Pattern Recognition 47 (2014) 1011–1020

Contents lists available at ScienceDirect

Pattern Recognition

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

Semi-supervised learning for character recognition in historical archive documents

Jan Richarz, Szilard Vajda, Rene Grzeszick*, Gernot A. Fink

TU Dortmund University, Department of Computer Science, Dortmund, Germany

ARTICLE INFO

Available online 31 July 2013

Keywords:
Character recognition
Semi-supervised learning
Historical documents

ABSTRACT

Training recognizers for handwritten characters is still a very time consuming task involving tremendous amounts of manual annotations by experts. In this paper we present semi-supervised labeling strategies that are able to considerably reduce the human effort. We propose two different methods to label and later recognize characters in collections of historical archive documents. The first one is based on clustering of different feature representations and the second one incorporates a simultaneous retrieval on different representations. Hence, both approaches are based on multi-view learning and later apply a voting procedure for reliably propagating annotations to unlabeled data. We evaluate our methods on the MNIST database of handwritten digits and introduce a realistic application in form of a database of handwritten historical weather reports. The experiments show that our method is able to significantly reduce the human effort that is required to build a character recognizer for the data collection considered while still achieving recognition rates that are close to a supervised classification experiment.

1. Introduction

After several thousand years of human history and roughly 2000 years after the invention of paper, historical archives and museums store tremendous amounts of handwritten documents. They contain information of great value for historians and the wide public.

Accessing this knowledge is typically not an easy task. It is necessary to browse through either printed or digital copies of these documents, which is a very tiresome and time consuming process. With digital copies browsing through those documents became much easier, but it is even more comfortable if they are indexed or transcribed. However, nowadays the transcription of documents is still done manually by experts. Much research has been dedicated to the task of transcribing documents automatically by recognizers, which is a very difficult problem. Scans of old documents are often of bad quality and show various artifacts. In addition, handwritten text shows a very high variability that is dependent on the writer. Typically a recognizer needs to be trained for different scripts and writers, which again requires a tremendous amount of training material that has been annotated before.

So far research has not been able to completely remove the process of manually annotating documents, but in the following we will give an overview of methods for reducing the required manual labeling operations for training a recognizer. The annotation task is executed in a machine-aided manner. Clustering and retrieval operations are used in order to choose representatives that are labeled by an expert annotator.

For evaluation we consider the well known MNIST dataset as well as a realistic set of historical weather reports. In both cases the methods are able to considerably reduce the amount of labeling operations to less than one percent of the original training data. We will show that it is possible to perform labeling with high precision, so that high recognition rates can be achieved with data that has been labeled in a semi-supervised manner.

2. Related work

The general idea of semi-supervised learning is to reduce the required manual work by combining labeled and unlabeled data (cf. [34]). Typically, in such scenarios the vast majority of data is unlabeled. The known labels must be highly reliable and robustly be propagated to the unknown data.

High reliability of the labels can only be ensured by presenting selected samples to an expert annotator. Additionally, the labeled subset should be representative for the remaining data since propagation is typically achieved by analyzing sample similarity. Consequently, random selection is generally not advisable. We utilize two different approaches for selecting a representative subset. The

* Corresponding author. Tel.: +49 231 755 6151; fax: +49 231 755 6116.
E-mail addresses: jan.richarz@udo.edu (J. Richarz), szilard.vajda@nih.gov (S. Vajda), rene.grzeszick@udo.edu (R. Grzeszick), gernot.fink@udo.edu (G.A. Fink).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.07.013

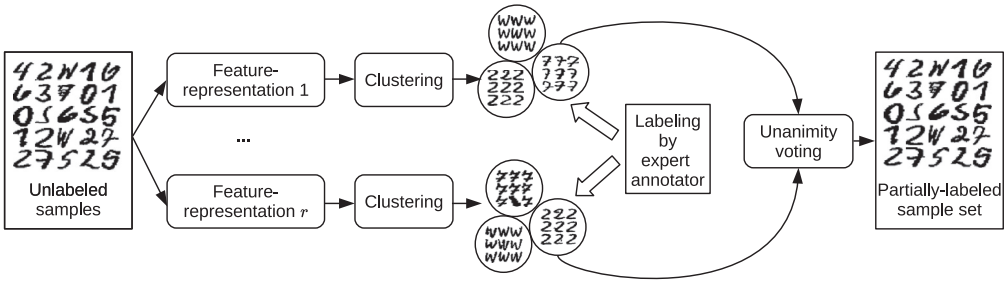

Fig. 1. Illustration of the clustering-based labeling approach. Given a set of characters, r different feature representations of the data are computed. Each of those representations is clustered and then a label is assigned to each of those clusters by an expert annotator. Finally, a unanimity voting scheme is applied to assign a label to each character, which results in a partially labeled sample-set.

first (cf. Section 3.1) relies on clustering and selects the cluster centroids as representatives that are labeled. The second (cf. Section 3.2) uses a realization of the *active-learning* concept where the system actively selects the data that should get annotated based on its current knowledge in a feedback loop (cf. [26]).

