https://doi.org/10.2352/ISSN.2470-1173.2018.2.VIPC-174

Generative Adversarial Networks for Open Set Historical Chinese Character Recognition

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

Abstract

Historical Chinese character recognition has been suffering from the problem of samples labeling, not only the problem of lacking sufficient labeled training samples, but also of sample classes. So the scenario for Historical Chinese character recognition is "open set" recognition, where incomplete labeling of sample classes is present at training time, and unknown classes can be submitted to the system during testing. This paper proposes a method for open set Historical Chinese Character Recognition. For open set recognition, the features available in the training data cannot effectively characterize different kinds of unknown classes. We assume that the features which characterize unknown classes can be derived or learned from other similar data sets. We utilize an auxiliary data set combined with the open set training data set to learn good features to represent historical Chinese characters. The auxiliary data set is translated using Generative Adversarial Networks (GAN) to make sure that the translated data set is as close to the historical Chinese character dataset as possible. Then we construct a neural network for features extraction. The neural network is trained using an alternative training method with the translated auxiliary dataset and incomplete labeled historical Chinese character data set. Last, features are extracted from certain layer of the trained neural network. Unknown samples are detected using statistical modelling of the Euclidean metric between samples. 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. In our previous work, we have proposed a semi-supervised learning method using unlabeled training samples to improve the recognition accuracy [1]. Although semi-supervised learning algorithm works, it is essentially based on the closed world assumption [2]. In other words, it is assumed that the testing data pertains to one of K classes that are used during training. But in practice, training data for historical Chinese character recognition are labeled sequentially (e.g. page by page, book by book). Using the labeled data to train a model, during testing time, testing data may come from a class that is not necessarily seen in training. This problem where the testing data corresponds to a class that is not seen during training is known as open set recognition [2]. As for historical Chinese character recognition, if the training samples correspond to K different historical Chinese character classes (manually label characters from several historical books to form a training set), then given a test image corresponding to a character from one of the K classes, the algorithm should be able to recognize the character. However, if the test image corresponds to a character which does not match one of the K character classes seen during training, then the algorithm should have the capability to ignore or reject the test sample [2]. The reject test samples can be labeled manually (as a new class) and putted into the training set again to train a new model. Figure 1 shows the scenario of our problem to be solved.

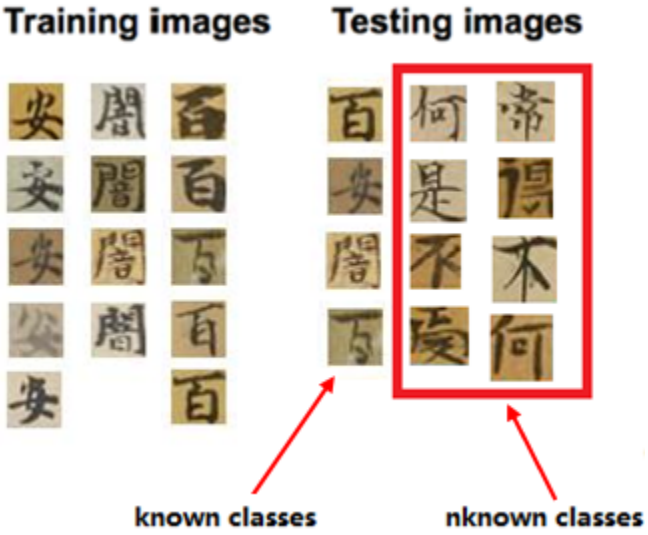

Figure 1. Open Set Historical Chinese Character Recognition

For **Closed Set Recognition**, all testing classes are known at training time. For **Open Set Recognition**, incomplete knowledge of the world is present at training time, and unknown classes can be submitted to an algorithm during testing. The goal of an open set recognition algorithm is to learn a predictive model that classifies the known data into correct class and rejects the data from open class [2]. A number of approaches have been proposed in the literature for open set recognition. For instance, [3] introduced a concept of open space risk and developed a 1-vs-Set Machine formulation using linear SVMs for open set recognition.

