Chinese Journal of Electronics
Vol.25, No.3, May 2016

Rejecting Character Recognition Errors Using CNN Based Confidence Estimation*

LI Pengchao[1], PENG Liangrui[1] and WEN Juan[2]

(1. *Tsinghua National Laboratory for Information Science and Technology, Dept. of Electronic Engineering, Tsinghua University, Beijing 100084, China*)

(2. *Equipment Academy, Beijing 101416, China*)

Abstract — Although Optical character recognition (OCR) technology has achieved huge progress in recent years, character misrecognition is inevitable. In order to realize high fidelity content of document digitalization, we propose a new Convolutional neural networks (CNN) based confidence estimation method. We detect the misrecognized characters through comparing the confidence value with a preset threshold, so as to leave the recognition errors as embedded images in the output digital documents. We adopted sofmax as the estimation of posteriori probability, overlap pooling and maxout with dropout technologies in CNN architecture design. Experimental results show that our method has achieved an explicit improvement compared to baseline system.

Key words — Optical character recognition (OCR), Confidence estimation, Convolutional neural networks (CNN).

I. Introduction

Although the state-of-art Optical character recognition (OCR) technologies have achieved a high performance with the total accuracy of above 95%[1] when applying to the scanned document images, some words or characters are still difficult to recognize due to degradation of printed document images, *e.g.* noises, broken or touched strokes. Specially, according to statistic data[2], more than half of the residual errors is mainly caused by the segmentation errors. Character misrecognition is inevitable due to failures in layout analysis, segmentation and classification in OCR systems.

Based on current OCR technologies, a conventional method to reduce recognition errors has been developed. Full information of a document is recorded including: image level, text level, physical layout level, and logical structure level. The final document is reconstructed in formatted documents with the full information, and output as the standard PDF, RTF, HTML, and XML, *etc.* The misrecognized text is proofreaded manually.[3] This task is usually tedious and error-prone. In order to digitalize documents accurately, we try to propose a meaningful solution, which is, to leave the suspicious misrecognized characters as embedded character images in the output electronic document. To achieve this goal, the suspicious misrecognized characters should be labeled without human aid, which means the confidence value of each text processed by OCR system should be estimated. If the confidence value is above a preset threshold, it is conceived as positive true; otherwise, the recognition of suspicious candidates will be rejected. The key question is how to determine the confidence value for the success of the misrecognized characters labeling.

Regret to say, most of the current OCR systems adopt distance based classifier whose output has been confirmed not reflecting confidence value[4]. Xiaofan Lin *et al.* proposed a method to transform the output of a distance based classifier into a confidence-like output[5]. The goal of their method is to combine different weak classifiers into a powerful one. Inspired by the idea of Xiaofan Lin *et al.* in Ref.[5], we use confidence estimation to reject recognition errors. The main contribution of this paper is described as follows. First, deep learning which has been applied to a number of challenging tasks in computer vision with breakthrough improvements achieved in last few years[6,7], is adopted to realize confidence estimation. Second, to be more precise, we optimize the architecture of CNN[8] to achieve optimal performance on confidence estimation. CNN is capable to refine fea-

*Manuscript Received Feb. 13, 2015; Accepted May 8, 2015. This work is supported by the National Basic Research Program of China (973 Program) (No.2014CB340506), National Natural Science Foundation of China (No.61261130590, No.61032008), and Tsinghua National Laboratory for Information Science and Technology (TNList) Cross-discipline Foundation.

 DOI:10.1049/cje.2016.05.018

ture representation and abstract meaning gradually. CNN has been successfully deployed in many commercial applications from hand writing recognition[9] to scene text detection[10,11,12,13]. Specially, as a kind of neural network, the output of CNN has been proved as a concrete fitting of posteriori probabilities[4].

II. Basic Theory

We all know that human being could easily distinguish text which is hard to recognize. The uncertainty of recognition should be measured in a way like the human being judging process. CNN is a biologically inspired trainable architecture that can learn features automatically while no need of human prior knowledge and interference[14]. The most meaningful advantage of CNN is that the output of softmax layer is natural generalized confidence.

