Published in final edited form as:

Pattern Recognit Lett. 2015 June 1; 58: 23–28. doi:10.1016/j.patrec.2015.02.001.

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

Szilárd Vajda^{a,**}, Yves Rangoni^b, and Hubert Cecotti^c

^aNational Library of Medicine, National Institutes of Health, Bethesda, MD, 20894, USA

^bHenri Tudor Public Research Center, Kirchberg, L-1855, Luxembourg

^cFaculty of Computing and Engineering, University of Ulster, Londonderry, BT48 7JL, Northern Ireland, UK

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.

1. Introduction

The exponential increase of images to be processed and analyzed nowadays opens new challenges in the field of document recognition (Ghosh et al., 2010; Jain et al., 2000). 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

© 2015 Elsevier Ltd. All rights reserved.

^{**}Corresponding author: Tel.: +1-301-594-7811; fax: +0-000-000-0000; szilard.vajda@nih.gov (Szilárd Vajda).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

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 (Bai et al., 2010; Hu et al., 2014). 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 semiautomatic, high accuracy labeling mechanisms – involving mostly unsupervised learning strategies, e.g., k-means (Lloyd, 2006), self organizing maps (SOM) (Kohonen et al., 2001), growing neural gas (GNG) (Fritzke, 1994). 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 (Richarz et al., 2014) 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 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 (Vajda et al., 2011), more sophisticated and lower dimensionality features such as profiles, local binary patterns (Pietikäinen et al., 2011), and Radon transform (Miciak, 2010; Cecotti and Vajda, 2013) were considered to better exploit the advantage of the original method (Vajda et al., 2011). 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. Section 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 (Hirata and Honda, 2011), 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., 2012) 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 (Li et al., 2014).

A similar attempt is proposed in (Le et al., 2013), 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 (Vajda et al., 2011; Richarz et al., 2014), 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 (Verlinde et al., 2000; Rahman and Fairhurst, 2003). 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 (LeCun et al., 2001; Vajda et al., 2006, 2011). The best performances were achieved for handwritten digits using raw images, in particular for classifiers using deep architectures (Ciresan et al., 2012).

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 (Heutte et al., 1998). 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) (Pietikäinen et al., 2011) 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 (Miciak, 2010). 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.

3.5. Encoder network

Encoder networks are a special kind of deep learning architectures proposed originally by Hinton (Hinton and Salakhutdinov, 2006). Unlike standard statistical data reduction method (e.g. Principal Component Analysis), such encoder network is data-driven. Thus, the new space projection is learnt using supervised learning. Figure 1 provides the schematic of such an encoder process. The features considered are the outputs of the bottleneck layer. The size of this layer controls the dimensionality of the feature. Such features have been successfully applied in our previous work (Vajda et al., 2011).

4. Classifiers

4.1. Definitions

We consider the problem of classification with $N_c > 2$ classes. Each class is denoted by C_i , $i \in \{1, \dots, N_c\}$. Let's consider a classifier E and an input pattern $\mathbf{x}$ of dimension N_f features; $\mathbf{x}$ belongs to a set of N_p patterns. N_p^i represents the number of patterns that should be assigned to the class C_i , and N_+^i the number of patterns that are assigned to the class C_i after classification. The accuracy of the classifier is then defined by:

$$\text{Acc} = \frac{\sum_{i=1}^{N_c} N_+^i}{N_p} \quad \text{with} \quad N_p = \sum_{i=1}^{N_c} N_p^i \quad (1)$$

For an ensemble of classifiers E_m , $m \in \{1, \dots, M\}$, M stands for the total number of classifiers, we define combination approaches such as consensus voting and majority voting. An input pattern $\mathbf{x}$ is assigned to the class for which there is a consensus, i.e., when at least m of the experts agree. We denote by consensus voting, where $m = M$, and by majority voting the solution where m is defined as:

$$m = \begin{cases} M/2 + 1 & \text{if } M \text{ is even} \\ (M+1)/2 & \text{if } M \text{ is odd} \end{cases} \quad (2)$$

