A Further Step to Perfect Accuracy by Training CNN with Larger Data

Seiichi Uchida, Shota Ide, Brian Kenji Iwana, Anna Zhu
ISEE-AIT, Kyushu University, Fukuoka, 819-0395, Japan
{*uchida, ide, brian, anna*}*@human.ait.kyushu-u.ac.jp*

***Abstract*—Convolutional Neural Networks (CNN) are on the forefront of accurate character recognition. This paper explores CNNs at their maximum capacity by implementing the use of large datasets. We show a near-perfect performance by using a dataset of about 820,000 real samples of isolated handwritten digits, much larger than the conventional MNIST database. In addition, we report a near-perfect performance on the recognition of machine-printed digits and multi-font digital born digits. Also, in order to progress toward a universal OCR, we propose methods of combining the datasets into one classifier. This paper reveals the effects of combining the datasets prior to training and the effects of transfer learning during training. The results of the proposed methods also show an almost perfect accuracy suggesting the ability of the network to generalize all forms of text.**

***Keywords*-convolutional neural network; digit classification; large data; universal OCR; transfer learning**

I. Introduction

Nowadays, most pattern recognition researchers know that deep Convolutional Neural Networks (CNN) have had a significant impact on recognition performance. For example, He et al. [1] proved that CNNs can outperform human beings in a 1000-class visual object recognition task. It is true that the performance by CNNs is beyond our past expectations, although we still can say that human beings have better ability in some more general and practical recognition tasks.

Not surprisingly, CNNs also have had a dramatic impact on the most traditional recognition task, i.e., character recognition. As reviewed in Section II, character recognition performance has been drastically improved by CNNs. A recent trial by Jaderberg et al. [2] achieved state-of-the-art results when using a CNN for end-to-end text recognition in natural scene images. Many methods featured in the ICDAR Robust reading competition [3] were also based on CNNs or their extension. These breakthroughs raise our natural and scientific curiosity about the best possible performance of a CNN on character recognition. It inspires us to explore the maximum capabilities of CNNs in accuracy and flexibility.

The purpose of this paper is to observe the near-perfect recognition performance of CNNs trained with very large datasets. For instance, we use an original handwritten digit dataset with 822,714 samples, which is ten times or larger than MNIST. This very large dataset allows to prepare more than 50,000 ground-truthed samples *for each class*. In addition to the handwritten digit dataset, we use a very large machine-printed digit dataset and a large multi-font digit dataset. Since each dataset has a different distribution complexity, we can observe how the performance of the CNN is affected by the difference.

In addition, we will observe the performance of CNNs in various combinations of these datasets. The first is a trial of *universal OCR*, which can recognize digits in any printing type. In conventional systems, handwritten characters and machine-printed characters are separated first and then classified by their respective OCR systems. In contrast, we use a CNN for recognizing handwritten and machine-printed (and also multi-font) digits without any pre-separation. Another trial is to use different datasets in a scenario of *transfer learning*, where multi-font digits are utilized for recognizing handwritten digits and vice versa.

A. Contributions

The main contributions of the paper are summarized as follows.

1) We report a near-perfect performance on handwritten digit recognition by a CNN (LeNet) trained with a very large dataset comprised of about 820,000 *real* samples. Note that we will use only those real samples for training and not use any artificial samples generated by data augmentation.
2) In addition, we also report a near-perfect performance on machine-printed digit recognition and multi-font digit recognition by CNN trained with large datasets. The performance of the latter recognition task is surprising because the CNN recognizes digits printed in heavily decorated and barely legible fonts – this suggests that the CNN may outperform human beings in its character reading ability. It should be noted that the distribution of multi-font digits is neither a simple Gaussian nor its mixture and thus this proves CNN's discrimination ability on the complicated class distribution.
3) We report the first result of using a CNN for universal OCR. The result shows that the CNN recognizes handwritten and machine-printed digits without serious accuracy degradation from CNNs for each printing type.
4) We observe how multi-font digit dataset is useful on training CNNs for handwritten digit and vice versa. This is a trial of transfer learning between two different printing types.

The most important claim is that we do not need to be serious about isolated digit recognition any more. Even a

Table I
TRIALS OF DIGIT RECOGNITION BY CNN.