1. CNN output reflects confidence value

CNN model could be divided into two parts, convolutional networks and neural networks. Both of them are usually connected back-to-back. Convolutional networks have several customized stages. The input and output of each stage are sets of arrays called feature maps. The highest layer is connected to some kind of classifier, one of which is softmax classifier.

Softmax model generalizes logistic regression to classification problems where the class label y can take on more than two possible values. Softmax regression is a supervised learning algorithm, and we will later use it in conjunction with deep learning methods. Thus, in our training set$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots (x^{(m)}, y^{(m)})\}$, we now have that $y^{(i)} \in \{1, 2, \dots k\}$. According to MAP regulation, we want our hypothesis to estimate the probability that $p(y = j|x)$ for each value of $j = 1, 2, \dots k$. Thus, hypothesis function will output a k dimensional vector giving us our k estimated probabilities.

$$\boldsymbol{h_\theta}(x^{(i)}) = \begin{bmatrix} P(y^{(i)} = 1|x^{(i)}; \boldsymbol{\theta}) \\ P(y^{(i)} = 2|x^{(i)}; \boldsymbol{\theta}) \\ \vdots \\ P(y^{(i)} = k|x^{(i)}; \boldsymbol{\theta}) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} \mathrm{e}^{\boldsymbol{\theta}_j^{\mathrm{T}} x^{(i)}}} \begin{bmatrix} \mathrm{e}^{\boldsymbol{\theta}_1^{\mathrm{T}} x^{(i)}} \\ \mathrm{e}^{\boldsymbol{\theta}_2^{\mathrm{T}} x^{(i)}} \\ \vdots \\ \mathrm{e}^{\boldsymbol{\theta}_k^{\mathrm{T}} x^{(i)}} \end{bmatrix} \tag{1}$$

where $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_k \in \mathbb{R}^{n-1}$ are the parameters. Just like logistic regression, cost function is defined to be optimized.

$$J(\boldsymbol{\theta}) = -(1/m)[\sum_{i=1}^{m} \sum_{j=1}^{k} \delta(y^{(i)} - j) \log \frac{\mathrm{e}^{\boldsymbol{\theta}_j^{\mathrm{T}}} x^{(i)}}{\sum_{l=1}^{k} \mathrm{e}^{\boldsymbol{\theta}_l^{\mathrm{T}} x^{(i)}}}] \tag{2}$$

where δ represents the famous Kronecker delta function. There is no known closed-form way to solve for the minimum of $J(\boldsymbol{\theta})$, and thus as usual an iterative optimization algorithm is applied such as gradient descent or Limited-memory Broyden Fletcher Goldfarb Shanno (L-BFGS)[15].

The gradient descent algorithm could adopt $\nabla J(\boldsymbol{\theta})$ to minimize $J(\boldsymbol{\theta})$, *e.g.* for each $j = 1$ to k each iteration we update $\boldsymbol{\theta}_{j,n} = \boldsymbol{\theta}_{j,n-1} - \eta \nabla_{\boldsymbol{\theta}_j} J(\boldsymbol{\theta})$ where parameter is a constant study rate, according to standard implementation. Moreover, gradient descent is also adopted to minimize the discrepancy between the desired output and the actual output of the whole network. All the coefficients of all the filters in all the layers are updated simultaneously by the learning procedure. The optimization process follows back-propagation method.

2. Confidence value and rejection criterion

A general theory of recognition confidence estimation is as follows. If there exists a function $c(\omega_i|X)$ and a monotonically increasing function $g(\cdot)$, which satisfy:

$$c(\omega_i|X) = g(p(\omega_i|X)) \tag{3}$$

where $p(\omega_i|X)$ represents the posteriori probability. Then $c(\omega_i|X)$ is called generalized confidence that sample X belongs to class ω_i[5]. The truth is that the output of an optimized multi-level neural network could reflect generalized confidence[4]. If we pre-train a properly designed neural network so as to achieve high recognition accuracy, the output of the highest layer could be chosen as a confidence parameter.