For the performance analysis, we denote by N_{dec} , the number of patterns that have a class assigned by the combination of expert, i.e., when a decision was made, and N_{rej} the number of patterns that have no assigned patterns ($N_{dec} + N_{rej} = N_p$). Among the N_{dec} patterns, we define by N_+ and N_- , the number of patterns that have been correctly and incorrectly classified, respectively. For the performance evaluation of the classifiers combination, we consider the following measures:

$$\text{Dec} = \frac{N_{dec}}{N_p} \quad \text{Acc}_c = \frac{N_+}{N_p} \quad \text{Acc}_r = \frac{N_+}{N_{dec}} \quad (3)$$

Dec corresponds to the number of decisions that were taken by the combination of classifiers. Acc_c represents the accuracy of the classifier combination, and Acc_r is the number of correct detections among the N_{dec} decisions.

4.2. Unsupervised clustering methods

We compare three state-of-the art unsupervised clustering techniques. First, the general Lloyd algorithm¹ is a method to find k well-shaped and uniformly distributed partitions among data points (Jain, 2010). Each iteration estimates new centroids and re-partitions the points according to the closest centroid. Second, we use Self Organizing Map (SOM), which is a special type of neural network trained in an unsupervised fashion, to produce a two-

¹We use the name "Lloyd algorithm" to refer to k-means clustering

dimensional mapping of the input data (Kohonen et al., 2001). Third, the growing neural gas (GNG) method (Fritzke, 1994), successfully applied on character recognition problems (Marinai et al., 2008), which does not have constraints on the topology contrary to the SOM.

4.3. Supervised classification

The k-nearest neighbor (k-nn) classifier can be considered as one of the pioneer supervised classification methods (Fix and Hodges, 1951). Considering an annotated data collection, for an unknown data sample the class label is assigned based on the majority of its k nearest neighbors. The nearest neighbor classifier is a special case of the previously mentioned one for $k = 1$. The method has several advantages such as: simplicity, effectiveness, non-parametric, and high performance for classification tasks (Torralba et al., 2008). Its parameters are the major challenges; the proper selection of the neighborhood size (k), and the distance function responsible for the quality of the topological representation can have a significant impact on the results (Gou et al., 2012).

4.4. Evaluation method

To objectively evaluate different clustering methods, various measurements were proposed based on quality and quantity (He et al., 2003). For the current purpose, a measurement based on the cluster distribution was considered, namely the cluster compactness, which is computed using the generalized definition of variance of a vector:

$$v(X) = \sqrt{\frac{1}{N} \sum_{i=1}^N d^2(x_i, \bar{x})} \quad (4)$$

where $d(x_i, x_j)$ is a metric defined between the vectors x_i and x_j . N stands for the number of vectors in X , while $\bar{x} = \frac{1}{N} \sum_i x_i$ is the mean of X (total numbers of vectors to be clustered). Using Eq. 4, the cluster compactness for $C_1, C_2, \dots, C_K$ is defined as:

$$\text{Compactness} = \frac{1}{K} \sum_i^K \frac{v(C_i)}{v(X)} \quad (5)$$

Due to the fact, that these compactness measures combine inter-class and intra-class variances, the selection of one measure or another is not influencing much the final results. The more compact the clusters are, the more reliable the labeling strategy would be.

5. Labeling strategy

5.1. Method

The labeling strategy algorithm is presented in pseudo code (Alg. 1). Considering a set of patterns $\mathbf{x}_i, 1 \leq i \leq N_p$, which are represented by M set of different features $F_m, 1 \leq m \leq M$, the technique provides a deterministic way to label the patterns by minimizing human involvement, and therefore maximizing the automatic process of clustering M representations of $\mathbf{x}_i$ by using unsupervised clustering strategies with K clusters.

