https://doi.org/10.2352/ISSN.2470-1173.2017.2.VIPC-411


# Semi-supervised Learning Feature Representation for Historical Chinese Character Recognition

*Xiaoyi Yu, Wei Fan, Jun Sun and Satoshi Naoi; Fujitsu Research & Development Center, Beijing, China*

## Abstract

*Historical Chinese character recognition has been suffering from the problem of lacking sufficient labeled training samples. An Semi-supervised learning method based on Convolutional Neural Network (CNN) for historical Chinese character recognition is proposed in this paper. We use traditional feature extraction method to extract features from the unlabeled sample sets at first; then according to the distance between the extracted features, samples pairs are constructed; With the constructed pairs, a Siamese network S is trained; The network structure and weights of model S are used to initialize another CNN model T. The model T is then fine-tuned by a few labeled historical Chinese character samples, and used for final evaluation. Experimental results show that the proposed method is effective.*

## Introduction

Historical Chinese character recognition is very important for classical literature digitization, ancient documents collation and culture preserving. However, historical Chinese character recognition is a very challenge problem compared with modern character recognition. First, the number of historical Chinese characters is much larger than modern Chinese characters; second, the structure of historical Chinese characters is much more complex than modern simplified Chinese characters; third, the historical Chinese characters are much more polymorphic, i.e. a certain number of historical Chinese characters have many variant forms; fourth, the writing style is different because of the use of pen-brushes or woodblock printing; and last, the image degradation of photographed or scanned ancient documents is worse than that of modern documents.

In recent years, deep learning methods, e.g. Convolutional Neural Network (CNN) outperformed traditional methods in OCR research field. Currently dominant CNN based supervised learning methods typically require thousands of millions samples of training material which needs to be explicitly labeled by human. Although there are millions of natural image data available for training, labeling all of such data followed by supervised learning is simply not feasible. As for historical Chinese character recognition, the problem, which it has been suffering from the problem of lacking sufficient labeled training samples, is even worse. We can obtain a huge amount of unlabeled samples from ancient literature by scanning, or photographing, and then segmenting using automatic character segmentation method. Manually labeling such a data set is time consuming, labor consuming and expensive. To make effective use of such huge amounts of data in character recognition, the practical unsupervised or semi-supervised learning approach would be called for.

In this paper, we take a practical approach to propose a semi-supervised learning method using unlabeled training samples. If successful, the benefit of such semi-supervised learning would be tremendous. Through semi-supervised learning, we can utilize the unlabeled samples to benefit the training process using labeled samples.

The early work on completely unsupervised training has been done in the field of the machine-printed text recognition(OCR) and utilized cipher breaking algorithms [7]. Recent research works on semi-supervised/unsupervised learning for character recognition in literature includes [1-2]. Coates et.al. [1-2] mimic the CNN procedure to use convolution process to filter robust image features, and pooling process to reduce the feature dimension, and the K-means are used to perform clustering over training data to discover semantic classes for semi-supervised/unsupervised learning. Most semi-supervised/unsupervised methods in literature are not specifically designed for character recognition. We are not intending to give detailed survey of semi-supervised/unsupervised learning, just list some typical method in this area. The above mentioned method can be grouped into clustering method. Other popular direction of semi-supervised/unsupervised learning: generative model. Most generative models have this basic setup, but differ in the details. Earlier work on generative mode includes Hinton et.al 's famous" wake-sleep" algorithm [9]. Typical method includes Generative Adversarial Networks (GANs) [3-4], Variational autoencoders (VAEs) [5] and Autoregressive models[6]. Kozielski [8] relies on aprior language model.

There is another popular direction on semi-supervised/unsupervised learning: feature learning directly from samples using deep neural network [10]. In [11], visual tracking is used to provide a weak supervision of unlabeled image patches. That is, two patches connected by a track should have similar feature since they probably belong to the same object or object part. Then a Siamese-triplet network is applied to train a CNN model. The network structure and weights of the trained Siamese model is fine-tuned for image classification and other task. The last step of [11] is very similar to deep CNN based transfer learning [12].