Theorem 1 As for a pattern classifier, given a certain rejection rate P_r, when different rejection regions are selected, if rejection region R is subjected to

$$R = \{x|c_s(x_j) < TH(P_r)\} \tag{4}$$

where $TH(\cdot)$ is a threshold function correlated with P_r, and x_j represents jth input of the estimator. The error rate of the classifier will have the minimized value.

The detailed proof of this theorem could be found in Ref. [5]. Theorem 1 is meaningful to inform us that given a certain rejection rate Pr, if we choose a threshold properly and make unbiased estimate of c_s, the error rate of our recognition system will be minimized. Actually, the theorem show us the way of determining whether a character candidate should be labeled as misrecognition so as to minimize the error rate, which is, just simply to obtain the confidence value and compare it with a pre-set threshold.

Given a character image processed by segmentation module, the output of recognition rejection module is as follows:

$$y(x_j) = \begin{cases} 1, & c(x_j) < TH(Pr) \\ 0, & \text{elsewise} \end{cases} \tag{5}$$

The output is “1” if the generalized confidence is below a certain value; otherwise the output is “0” which means the result is accepted as a correct recognition. Adding this module yields some embedded pictures in a document when the whole recognition process is accomplished.

3. Character recognition system framework

To validate the effectiveness of confidence estimator in rejecting character recognition errors. We propose a document digitalization system. TH-OCR is a document recognition system, which is able to analyze, recognize and reconstruct multilingual documents, such as Chinese, Japanese, Korean, Tibetan, Uyghur, English and mixed-script documents[3,16]. TH-OCR could convert scanned document image into digital document, while preserving its original layout meanwhile achieving high characters recognition accuracy. Scan-based document is first input into TH-OCR system to do first stage classification. Afterwards, the segmented character candidates are transmitted into a CNN based confidence estimator to determine whether a candidate should be classified as a character or not. A confidence value is generated to evaluate the reliability of recognition result.

III. Methodology Details

There are two key design methodology points we conceive as novel and main practical contribution of this paper. First, considering recent published research reports on off-line handwritten recognition or scene text detection scenarios[17,18,19] based on deep learning principle, the deeper learning architecture is, the more generalization ability learning architecture possesses. Second, as the hierarchy growing deeper, the amount of parameter may reach to larger than that of training samples, which may cause overfitting of the network. There are two solutions dealing with overfitting, named overlap pooling and dropout. In this paper we adopt maxout technique[20] in conjunction with dropout[21] to avoid training saturation and overfitting.

1. Overlap pooling

Although CNNs have much fewer connections and parameters compare to standard feedforward neural networks with similar layers, the deep architecture still has a large learning capacity. Compare to the huge amount of parameters of the whole CNN architecture, the training dataset of labeled images is relatively small. Pooling layers in CNN summarize the neighboring neurons in the same kernel map. We adopt overlapping pooling as it is more difficult to overfit.

2. Maxout in conjunction with Drop out

Maxout network is a new type of deep neural network that achieves outstanding result in a range of machine learning tasks.[20] The neurons in maxout network select activations by max pooling across a group of linear pieces, and then pass the activations to the next layer without any nonlinear transformations. Maxout network has two attractive features compared to traditional sigmoid networks. First, as the gradient of maxout network is always equal to one during backpropagation training, this kind of architecture achieves better optimization performance. Second, although the representation maxout neuron produces is not sparse at all, the gradient is highly sparse and dropout will artificially sparsify the effective representation during training, which also facilitates optimization. One has to use regularization methods such as dropout to control overfitting.

3. Our learning architecutre

The architecture of our network is summarized in Fig.1. Take LeNet-5[14] as example, when applying to MNIST**. dataset, the experimental results show that the claimed error rate achieves as low as 0.85%. Euclidean Radial base function (RBF), one for each class is added to the output layer. Such Euclidean distance based classifier is unable to reflect the recognition confidence. Plus, the network is not deep enough to imitate human recognition process. Considering these factors, our network architecture design is similar to the one Krizhevsky *et al.* proposed in Ref.[22], which has accomplished state of art performance in ImageNet***. dataset classification competition. The whole net contains 5 trunk layers, including 3 convolutional layers with max pooling layers and 2 full connected layers. The output of the last fully-connected layer is fed to a 3755-way softmax that plays a classifier role.