For each F_j representation, the points are clustered in k clusters, each cluster being represented by its centroid c_t^j . As the centroids represent features, they cannot be presented directly to a human expert to determine the corresponding label. Therefore, the closest real data point for any c_t^j is selected, and each of those original data points is annotated by the human expert. The resulting number of points to label is $m \times k$: one point for each cluster and each feature representation. Then, for each feature representation, each point inherits a label from its associated cluster.

The label corresponding to a data point is obtained by voting. It is determined by the number of winning votes accumulated over the different F_j feature representations (Cecotti and Beläid, 2005; Kittler and Alkoot, 2003; Kuncheva, 2004). Two voting approaches are considered: consensus and majority.

Algorithm 1

Labeling strategy using complete majority vote.

```

1:  Points:  $\forall i \in [1, n], P_i \in \mathbb{R}^q$ 
2:  Features:  $\{F_j, \forall j \in [1, M]\}$  represents  $P_i$ , any  $\|F_j\| \ll q$ 
3:   $k$  the number of clusters, and the clustering method  $C_k$ 
4:
5:  for  $j \leftarrow 1, m$  do
6:    Cluster  $C_k(F_j)$  producing cluster centers  $c_t, t \in [1, k]$ 
7:    Label cluster centers  $c_t^j \in F_j$ 
8:    Expand the  $c_t^j$  label to all cluster elements
9:    for  $i \leftarrow 1, n$  do
10:     for  $j \leftarrow 1, M$  do
11:       Label( $P_i^j$ )  $\leftarrow$  Label( $P_i$ ) in  $F_j$  representation
12:     if Label( $P_i^j$ ) = M then
13:       Label( $P_i$ )  $\leftarrow$  argmax(Label( $P_i^j$ ))
14:     else
15:       Label( $P_i$ )  $\leftarrow \emptyset$ 
16: return Each data point  $P_i$  is labeled, or rejected

```

While the clustering process could be time consuming, the number of manual annotations is insignificant compared to the number of total data points to be labeled. The more cluster centers are to be labeled, the more precise labels are to be propagated over the data points. In order to lower human expertise involvement, but still keeping high the annotation accuracy, a certain balance should be maintained between the number of feature sets (M) and the number of clusters (K) used in the process. In order to maintain complete control over the number of clusters in our experimental setup, k-means, SOM, and GNG algorithms were considered (see Section 4.2).

6. Databases

MNIST (LeCun et al., 2001) is a benchmark data-set² containing separated Arabic digits assigned to 10 classes. The images stemming mainly from census forms are size normalized and centered to 28×28 gray level images. The data set contains 60,000 training and 10,000 test images.

The Lampung characters³ used in our experiments were extracted from a multi-writer handwritten collection produced by 82 high school students from Bandar Lampung, Indonesia. Samples of the alphabet are presented in Fig. 2.

For the experiments, 23, 447 characters were used as training set and 7, 853 characters for the test. Altogether 20 different character classes were identified. Each character is represented by a centered and normalized 32×32 gray scale image. More details about the data set can be found in (Vajda et al., 2011; Junaedi et al., 2011).

7. Experiments and results

7.1. Clustering performance

The accuracy (in %) and the input feature size by considering the information of the label are presented in Table 1. Those results determine the best results that can be achieved by considering the label information with a k-nn classifier ($k = 1$) using the Euclidean distance (Norm L2). The goal of the proposed method is to prove that with considerably less labels (involvement of human annotators) similar or even better scores can be achieved.

We can observe that the raw feature (see pixels) outperforms the rest of the features in the case of MNIST (96.91%). Interestingly, for the Lampung collection the Radon transform performs the best (90.61%). While the results provided by the encoder network and Radon transform perform in the same range for MNIST, the local binary pattern is slightly less efficient (95.24%). For the Lampung, the encoded results are followed by the pixel based scores. The LBP provides average scores for both data collections. The profiles provide the lower scores, a fact somehow understandable, as the profile encodes only the outer shape of the characters, thus a lot of information is discarded. While for larger k values the nearest neighbor performs slightly better, however, we consider as baseline the results provided by the raw image (96.91%) and the Radon transform (90.61%) for both MNIST and Lampung.