[2] proposed a generalized Sparse Representation based class reconstruction errors algorithm for open set recognition. The tail of those error distributions are modeled using the statistical Extreme Value Theory (EVT). In [4], the concept of Compact Abating Probability (CAP) was introduced for open set recognition. In particular, Weibull calibrated SVM (W-SVM) algorithm was developed which essentially combines the statistical EVT with binary SVMs for open set recognition. Also, the W-SVM framework was recently used in [5] for fingerprint spoof detection. We are not intending to give detailed survey of open set recognition, just list some typical method in this area.

We found that most algorithms of open set recognition in literature assume that the features available in the data can effectively characterize different kinds of unknown classes (open set classes). If such features were not available in the data, it would not be possible to utilize the characteristics of previously discovered unknown unknowns to find new ones. But in many scenario, the assumption did not hold or partially hold, e.g. the end to end CNN based method [3] prefers to features of known classes. So learning good features is the key to unknown class detection.

In this paper, we take a practical approach to propose a representation learning (feature learning) method using auxiliary data set to augment the features for open set recognition. If successful, the benefit of such open set recognition method would be tremendous.

The rest of this paper is organized as following. Section 2 describes the principle of our proposed method. Experimental results are given in Section 3. The final section is the conclusion.

Proposed Method

Most open set recognition methods proposed in the literature focus on identifying unknown classes based on the assumption, that the features available in the data can effectively characterize different kinds of unknown classes. In reality, such good features not always exist. Our goal is to learn a good representation (features) from training data set of open set recognition problem and/or auxiliary outside data set, then apply the learned features for open set recognition. Figure 2 shows the block diagram of our proposed method.

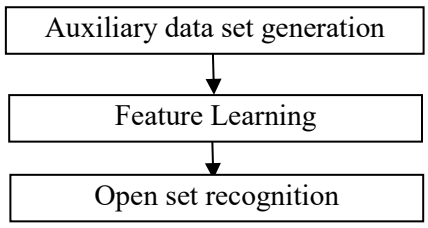

Figure 2. System diagram

As we mentioned in the introduction section, we want to utilize an auxiliary data set combined with the open set training data set to learn good features to represent historical Chinese characters. The features then can be applied for the detection of unknown classes. In principle, any data set can be used for the feature learning. In reality, if the auxiliary data set is totally different from the training data set, the learned features may be unrelated to the training data. The learned features are useless for unknown class detection. A proper auxiliary data set should be chosen at first. Furthermore, we hope the auxiliary data set is as closer to the training data set as possible. A translation to the original auxiliary data set is necessary, so the translated auxiliary data set can be used to extract features which are not only strong for auxiliary data expression of itself, but also for the detection unknown classes in historical Chinese character recognition. The translated auxiliary data set combined the training data set are feed into a CNN to learn features. The learned features can be used to improve the open set recognition performance.

Our system proceeds in several stages:

1. We choose an auxiliary data set and apply translation operation on the data set;
2. We try to utilize a CNN based method and alternative training to learn good features from the training data set and translated auxiliary data sets;
3. Extract the learned features from the training data set for unknown class detection.

The key steps are described as below.

Auxiliary data set translation

Since our goal is to detect unknown classes of historical Chinese characters. Many character recognition dataset can be used as auxiliary dataset. In this research, CASIA handwriting dataset [7], containing samples of isolated characters and handwritten texts, is chosen as auxiliary data set in this research. Figure 3 shows examples of historical Chinese character (bottom row) and CASIA handwriting dataset (top row). The CASIA handwriting dataset is close to our historical Chinese character training data set, but the writing style is quite different. A translation is need to apply in the CASIA handwriting dataset to ensure that these two data sets are close enough.

