Annotating handwritten characters with minimal human involvement in a semi-supervised learning strategy

Jan Richarz, Szilárd Vajda and Gernot A. Fink
Faculty of Computer Science XII
TU Dortmund University
Dortmund, Germany
{*jan.richarz, szilard.vajda, gernot.fink*}*@udo.edu*

***Abstract*—One obstacle in the automatic analysis of handwritten documents is the huge amount of labeled data typically needed for classifier training. This is especially true when the document scans are of bad quality and different writers and writing styles have to be covered. Consequently, the considerable human effort required in the process currently prohibits the automatic transcription of large document collections. In this paper, two semi-supervised multiview learning approaches are presented, reducing the manual burden by robustly deriving a large number of labels from relatively few manual annotations. The first is based on cluster-level annotation followed by a majority decision, whereas the second casts the labeling process as a retrieval task and derives labels by voting among ranked lists. Both methods are thoroughly evaluated in a handwritten character recognition scenario using realistic document data. It is demonstrated that competitive recognition performance can be maintained by labeling only a fraction of the data.**

***Keywords*-document analysis; semi-supervised annotation; multiview learning; handwritten character recognition;**

I. Introduction

Training pattern recognizers typically requires large amounts of manually annotated samples in order to capture the characteristics of the data. In offline handwriting recognition, where characters may exhibit large variations in appearance due to different writers and writing styles, and no temporal information is available to aid the task, this problem is especially severe. Therefore, offline recognizers typically need to be adapted or retrained to facilitate new writers and scripts. Consequently, the burden involved in labeling large sample sets is recurring.

In museums and archives, large collections of handwritten documents exist that are of great potential interest for historians and scientists. Thus, there has been considerable effort to digitize them. However, ideally, the scanned document images should be fully transcribed in order to enable fast searching, browsing, efficient storage, and complete automatic analysis. Today, the transcription is still mostly done manually, which is a very laborious and tiresome task, and only insignificant numbers of documents can be processed this way. Nevertheless, the lack of sufficiently reliable general offline handwriting recognizers and the amount of effort that has to be spent in training them currently prohibits fully automatic processing of such collections.

The focus of this paper is thus not primarily to investigate features or techniques for handwritten character recognition, but to develop methods that help in annotating large amounts of data efficiently. Suppose it is possible to label large databases with sufficiently small manual effort, say, a few hundred manual operations to label several thousand samples. Then, even the repeated training of classifiers specialized on some specific writer or script may become feasible. It can be expected that such a specifically trained recognizer will perform better than a general one, that has to cover the variations of largely differing writing styles.

Therefore, two methods are presented that greatly reduce the required manual effort in labeling by adopting approaches from the field of semi-supervised and multiview learning (cf. e.g. [1]), utilizing a voting scheme over different feature representations of the data in order to increase reliability. The first method is based on clustering. Clusters are annotated as a whole and reliable labels are determined by performing a majority voting procedure. The second casts the labeling problem as an interactive retrieval task, and also relies on a majority vote between ranked lists in order to derive the final labels.

The remainder of this paper is organized as follows: First, the problem statement is provided and related literature is shortly reviewed in the process. Afterwards, the proposed semi-supervised labeling methods are explained in detail. The paper concludes with a detailed experimental evaluation, where the applicability of both methods is demonstrated on a set of handwritten documents.

II. Problem statement

In offline handwriting recognition, the large variability of character appearances over different writers and styles poses a challenging problem. Consequently, the error rates obtained in multi-writer recognition scenarios are typically high (cf. e.g. [2]). Existing recognizers thus are often restrained to a single writer (cf. e.g. [3], [4]). While there exist successful approaches for automatic keyword indexing in document collections (cf. e.g. [5], [6]), the automatic

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.181

transcription of handwritten documents still remains an open problem. Generally, recognizers can only perform reliably on data similar to the training data, and have to be adapted or retrained when applied to a new data domain. Consequently, the amount of human effort required to collect sufficient ground truth data for training becomes prohibitive for applications outside the academic field.

Thus, it is worthwile to develop methods that reduce this labeling effort, making the recurring training of writer- or collection-specific – and, thus, more reliable – classifiers feasible for large and diverse collections. We seek to contribute in this field by presenting approaches for efficient labeling of large data sets, adopting ideas from *semi-supervised learning* to alleviate the required manual effort.