Method(Year)	Type / Dataset	#Training + #Test	Accuracy (%)
Neocognitron(2003) [5]	HW / ETL-1	3,000 + 3,000	98.6
LeNet-5(1989) [6]	HW / MNIST	60,000 + 10,000	99.05
Ibid.	HW / MNIST	(†) 60,000 + 10,000	99.20
CNN(2003) [7]	HW / MNIST	(†) 60,000 + 10,000	99.6
CNN committee(2011) [8]	HW / MNIST	(†) 60,000 + 10,000	99.73
Multi-column DNN(2012) [9]	HW / MNIST	(†) 60,000 + 10,000	99.77
CNN+SVM(2012) [10]	HW / MNIST	(†) 60,000 + 10,000	99.81
Cascaded CNN(2012) [11]	HW / MNIST	60,000 + 10,000	99.77
Spatially-sparse CNN(2014) [12]	HW / MNIST	(†) 60,000 + 10,000	99.69
Ibid.	HW / PenDigits	(†) 17,494 + 3,498	99.31
Discriminative cascaded CNN(2015) [13]	HW / MNIST	(†) 60,000 + 10,000	99.82
Stacked sparse auto encoder(2011) [14]	MF / SVHN	604,388 + 26,032	89.7
CNN + Maxout(2013) [15]	MF / SVHN	(†) 604,388 + 26,032	(⋆) 97.84
CNN + original large dataset (This paper)	HW / original	(822,714 →) 90% + 10%	99.88
	MP / original	(622,678 →) 90% + 10%	99.99
	MF / original	(66,470 →) 90% + 10%	95.7
	HW+MP / original	(1,445,392 →) 90% + 10%	99.91
	HW+MP+MF / original	(1,511,862 →) 90% + 10%	99.69

HW: handwritten, MP: machine-printed, MF: multi-font.
(†) Training with augmented data. (⋆) Performance on non-isolated digits.

very simple CNN, LeNet, could achieve near-perfect performance in a very straightforward way to feed a large training dataset. This fact could disappoint some researchers who want to use a more complicated CNN or another recognition scheme with some elaborated data augmentation technique.

II. RELATED WORK

A. CNN for Digit Recognition

It is interesting to note that the pioneering trials of CNNs, such as Neocognitron [4], [5] and LeNet [6], used handwritten digits as their targets. After that, various CNNs have been tested on handwritten digit datasets as reviewed in Table I. In most cases, MNIST has been used for the dataset. MNIST is comprised of 60,000 training samples and 10,000 test samples. Since MNIST is not large enough to train CNN until it can show a near-perfect performance, data augmentation is often employed to generate artificial training samples from the original training samples. As noted in Section I-A, one of the purposes of this paper is to observe near-perfect performance of CNNs with a very large amount of *real* samples. We will see that 800,000 real samples could show a much better accuracy over the past trials and thus can say CNNs will have potential of achieving human character reading ability by a larger dataset.

In [15], it is reported that CNN could recognize hard Latin-alphabet CAPTCHA with 99.8% accuracy. It has been known that computer can have "better" reading ability through breaking-CAPTCHA trials such as [16]. Our trial on multi-font digit recognition also confirms this fact; if a computer is trained with samples which are almost-unreadable to human beings, the computer begins to read similar almost-unreadable samples.

B. Universal OCR

To the authors' best knowledge, there is no prominent trial on universal OCR. In most cases, a two-class classifier is employed in the preprocessing step for separating input characters (or words) into handwritten characters and machine-printed characters and then each character is fed into an OCR module for handwritten characters or machine-printed characters according to its type. One of recent trials on this separation task is Zagoris et al. [17].

A possible reason why universal OCR has not been tried is the difference between the distributions of handwritten characters and machine-printed characters. Roughly speaking, handwritten characters will have a rather anisotropic and wider Gaussian distribution (or Gaussian mixture) whereas machine-printed characters more isotropic and narrower. This difference might lead the use of different recognition techniques (such as quadratic discrimination for handwritten characters and nearest neighbor classification for machine-printed characters).

III. CONVOLUTIONAL NEURAL NETWORKS

A simple CNN shown in Fig. 1 is used for all experiments in this paper. It is similar to a LeNet-5 [6] but uses ReLU as the activation function and max-pooling as subsampling[1]. This is a rather shallow CNN compared to the recent CNNs, such as ResNet [18]. However, it still performed with an almost perfect recognition accuracy, as shown in the following sections. The network is initialized with random values (i.e., without any pre-training) and then trained with back-propagation for 10 epochs.

IV. ORIGINAL DATASETS

Three original datasets are prepared for training and testing CNNs. The size of each dataset is very large enough

[1]This is default of LeNet implemented in `Caffe` library.