The compactness measures (see Eq. 5) for different methods are reported in Table 2. The results for MNIST and Lampung reported for raw pixels – considering 50, 100, 150 cluster centers show similar a similar trend. The more cluster are available, the more compact the clusters are. Similar trends can be observed for other features as well. While k-means outperform the SOM and GNG, the GNG perform better for larger cluster numbers. This effect can be explained with the rigidity of the SOM, as the neurons in the map cannot modify their spatial organization, while for GNG neuron connections can disappear. These results strongly support the upcoming results (see Table 3 and Table 4), as the more compact

²<http://yann.lecun.com/exdb/mnist/>

³<http://patrec.cs.tu-dortmund.de/files/Lampung-Dataset.zip>

the clusters are, the more chances we have to have pure clusters, a necessary condition for a successful label extension and voting.

7.2. Labeling performance

In order to report results on the labeling accuracy, we considered two voting schemes. First we considered a simple majority vote (at least 3 labels out of the possible 5 should vote for the same label), while for second choice the complete majority vote was used (in order to accept a label as being “valid” all 5 labels should be the same). While the vote type can be considered as a quality measure for the label acceptance, this still does not guarantee that the voted label will be the correct one. It might happen that even though the vote is indicating a certain label and the sample is accepted in the training set with the corresponding label, the choice is wrong. To measure the quantity of those samples where the vote is valid (3 methods provide the same label for an image: majority voting, or the 5 methods provide the same label for an image: consensus voting), we report the percentage of the original data, where the labels agree based on the votes, while the accuracy (Acc.) stands for accuracy among those who have been selected (voted) in the training samples.

In Table 4, it can be observed that for each feature and clustering method the simple majority criterion is met much more often than the complete majority, which was expected due to the strong constraint of the consensus voting. However, once the samples selected in the new training set are defined (based on the voting schemes), we can observe that the accuracy is much higher in the case of samples selected based on majority vote. The majority vote assures that the selected samples with the corresponding labels can be considered as valid labels with higher certainty. However, for underrepresented classes, due to the nature of clustering methods, some classes might be attracted by larger ones, therefore, these classes might be wrongly labeled, thus for the voting these classes might not be selected at all.

Once the new training sets are built based on the votes, we performed a k-nn ($k = 1$) classification to compare our results with the original ones reported in Table 1. For the simple majority vote, using k-means and a maximum number of 750 human labeled samples the score reach 96.77% which is similar to 96.91% reported for considering 60,000 samples. For the complete majority the 750 human labeled samples (clustered originally by GNG) performs up to 93.28%, which is lower than the scores reported in Table 1, but the scores could be considered with high confidence.

7.3. Classification performance

In Table 1, a k-nn classification scheme was considered using all the available labels. The accuracies can reach 96.91% and 90.61% with 10,000 and 7,853 test images for MNIST and Lampung, respectively. Using the same setup, but using only a few labels, the scores reported in Table 4 can go up to 96.77% and 89.53% respectively. The results show that competitive recognition scores can be obtained with few human-labeled samples.

To prove the efficiency of the method, we tested it against several Monte Carlo simulations which pick random samples from the complete training and their corresponding labels.

Different set sizes are tested (i.e. 250, 500, 750) and k-nn ($k = 1$) is applied on pixel values (MNIST) and Radon (Lampung) to provide a one-to-one comparison. Each experiment was repeated 100 times, the average results and standard deviations are reported in Table 5.

It can be observed that the scores reported in Table 5 for MNIST are up to 9.79% less than ours using the same number of labels. A similar trend can be observed for the Lampung data set, where the net gain is 15.94%, proving the superiority of the proposed method over random solutions for two different data collections containing 10 and 18 classes, respectively. The feature diversity and clustering strategies can improve radically the performances, while the huge advantage is the low number of samples to be annotated by the human expert.