The general idea of semi-supervised learning (cf. e.g. [1]) is to operate on both labeled and unlabeled data. Specifically, *semi-supervised classification* is the problem of training a classifier when only a small part of the data is annotated and the (typically) vast majority of data labels is unknown. Consequently, known labels must be robustly propagated to the unknown data by using unsupervised techniques. In order to achieve this, we adopt the concept of *multiview learning*. Here, an *ensemble* of learners is trained, each having a different view on the data (for an overview of ensemble methods, cf. e.g. [7]). Decisions are then obtained by combining the outputs of the different learners, e.g. by majority voting. Some concepts used in this work are also related to *active learning* (cf. e.g. [8]), where the learner actively selects the data that should get annotated based on its current knowledge in a feedback loop.

The problem of propagating labels to large corpora from just a few annotated instances has been studied extensively in the field of semantic image retrieval (cf. e.g. [9], [10]). In [11], it is shown that the recognition rate of a handwriting recognizer can be improved using a self-learning strategy on unlabeled adaptation data. In [12], character annotations are derived from word-level ground truth by optimizing segmentation hypotheses in an unsupervised manner. However, the initial set of word annotations must be provided manually.

III. Proposed methods

In the following, we outline the proposed semi-supervised multiview methods for efficient labeling in a general manner. Concrete realizations and parametrizations will be presented in section IV.

A. Clustering-based annotation (CBA)

In our previous work [13], a multiview labeling method requiring minimal human effort was proposed. We slightly enhance this method in the following, and provide a more throrough evaluation. The idea behind the method is simple: Label as few samples as possible and automatically infer the labels for other samples that are similar. Here, similarity is defined by applying a clustering algorithm.

The labeling process consists of three major steps. First, an ensemble of data representations is created, providing alternative views of the data by using different types of features. Then, each representation is clustered into k_j clusters, where k_j may differ for each representation. Further diversification is achieved by applying different clustering algorithms. The result is a set of r different data setups $\mathcal{R}_j, j = 1...r$, i.e. alternative combinations of features and clustering algorithms.

Given a set of clusters, only the centroids are labeled manually, and the rest of the samples in the cluster inherit the label. This implies $\sum_j k_j$ manual operations, and yields r independent labels for each sample.

Inheriting labels from cluster centroids will result in some incorrectly labeled samples, since, generally, the clusters are not pure. Thus, the final step of the procedure is to determine which labels are reliable. Assume that the labels are given as d-dimensional binary vectors $[l_{i,1}, \ldots, l_{i,d}]^T \in \{0,1\}^d$, $i = 1, \ldots r$, where $l_{i,j} = 1$ if a sample p is assigned to class ω_j in setup $\mathcal{R}_i$, and 0 otherwise. Applying a majority voting procedure results in an ensemble decision for a specific class label ω_k^{max}. A threshold κ_c on the ensemble decision is used to select only those samples where the class membership is determined with high agreement:

$$\omega_k^{max} = \max_k \sum_{i=1}^{r} l_{i,k} \geq \kappa_c. \tag{1}$$

In the following, we only retain samples for which all votes agree on the same label (*unanimity vote*), i.e. $\kappa_c = r$. Finally, the subset of samples and assigned labels that is retained from the above procedure is used to train a classifier. This classifier can then be used to either re-evaluate the training data (*inductive learning*) or classify a test set of unknown data (*transductive learning*).

B. Retrieval-based annotation (RBA)

The second proposed method is based on interactive retrieval, and is related to pool-based active learning with relevance feedback (cf. e.g. [14]). However, it differs in a few important aspects from this paradigm. Most importantly, we want to retrieve labels for all possible classes (quasi-) simultaneously. Additionally, selecting relevant samples manually from a potentially very large retrieval list counteracts the goal of lessening the burden for the annotator. Consequently, the manual relevance feedback step is replaced by a simple automatic selection rule on the retrieval list, propagating the annotation to unlabeled samples. Since incorrectly assigned labels will occur in this stage due to irrelevant samples in the list, a multiview voting concept taking into account several retrieval runs is integrated. The intuition behind is that, if multiple runs for the same query in several data representations agree on a subset of samples, then those belong to the query class with high confidence. Figure 1 gives an overview of the proposed procedure.

