

Unsupervised writer style adaptation for handwritten word spotting

José A. Rodríguez^{+,*}, Florent Perronnin⁺, Gemma Sánchez^{*}, Josep Lladós^{*}

⁺*Textual and Visual Pattern Analysis (TVPA), Xerox Research Centre Europe, France*

^{*}*Computer Vision Center (CVC), Universitat Autònoma de Barcelona, Spain*

jrodriguez@cvc.uab.es, Florent.Perronnin@xerox.com, gemma@cvc.uab.es, josep@cvc.uab.es

Abstract

We propose a novel approach for writer adaptation in a word spotting task. The method exploits the fact that a semi-continuous hidden Markov model separates the word model parameters into (i) a shared codebook of shapes and (ii) a set of word-specific parameters. Our main contribution is to derive writer-specific word models by statistically adapting an initial universal codebook to each document. This process is unsupervised and does not even require the appearance of the keyword(s) in the searched document. Experimental results show an increase in performance when this adaptation technique is applied. To the best knowledge of the authors, this is the first work dealing with adaptation for word spotting.

1. Introduction

Handwritten word spotting (HWS) is the pattern classification task which consists in detecting keywords in handwritten document images [10]. Two main approaches can be identified. The first approach is the traditional “query-by-example”: given one or multiple examples of a word, find the word images with high similarity to the query [11].

The second approach uses handwritten word recognition (HWR) systems which model words as concatenations of character models. The main advantage of the latter approach is that one can search for any keyword once the character models are trained. However, it requires a large amount of training samples and does not usually work well in degraded or unconstrained documents. For this reason, we chose the former approach and, in the remainder of the paper, we adopt the usual convention that “HWS” only refers to the first type of techniques.

The problem of adapting to the writer style of a document for improving the results has not yet been consid-

ered in HWS. One reason might be that several works are applied to historical documents [10], which are usually produced by a single writer. Another reason is that adaptation is best formulated in a statistical hidden Markov model (HMM) based framework, which is not the most common approach in HWS.

Furthermore, the techniques employed for adaptation in HWR cannot be directly applied to word spotting. In HWR, it is possible to adapt character models on a set of word samples which is different from the samples that will be recognized at test time. Even if such an adaptation set is not available, a first recognition pass can be performed and the transcription can be used as the adaptation set [6]. In contrast, in a spotting application without character models, only labels are available for the keywords to spot. For this reason, we have to introduce an adaptation strategy that does not rely on keyword examples but that can learn from unlabeled data of the writer and incorporate this information in the keyword models.

In handwritten character and word recognition, some simple techniques to cope with writer styles have been experimented, for instance by complementing the training set with samples from the new writer [2] or by training models for each allograph and choosing the best one at test time [5]. However, a more powerful approach in a statistical context is HMM adaptation, successful in speech recognition [7, 9] and handwriting recognition [16, 4]. Here, the speaker/writer-independent set of parameters θ is transformed to θ_{ad} using a (relatively) small amount of data from the corresponding speaker/writer. We work using this formulation because it is suitable to our HMM modeling of words.

Therefore, we propose a novel unsupervised adaptation method in which keywords are modeled with a semi-continuous HMM (SC-HMM) [3]. This choice is due to the fact that the SC-HMM parameters can be separated in two sets: (i) the parameters of a common Gaussian mixture model (GMM) representing a codebook of shapes, and (ii) word-dependent mixture

weights. Then, a writer-independent SC-HMM can be specialized to the writer style of a document by applying statistical adaptation techniques to the GMM in an unsupervised manner.

The rest of the article is structured as follows. Section 2 describes the keyword models. Section 3 details the adaptation techniques. Section 4 reports the experimental validation, and in section 5 conclusions are drawn.

2. Keyword models

We use a statistical approach that builds on [14] to model handwritten words. A word image is described as a sequence $X = x_1x_2 \dots x_T$ of feature vectors or *frames* x_t , extracted using a sliding window. Each keyword to search is modeled using a SC-HMM [3]. Its main property is that all the states of all keywords share a common pool of Gaussians with parameters $\theta = \{(\mu_k, \Sigma_k)\}_{k=1}^K$. Let p_k denote the probability density of Gaussian k in the pool and $p_{n,s}$ the emission probability in state s of keyword W_n . Then, the probability of emitting the frame x in this state can be written as:

$$p_{n,s}(x) = \sum_{k=1}^K w_{n,s,k} p_k(x). \quad (1)$$

The mixture weights $w_{n,s,k}$ are the only word- and state-specific parameters.

The pool of Gaussians is obtained by training a GMM $p(x|\theta)$ in an unsupervised way on a large set of frames appearing in many different word samples. Here, the Gaussians represent soft clusters of similar frames, so they can be interpreted a 'codewords' of a vocabulary, like in the computer vision literature. We should refer to this GMM as "shape vocabulary", since each codeword represents a part of a character, a connector, a whole character, etc. Then, the SC-HMM for a keyword W_n is composed by the parameters θ of the GMM and the set of mixture weights denoted compactly λ_n . A weight $w_{n,s,k}$ is the frequency of the codeword shape k in state s of word n .