Our method is also related to above described feature learning framework [11] and fine-tuning style[12], but we don't directly apply the method [11] in our traditional Chinese character recognition. First, we don't have videos with a lot of traditional Chinese characters to train a Siamese network. Second, even we have such videos; the variation of historical Chinese character (different people have different writing style with different writing method such as pen-brushes or woodblock printing) is quite different from image blocks in video (usually same object in different view angle) .

Inspired by these methods, we try to apply feature learning method in an alternative way for historical Chinese character recognition problems. Specifically, we propose a semi-supervised learning method based on Convolution Neural Network (CNN) for historical Chinese character recognition.

The rest of this paper is organized as following. Section 2 describes the principle of CNN based semi-supervised learning. Experimental results are given in Section 3. The final section is the conclusion.

## Proposed Method

Semi-supervised/unsupervised learning, one major branch of machine learning involving learning without labeled data, has been a long standing research over decades. But it has achieved much less success compared with supervised learning that requires labeled training data. Our goal is to train a convolutional neural network using hundreds of thousands of unlabeled character samples. The network is then fine-tuned using labeled character samples to improve the recognition rate .Figure 1 shows the block diagram of our proposed method.

As we mentioned in the introduction section, we can't apply the video track method to connect two image patches to set up a weak supervise information. We explore the alternative: traditional character recognition method. Traditional character recognition methods usually consist of two steps：Manually design a feature extractor to extract features, and then train a classifier to classify the features. The two steps are performed separately, so the parameters in each step are optimized independently. Although traditional methods are not globally optimized, they still set up a weak connection between samples. That is, two samples with a very close feature distance should be the same character with a certain probability, while a large distance of two samples must belong to different character. A Siamese network can encode this kind of information. The weights of the network should be close to the global optimized parameters for character classification. The fine-tuning of the weights using labeled samples then can improve the recognition performance.

Our system proceeds in several stages:

1. We try to utilize the traditional or CNN based method to extract features from the unlabeled sample sets;
2. According to the distance between the extracted features, samples pairs are constructed;
3. With the constructed pairs, a Siamese network is trained;
4. The labeled samples are utilized to fine-tune the trained Siamese network for final classification.

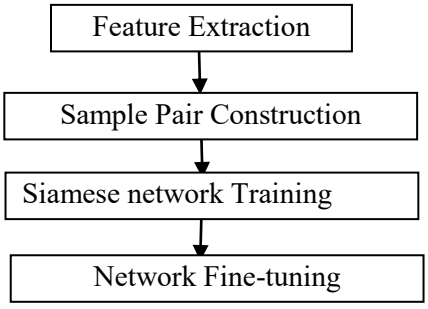


Figure 1 System diagram

The key steps are described as below.

### Feature Extraction

Given an unlabeled sample set, we want to choose sample pairs to create training instances. One obvious way to find pairs of interest is to compute distances between them. Sample pairs with a small distance are labeled as similar, and large distances for pairs are labeled dissimilar. However, since sample pairs are raw images and noisy, it is hard to calculate the distance directly. Thus we should extract the features from raw samples at first. We use traditional feature extraction method for character recognition such as discriminative feature extraction method - modified quadratic discriminant function (MQDF) to extract a feature for each sample in a subspace. Some other techniques, including zoning, projections and profiles, and crossings and distances can be used as statistics of feature too. Alternative method is to train an end to end Siamese network to measure the similarity directly using labeled samples, then the Siamese network is used as a feature extractor to extract features.

Both of the method are used in our proposed method.

### Sample Pair Construction

For a labeled dataset, the data samples with a same label can be grouped to similar pairs and different labels to dissimilar pairs. We assume that the total number of samples is N, and the sample number in each class is M. So the number of similar pairs is $N(M-1)/2$, and the number of dissimilar pairs is $N(N-M)$. Usually N >>M, thus the number of similar pairs is much less than that of dissimilar pairs. For a database with hundreds thousands samples, the unbalance of the similar pairs and dissimilar pairs affect the training greatly.

