Pattern Recognition Letters 58 (2015) 23–28

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Semi-automatic ground truth generation using unsupervised clustering and limited manual labeling: Application to handwritten character recognition ☆

Szilárd Vajda [a,*], Yves Rangoni [b], Hubert Cecotti [c]

[a] National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
[b] Henri Tudor Public Research Center, Kirchberg, L-1855, Luxembourg
[c] Faculty of Computing and Engineering, University of Ulster, Londonderry BT48 7JL, Northern Ireland, UK

ARTICLE INFO

Article history:
Received 11 June 2014
Available online 2 March 2015

Keywords:
Character recognition
Classifier combination
Clustering
Feature selection

ABSTRACT

For training supervised classifiers to recognize different patterns, large data collections with accurate labels are necessary. In this paper, we propose a generic, semi-automatic labeling technique for large handwritten character collections. In order to speed up the creation of a large scale ground truth, the method combines unsupervised clustering and minimal expert knowledge. To exploit the potential discriminant complementarities across features, each character is projected into five different feature spaces. After clustering the images in each feature space, the human expert labels the cluster centers. Each data point inherits the label of its cluster's center. A majority (or unanimity) vote decides the label of each character image. The amount of human involvement (labeling) is strictly controlled by the number of clusters – produced by the chosen clustering approach. To test the efficiency of the proposed approach, we have compared, and evaluated three state-of-the art clustering methods (*k*-means, self-organizing maps, and growing neural gas) on the MNIST digit data set, and a Lampung Indonesian character data set, respectively. Considering a *k*-nn classifier, we show that labeling manually only 1.3% (MNIST), and 3.2% (Lampung) of the training data, provides the same range of performance than a completely labeled data set would.

Published by Elsevier B.V.

1. Introduction

The exponential increase of images to be processed and analyzed nowadays opens new challenges in the field of document recognition [1,2]. All the images can be acquired by cheap devices such as cell phones, tablets, and digital cameras. With the increase of data volume and types to be classified, pattern recognition techniques cannot easily cope with all the possible classification efforts. We can distinguish three types of multiclass classification tasks, where the goal is to assign a label to a certain image. In the first type, the images to be processed are too variable, and the number of samples may be too small to use supervised classification techniques. In this case, image retrieval methods are typically used [3,4]. In the second type, the training data is well identified, and a ground truth is available, therefore supervised classification techniques can be used. In the third type, the difficulty of the problem may not allow the use of shape retrieval techniques. It will require supervised classification techniques. However, because the images can belong to a new type of problem, an efficient technique has to be provided to facilitate data labeling, i.e., the creation of the ground truth. The estimation of a ground truth is an important aspect, because providing accurate labels is a tedious process, involving a lot of human resources and expert knowledge. As a consequence, such labeling initiatives are very costly, and time consuming.

One of the major goals in large data collections classification paradigm is to provide fully automatic, or at least semi-automatic, high accuracy labeling mechanisms – involving mostly unsupervised learning strategies, e.g., *k*-means [5], self organizing maps (SOMs) [6], growing neural gas (GNG) [7]. Such hybrid labeling strategies involve data driven clustering algorithms and human expertise. The more label discovery is made automatically, the better the method can be applied to different fields – without using any type of data specificity or metric related prior knowledge.

In this paper, we propose to extend our previous work [8] on semi-automatic character labeling by including five types of features, and by comparing three state-of-the art clustering methods against each other. In addition, they are evaluated at two levels: the clustering

☆ This paper has been recommended for acceptance by S.K. Pal.

* Corresponding author. Tel.: +1 301 594 7811. Fax. +1 3014020341.
E-mail address: szilard.vajda@nih.gov (S. Vajda).

http://dx.doi.org/10.1016/j.patrec.2015.02.001
0167-8655/Published by Elsevier B.V.

method performance, and the effect of this performance on the classification of the test data set using k-nn. Instead of limiting the input features to the pixel values of the raw images in gray level [9], more sophisticated and lower dimensionality features such as profiles, local binary patterns [10], and Radon transform [11,12] were considered to better exploit the advantage of the original method [9]. Currently, each image is projected in five different feature spaces. Each feature space is clustered in an unsupervised manner. The cluster centers are then labeled by a human expert, and the images belonging to the cluster are labeled with the cluster's label. The final label of an image is decided based on a voting mechanism, using the label obtained from each feature set.

The goal of the paper is: (i) to determine the relevance of the proposed sets of features and their complementarity during the vote, (ii) to evaluate the control of the labels to be accepted, and (iii) to determine the best clustering method.