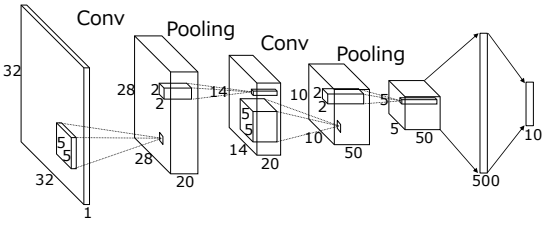

Figure 1. The architecture of the CNN used for our experiments.

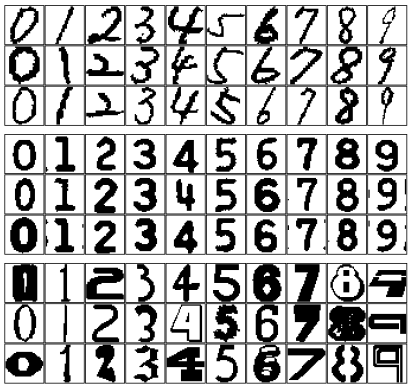

Figure 2. Examples of digit images. From top to bottom, handwritten digits, machine-printed digits, and multi-font digits.

to observe the near-perfect performance of CNNs. Note that all digit samples in those datasets are 32×32 binary images.

- **Handwritten digit dataset**[2] (HW/original) contains 822,714 images collected from forms written by multiple people. Each sample is carefully ground-truthed by manual inspection of several experts. The number of the samples of each class is different but almost the same. (The samples of "0" are slightly more than the other classes.)
- **Machine-printed digit dataset** (MP/original) contains 622,678 images collected from documents printed with two (standard) font types. Similar to HW/original, each sample is carefully ground-truthed by manual inspection of several experts. As shown in Fig. 2, images are often degraded by scanning noise and thus show some difference even on the same font.
- **Multi-font digit dataset** (MF/original) contains 66,470 images generated from 6,647 different fonts. The font data is collected from Ultimate Font Download[3] as TrueType font data. From about 11,000 downloaded font data, 6,647 fonts are selected by removing "ornamental fonts" like ⚠ and ♆ and initial cap fonts. As shown in Fig. 2, fonts in this dataset can show heavy decorations and are often difficult to read even by human. The distribution of multi-font digits is neither a simple Gaussian nor its mixture and thus this task will evaluate CNN's discrimination ability on the complicated class distributions.

[2]The handwritten digit dataset and the machine-printed digit datasets were already used in [19] for analyzing the difference between handwritten patterns and machine-printed patters in their distribution.

[3]http://www.ultimatefontdownload.com/

V. Recognition experiments on individual datasets

A. Accuracy Achieved by CNN

The recognition accuracy of the CNNs for each digit dataset was evaluated. In every evaluation, the CNN was trained with 90% of the dataset and tested with the remaining 10%. The dataset separation was done randomly. This evaluation was repeated three times with different separation and then the average accuracy was derived.

The lower section of Table I shows the accuracy by CNN on individual datasets (i.e., "HW/original", "MP/original", and "MF/original"). For HW/original and MP/original, the CNN could achieve an almost perfect performance in accuracy (99.88% and 99.99%). For MF/original, the accuracy (95.7%) was not as high as the others but still surprising by considering the fact that the dataset contains many heavily decorated fonts.

B. Misrecognized Samples

Figures 3-5 show misrecognized samples and a part of the correctly-recognized samples for HW/original, MP/original, and MF/original, respectively. (Note that those samples are taken from one of the three trials on accuracy evaluation.)

- **Handwritten digits** (Fig. 3): Roughly speaking, 20-30% of all the 92 misrecognized samples are ambiguous about their class. For example, several misrecognized samples of "0" may also be read as "6" even by human. In addition, many misrecognized samples are heavily deformed and have low legibility. Consequently, it is possible to say that they are often "convincing misrecognition."
- **Machine-printed digits** (Fig. 4): There are only two misrecognized samples and both of them shows heavy distortion by binarization.
- **Multi-font digits** (Fig. 5): The misrecognized samples are often hardly legible. For example, the two samples of "1" which are misrecognized as "0" might be (mis)read as "0" unless we observe them together with the real "0" of the same font. Some samples are just a bold black bar.

  What is more remarkable is that the CNN could recognize many hardly legible samples. Considering those shape variations and decorations, it is possible to say that the recognition accuracy 95.7% is very high. Moreover, it is even possible to say that CNN can have high reading ability like human beings.