Figure 3. examples of CASIA handwriting dataset and historical Chinese characters

Many image processing and computer vision tasks, e.g., image segmentation, stylization, and abstraction, can be posed as image-to-image translation problems [8], which convert one visual representation of an object or scene into another. Generative Adversarial Networks (GANs) for cross-domain image-to-image translation have made much progress recently [9]. Many methods need thousands to millions of labeled image pairs to train a GAN. However, human labeling is expensive, even impractical, and large quantities of data may not always be available. In [8], authors developed a dual-GAN mechanism, which enables image translators to be trained from two sets of unlabeled images from two domains. We applied this method for our character set translation.

In our architecture, the primal GAN learns to translate images from CASIA handwriting dataset to those in historical Chinese character training data set, while the dual GAN learns to invert the task. As noted in [8], the closed loop made by the primal and dual tasks allows images from either domain to be translated and then reconstructed. Figure 4 shows the network architecture and data flow chart of the translation. The figure is revised from Figure 1 in [8].

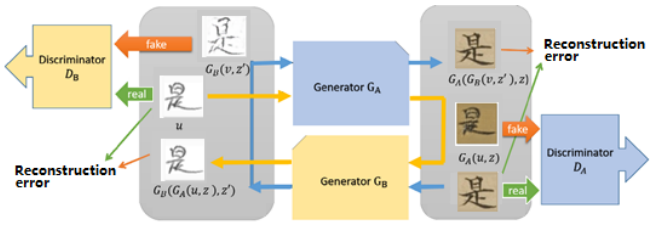

Figure 4. Network architecture of Auxiliary data translation

Alternative learning for feature representation

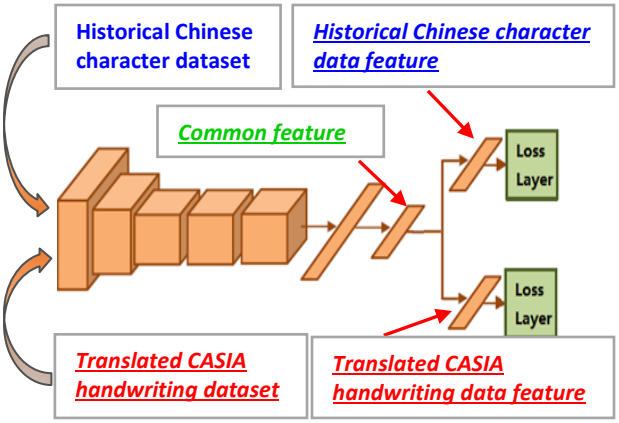

Figure 5. Alternative learning for feature representation

After the translation operation on the auxiliary dataset, we obtained a translated auxiliary dataset which is very close to the historical Chinese character set, and we can learn powerful features from these two datasets (historical Chinese character and translated CASIA handwriting dataset).

Our method is somewhat related to transfer learning framework [6], but we don't directly apply transfer learning [6] for our open set traditional Chinese character recognition problem. The transfer learning algorithms learn tasks in a sequential fashion and it has limitation of catastrophic forgetting, that is an inevitable feature of connectionist models. Our approach is similar to multi-task learning, which can remember all tasks by an alternative or parallel training procedure. Specifically, we demonstrate our approach is scalable and effective by applying feature learning method in an alternative or parallel way for open set historical Chinese character recognition problems.

Figure 5 shows the learning framework. Mini-batches of training samples from historical Chinese character training data set and translated CASIA handwriting dataset are feed into a CNN alternatively. The network weights below the common layer are updated using the loss of both the historical Chinese character training data and translated CASIA handwriting data. The specific feature layers (historical Chinese character feature layer and translated CASIA handwriting feature layer) are updated by the loss of corresponding data.

The training procedure is shown in Figure 6 and described as follows:

a. Feed the model using min-batch data from translated CASIA handwriting dataset and calculate the loss using the corresponding loss layer;
b. Update the weights of below the common layer and the corresponding feature layer;
c. Feed the model using min-batch data from historical Chinese character dataset and calculate the loss using the corresponding loss layer;
d. Update the weights of below the common layer and the corresponding feature layer;
e. Go to step a. until convergence.