A new sequence X is scored against the keyword model of W_n by using the likelihood ratio $p(X|\lambda_n, \theta)/p(X|\theta)$, where $p(X|\lambda_n, \theta)$ is the likelihood on the SC-HMM and $p(X|\theta) = \prod_{t=1}^T p(x_t|\theta)$ is the likelihood computed on the shape vocabulary GMM. In [14] we showed that the use of $p(X|\theta)$ is crucial for good performance.

3. Proposed writer style adaptation

3.1. Formulation

The main contribution of this work is to provide an adaptation method that exploits the separation of the parameters λ_n and θ in the SC-HMM. First, a *universal* shape vocabulary is built by training the GMM $p(\cdot|\theta)$ using frames from a large amount of samples of different writers. Then, during training and testing of the SC-HMMs, when a document i is processed, we apply a statistical adaptation technique to obtain a document-specific vocabulary from the universal shape vocabulary. The mixture weights remain unchanged. The particular details of the method are given below.

Training process:

- Learn the parameters θ of a universal GMM (shape vocabulary) on a large set of varied data containing many words and writing styles.
- For each document i , adapt the universal GMM using all the data available in the document to obtain the writer-dependent parameters θ_i . The particular adaptation methods are discussed below.
- Estimate the mixture weights λ_n of the SC-HMM of word W_n by maximizing the likelihood function $\sum_j \log p(X_j|\lambda_n, \theta_{d(j)})$ over all training samples $\{X_j\}$ where $d(j)$ indicates the index of the document X_j belongs to.

Testing process:

- Adapt the universal GMM parameters θ to the current document i using all samples of the document to obtain θ_i .
- Score each sample X using the likelihood ratio $p(X|\lambda_n, \theta_i)/p(X|\theta_i)$.

While we believe this adaptation method is novel, we identified some common points with the work of Fink and Plötz [6]. These authors also apply adaptation to a global GMM to obtain document-dependent GMMs. But their goal is to cluster these GMMs to obtain a partition of the training set into different writing styles. Then all *labeled* data in a training partition can be used to adapt the HWR model to this writer style. As we have discussed in the introduction, in a word spotting problem the lack of labeled data does not allow such an approach.

It is also interesting to notice that the adaptation of a universal GMM is used by Reynolds et al. [12] for

speaker identification. Recently, the Reynolds method has been applied as is to writer identification [15] with success. This shows adaptation and identification/verification are closely related problems.

3.2. Adaptation techniques

We have experimented with the two most popular statistical adaptation techniques [17] for obtaining a set of source-dependent (SD) parameters θ_{ad} from a source-independent (SI) set θ when few SD data is available.

Maximum A Posteriori (MAP) In MAP adaptation [7, 12], one assumes the existence of a prior distribution $p(\theta)$ over the SI parameters θ , and so the adapted parameters θ_{MAP} are given by:

$$\theta_{MAP} = \arg \max_{\theta} p(\mathcal{D}|\theta)p(\theta), \quad (2)$$

where \mathcal{D} are the SD data. It was shown that a convenient form of the prior $p(\theta)$ is the product of a Dirichlet density (accounting for the mixture weight parameters) and a Normal-Wishart density (accounting for the Gaussian parameters). One can then apply the EM algorithm as is the case of MLE, but substituting the auxiliary function $Q(\theta, \hat{\theta})$ by $R(\theta, \hat{\theta}) = Q(\theta, \hat{\theta}) + \log p(\theta)$. The maximization of this expression leads to the following equation for adapting the means:

$$\mu_k^{MAP} = \alpha_k \mu_k + (1 - \alpha_k) \mu_k^{MLE}. \quad (3)$$

Since α_k depends on the occupancy of Gaussian k , MAP does not adapt “unseen” Gaussians, and therefore a large amount of adaptation data is usually needed.

Maximum likelihood linear regression (MLLR) In MLLR adaptation [9], the adapted means are linearly transformed versions of the original ones:

$$\mu_k^{MLLR} = A\mu_k + b. \quad (4)$$

The EM algorithm can be employed to estimate the matrix A and offset vector b , for which multiple systems of linear equations have to be solved. Usually, the means are the only adapted parameters. Since MLLR transforms all the means even if the corresponding Gaussians are “unseen”, it tends to give better results than MAP with fewer SD data.

Eq. 4 assumes that there is a single transform applied to all Gaussians. However, this expression can be extended to apply different transforms to different

groups of Gaussians. These are obtained through clustering using a suitable distance measure between Gaussians (e.g. the log-likelihood loss). This technique increases the flexibility, which may improve performance but may also lead to overfitting.

4. Experiments

4.1. Experimental conditions

Adaptation is evaluated in the context of a word detection task. Experiments are conducted on 630 scanned letters (written in French), containing unconstrained handwriting from approximately the same amount of writers. First, a segmentation process extracts word images from the documents. The initial set of words is pruned using holistic features and a linear classifier. The surviving samples are normalized with respect to slant, skew and text height. At this point features are extracted using a sliding window. At each position, the sliding window is adjusted to the area containing pixels and split into a 4×4 grid. We report results with the local gradient histogram (LGH) features [13].