Figure 3. (Above) All misrecognized samples in HW/original. (Below) Examples of correctly-recognized samples.

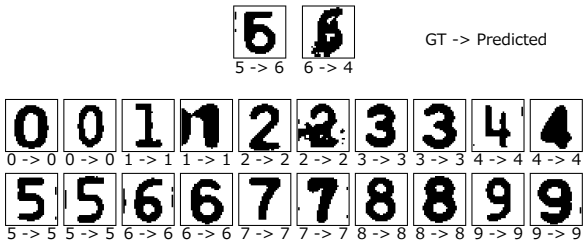

Figure 4. (Above) All misrecognized samples in MP/original . (Below) Examples of correctly-recognized samples.

Table II
COMPARISON WITH 1-NEAREST NEIGHBOR CLASSIFICATION WITH HAMMING DISTANCE.

	Accuracy(%)	
Type / Dataset	1-Nearest Neighbor	CNN
HW / original	98.94	99.88
MP / original	100.00	99.99
MF / original	89.4	95.7
HW+MP / original	99.40	99.91
HW+MP+MF / original	98.92	99.69

HW: handwritten, MP: machine-printed, MF: multi-font.

C. Comparison with 1-Nearest Neighbor Classification

A comparative evaluation was done with 1-nearest neighbor (1-NN) classification with Hamming distance under the same evaluation condition. As proved in [20], we can expect a high accuracy by the simple 1-NN classification when we have a very large amount of training samples.

Table II, however, shows that 1-NN classification could not outperform the CNN for HW/original and MF/original. Even though both classification methods use exactly the same training samples, the CNN shows a drastic improvement from 1-NN, except for MP/original where the accuracy is almost 100%[4]. For MF/original, the CNN outperforms 1-NN by reducing misrecognitions from 10.6% to 4.3%. This reduction proves that a CNN could recognize more than half of the very difficult samples which are misrecognized by 1-NN. The difference between 98.94% and 99.88% on HW/original is far more significant — CNN could reduce the misrecognized samples to 10% of 1-NN.

It is necessary to emphasize that the CNN is much faster than 1-NN if the training set size becomes larger. the computations required by 1-NN are huge for a nearest neighbor search, whereas the computations required by CNN are far less and independent of the training set size. In fact, 1-NN and CNN required 5,300s and 5s for recognizing all test samples of the MP/original, respectively, on a computer with GPU. Thus, for recognizing machine-printed digits, CNNs are a better choice compared to 1-NN, even though 1-NN could achieve very slightly better recognition accuracy (100.00%) than a CNN (99.99%).

D. Effect of Dataset Size

Figure 6 shows the recognition accuracies by the CNN and 1-NN under different training dataset sizes. A general tendency in the result by the CNN is that "more is better" as expected. Except for saturation on MP/original after about 10^5 samples, more data achieves better accuracy. For HW/original, the accuracy is improved from 99.63% to 99.88% by increasing training samples 10 times (i.e., from

[4]Amoung three trials, 1-NN achieved 100% (i.e., perfect result) on two trials and 99.99% on the other trial. Thus, the average of the three trials is 99.996... and becomes 100.00% after numerical rounding.