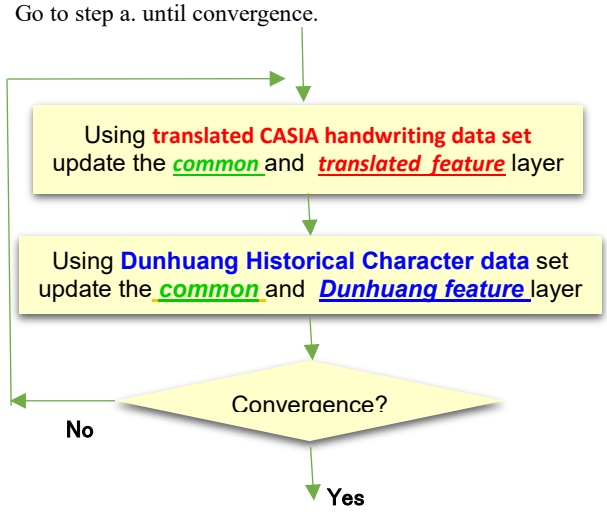

Figure 6. Training procedure

Open set historical Chinese character recognition

For open set recognition, in order to reject invalid samples, a score or value for matched or non-matched samples are defined first. For example, the notion of Sparsity Concentration Index (SCI) was proposed in [11]. Similarly, a rejection rule based on the ratio of the first two highest projection scores was developed for rejecting non-face images in [10]. The score distributions of the matched samples and that of non-matched samples are modeled using some statistical distraction such as Gaussian distribution, the statistical Extreme Value Theory (EVT) [12] etc.

In this paper, we use a simple score (Euclidean metric between sample features) and Gaussian distribution to model the scores for open set recognition. We believe that features are the much more important than the definition of the score and distribution of the score. To verify our proposed method, we use the widely and frequently used metric and statistical distribution in this paper. However; other metrics such as cosine distance metric, joint Bayesian metric, their combinations and others can applied in our proposed method too. Although Gaussian distribution is applied in this research, the statistical EVT [12] is also applicable in our research. Our method consists of two main stages. In the first stage, the distributions of the matched Euclidean metrics and the non-matched Euclidean metrics are modeled using the Gaussian distribution to simplify the open set recognition problem into two hypothesis testing problems (see [2] for the detailed hypothesis). In the second stage, the Euclidean metrics corresponding to a test sample from each class are calculated and the confidence scores based on the two distributions are fused to determine the identity of the test sample.

Experimental Results

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

Open Set Historical Chinese Character Recognition Training (abbreviation: Open Set) Dataset: Our first dataset consists of 206375 historical Chinese characters samples collected from Dunhuang historical Chinese documents. Among these samples, we randomly choose 41381 samples (22373 samples, which are from 300 character classes, for training and 19008 samples, which are 300 known classes and 1053 unknown classes, for testing in this paper) for our open set historical Chinese character recognition training dataset.

Auxiliary dataset: as described in Section of proposed method, CASIA handwriting dataset is chosen as auxiliary dataset. The dataset consists of 3755 character classes and 897910 samples in total.

Some Chinese character examples are shown in Figure 3. In Open Set 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. In auxiliary dataset, the character numbers in each class is roughly equally distributed. There are about 240 characters in each class

Based on these dataset, we implemented experiments shown as follows.

Auxiliary dataset translation

First, we demonstrate that writing style translation from CASIA handwriting style to Historical Chinese Character style is reasonable. We use the method [8] to train a dual-GAN model, which enables image translators to be trained from two sets of unlabeled images from two domains. The training procedure is the same procedure as that described in [8].

In our experiment, the image set A is our Auxiliary dataset, i.e. CASIA handwriting dataset, and the image set B is the Open Set Dataset. Total 897910 samples from CASIA handwriting dataset and total 41381 samples of Open Set dataset have been used in our experiment. We show some auxiliary data translation results in Figure 7. The odd number columns show original samples of original CASIA handwriting dataset and the even number columns show the translated results. From the figure, we can see the translation operation can translate the original CASIA handwriting dataset to a dataset which is very close to the historical Chinese character dataset.