For an unlabeled dataset, there is no label information to group a similar pair or a dissimilar pair. But we have extracted features for all the unlabeled samples. The distances between the extracted features carry such information, which can be used to construct pairs. We can use distance measurement to generate pairs of samples. To utilize any amount of pairwise constraints from a dataset to train a pre-training neural network, we try to samples pairs according to the similar pairs and dissimilar pairs ratio, which is $\frac{2(N-M)}{M-1}$. This ratio can be obtained using labeled dataset. We use the following procedure to generate millions of such pairs for a Siamese network.

```
for i = 1 to all
        for j = i+1 to all-1
            calculate the distance between sample i and j
            if distance is less than the threshold
                the sample pair i and j are labeled as simillar
                class (class A)
            end
            if distance is larger than the threshold
                the sample pair i and j are labeled as different
                class (class B)
            end
        end
    Randomly choose a certain number of pairs from class A
    and a corresponding number of pairs from class B to form
    the training set and testing set.
end
```

Figure 2 Sample pairs construction

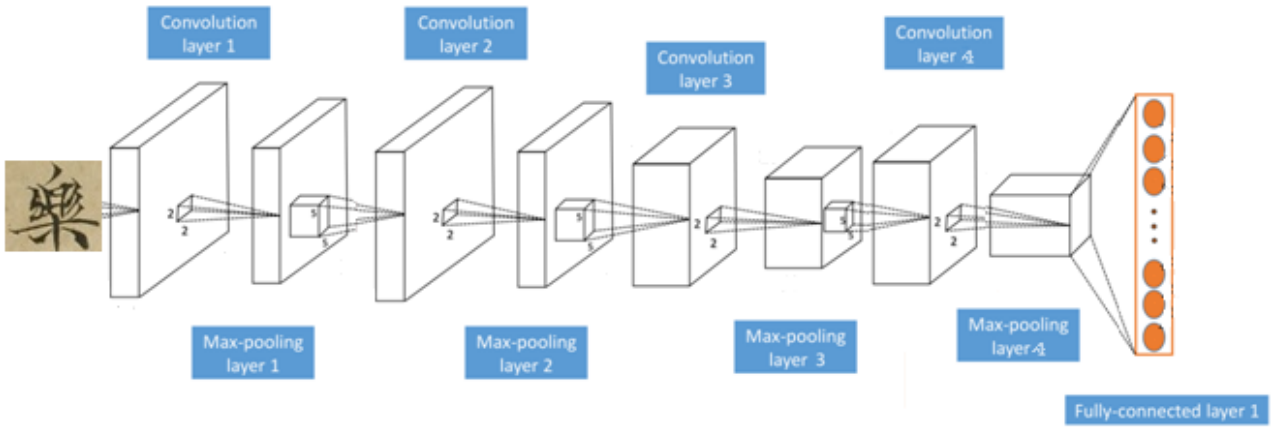


Figure 3 Base network

### Siamese network Training

Our goal in this step is to learn a pre-trained network using unlabeled samples such that a sample pair with small feature distance, which is calculated in the previous step, is closer than that with a larger distance. This learned network is then fine-tuned from the learned model weights. To learn this pre-trained model we design a Siamese network. A Siamese network consist of 2 base networks which share the same parameters. The base network is shown as Figure 3 for the convolutional and Max-pooling layers. Then we stack one fully connected layers on the pool4 outputs, whose neuron numbers is 1024. Thus the final output of each single network is 1024 dimensional feature space $f(\cdot)$. We define the loss function on this feature space. Other architecture can also be used to construct the Siamese network structure. We will discuss in the experimental section.

Loss function: Given the set of pairs sampled from the image set, we propose to learn a similarity model in the form of CNN. Specifically, Given a pair of features f(X1) and f(X2), obtained by feeding sample pair X1 and X2 into network f, we define the distance of two image X1, X2 based on the distance in the feature space as, $D(X1, X2) = \| f(X1) - f(X2) \|$. Based on the distance, the loss is calculated using the contrastive loss [13].