The set of letters is divided into 6 folds (0-5). Fold 0 is employed to construct the GMM corresponding to the universal shape vocabulary, as explained in section 2. Ten keywords are selected to be searched (e.g. *résiliation*, *contrat*, *demande*¹) and the corresponding 10 SC-HMMs are trained, using 10 states per letter and 512 Gaussians. Training and test is carried out using 5-fold cross validation with folds 1-5.

For each word to detect, the average precision (AP) of the detection task is computed.

4.2. Results

We experimented with MAP and 3 variations of MLLR, namely: offset-only transform (MLLR-b), diagonal matrix A (MLLR-diag) and full matrix A (MLLR-full). Table 1 shows the mean of the AP across the 10 keywords, for each adaptation method. Adaptation improves the AP in all cases, by 1.2% on average using MLLR-full. The largest observed increase for a single word is 2.8 %.

Our baseline system works at a point of fixed false rejection rate of FR=40% and average false acceptance rate of FA=0.32%. Using the MLLR-full adaptation, the average FA is reduced to 0.26%. This means, the number of errors is reduced by about 19%, a significant reduction of the number of misclassified words.

Regarding the MLLR-b and MLLR-diag methods, the best results are obtained when using 32 Gaussian

¹ translated as cancellation, contract and request, respectively

Table 1. Mean average precision for the different adaptation techniques

Adaptation method	Mean AP (%)
None	79.3
MAP	80.3
MLLR-b	80.2
MLLR-diag	80.4
MLLR-full	80.5

clusters. For the case of MLLR-full, the introduction of multiple clusters degrades performance, which is probably due to the higher number of parameters with a full matrix.

5. Conclusions

To the best of our knowledge, this is the first work to apply writer adaptation in a HWS task. We have presented a novel adaptation method based on a separation of the word-dependent parameters (mixture weights) from the writer-dependent (shape vocabulary) parameters in a SC-HMM and the adaptation of the latter for each writer at training and test time. Experiments show that this adaptation technique improves the performance of a detection task.

Results could be further improved by considering the following facts: (i) the adaptation material contained in a single document might be insufficient and (ii) the universal shape vocabulary may contain writer-dependent shapes while ideally the Gaussians should be independent of the writer style. The first might be improved using very fast adaptation techniques such as eigen-voices [8]. For the second one, iterative techniques such as speaker-adaptive training (SAT) [1] can be applied to remove the writer variability from the GMM shape vocabulary.

Acknowledgements

The work of the CVC authors is partially supported by the Spanish projects TIN2006-15694-C02-02 and CONSOLIDER-INGENIO 2010 (CSD2007-00018).

References

[1] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul. A compact model for speaker-adaptive training. In *Proc. ICSLP '96*, volume 2, pages 1137–1140, 1996.

[2] G. Ball and S. R. Srihari. Writer adaptation in off-line Arabic handwriting recognition. In *Document Recognition and Retrieval XV*, 2008.

[3] J. R. Bellegarda and D. Nahamoo. Tied mixture continuous parameter modeling for speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38:2033–2045, 1990.

[4] A. Brakensiek, A. Kosmala, and G. Rigoll. *Writer Adaptation for Online Handwriting Recognition*, page 32. Pattern Recognition: 23rd DAGM Symposium, 2001.

[5] S. D. Connell and A. K. Jain. Writer adaptation for online handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):329–346, 2002.

[6] G. A. Fink and T. Plötz. Unsupervised estimation of writing style models for improved unconstrained off-line handwriting recognition. In *Proc. 10th Int. Workshop on Frontiers in Handwriting Recognition*, 2006.

[7] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, Apr 1994.

[8] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski. Rapid speaker adaptation in the eigenvoice space. *IEEE Transactions on Speech and Audio Processing*, 8, 2000.

[9] C. Leggetter and P. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9:171–185, 1995.

[10] R. Manmatha, C. Han, and E. M. Riseman. Word spotting: A new approach to indexing handwriting. In *CVPR*, page 631, 1996.

[11] T. M. Rath and R. Manmatha. Word image matching using dynamic time warping. In *CVPR*, pages 521–527, 2003.

[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10:19–41, 2000.

[13] J. A. Rodríguez and F. Perronnin. Local gradient histogram features for word spotting in unconstrained handwritten documents. 2008. 1st International Conference on Frontiers in Handwriting Recognition.

[14] J. A. Rodríguez and F. Perronnin. Score normalization for HMM-based handwritten word spotting using a universal background model. 2008. 1st International Conference on Frontiers in Handwriting Recognition.

[15] A. Schlapbach, M. Liwicki, and H. Bunke. A writer identification system for on-line whiteboard data. *Pattern Recognition*, 2008. in press.

[16] A. Vinciarelli and S. Bengio. Writer adaptation techniques in HMM based off-line cursive script recognition. *Pattern Recognition Letters*, 23(8):905–916, 2002.

[17] P. Woodland. Speaker adaptation: techniques and challenges. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, 1999.