Figure 7. Auxiliary data translation results

Open set recognition

After the translation operation on the CASIA handwriting dataset, we obtained a translated CASIA handwriting dataset which is very close to the historical Chinese character set, and we can learn powerful features from these two datasets (historical Chinese character and translated CASIA handwriting dataset). The training procedure is detailed described in Section of "proposed method". The neural network consists of 2 convolution layers and 2 pooling layers, followed by paralleled full connection layers corresponding to open set loss and translated auxiliary loss. We train the network using all the training samples of 300 classes in open set dataset and samples of 100 classes, 200 classes and 464 classes from translated auxiliary dataset respectively.

After the training procedure using alternative training method, we can extract the features from training data set (open set dataset). For training dataset, each class is then represented as a point, a mean feature (MF) with the mean computed over only the correctly classified training examples. Given the MF and an input image, we measure distance between them. We directly threshold distance to determine an overall maximum distance threshold. Of course, other complicated method such as per class meta-recognition model [13] can be applied here. As we mentioned in the Section of Open set historical Chinese character recognition, we believe that the feature is much more important than distance metrics and statistical modeling of the metrics. If we don't consider statistical modeling (EVT) of the distance metrics, the method [13] is just the method of extracting features from only open set dataset. Hence, we treat the method of extracting features from only open set dataset as a benchmark for comparisons in this paper.

For a given threshold on features distance values, we compute true positives, false positives and false negatives over the entire testing samples of the Open Set dataset. For example, when testing the system with images from testing samples of open set dataset, true positives are defined as the correctly classified as unknown classes on the testing samples, false positives are incorrect classified as unknown classes on the testing set and false negatives are images from the unknown classes that the system incorrectly classified as known examples. Figure. 8-11 shows performance of our prosed method for varying thresholds. In Figure 8-11, the horizontal axis is the false positive ratio of open set recognition, i.e. $FP/N = FP/(FP+TN)$ where FP is the number of false positives, TN is the number of true negatives and $N=FP+TN$ is the total number of negatives. The vertical axis shows true positive ratio of open set recognition, i.e. $TP/P = TP/(TP+FN)$ where FN is the number of false negatives, TP is the number of true negatives and $N=FP+TN$ is the total number of positives.

In Figure 8, the ROC curves show open set recognition results using features extracted from the last pooling layer. Different color represents different network training methods which are denoted in figure legends. From Figure 8, we can see that the proposed method can find much more unknown classes with the help of translated auxiliary dataset, which shows that the effectiveness of the proposed system.

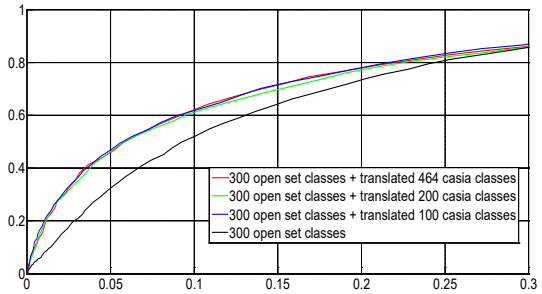

Figure 8. ROC curve

Figure 9 shows the open set recognition results using features extracted from the full connection layer corresponding to the open set loss. The performance is not robust. So the features from the full connection layer is not appropriate for open set recognition.

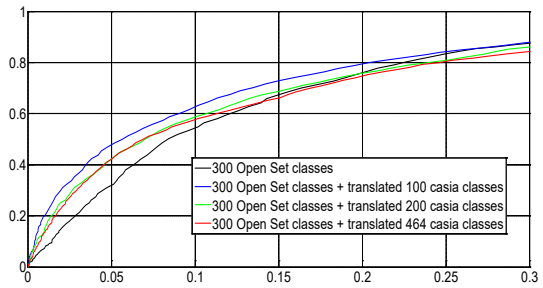

Figure 9 ROC curve

Figure 10, 11 compares performance using different training numbers of auxiliary dataset. In Figure 10 and 11, blue curves show the results using features from network trained using 100 classes from auxiliary dataset, green and red curves show results from 200 and 464 classes respectively. From the figures, the number of training samples have impact on the features from the full connection layer.