Our Siamese model and final CNN classification mode, which will described in next subsection, are based on Caffe open source tools created by Jia et al.[13]. In the preprocessing step, the samples are normalized and converted to the input data format required by Caffe. During the training process, Siamese model S is trained by sample pairs obtained in Sample Pair Construction stage. All the training data in this stage are unlabeled historical Chinese character image samples.

The parameters of the base network are shown Table 1.

Table 1. Parameters of the base network

| | conv1 | pool1 | conv2 | pool2 | conv3 | Pool3 | conv4 | Pool4 |
|---|---|---|---|---|---|---|---|---|
| num_output | 64 | - | 128 | - | 256 | - | 256 | - |
| kernel_size | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| stride | 1 | 2 | 1 | 2 | 1 | 1 | 1 | 1 |
| pad | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |

### Fine tuning

Given the Siamese model learned by using unlabeled data, we want to transfer the learned model to the final historical Chinese character recognition with labeled data. In our experiments, we directly apply our trained Siamese model as a pre-trained network for the final historical Chinese character recogntion. The network architecture consist of based Siamese net followed by two fully connected layers with a final softmax corresponding to the class number of the character set. In the fine-tuning stage, we use the parameters of the convolutional layers in the base network of the learned siamese architecture as initialization for the final historical Chinese character recogntion. For the fully connected layers, we initialize them randomly. This method of transferring feature representation is very similar to the approach applied in [12]. The fine-tuning stage is controlled by setting the learning rate of the corresponding layer. If the learning rate of a layer is set to 0, the weights of this layer will not be changed in the fine-tuning process. In our experiments, the learning rates of the convolutional layers are set to normal values or larger values, so the network can correct the errors caused by the weak connection in pair construction stage.

## Experimental Results

We first introduce the datasets used in our experiments, and then the experimental results are described.

Dataset: Our first dataset consists of 206375 historical Chinese characters samples collected from Dunhuang historical Chinese documents. Among these samples, 38050 samples (19042 samples for training and 19008 for testing in this paper) are labeled, and 15835 samples are unlabeled data.

Figure 4 database exmaples

Some Chinese character examples are shown in Figure 4. In this dataset, the character numbers in each class is not equally distributed. There are 138 characters class which have samples larger than 100, and 504 character class with samples larger than 50. The figure 5 shows part character class distribution.

To compare with the method of [12], we also construct a printed Chinese character set, which consists of 150000 samples with about 4808 character class. In this set, all the samples are labeled.