Given a 48×48 pixels gray image, we realize maxout pooling in the first convolution layer. First we compute the convolution by 100 kernels group by 2 with size 5×5, with a stride of 2 pixels. Then the output of the convolution layer is connected to maxout neurons. Max pooling layer mentioned in Section II is connected on top of maxout layer. The second convolutional layer takes as input the output of max pooing layer and filters it with 100 kernels of size $3 \times 3 \times 50$. The third convolutional layer has 100 kernels of size $3 \times 3 \times 100$ connected to the (max pooled mentioned above) outputs of the second convolutional layer. The first fully-connected layer have 5000 neurons, and the second fully-connected layer consists of 3755 neurons, corresponding to 3755 GB2312-80 Level-I Chinese characters. Finally a softmax loss layer is connected to the top layer in order to realize classifier. Fig.1 illustrates the full architecture of CNN based confidence estimator.

As is developed by Berkeley vision and learning center (BVLC), Caffe is claimed to provide C++ library with Python and MATLAB bindings for training and deploy-

** MNIST has a training set of 60,000 samples and a test set of 10,000 samples, both of which consist of digits written by different people.

*** ImageNet is a dataset of over 15 million labeled high-resolution images belonging to roughly 22,000 categories. ImageNet large-scale visual recognition challenge (ILSVRC) uses a subset of ImageNet with roughly 1000 images in each of 1000 categories. Hinton team perform experiments on ILSVRC 2010.

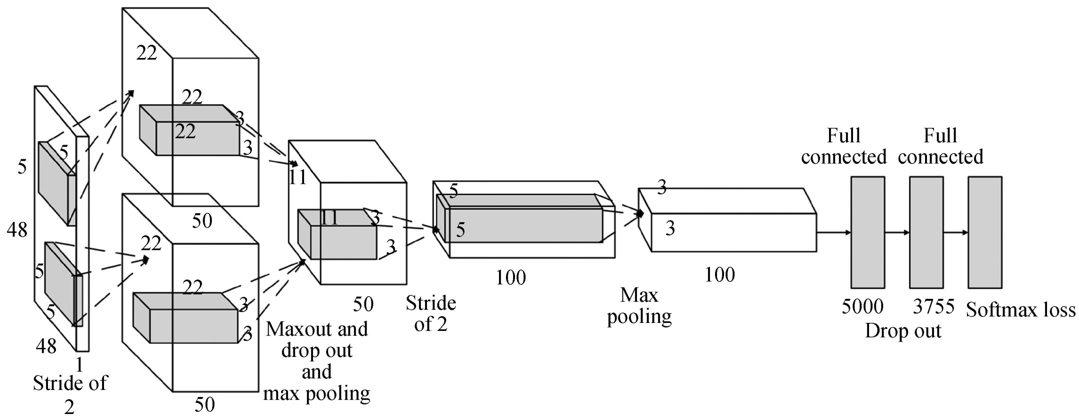


Fig. 1. Basic structure of our CNN model. (Notice that pooling layer is integrated into convolution layer)

ing general purpose convolutional neural networks and other deep models efficiently[23]. Caffe has been proofed as clean, readable, and high speed deep learning model, by which we customize our confidence estimation framework.

IV. Experimental Results

In this section, we present a particular evaluation of the confidence estimator. We perform analysis to evaluate the significance of the confidence estimator independently. In particular, additional experiments are made to examine the effectiveness on segmentation error dataset. Apart from that, the rejection-error rate histogram of the whole system at different confidence thresholds is described to show the balance between them.

1. Experiments with independent TH-OCR

First we put the test samples into TH-OCR to do the basic recognition task without help of confidence estimator. The single character recognition accuracy of TH-OCR system is above 98%. In detail, characters in fonts Song and Fangsong outperform other fonts for about 1%. Table 1 shows quantitative result.