To compare different classification schemes, we implemented a fully-connected multi-layer perceptron classification scheme too, using the same inputs as mentioned in Table 4. The input layer size was defined by the dimensionality of the input, 784 and 343, respectively. For the output layer 10 and 20 neurons were used. The hidden layer had 100 neurons for both networks. The results can be seen in Table 6. The results are slightly lower than the results reported in Table 4, which is due to the fact that, the network is more sensitive to erroneously labeled samples, which admittedly appear in the training data.

If we consider that 0.5 second is necessary to label one character, the speedup is around $80\times$ for MNIST and $31.26\times$ for Lampung. In addition, the decrease of the number of images to label can also involve a decrease of the cost required by an expert to label the images.

8. Conclusion

In this paper, we presented an extension of our method for semi-automatic labeling characters with minimal human involvement. For this purpose each character from the data collection was projected in five different feature spaces. Each feature space was clustered using k-means, SOM, and GNG with strict control on the number of clusters. The cluster elements inherited the label of the element closest to the center, annotated by a human expert. Once the clustering was performed in each space, different voting schemes were invoked to decide on the final labels. These newly discovered labels were compared in a k-nearest neighbor scheme, first with randomly selected samples (same number of labels to be known) where up to 15.94% net superiority could have been observed. The second comparison was meant to compare our method (discovered labels) with the complete data collection (all labels are known) and the results are almost similar. While labeling the whole data collection can be a tremendous effort both in cost and time, our method provides results in the same range, but using considerably less human expertise. It is therefore more cost effective, and definitely way faster.

References

- Bai X, Yang X, Latecki LJ, Liu W, Tu Z. Learning context-sensitive shape similarity by graph transduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2010; 32:861–874. [PubMed: 20299710]

- Cecotti, H.; Beläid, A. Hybrid OCR combination approach complemented by a specialized ICR applied on ancient documents. *International Conference on Document Analysis and Recognition*; 2005. p. 1045-1049.
- Cecotti, H.; Vajda, S. A radial neural convolutional layer for multi-oriented character recognition. *International Conference on Document Analysis and Recognition*; 2013. p. 668-672.
- Ciresan, DC.; Meier, U.; Gambardella, LM.; Schmidhuber, J. *Neural Networks: Tricks of the Trade*. 2. 2012. Deep big multilayer perceptrons for digit recognition; p. 581-598.
- Fix, E.; Hodges, JL. USAF School of Aviation Medicine. 1951. *Discriminatory Analysis: Nonparametric Discrimination, Consistency Properties*.
- Fritzke, B. *Advances in Neural Information Processing Systems*. 1994. A growing neural gas network learns topologies; p. 625-632.
- Ghosh D, Dube T, Shivaprasad A. Script recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2010; 32:2142–2161. [PubMed: 20975114]
- Gou J, Du L, Zhang Y, Xiaong T. A new distance-weighted k-nearest neighbor classifier. *Journal of Information and Computational Science*. 2012; 9:1429–1436.
- He, J.; Tan, AH.; Tan, CL.; Sung, SY. *On Quantitative Evaluation of Clustering Systems*. Kluwer Academic Publishers; 2003.
- Heutte L, Paquet T, Moreau JV, Lecourtier Y, Olivier C. A structural/statistical feature based vector for handwritten character recognition. *Pattern Recognition Letters*. 1998; 19:629–641.
- Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science*. 2006; 313:504–507. 10.1126/science.1127647 [PubMed: 16873662]
- Hirata, NST.; Honda, WY. Automatic labeling of handwritten mathematical symbols via expression matching; *Proceedings of the 8th International Conference on Graph-based Representations in Pattern Recognition*; Berlin, Heidelberg: Springer-Verlag; 2011. p. 295-304.
- Hu RX, Jia W, Ling H, Zhao Y, Gui J. Angular pattern and binary angular pattern for shape retrieval. *IEEE Transactions on Image Processing*. 2014; 23:1118–1127. [PubMed: 24144665]
- Jain AK. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*. 2010; 31:651–666.
- Jain AK, Duin RP, Mao J. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2000; 22:4–37.
- Junaidi, A.; Vajda, S.; Fink, GA. Lampung - a new handwritten character benchmark: Database, labeling and recognition; *International Workshop on Multilingual OCR*; Beijing, China: ACM; 2011. p. 105-112.
- Kittler J, Alkoot FM. Sum versus vote fusion in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2003; 25:110–115.
- Kohonen, T.; Schroeder, MR.; Huang, TS., editors. *Self-Organizing Maps*. 3. Springer-Verlag New York, Inc; Secaucus, NJ, USA: 2001.
- Kuncheva, LI. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley; 2004.
- Le TN, Ho TB, Kawasaki S, Kanda T, Takabayashi K, Wu S, Yokosuka O. A semi-supervised ensemble learning method for finding discriminative motifs and its application. *Journal of Universal Computer Science*. 2013; 19:563–580.
- LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. *Intelligent Signal Processing*. IEEE Press; 2001. Gradient-based learning applied to document recognition; p. 306-351.
- Li, J.; Mouchère, H.; Viard-Gaudin, C. Reducing annotation workload using a codebook mapping and its evaluation in on-line handwriting. *International Conference on Frontiers in Handwriting Recognition*; 2012. p. 752-757.
- Li J, Mouchère H, Viard-Gaudin C. An annotation assistance system using an unsupervised codebook composed of handwritten graphical multi-stroke symbols. *Pattern Recognition Letters*. 2014; 35:46–57.
- Lloyd S. Least squares quantization in pcm. *IEEE Transactions on Information Theory*. 2006; 28:129–137.
- Marinai S, Marino E, Soda G. Self-organizing maps for clustering in document image analysis. *Machine Learning in Document Analysis and Recognition*. 2008:193–219.