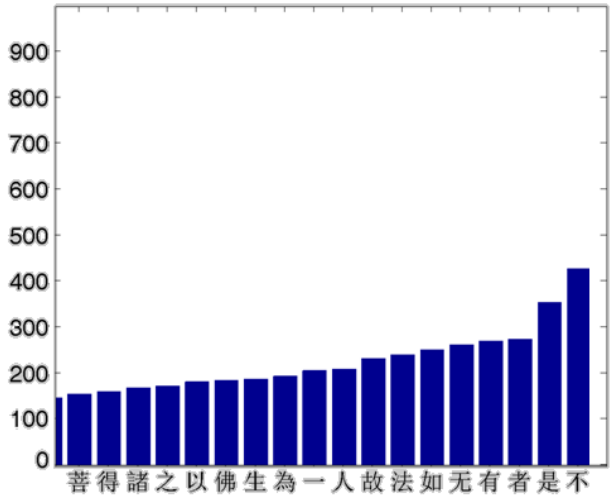


Figure 5 Part character class distribution

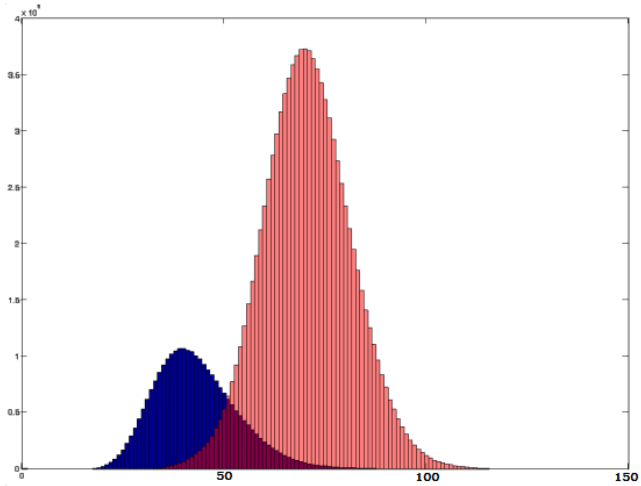


Figure 6. Semi-supervised CNNs without Finetuning

## Semi-supervised CNNs without Finetuning

First, we demonstrate that the semi-supervised-CNN feature learned using pair construction (without fine-tuning) is reasonable. We use the method [1] to extract the features of 158325 unlabeled samples. According to the distances between the extracted features, sample pairs are constructed. The constructed pairs are feed to the Siamese network using method described in previous section. To verify the effectiveness of the proposed method, we feed the trained Siamese network with sample pairs of labeled data. The Siamese network can output the distance of the sample pairs. Since we have label information of each sample feed to the network, we can annotate pair as similar or dissimilar easily. The distance distributions calculated using Siamese network are shown in Figure 6. The red bars show the distance distribution of similar pairs and blue ones are the distribution of dissimilar pairs. From the Figure 6, we can see the distribution is quite different and can be partly separated.

.

## Semi-supervised CNNs with Finetuning

Next, we evaluate our approach by transferring the feature representation learned in semi-supervised manner to the tasks with labeled data for character recognition. First, the Siamese net is trained using 15835 unlabeled samples; then the based Siamese net, two fully connected layers and the softmax layer form the final classification network. The weights of the convolutional layers are initialized using the weights of learned Siamese network base net. For the fully connected layers, we initialize them randomly. Following the method in [12], the maximum number of selected samples per character class is set to 10, 20, 30, 40 and 50 respectively. Thus we also form a similar table like the Table V in [12].

Table 2. Training samples setup

| Total samples | 19042 | | | | |
|---|---|---|---|---|---|
| Maximum number per character class for fine-tuning | 10 | 20 | 30 | 40 | 50 |
| Number of labeled samples for fine-tuning | 6707 | 9524 | 12363 | 13632 | 13558 |

We first train a Siamese network with 3 convolutional layers and 3 pooling layer interlaced. Due to the limited space, we won't list the network parameters here. However, the parameters are very similar to Table 1. The traditional method used for sample pair construction is the method [1]. The dimension of the features extracted by the Siamese network is 256. The based network parameters are then initialize a CNN classifier with two adaptive layers. The adaptive layers are randomly initialized. All the weights are then updated using the labeled samples. We used fixed samples in Dunhuang historical Chinese documents excluding training samples as the testing set (the number of testing samples is 19008). To compare the performance of the proposed semi-supervised pre-training method with no pre-training method, we also conduct the training the same CNN classifier using only labeled samples. All the parameters are randomly initialized for no pre-training method. The experimental results are shown in Table 3. The first row shows the Maximum number per character class for fine-tuning. The last row shows the experimental results (the accuracy of historical Chinese character recognition) of the proposed method (use unlabeled samples to do pre-training and then using labeled samples to fine-tune). The middle row shows the results only use labeled data samples to training. From the table, we can see the unlabeled sample improve the classification performance greatly.

Table 3. Experimental Results

| number per character class | 10 | 20 | 30 | 40 | 50 |
|---|---|---|---|---|---|
| Direct Training | 0.66593 | 0.763468 | 0.791509 | 0.805135 | 0.810869 |
| Proposed method | **0.731745** | **0.817708** | **0.844329** | **0.848853** | **0.854851** |

The second experiment is to train a Siamese network with 4 convolutional layers and 4 pooling layer interlaced. The network architecture and network parameters are described in the previous section and in Table 1. For the second experiment, the method to construct sample pairs is a powerful method. The method itself is a CNN based feature extractor trained by handwriting character samples [14]. The dimension of the features extracted by the Siamese network is 1024. We also compare the performance of the proposed semi-supervised pre-training method with transfer learning method [12]. For transfer learning, we use the dataset (printed Chinese character set) described above in this section to train a CNN classifier, then the CNN is fun-tuned using the labeled samples. The experimental results are shown in Table 4. The second row shows the results of direct training method. The second row shows the results of transfer learning. The last row shows the experimental results of the proposed method. From the table, we can see the unlabeled sample improve the classification performance greatly. With pairs constructed by method of [14], our method outperforms the transfer learning method [12].