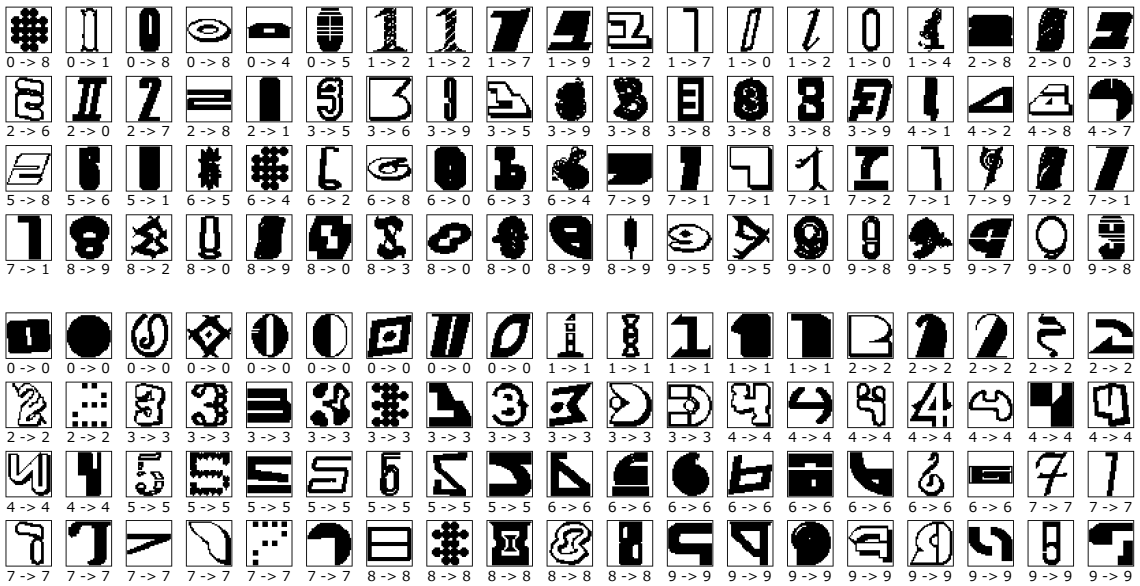

Figure 5. (Above) Examples of misrecognized samples in MF/original. (Below) Examples of correctly-recognized samples.

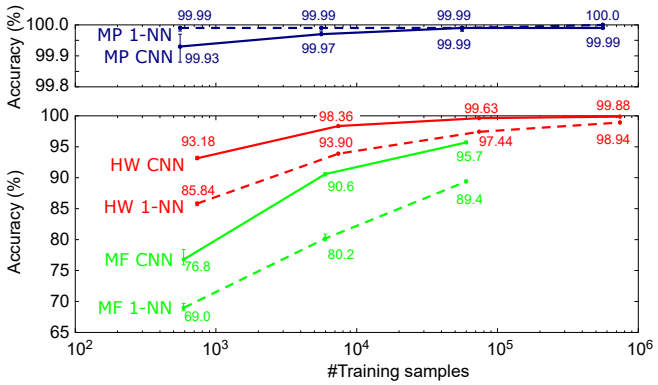

Figure 6. The effect of training set size on the recognition accuracy for the individual datasets. (HW: handwritten, MP: machine-printed, MF: multi-font.)

about 70,000 to 700,000). This fact suggests the past results on MNIST (with 60,000 samples plus augmented samples) were not the near-perfect performance regardless of their elaborated usage of a CNN. The result also suggests that there is still a room for a further improvement from our result because the accuracy is not saturated yet.

Another tendency is that the CNN outperforms 1-NN on HW/original and MF/original even with less training samples. Roughly speaking, 1-NN requires 10 times more training samples to achieve the same accuracy as a CNN. The tendency suggests the better generalization ability of CNN.

VI. Universal OCR by CNN

Table I shows the recognition accuracies of universal OCR by CNN in two scenarios (HW+MP/original and HW+MP+MF/original). In the scenario for HW+MP/original, the universal OCR was trained with 90% of the mixed dataset of HW/orignal and MP/original and tested with the remaining 10%. In the scenario for HW+MP+MF/original, the MF/original dataset was also mixed. It should be noted that the number of the output nodes is still 10. This means that the CNN needs to recognize digits without special treatment for dealing with multiple printing types.

Table III
The number of misrecognized samples by OCR for single printing type and universal OCR.

Scenario	Type of the test sample	Misrecognized by single-type OCR only	both OCRs	universal OCR only
HW+MP	HW	13	71	37
	MP	1	2	6
HW+MP+MF	HW	25	59	51
	MP	2	1	3
	MF	76	179	140

(HW: handwritten, MP: machine-printed, MF: multi-font.)

The accuracies in Table I indicate that there are no serious performance degradations by mixing the multiple printing types. In addition, the lower section of Table II shows that CNN keeps its superiority over 1-NN in the universal OCR scenarios. CNNs, therefore, are promising classifiers for realizing universal OCR. Table III shows a more detailed analysis of the misrecognized samples. The results show that some additional samples were misrecognized, however, there are samples that are now correctly recognized. Although, the amount of the newly correct samples is about half or less of the newly failed samples, it is interesting to confirm positive effect by the mixture.

Table IV
THE EFFECT OF TRANSFER LEARNING BETWEEN HANDWRITTEN AND MULTI-FONT DIGIT DATASETS.

Training/Test	HW/HW	MF→HW/HW	MF/MF	HW→MF/MF
Accuracy(%)	99.88	99.90	95.7	96.1

(HW: handwritten, MF: multi-font.)