Table 1. Test results of character recognition

Font	Character number	Recognition accuracy
Song	7615	99.01%
Fangsong	1191	99.75%
Kai	7541	98.23%
Hei	6840	98.34%
Youyuan	5583	98.69%
Average	28770	98.61%

2. Experiments with independent CNN-based confidence estimator

A single NVidia TESLA K20 GPU is deployed to do the training task. The training set contains 743 $\times$ 3,755=2,789,965 samples of GB2312-80 Level-I standard Chinese character database. The fonts of the training set consists of 5 types, including Song, Fangsong, Kai, Hei and Youyuan, which are commonly used in Chinese documents. It turns out that 2.7 million training examples are enough to train networks. The error rate of recognizing single character is below 2%. Test set consists of correct and misrecognized samples labeled with human aid. These samples could be conceived as ground truth of the confidence estimator. It means that ideally a perfect confidence estimator should detect all the misrecognized samples. Basically, to validate whether the output of softmax classifier could reflect the generalized confidence value, we depict the output distribution of correctly recognized character set and misrecognized character set respectively.

We could learn from Fig.2 that most of the output values of correct recognized characters are centralized between 0.95 and 1, while most of output values of the misrecognized characters are concentrated below 0.95. The experimental result confirm the fact we mentioned in Sections II and III, *i.e.* the output of the CNN softmax layer reflect the recognition confidence.

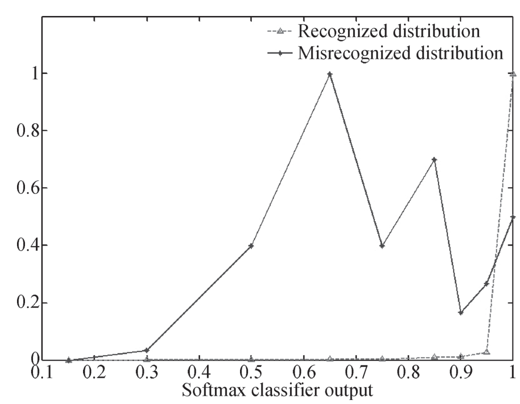


Fig. 2. Softmax output distribution of recognized character set and misrecognized character set

In order to validate our CNN-based confidence estimator precisely, misrecognized characters recall rate and precision rate broken line are drawn on different threshold of the softmax output. The experiment process is as follows. The misrecognized character set is transmitted into the CNN based confidence estimator. For different softmax output threshold, the misrecognized recall rate is defined as:

$$TPR(TH) = \frac{count(\{x|c_s(x) < TH, x \in T\})}{count(T)} \tag{6}$$

where count means the number of a set, x represents the misrecognized samples by TH-OCR, T is misrecognized set processed by TH-OCR. TPR means true positive rate (or recall rate). Precision is defined as the number of misrecognized samples labeled by the confidence estimator divided by the number of all returned results:

$$Pr(TH) = \frac{count(\{x|c_s(x) < TH, x \in T\})}{count(\{x|c_s(x) < TH, x \in Y\})} \tag{7}$$

where Y represents all the samples processed by TH-OCR neglecting whether it is true positive or not. The precision rate reflects the confidence estimator's capability of detecting recognition error without false alarm. Fig.3 shows our results on TPR broken line tested on different CNN architecture.

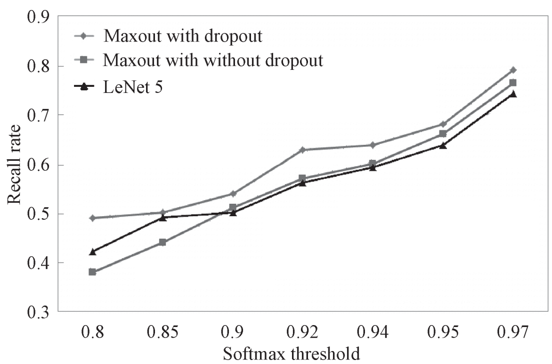


Fig. 3. Recall rate of misrecognized characters

Recall rate subjects to a positive correlation to the output threshold. The performance of maxout network with dropout is 0.7899 when the threshold is set to 0.97. We adopt LeNet-5 as our baseline confidence estimator system. One could deduce from the experimental result that the architecture we designed has accomplished a notable improvement compare to the baseline system.