Table 4. Experimental Results

| Samples/class | 10 | 20 | 30 | 40 | 50 |
|---|---|---|---|---|---|
| Direct Training | 0.59680 | 0.73884 | 0.77135 | 0.78115 | 0.80314 |
| Transfer Learning | 0.75710 | 0.82286 | 0.84412 | 0.84945 | 0.85711 |
| Proposed Method | **0.83585** | **0.86600** | **0.87615** | **0.87905** | **0.88331** |

From table 3 and 4, we can also find that the sample number and the depth of base network affect the performance. For example, when samples per class is less than 10, the performance of deep network is worse than the shallow network, just because of training samples are not enough. With the help of unlabeled sample for pre-training, the performance is improved largely.

From table 3 and 4, we can also see the traditional method used for sample pairs construction affect the proposed method greatly. A strong feature extraction method can construct an accurate similar pairs and dissimilar pairs, which is very important for unlabeled pre-training.

## Conclusion

This paper presents a CNN-based semi-supervised learning method which is useful for historical Chinese character recognition. Experiments show that the proposed semi-supervised method can help to improve the supervised classification, but several factors affect the final performance, i.e., traditional feature extraction method (which affects the pairs construction for pre-training), the base network architecture and the schemes of updating network parameters. In future research, semi-supervised learning methods based on generative models will be explored.

## References


[1] Coates A, Carpenter B, Case C, et al. Text detection and character recognition in scene images with unsupervised feature learning[C]//2011 International Conference on Document Analysis and Recognition. IEEE, 2011: 440-445.

[2] Netzer Y, Wang T, Coates A, et al. Reading digits in natural images with unsupervised feature learning[J]. 2011.

[3] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems. 2014: 2672-2680.

[4] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv preprint arXiv:1511.06434, 2015.

[5] Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.

[6] van den Oord A, Kalchbrenner N, Kavukcuoglu K. Pixel Recurrent Neural Networks[J]. arXiv preprint arXiv:1601.06759, 2016.

[7] G. Huang, E. Learned-Miller, and A. McCallum, “Cryptogram decoding for optical character recognition,” University of Massachusetts, Amherst, MA 01003, Tech. Rep., 2006.

[8] Kozielski M, Nuhn M, Doetsch P, et al. Towards Unsupervised Learning for Handwriting Recognition[C]//Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on. IEEE, 2014: 549-554.

[9] G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal. The” wake-sleep” algorithm for unsupervised neural networks. Science, 268(5214):1158–1161, 1995.

[10] E. Hoffer and N. Ailon. Deep metric learning using triplet network. CoRR, /abs/1412.6622, 2015.

[11] Wang X, Gupta A. Unsupervised learning of visual representations using videos[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 2794-2802.

[12] Tang Y, Peng L, Xu Q, et al. CNN Based Transfer Learning for Historical Chinese Character Recognition[C]//2016 12th IAPR Workshop on Document Analysis Systems (DAS). IEEE, 2016: 25-29.

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional Architecture for Fast Feature Embedding,” arXiv preprint arXiv:1408.5093, 2014.

[14] Chen L, Wang S, Fan W, et al. Beyond human recognition: A CNN-based framework for handwritten character recognition[C]//2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). IEEE, 2015: 695-699.


## Author Biography

*Xiaoyi Yu received her BS in Industry Automation from Hunan University, China (1995) and her PhD in Pattern Recognition from Institute of Automation, Chinese Academy of Science, China (2005). Since then he has worked in Tokyo University, Tokyo, Japan, Osaka University, Osaka, Japan , Peking University, Beijing, China and Fujitsu R&D Center, Beijing, China. His work has focused on image processing, computer vision.*