VII. HOW ARE MULTI-FONT DIGITS USEFUL FOR RECOGNIZING HANDWRITTEN DIGITS BY CNN?

Transfer learning was performed to see how multi-font digit samples enhances the performance of handwritten digit recognition and vice versa. Since both multi-font digits and handwritten digits are deformed versions of digits, we can expect an enhancement. The procedure of our transfer learning technique was simple; for example, a CNN was first trained with the multi-font digit dataset and then further trained with the handwritten digit dataset.

Table IV shows the results, where "MF→HW" and "HW→MF" suggest the results by the transfer learning. This table proves that transfer learning can enhance recognition accuracy. Although the improvement by "MF→HW" is not drastic, the improvement by "HW→MF" is meaningful. As shown in Fig. 5, most of the misrecognized samples without transfer learning were severely decorated and often barely legible. Transfer learning could recover 10% of them.

VIII. CONCLUSION

A near-perfect performance of CNNs on a handwritten digit recognition task was observed by using a dataset which is 10 times larger than MNIST. The effect of using the very large dataset was significant and the performance of the CNN almost achieves human reading ability. The performance of the CNN was also impressive for a large machine-printed digit dataset and a multi-font digit dataset. A CNN could show a promising performance in universal OCR scenarios and thus we do not need to separate input samples according to their printing type before recognition.

Our result suggests that the task of digit recognition is almost terminated. Despite isolated digit recognition being one of the easiest image pattern recognition tasks, this fact might be somewhat disappointing for pattern recognition researchers. On the other hand, we need not to be disappointed at the result because we can start considering new research directions *beyond 100%* – what can we do if we have OCRs with perfect accuracy?

Acknowledgment: This research was partially supported by MEXT-Japan (Grant No. 26240024).

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv*, 2015.

[2] M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman, "Reading Text in the Wild with Convolutional Neural Networks", *IJCV*, 116(1):1-20, 2016.

[3] D. Karatzas, et al., "ICDAR 2015 Competition on Robust Reading", *Proc. ICDAR*, 2015.

[4] K. Fukushima. "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position", *Biological Cybernetics*, 36(4):193-202, 1980.

[5] K. Fukushima, "Neocognitron for handwritten digit recognition", *Neurocomputing*, 51, 161-180, 2003.

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition", *Proceedings of the IEEE*, 86(11):2278-2324, 1998.

[7] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis," *Proc. ICDAR*, 2003.

[8] D. Cireşan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Convolutional Neural Network Committees for Handwritten Character Classification", *Proc. ICDAR*, 2011.

[9] D. Cireşan, U. Meier, and J. Schmidhuber, "Multi-column Deep Neural Networks for Image Classification", *Proc. CVPR*, 2012.

[10] X. -X. Niu and C. Y. Suen, "A Novel Hybrid CNN-SVM Classifier for Recognizing Handwritten Digits", *Pattern Recognition*, 45(4):1318-1325, 2012.

[11] C. Wu, W. Fan, Y. He, J. Sun, S. Naoi, "Cascaded Heterogeneous Convolutional Neural Networks for Handwritten Digit Recognition", *Proc. ICPR*, 2012.

[12] B. Graham, "Spatially-Sparse Convolutional Neural Networks", *arXiv*, 2014.

[13] S. Pan, Y. Wang, C. Liu, X. Ding, "A Discriminative Cascade CNN Model for Offline Handwritten Digit Recognition", *Proc. MVA*, 2015.

[14] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading Digits in Natural Images with Unsupervised Feature Learning", *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[15] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet, "Multi-Digit Number Recognition from Street View Imagery Using Deep Convolutional Neural Networks", *arXiv*, 2013.

[16] K. Chellapilla, K. Larson, P. Simard, and M. Czerwinski, "Computers beat humans at single character recognition in reading based human interaction proofs (HIPs)", *Proc. Email and Anti-Spam*, 2005.

[17] K. Zagoris, I. Pratikakis, A. Antonacopoulos, B. Gatos, and N. Papamarkos, "Handwritten and Machine Printed Text Separation in Document Images Using the Bag of Visual Words Paradigm", *Proc. ICFHR*, 2012.

[18] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition", *arXiv*, 2015.

[19] M. Goto, R. Ishida, Y. Feng, and S. Uchida, "Analyzing the Distribution of a Large-scale Character Pattern Set Using Relative Neighborhood Graph", *Proc. ICDAR*, 2013.

[20] S. Uchida, R. Ishida, A. Yoshida, W. Cai and Y. Feng, "Character Image Patterns as Big Data", *Proc. ICFHR*, 2012.