text, respectively, which confirms that SSM-MCE is a better objective function to learn the feature transform.

Fig. 1 illustrates an overall system development flow of our work in this paper. In training stage, after feature extraction of training samples, an LBG clustering algorithm [32] is used to construct multiple prototypes for each character class. Then a baseline classifier is constructed by using the SSM-MCE training. For model-space adaptation, an adapted classifier is constructed by using the linear regression transform(s) estimated from the adaptation data and the baseline classifier, which will be used to recognize unknown characters in the target scenario. For feature-space adaptation, a global feature transform is estimated from the adaptation data, which will be used to transform the feature vector of the unknown character back into the feature space of the training data so that the baseline classifier can be used in recognition stage.

It is noted that the preliminary results of this study have been published in [33]. The current paper is an extended version of the above report by including more detailed descriptions of relevant procedures, reporting additional experimental results and findings, and adding new figures and references to make the presentation more readable and accessible.

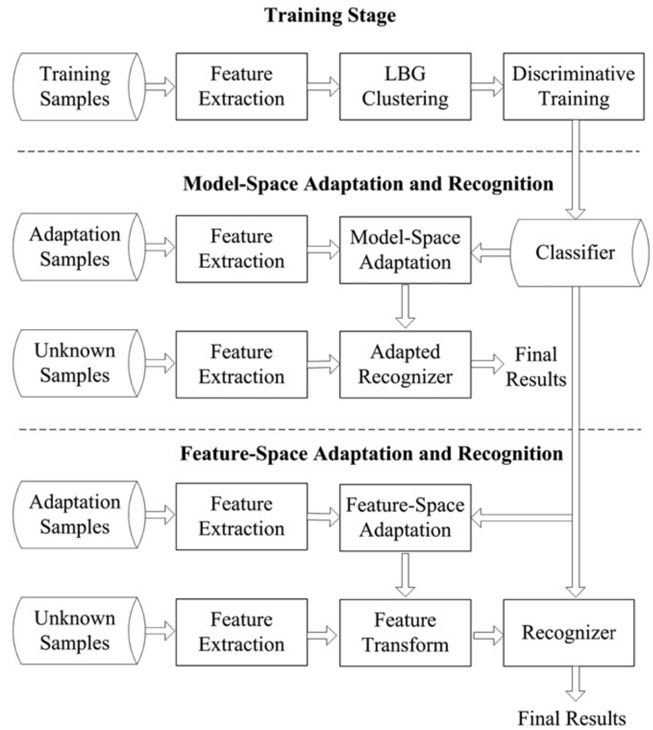

Fig. 1. Overall flow of system development.

The remainder of the paper is organized as follows. In Section 2, we describe briefly how to construct a multi-prototype based classifier by using the SSM-MCE training. In Section 3, we present formulations of SSM-MCE LR for both model-space and feature-space adaptation. Several important implementation issues are discussed in Section 4. In Section 5, we report experimental results for adaptation of font styles and low-quality text, respectively. Finally we conclude the paper in Section 6.

2. SSM-MCE training of a multi-prototype based classifier

Suppose our classifier can recognize M character classes denoted as $\{C_i | i=1,\ldots,M\}$. For a multi-prototype based classifier, each class C_i is represented by K_i prototypes, $\boldsymbol{\lambda}_i=\{\boldsymbol{m}_{ik}\in\mathcal{R}^D | k=1,\ldots,K_i\}$, where $\boldsymbol{m}_{ik}$ is the kth prototype of the ith class. Let us use $\boldsymbol{\Lambda}=\{\boldsymbol{\lambda}_i\}$ to denote the set of prototypes. In the classification stage, a feature vector $\boldsymbol{x}\in\mathcal{R}^D$ is first extracted. Then $\boldsymbol{x}$ is compared with each of the M classes by evaluating a Euclidean distance based discriminant function for each class C_i as follows:

$$g_i(\boldsymbol{x};\boldsymbol{\lambda}_i)=-\min_k\|\boldsymbol{x}-\boldsymbol{m}_{ik}\|^2. \quad (1)$$

The class with the maximum discriminant function score is chosen as the recognized class $r(\boldsymbol{x};\boldsymbol{\Lambda})$, i.e.,

$$r(\boldsymbol{x};\boldsymbol{\Lambda})=\arg\max_i g_i(\boldsymbol{x};\boldsymbol{\lambda}_i). \quad (2)$$

In the training stage, given a set of training data $\boldsymbol{X}=\{\boldsymbol{x}_r\in\mathcal{R}^D | r=1,\ldots,R_1\}$, first we initialize $\boldsymbol{\Lambda}$ by LBG clustering [32]. Then $\boldsymbol{\Lambda}$ can be re-estimated by minimizing the following MCE objective function:

$$l(\boldsymbol{X};\boldsymbol{\Lambda})=\frac{1}{R_1}\sum_{r=1}^{R_1}\frac{1}{1+\exp[-\alpha d(\boldsymbol{x}_r;\boldsymbol{\Lambda})+\beta]}, \quad (3)$$

where α, β are two control parameters, and $d(\boldsymbol{x}_r;\boldsymbol{\Lambda})$ is a misclassification measure defined by using a so-called sample