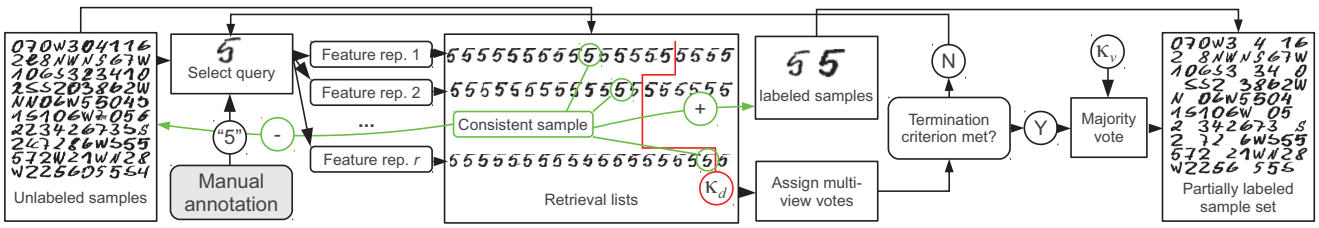

Figure 1. Overview of the proposed retrieval-based annotation process.

Let $\mathcal{X} = \{\mathbf{X}_i, i = 1...n\}$ be a pool of n unlabeled data samples , $\mathbf{X}_i = \{\mathbf{x}_{ij}, \mathbf{u}_i, \mathbf{v}_i, n_i, j = 1...r\}$, where $\mathbf{x}_{ij}$ is the i-th data sample's j-th feature representation, $\mathbf{u}_i$ is a variable-length vector of class labels (empty on initialization), $\mathbf{v}_i$ is a vector of confidence scores associated with each label in $\mathbf{u}_i$, and n_i is the number of times the sample has been considered. Also, let $\mathcal{T} = \{\mathbf{X}_k, y_k, k = 1...m\}$ be an initially empty pool of trusted samples, with the k-th sample's final label y_k.

First, a sample $\mathbf{X}_s$ is selected from $\mathcal{X}$. Since no information is available in the first iteration, this selection is random. In further iterations, only the subset $\hat{\mathcal{X}}$ of samples with minimum value of n_i is considered, i.e

$$\hat{\mathcal{X}} = \{\mathbf{X}_l : n_l = \min_i(n_i)\}, \tag{2}$$

and $\mathbf{X}_s$ is selected randomly from this subset. The rationale behind is that those samples should be selected that have not been considered (often) before, thus exploring the data set. Then, a class annotation ω_s for $\mathbf{X}_s$ is requested from the annotator, which can also be a rejection label. This manual assignment is trusted, hence $\mathbf{X}_s$ is removed from $\mathcal{X}$ and appended to $\mathcal{T}$, with $y_s = \omega_s$. Note that, in "real" active learning, the sample selection would be based on the current state of some classifier. In omitting this, we are able to start with a completely unlabeled data set and avoid frequent retraining with potentially unreliable ground truth.

Afterwards, r retrieval tasks are carried out on the remaining set $\mathcal{X}\backslash\mathbf{X}_s$ using the r different feature representations $\mathbf{x}_{sj}$ as queries. This results in r retrieval lists $\mathcal{L}_j$, ranked according to some distance $d(\mathbf{x}_{ij}, \mathbf{x}_{sj})$, and thresholded with a common threshold κ_d on the distances. In the following, the cosine distance is used for $d(...)$. Assume that a sample $\mathbf{X}_p$ is present in N_r of the r thresholded lists. Then, the confidence γ_{ps} of $\mathbf{X}_p$ belonging to the query class ω_s is given by $\gamma_{ps} = \frac{N_r}{r}$. If γ_{ps} is 1, i.e. if $\mathbf{X}_p$ appears in *all* thresholded lists, then $\mathbf{X}_p$ is assigned the label ω_s, removed from the sample pool $\mathcal{X}$ and appended to the pool of trusted samples $\mathcal{T}$, i.e: $\mathcal{X} = \mathcal{X}\backslash\mathbf{X}_p$, $y_p = \omega_s$, $\mathcal{T} = \mathcal{T} \cup \{\mathbf{X}_p, y_p\}$. Otherwise, ω_s is appended to $\mathbf{u}_p$, γ_{ps} is appended to $\mathbf{v}_p$, and n_p is incremented, i.e: $\mathbf{u}_p = \mathbf{u}_p \cup \omega_s$, $\mathbf{v}_p = \mathbf{v}_p \cup \gamma_{ps}$, $n_p = n_p + 1$. The sample remains in the pool $\mathcal{X}$ in this case, and may be considered again in further iterations. Samples with $d(...) > \kappa_d$ in all retrieval lists remain unchanged.