- Miciak M. Radon transformation and principal component analysis method applied in postal address recognition task. *International Journal of Computer Science and Application*. 2010; 7:33–44.
- Pietikäinen, M.; Hadid, A.; Zhao, G.; Ahonen, T. *Computer Vision Using Local Binary Patterns*. Springer-Verlag London Limited; 2011.
- Rahman AFR, Fairhurst MC. Multiple classifier decision combination strategies for character recognition: A review. *International Journal on Document Analysis and Recognition*. 2003; 5:166–194.
- Richarz J, Vajda S, Grzeszick R, Fink GA. Semi-supervised learning for character recognition in historical archive documents. *Pattern Recognition*. 2014; 47:1011–1020.
- Torralba A, Fergus R, Freeman WT. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2008; 30:1958–1970. [PubMed: 18787244]
- Vajda, S.; Junaidi, A.; Fink, GA. A semi-supervised ensemble learning approach for character labeling with minimal human effort. *International Conference on Document Analysis and Recognition*; 2011. p. 259-263.
- Vajda, S.; Rangoni, Y.; Cecotti, H.; Beläid, A. A Fast Learning Strategy Using Pattern Selection for Feedforward Neural Networks. *International Workshop on Frontiers in Handwriting Recognition*; 2006. p. 56-61.
- Verlinde P, Chollet G, Acheroy M. Multi-modal identity verification using expert fusion. *Information Fusion*. 2000; 1:17– 33.

Highlights

- We propose a fast and accurate semi-automatic labeling strategy.
- Human expert knowledge (time and cost) is reduced to minimum.
- Unsupervised clustering and voting mechanism decide for the labels.
- The method is generic, can be applied to other type of data.

Fig. 1.
Encoder network considered for feature dimensionality reduction.

Fig. 2.
Samples for all the Lampung script characters.

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

Table 2

Compactness results for k-means, GNG, and SOM for MNIST and Lampung, considering raw pixels as input for multiple cluster setups.