The precision-threshold curve is to reflect the balance between recall and precision of different thresholds. Fig.4 shows us that more than 98% of the samples with generalized confidence below 0.97 are correct recognized characters, regardless the network architecture. Nevertheless, our network design outperforms baseline architectures in precision. The fact that most rejected characters are correct ones is mainly because that more than 98% of the characters processed by TH-OCR are correct recognized.

The proposed method can also detect character segmentation errors with the accuracy of 93%. Fig.5 shows an example of the confidence estimation results of typical segmentation errors for over-segmentation and touched characters.

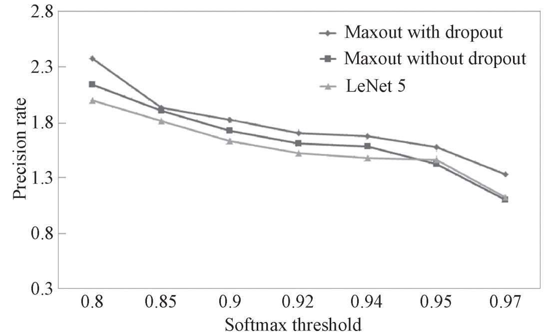


Fig. 4. Precision rate of misrecognized characters

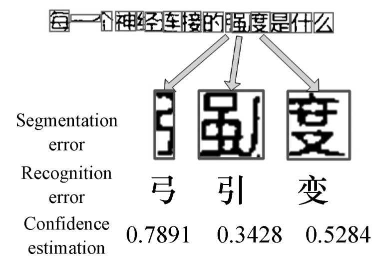


Fig. 5. Segmentation error samples and their confidence value

3. Overal performance of the whole recognition system when applying confidence estimation

To evaluate the importance of confidence estimator as part of document digitalization pipeline, let's examine the confidence estimator in practical situation. Two definitions are described below to express the experimental results. Let N_a be the number of the accurately recognized characters, N_r be the number of rejected samples, N_e be the number of mis-recognized but accepted by the confidence estimator samples, the rejection rate P_r is defined as:

$$P_r = \frac{N_r}{N_a + N_r + N_e} \times 100\% \tag{8}$$

The error rate of the whole system is defined as:

$$P_e = \frac{N_e}{N_a + N_r + N_e} \times 100\% \tag{9}$$

The rejection rate reflects the digitalizing ratio of a document image directly. The higher rejection rate is, the more character pictures will be embedded into the reconstructed document. Error rate reflects effectiveness of the confidence estimation module. We build up test set consists of character samples with a total number of 35,401. Fig.6 illustrate the relationship between rejection rate and error rate. The curve shows that if we reject about 19% character candidates processed by segmentation stage, the accuracy of the whole system will rise up to 99.90%. This is a tremendous stride toward accurate document digitalization with no error.

Output examples of the integrated pipeline recognition system are shown in Fig.7. The suspicious misrecognized characters are shown as embedded character images surrounded by small rectangle. The restoration of documents is accurate although some characters are rejected.

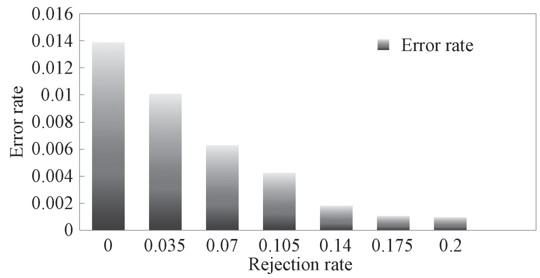
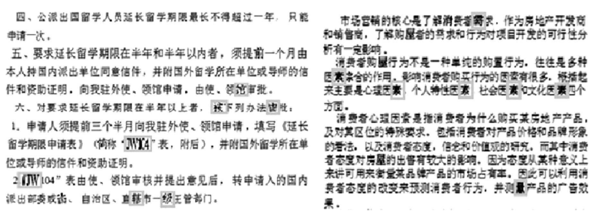


Fig. 6. Rejection-error histogram



四、公派出国留学人员延长留学期限最长不得超过一年，只能申请一次。

五、要求延长留学期限在半年和半年以内者，须提前一个月由本人持国内派出单位同意信件，并附国外留学所在单位或导师的信件和资助证明，向我驻外使、领馆申请，由使、领馆审批。