This interactive procedure is repeated until a termination criterion is met. In practice, we simply abort after a fixed number I_m of manual operations. Each remaining sample $\mathbf{X}_i \in \mathcal{X} : n_i > 0$ now has a list $\mathbf{u}_i = \{u_{it}, t = 1...n_i\}$ of assigned labels with associated confidences $\mathbf{v}_i = \{v_{it}\}$. Furthermore, the initially unknown set of available classes $\Omega = \{\omega_k, k = 1...c\}$ has evolved based on the manual annotations provided in the process. For each $\mathbf{X}_i$, the accumulated class confidences

$$\sigma_i(\omega_k) = \sum_t \sum_k v_{it}\delta(u_{it} - \omega_k) \tag{3}$$

are calculated and then normalized to $[0, 1]$ by dividing by the maximum possible accumulated confidence:

$$\tilde{\sigma}_i(\omega_k) = \frac{r}{(r-1)\cdot n_i} \cdot \sigma_i(\omega_k). \tag{4}$$

Finally, the final class label y_i of $\mathbf{X}_i$ is determined as the one having the maximum accumulated confidence:

$$y_i = \operatorname*{argmax}_k(\tilde{\sigma}_i(\omega_k)). \tag{5}$$

If the associated maximum confidence $\tilde{\sigma}_i(y_i)$ is above a threshold κ_v, the sample is added to $\mathcal{T}$. Otherwise, the sample is rejected because the assigned label would be too unreliable. The classifier is then trained on the retained set of trusted samples $\mathcal{T}$.

Compared to CBA presented in the previous section, the above procedure offers several advantages. No prior knowledge or assumption about the number of classes is required because they will evolve implicitly based on the labels assigned by the annotator. Also, it is possible to manually reject "bad" samples, and the required manual effort does not depend on the number of different representations r since they are evaluated simultaneously. On the other hand, the impact of errors in the manual annotation can be expected to be higher. Additionally, values for the parameters κ_d, κ_v and I_m have to be selected heuristically. Suitable values will be determined experimentally in section IV-D.

IV. Experiments

In order to assess the performance of the proposed methods, we first derive suitable parameters on the MNIST handwritten digit datset [15], using the original division

of samples in training $(60,000)$ and test set $(10,000)$. Afterwards, a realistic transductive recognition experiment is conducted on a set of historical documents.

A. Data description and experimental setup

Our data set consists of historical official weather reports, kindly provided by the German Weather Service ("Deutscher Wetterdienst", DWD). Detailed characteristics of the data are given in [16]. For the experiments reported here, a set of 106 document images was provided, scanned at approximately 200 dpi. The entire collection comprises several $10,000$ pages, but only a small subset is currently available in digitized form.

In total, the data contains 13,331 samples (5,140 characters, 8,191 digits) in 17 classes (digits 0–8, characters N, S, W, O, E, T, I, L). Ground truth labels are available for all samples. The data set is highly unbalanced, and several classes occur very rarely. Since the documents are tabular and each table field should exclusively contain either characters or digits, we assume that knowledge about the type of each sample is known from a document template (cf. [16]). This information will be used in the retrieval and classification to limit the set of candidates. Documents are subdivided in a 3-fold cross-validation setup, where, in each validation set, approximately $2/3$ of the documents constitute the training set and the remaining the test set, respectively. Thus, training and test set are disjunct, but overall, all documents are considered once for testing.

B. Features

The proposed labeling methods rely on a multiview approach using different feature representations. In principle, the methodology is independent of the types of features used. Obviously, the better the discriminative power of the features, the better the results will be. Especially for the retrieval-based approach, compact features are desirable in order to keep the process efficient and avoid latencies in the interactive loop.

In the following, we focus on a set of features that yielded good performance in our previous work. Specifically, we consider the raw image of a character (RAW), normalized to size 28×28 pixels, its PCA representation using the first 80 PC [16], and higher-level structural features based on contour chain codes (CC), skeletons (SKEL) and character reservoirs (RES) [17]. The latter were modified by considering 5 types of reservoirs (top, bottom, left, right, loop) and using just a soft assignment of the positions of their gravity centers to image cells. All features are calculated in a 4×4 regular grid on the normalized character image. Feature dimensionalities range from 36 (SKEL) to 80 (PCA).