For achieving a robust propagation the concept of *multi-view-learning* is adapted by training an *ensemble* of learners (cf. [11,23]). Each of these learners has a different view on the data, e.g., by using different features. Decisions are made by combining the outputs of different learners. A common concept is using a majority vote [11]. The advantages of incorporating ensembles in semi-supervised learning approaches for robust propagation are, for example, discussed in [33].

The problem of propagating a small set of labels to a large dataset has been studied in different fields of research. Applications include, for example, the clusters of text documents [31], image retrieval [3,27] or the active learning of gesture trajectories [25]. Semi-supervised approaches have also been studied in the field of character annotation [1,24]. In [1] it has been shown that the recognition rate of a handwriting recognizer can be improved using self-learning strategies on unlabeled data. However, in all cases an initial set of annotations must be provided manually.

For handwritten graphical multi-stroke symbols an annotation assistance is proposed by Li et al. [13], where the annotation of the symbols is reduced to finding sub-graphs in a relation graph built from different segments. In the graph the nodes are the segments and the arcs represent the spatial relationships between them. The authors show that only 58.2% of the strokes need to be labeled.

With respect to the goal of reducing the manual effort in the transcription of historical documents, the work introduced by Toselli et al. in [29,30] has a similar goal than ours. However, the principle differs from our approach. We propose using a semi-supervised approach to label the data and train a new recognizer for a given document collection, while they rather refine an existing recognizer with feedback from the annotator.

Our own contributions to semi-supervised learning strategies for character labeling have been introduced in [32] for characters of the Lampung script, written in Indonesia and in [21,22] for Latin characters of the dataset of historical weather reports that is also considered in this paper. In the following we present an extended comprehensive overview of our semi-supervised learning methods for character recognition as well as a detailed evaluation of the clustering- and retrieval-based methods on two handwritten character databases.

3. Semi-supervised labeling approaches

In the upcoming sections we present two different methods that allow labeling training data on character level with a minimum amount of manual work: (a) clustering-based labeling (CBL), and (b) retrieval-based labeling (RBL). Our main goal is not to achieve the best possible classification scores, hence not concentrating on the most appropriate classifier selection and tuning, but rather to show that competitive results can be achieved with semi-supervised approaches using minimal human effort for the labeling process. To achieve this goal a high labeling accuracy is crucial since it strongly influences the subsequent recognition process.

3.1. Clustering-based labeling

In [21,32] we introduced a preliminary version of the clustering based multi-view labeling algorithm for handwritten characters that requires only minimal human effort for labeling the unknown data. The method is illustrated in Fig. 1 and can be described by four major steps:

(1) An ensemble of different views of unlabeled data is created using a set of different feature representations.
(2) In all representations the features are clustered in an unsupervised manner.
(3) A single label is assigned to each cluster center by the expert annotator.
(4) Unanimity voting among the different views is used for determining the label for each data point.

In order to implement an ensemble of representations that have a different view on the data we compute r different setups R_i. A setup is defined as a combination of a feature representation and a clustering method.

In every setup R_i the clustering is computed independently, creating k_i partitions of the data. Note that the number of clusters may vary for each feature representation. Usually the partitions are generated using a vector quantization algorithm, like k-Means clustering [15] or the generalized Lloyd algorithm [14], but other unsupervised methods like Self Organizing Map [10], Growing Neural Gas [8] or Affinity Propagation [7] can also be considered to separate the input space into separate regions.

Once the partitioning is performed, each cluster is labeled manually by an expert annotator. Only the cluster centroids are labeled, all other samples belonging to the same cluster will automatically inherit the label from the centroid. This way, the number of required manual annotations is reduced to $\sum_{i=1}^{r}(k_i)$. Hence, it depends only on the total number of clusters for all feature representations. Depending on the number of expected classes, large datasets of several thousand samples can easily be labeled using only a few hundred manual annotations.

Considering the number of clusters there are two factors that counteract: The smaller the number of clusters, the less manual work is required, but more clusters will represent the samples more accurately reducing considerably the intra-class and inter-class variances.