六、对要求延长留学期限在半年以上者，按下列办法审批：

1．申请人须提前三个半月向我驻外使、领馆申请，填写《延长留学期限申请表》（简称“JWX4”表，附后），并附国外留学所在单位或导师的信件和资助证明。

2．“JW104”表由使、领馆审核并提出意见后，转申请入的国内派出部委或省、自治区、直辖市一级主管部门。

市场营销的核心是了解消费者需求，作为房地产开发商和销售商，了解购屋者的需求和行为对项目开发的可行性分析有一定影响。

消费者购屋行为不是一种单纯的购置行为，往往是多种因素综合的作用，影响消费者购买行为的因素有很多，概括起来主要是心理因素、个人特性因素、社会因素和文化因素四个方面。

消费者心理因素是指消费者为什么购买某房地产产品，及对某区位的特殊要求，包括消费者对产品价格和品牌形象的看法，以及消费者态度，信念和价值观的研究，而其中消费者态度对房屋的出售有较大的影响，因为态度从某种意义上来讲可用来衡量某品牌产品的市场占有率，因此可以利用消费者态度的改变来预测消费者行为，并测量产品的广告效果。

Fig. 7. Output examples of the character recognition system

V. Conclusion & Future Work

Although the state of art OCR system has performed an extremely high performance at approaching 99% accuracy, the residual error should be evaded when applying to some document digitalizing circumstances. In this paper, we have proposed a novel measure to reach this goal. CNN based confidence estimator is to determine whether a character should be rejected as recognition error. This not only means a departure from systems which rely on elaborately hand-engineered feature design, but benefits us by obtaining confidence value through the output of CNN as well. Based on the elegantly designed CNN structure, the detected character recognition errors are converted to embedded character images in the final digital document. The experimental results present a detail evaluation of confidence estimator. With confidence estimator integrated into the recognition process, the accuracy has improved from about 99.0% to 99.90%, which means a state-of-art result.

Even more, the system still has space to improve. Several measures could be implemented to eliminate residual errors. We could try other types of deep learning models rather than CNN to realize confidence estimation[24]. Meanwhile, adopting language model will eliminate errors which could be detected and corrected by context checking.

References

[1] H. Fujisawa, "A view on the past and Future of Character and Document Recognition" *Proc. of ICDAR*, Paraná, Brazil, pp.3–7, 2007.

[2] Chi Fang, Changsong Liu, Liangrui Peng and Xiaoqing Ding, "Automatic performance evaluation of printed Chinese character recognition systems" *IJDAR*, Vol.4, No.3, pp.177–182, 2002.

[3] Ding Xiaoqing, Wen Di, Peng Liangrui and Liu Changsong, "Document digitization technology and its application for digital libraries in China", *Proc. of First Workshop on Document Image Analysis for Libraries (DIAL2004)*, Palo Alto, USA, pp.46–53, 2004.

[4] M.D. Richard and R.P. Lippmann, "Neural network classifiers estimate Bayesian a posteriori probabilities", *Neural Computation*, Vol.3, No.4, pp.461–483, 1991.

[5] Xiaofan Lin, Xiaoqing Ding, Ming Chen, Rui Zhang and Youshou Wu, "Adaptive confidence transform based classifier combination for Chinese character recognition", *Pattern Recognition Letters*, Vol.19, No.10, pp.975–988, 1998.

[6] Yoshua Bengio, "Learning deep architectures for AI", *Foundations and Trends in Machine Learning*, Vol.2, No.1, pp.1–127, 2009.

[7] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks", *Science*, Vol.313, No.5786, pp.504–507, 2006.

[8] Yann LeCun, Koray Kavukcuoglu and Cl' ement Farabet, "Convolutional networks and applications in vision", *Proc. ISCAS 2010*, Paris, France, pp.253–256, 2010.

[9] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard and L.D. Jackel, "Backpropagation applied to handwritten zip code recognition", *Neural Computation*, Vol.1, No.4, pp.541–551, 1989.