C. Parametrization of CBA

Even though using all features might provide a better discrimination for the different clusters, selection of a subset is advisable for efficiency reasons. In order to select the best setup in terms of combinations of feature and clustering method, an exhaustive search was performed. All different features extracted from the MNIST training material were clustered using k-means, SOM (Self Organizing Map) and GNG (Growing Neural Gas) [18]. The cluster centers were then manually annotated by an expert.

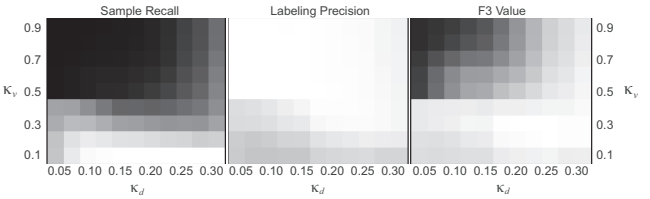

Figure 2. Sample recall $\mathcal{R}$, label precision $\mathcal{P}$ and $\mathcal{F}_3$ of RBA for different values of κ_d, κ_v using best feature set (CC, PCA, RES). Values are color-coded from 0 (black) to 100% (white).

As quality criteria, the sample recall $\mathcal{R}$ (percentage of retained trusted samples after voting) and label precision $\mathcal{P}$ (percentage of correctly labeled retained samples) obtained on MNIST were used. Ranking the different combinations, the best setup was: RAW/GNG, CC/GNG, and CC/k-means. In general, GNG and k-means outperform SOM clustering. With this setup and 54 cluster centers [13] for each combination, unanimity vote occured in $\mathcal{R} = 76.15\%$ of the cases with a precision $\mathcal{P} = 96.10\%$. Thus, 45,690 annotations were inferred using only 162 manual labeling operations, corresponding to a relative manual effort of 0.35%. Compared to [13], the sample recall increased substantially (approx. 21%) while the same labeling accuracy was obtained. This shows the benefit of incorporating not only different feature representations, but also different clustering methods resulting in a better diversification of the data views.

D. Parametrization of RBA

In order to find suitable values for the parameters κ_d and κ_v, a number of experiments was conducted on the MNIST data set, performing $I_m = 500$ labeling operations and averaging over 10 runs with identical parametrization in order to smooth the effects of the random selection. The manual labeling was simulated by assigning the respective ground truth label to the query example, i.e. error-free annotation is assumed. The goal is to find a range of parameters offering a good balance between sample recall and label precision. Numerous combinations of the different features were investigated, excluding RAW for efficiency reasons. Due to space restrictions, we will not provide details on all experiments, but just report results for the best feature combination (CC + PCA + RES, i.e. $r = 3$).

As can be seen in Fig. 2, the labeling precision is generally high, except for small values of κ_v. It also degrades for small values of κ_d, because then only few samples will be considered in each retrieval run, and the small overall

number of votes leads to an unreliable majority decision. In terms of sample recall, the method is more restrictive than CBA retaining large fractions of the data only for small values of κ_v. While CBA enforces exactly the same number of votes for all samples it varies in RBA. Consequently, samples at the boundary of class distributions may get very few or inconsistent votes and thus are rejected. This could possibly be improved by incorporating a more sophisticated sample selection method in the interactive annotation loop (instead of simple random selection), but investigation of this issue is left for future work.

In order to determine a suitable parametrization, we calculate the F3 score $\mathcal{F}_3 = \frac{10\mathcal{RP}}{9\mathcal{R}+\mathcal{P}}$, reflecting the assumption that it is more desirable to have correctly labeled samples than retaining large portions of the original data. The maximum score was $\mathcal{F}_3 = 91.54\%$ ($\mathcal{R} = 82.72\%$, $\mathcal{P} = 92.63\%$) for parameter values $\kappa_d = 0.25, \kappa_v = 0.20$. However, from Fig. 2, it becomes obvious that the quality in terms of $\mathcal{F}_3$ is comparable for a range of parameter values around this optimum, meaning that the method is not too sensitive against the concrete choice of values. In the following, again favoring high precision, we will use more restrictive parameter values of $\kappa_d = 0.2, \kappa_v = 0.30$, yielding $\mathcal{P} = 97.15\%$, $\mathcal{R} = 59.02\%$, $\mathcal{F}_3 = 91.26\%$ for the above experiment.