The remainder of the paper is organized as follows: Section 2 gives an overview of similar labeling initiatives, Section 3 focuses on describing the different feature representations, Section 4 gives a brief overview of the unsupervised technique used in the experiments, while Section 5 is dedicated to the description of the semi-automatic labeling process. Sections 6 and 7 describe the data sets used in the experimental setup and the obtained results. Finally, Section 8 concludes and elaborates on future work.

2. Related work

As more and more data is available, the big data paradigm becomes a reality. This tremendous amount of data has to be labeled properly, otherwise it becomes useless for all classification, regression, retrieval, identification, and recognition tasks.

In [13], an expression matching for mathematical expression transcription was proposed. The matching is performed as a graph matching, in which symbols of input instances of a manually labeled model expression are matched against symbols in the model. The pairwise matching cost considers both local and global features of the expression. For online handwritten digits, Li et al. [14] propose a codebook mapping to cluster strokes using an agglomerative clustering, followed by a mapping using Hausdorff distance of each stroke or stroke agglomeration to representative labels by a human annotator [15].

A similar attempt is proposed in [16], where resembling motifs have to be detected in medical sequences. First a so-called "self training" is applied, which can be seen as a boosting mechanism, followed by an ensemble learning with decision using majority vote rules as a linear combination, as a product and a vote. The results are rather promising, but all these methods share the same drawback. After the unsupervised clustering or boosting, the decision is made, and that data (label) is accepted as gold standard.

Our preliminary work [8,9] proposed an analogous scheme, but using much less feature spaces, and an unsupervised clustering mechanism, which relied only on k-means. In this paper, we extended the number of feature spaces considered for unsupervised clustering, and the clustering methods. Not only k-means but also SOM, and GNG were used. The main differences compared to other systems are: (i) our diversified feature space, which can help exploiting the complementarity between the features, (ii) the usage of three completely different unsupervised clustering methods, (iii) the voting mechanism, instead of accepting the labels discovered without proper judgment, (iv) the newly discovered labeled data is used in a supervised classification scenarios.

All these improvements allow us to discover labels with high accuracy, using only minimal human annotation effort, which for large handwritten character collections save tremendous human effort and costs.

3. Feature representations

To exploit the strength of the method, the different feature spaces should complement each other [17,18]. However, this complementarity is not available a priori. Therefore, we selected arbitrary different features among the used ones in the literature. Some of them being considered quite efficient, while some others less. Our features are as follows:

Raw pixel (F_1), profiles (F_2), local binary patterns (F_3), Radon transform (F_4), and Encoder network (F_5). We denote by I the gray level image of size $N_x \times N_y$.

3.1. Raw pixel

Pixel intensity was successfully considered in handwritten character recognition [9,19,20]. The best performances were achieved for handwritten digits using raw images, in particular for classifiers using deep architectures [21].

3.2. Profiles

Upper and lower profiles are computed considering the distance between the upper/lower horizontal line and the closest pixel to the upper/lower boundary of the character image. Similarly, we extracted the left/right profiles too.

This feature, a rather coarse representation of the character's outer shape, highly depending on the character's orientation and size, gives a less complex representation [22]. The representational power of this feature is much lower than all the others used in this experimental setup. A comparison among the different feature spaces is given in Table 1. For this purpose, we considered the two data sets used in our experiments. For each set a k-nn ($k = 1$) was performed to determine how these features discriminate the different digits and characters.

3.3. Local binary patterns

Local binary patterns (LBP) [10] were applied with success for face recognition, where the local texture can reveal differences. Even though characters are simpler shapes, local vicinity observed by the LBP provides a rather complex, and to some extent, rotation invariant representation of the characters. The discriminating power of this feature is similar to the raw pixels and the Radon transform.

3.4. Radon transform

The Radon transform computes projections of an image along some well-defined directions [11]. The Radon function computes the line integrals from multiple sources along parallel paths, or beams, in a certain direction. The beams are spaced one pixel unit apart. To represent an image, the Radon function takes multiple and parallel-beam projections of the image from different angles by rotating the source around the center of the image.

Table 1
k-nn classification accuracy (%) ($k = 1$) for the MNIST and Lampung test samples considering in the process all 60,000 labels and 23,447 labels, respectively.

Feature type/k-nn	MNIST		Lampung	
	Acc	# features	Acc	# features
Pixels	**96.91**	784	83.94	1024
Profiles	76.14	112	65.36	128
LBP	95.24	256	79.54	256
Radon	96.29	301	**90.61**	343
Encoder	96.76	200	88.57	200