[10] Y. Simard, Patrice, D. Steinkraus and J.C. Platt, "Best practices for convolutional neural networks applied to visual document analysis", *Proc. ICDAR*, pp.958–963, 2003.

[11] Tao Wang, D.J. Wu, Adam Coates and Andrew Y. Ng., "End-to-end text recognition with convolutional neural networks", *Proc. of ICPR*, pp.3304-3308, 2012.

[12] Weilin Huang, Yu Qiao and Xiaoou Tang, "Robust scene text detection with convolution neural network induced MSER trees", *Proc. of ECCV*, pp.497–511, 2014.

[13] K. Chellapilla, M. Shilman and P. Simard, "Optimally combining a cascade of classifiers", *Pro. Document Recognition and Retrieval 13, SPIE-IS&T Electronic Imaging*, San Jose, USA, pp.6067, 2006.

[14] Yann LeCun, Yoshua Bengio and Patrick Haffner, "Gradient-based learning applied to document recognition", *Proceeding of IEEE*, Vol.86, No.11, pp.2278–2324, 1998.

[15] Malouf, Robert, "A comparison of algorithms for maximum entropy parameter estimation", *Proc. Sixth Conf. on Natural Language Learning (CoNLL)*, pp.49–55, 2002.

[16] LIU Hailong DING Xiaoqing, "Handwritten character recognition using semi-tied full covariance gaussian mixture model", *Chinese Journal of Electronics*, Vol.14, No.4, pp.649–652, 2005.

[17] Md. Musfiqur Rahman Sazal, Sujan Kumar Biswas, Md. Faijul Amin and Kazuyuki Murase, "Bangla handwritten character recognition using deep belief network", *Proceeding of EICT*, Khulna, Bangladesh, pp.1–5, 2014.

[18] Utkarsh Porwal, Yingbo Zhou and Venu Govindaraju, "Handwritten arabic text recognition using deep belief networks", *Proc. ICPR*, Tsukuba, Japan, pp.302–305, 2012.

[19] Tao Wang, D.J. Wu, A. Coates and A.Y. Ng, "End-to-end text recognition with convolutional neural networks", *ICPR*, Tsukuba, Japan, pp.3304–3308, 2012.

[20] Ian J. Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville and Yoshua Bengio, "Maxout networks", *JMLR*, arXiv:1302.4389, 2013.

[21] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever and R.R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors", arXiv:1207.0580, 2012.

[22] A. Krizhevsky, I. Sutskever and G. Hinton, "ImageNet classification with deep convolutional neural networks", *NIPS*, 2012.

[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding", *Tech report for the Caffe software*, arXiv preprint arXiv:1408.5093, 2014.

[24] MIAO Fuyou, XIONG Yan, CHEN Huanhuan and WANG Xingfu, "A fuzzy quantum neural network and its application in pattern recognition", *Chinese Journal of Electronics*, Vol.14, No.3, pp.524–528, 2005.

LI Pengchao was born in 1983. He received the B.E. degree in information engineering and M.E. degree in signal and information processing from Beijing University of Post and Telecommunication. He is now a doctor candidate of Dept. of Electronic Engineering, Tsinghua University. His research interests include machine learning and character recognition. (Email: lpc13@mails.tsinghua.edu.cn)

PENG Liangrui received the B.S. degree and the M.S. degree from Huazhong University of Science and Technology, Wuhan, China, in 1994 and 1997 respectively. She received the Ph.D. degree from Tsinghua University, Beijing, China, in 2010. She has been a faculty member at the Department of Electronic Engineering, Tsinghua University, since 1997. She is currently an associate professor. Her research interests include multilingual document image recognition and understanding, machine learning, pattern recognition. She was one of the recipients of the National Awards for Science and Technology Progress (Second Class) in China in 1999, 2003 and 2008. She is a member of the IEEE and of the ACM. (Email: plr@ocrserv.ee.tsinghua.edu.cn)

WEN Juan was born in 1982, she received the B.E. degree in information engineering and Ph.D. degree in signal and information processing from Beijing University of Post and Telecommunication. She is now a instructor in Equipment Academy. Her research interests include artificial intelligence and natural language processing. (Email: pljuan0728@126.com)