E. Recognition experiments

The setups derived above are evaluated in a transductive classification experiment on the MNIST and DWD data, demonstrating that the methods perform reliably on different datasets and that tuning the parameters to specific data – which is not possible in real applications – is not necessary. We perform a realistic experiment, where the training sets are first labeled by an expert annotator using the proposed methods. Then, the classifiers are trained on the resulting sets of trusted samples, and evaluated on the disjunct test sets. For comparison, we also present results of oracle experiments, using all ground truth labels for training. These constitute an upper limit of the achievable performance.

In order to make the results comparable, an identical number of manual annotations (162) is used for both CBA and RBA. Statistics on MNIST for CBA using this setting were given in Sec. IV-C. For RBA, a recall of $\mathcal{R} = 57.91\%$ and precision of $\mathcal{P} = 90.02\%$ were obtained.

For the 3-fold cross-validation on the DWD data, also 162 labeling operations were performed for each validation set. With CBA, on average $\mathcal{R} = 92.00\%$ of the training material was retained, yielding an average label precision of $\mathcal{P} = 92.18\%$. With RBA, $\mathcal{R} = 87.74\%$ of samples were retained with $\mathcal{P} = 95.64\%$.

As discussed in [16], a drawback of CBA is that it tends to discard rare classes in the case of highly unbalanced data, since they do not form individual clusters and are eliminated by the unanimity voting. While this did not occur for the balanced MNIST set, only 14 out of 17 classes were recovered on average on the DWD data. With RBA, all 17 classes were retained. This shows a major advantage of the RBA approach: Because of the steered sample selection, under-represented classes are less likely to be discarded.

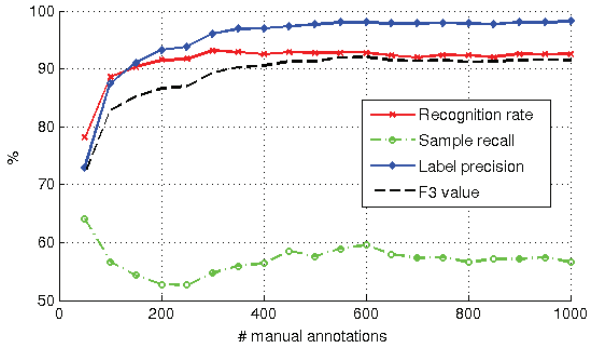

Figure 3. Labeling quality and classifier performance vs. number of manual annotations for RBA (MNIST, SVM classifier, $\kappa_d = 0.2, \kappa_v = 0.30$).

For recognition, a linear SVM classifier in a 1-vs-1 multiclass setup is used. The resulting character recognition rates are given in Table I. Applying the proposed labeling schemes to the MNIST data results in a substantial loss in recognition performance. The reason is that both methods rely on discovering clusters of similar samples, i.e. from the same writer or written in the same style. Since the MNIST data is very diverse and contains hundreds of writers, this assumption is violated, resulting in a loss of accuracy. RBA performs worse than CBA because propagating the labels based on the retrieval lists proceeds considerably slower than labeling the large portions of data contained within a cluster. As shown in Fig. 3, the performance of RBA keeps increasing until approximately 400–500 manual annotations were performed (still a relative effort of less than 1%). The saturated recognition score is comparable to CBA. annotation with a more achievable performance,

However, for the DWD data, which is much more homogeneous in terms of writing style, the results obtained with both CBA and RBA are close to the reference oracle experiment. This clearly shows the potential of both methods, provided that large portions of the data show similar characteristics. Only very little manual effort was required (performing the 162 labeling operations was a matter of a few minutes in both cases) to obtain competitive recognition rates, making the proposed methods especially promising for large single-writer collections.

V. Conclusion

Two semi-supervised methods for annotating data with minimum human effort were proposed. Both build on ideas from multiview learning in order to propagate labels to unknown data reliably incorporating a voting procedure over different feature representations. It was demonstrated that

Table I
OVERVIEW OF RECOGNITION RESULTS (IN %) AND CONFIDENCE INTERVALS FOR CONFIDENCE LEVEL 0.95.