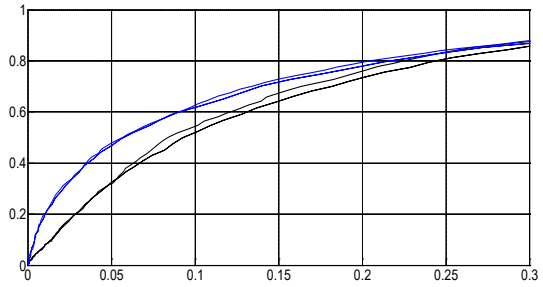

Figure 10 ROC curve

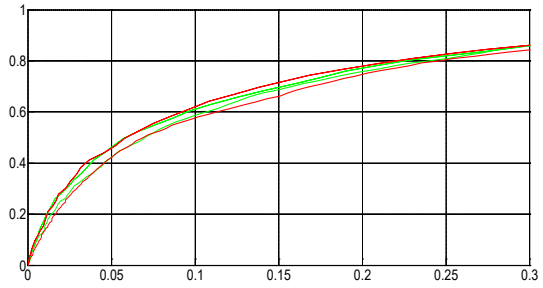

Figure 11. ROC curve

In all, our experiments show that the proposed approach consistently obtains better performance using features from pooling layer on open set testing

Conclusion

This paper presents a GAN-based open set recognition method which is useful for historical Chinese character recognition. Experiments show that the proposed method can help on improvement on unknown class detection, but several factors affect the final performance, i.e., layers where features are extracted. In future research, other feature enhanced methods will be explored.

References

[1] Yu, Xiaoyi, Wei Fan, Jun Sun, and Satoshi Naoi. "Semi-supervised Learning Feature Representation for Historical Chinese Character Recognition." Electronic Imaging 2017, no. 2 (2017): 73-77.

[2] Zhang, He, and Vishal M. Patel. "Sparse representation-based open set recognition." IEEE transactions on pattern analysis and machine intelligence39.8 (2017): 1690-1696.

[3] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, “Toward open set recognition,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 35, no. 7, pp. 1757–1772, 2013.

[4] W. J. Scheirer, L. P. Jain, and T. E. Boult, “Probability models for open set recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI), vol. 36, November 2014.

[5] A. Rattani, W. Scheirer, and A. Ross, “Open set fingerprint spoof detection across novel fabrication materials,” IEEE Transactions on Information Forensics and Security, vol. 10, no. 11, pp. 2447–2460, Nov 2015.

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

[7] Liu, Cheng-Lin, Fei Yin, Da-Han Wang, and Qiu-Feng Wang. "CASIA online and offline Chinese handwriting databases." In Document Analysis and Recognition (ICDAR), 2011 International Conference on, pp. 37-41. IEEE, 2011.

[8] Yi, Zili, Hao Zhang, and Ping Tan Gong. "DualGAN: Unsupervised Dual Learning for Image-to-Image Translation." arXiv preprint arXiv:1704.02510(2017).

[9] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In European Conference on Computer Vision (ECCV), pages 702–716. Springer, 2016.

[10] V. M. Patel, T. Wu, S. Biswas, P. J. Phillips, and R. Chellappa, “Dictionary-based face recognition under variable lighting and pose,” IEEE Transactions on Information Forensics and Security, vol. 7, no. 3, pp. 954–965, 2012.

[11] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp. 210–227, 2009.

[12] J. Pickands III, “Statistical inference using extreme order statistics,” the Annals of Statistics, pp. 119–131, 1975.

[13] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In CVPR, 2016.

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.