Method/#	MNIST			Lampung		
	50	100	150	50	100	150
k-means	0.87	0.85	0.81	0.86	0.85	0.83
GNG	0.99	0.87	0.83	0.93	0.87	0.85
SOM	0.92	0.89	0.83	0.95	0.9	0.87

For the different feature representations, different clustering methods and number of clusters (#) the performance accuracy (%) obtained on MNIST and Lampung training sets once the labels of the cluster centers were extended to the complete training set.

Table 3

Method	MNIST					Lampung					#
	Pixel	LBP	Pro.	Enc.	Radon	Pixel	LBP	Pro.	Enc.	Radon	
k-means	81.58	80.35	54.80	80.88	71.60	81.58	80.35	54.80	80.88	71.60	50
	86.81	84.96	58.76	85.61	76.96	86.81	84.96	58.76	85.61	76.96	100
	88.60	88.09	64.34	88.85	82.59	88.60	88.09	64.34	88.85	82.59	150
SOM	60.90	55.73	42.13	66.76	43.74	60.90	55.73	42.13	66.76	43.74	50
	79.11	75.98	63.27	83.18	70.11	79.11	75.98	63.27	83.18	70.11	100
	72.38	72.75	55.75	79.86	62.59	72.38	72.75	55.75	79.86	62.59	150
GNG	80.28	77.79	54.73	80.91	71.02	44.80	44.34	46.77	60.86	46.34	50
	87.68	86.13	62.02	86.75	77.75	48.29	51.70	52.27	64.64	58.80	100
	89.13	88.02	63.70	88.70	81.57	52.76	55.40	53.93	69.83	62.42	150

Table 4

For the different clustering methods (k-means, SOM, GNG) and data sets (MNIST, Lampung), the decision score and the accuracy of the unsupervised classifier decision combination based on majority voting or consensus is given as a function of the number of human annotations (#). Based on those results, the accuracy on the test database is reported with k-nn (k=1).

Method	MNIST						Lampung					
	Majority voting			Consensus			Majority voting			Consensus		
	Train	Test	Acc	Dec	Acc	Dec	Train	Test	Acc	Dec	Train	Test
k-means	93.19	86.93	96.48	37.79	97.10	89.33	82.74	58.60	88.53	26.77	77.22	62.37
	94.47	91.08	96.72	44.75	98.33	91.42	85.25	63.32	88.56	28.99	83.89	69.21
	95.55	93.10	96.77	53.40	98.53	92.51	88.27	66.76	89.53	33.38	88.96	70.58
SOM	72.95	78.27	95.49	16.35	96.04	78.45	66.85	48.00	86.13	16.37	69.62	49.20
	92.74	86.28	96.74	42.35	97.04	91.15	77.80	61.44	87.94	17.92	85.22	59.27
	85.41	86.90	96.41	31.19	98.03	85.41	76.34	60.27	88.48	17.27	81.16	60.33
GNG	92.18	85.60	96.69	39.20	96.18	90.54	82.55	55.73	88.48	25.49	77.56	65.43
	95.11	90.96	96.68	49.74	97.85	89.65	86.83	62.84	89.23	29.61	84.33	68.17
	96.73	92.27	96.73	52.79	98.32	93.28	88.40	66.44	89.35	32.67	89.50	72.43

Table 5

k-nn accuracy (k=1) (%) results with standard deviations using randomly generated training sets and training sets generated by our method for 250, 500 and 750 samples (#) for MNIST and Lampung data.

#	MNIST		Lampung	
	Rand.	Method	Rand.	Method
250	79.74 $\pm$ 0.99	96.69	62.88 $\pm$ 1.77	88.53
500	84.82 $\pm$ 0.64	91.08	70.02 $\pm$ 1.03	89.23
750	86.98 $\pm$ 0.51	96.77	73.59 $\pm$ 0.88	89.53

Table 6

Neural network classification accuracy (%) for the MNIST and Lampung test samples considering for training the samples selected with majority and consensus vote.

Vote/#	MNIST			Lampung		
	250	500	750	250	500	750
Majority	94.22	94.16	95.05	71.88	73.60	74.23
Consensus	82.37	86.71	86.80	46.31	59.13	52.89