Data	Anno. method	#Anno	RAW	PCA	CC	SKEL	RES
MNIST	Ground truth	60, 000	92.15±0.54	92.60±0.53	**95.30±0.43**	85.69±0.70	82.32±0.76
	CBA	162	88.13±0.65	88.64±0.64	**91.28±0.57**	83.83±0.73	81.58±0.77
	RBA	162	84.97±0.71	86.76±0.68	**89.50±0.62**	81.91±0.77	78.36±0.82
DWD	Ground truth	8,887	91.59±0.48	93.74±0.42	**95.54±0.36**	92.21±0.47	88.31±0.56
	CBA	162	90.26±0.52	92.05±0.47	**93.76±0.42**	91.07±0.50	87.43±0.57
	RBA	162	87.56±0.57	91.98±0.47	**94.64±0.40**	90.73±0.50	86.57±0.59

thousands of training labels can be inferred from very few manual annotations with high accuracy, therefore facilitating the annotation of large data sets within minutes. Using these labels a recognizer was trained that achieved good performance in a handwritten character recognition experiment on two different databases. Since both approaches rely on finding large clusters of similar data samples they are especially promising for collections containing large portions of data from the same or just a few different writer(s). In this case, the recognition performance achieved is close to the reference oracle experiment. Thus, the proposed methods can facilitate the efficient training of statistical recognizers specialized on specific document collections by massively reducing the prohibitive manual burden and associated cost that would normally be required.

ACKNOWLEDGMENTS

This work is supported by the German Federal Ministry of Economics and Technology on a basis of a decision by the German Bundestag within project **KF2442004LF0**.

REFERENCES

[1] X. Zhu and A. B. Goldberg, *Introduction to semi-supervised learning*. Morgan & Claypool, 2007.

[2] T. Plötz and G. A. Fink, *Markov Models for handwriting recognition*. Springer, 2011.

[3] M. Bulacu, A. Brink, T. van der Zant, and L. Schomaker, "Recognition of handwritten numerical fields in a large single-writer historical collection," in *Proc. Int. Conf. on Document Analysis and Recognition*, 2009.

[4] M. Wüthrich, M. Liwicki, A. Fischer, E. Indermühle, H. Bunke, G. Viehhauser, and M. Stolz, "Language model integration for the recognition of handwritten medieval documents," in *Proc. Int. Conf. on Document Analysis and Recognition*, 2009.

[5] T. M. Rath and R. Manmatha, "Word spotting for historical documents," *Int. Journal on Document Analysis and Recognition*, vol. 9, no. 2, pp. 139–152, 2007.

[6] M. Rusiñol, D. Aldavert, R. Toledo, and J. Lladós, "Browsing heterogeneous document collections by a segmentation-free word spotting method," in *Proc. Int. Conf. on Document Analysis and Recognition*, 2011, pp. 63–67.

[7] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[8] B. Settles, "Active learning literature survey," University of Wisconsin-Madison, Computer Sciences Technical Report 1648, 2009.

[9] E. Chang, G. Sychay, K. Goh, and G. Wu, "CBSA: Content-based soft annotation for multimodal image retrieval using bayes point machines," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 26–38, 2003.

[10] B. Settles, "Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances," in *Proc. Conf. on Empirical Methods in Natural Language Processing*, 2011, pp. 1467–1478.

[11] G. R. Ball and S. N. Srihari, "Semi-supervised learning for handwriting recognition," in *Proc. Int. Conf. on Document Analysis and Recognition*, 2009, pp. 26–30.

[12] J. Sas and U. Markowska-Kaczmar, "Semi-automatic training sets acquisition for handwriting recognition," in *Proc. Int. Conf. on Computer Analysis of Images and Patterns, LNCS 4673*. Springer, 2007, pp. 521–538.

[13] S. Vajda, A. Junaidi, and G. A. Fink, "A semi-supervised ensemble learning approach for character labeling with minimal human effort," in *Proc. Int. Conf. on Document Analysis and Recognition*, 2011, pp. 259–263.

[14] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proc. ACM Int. Conf. on Multimedia*, 2001.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Intelligent Signal Processing*. IEEE Press, 2001, pp. 306–351.

[16] J. Richarz, S. Vajda, and G. A. Fink, "Towards semi-supervised transcription of handwritten historical weather reports," in *Proc. IAPR Int. Workshop on Document Analysis Systems*, 2012, to appear.

[17] A. Junaidi, S. Vajda, and G. A. Fink, "Lampung - a new handwritten character benchmark: Database, labeling and recognition," in *Int. Workshop on Multilingual OCR*. ACM, 2011, pp. 105–112.

[18] B. Fritzke, "A growing neural gas network learns topologies." in *Neural Information Processing Systems*, G. Tesauro, D. S. Touretzky, and T. K. Leen, Eds. MIT Press, 1994, pp. 625–632.