Street View House Numbers Recognition using PCANet and SVM

Abstract—**Optical Character Recognition (OCR) in document images is considered to be a solved problem, however, text recognition in natural images is still a challenging problem. The ability to recognize text in natural images is crucially important for vision based applications of future. For instance, to build an autonomous robot having the capability to move around freely, it must be equipped with a natural images OCR system to read maps. This problem is very difficult as compared to document recognition due to the wide variety in background colors, text colors, font styles and noise. In this paper, we address a sub-problem, i.e. digits recognition. Research in this area has been accelerated with the availability of street view house numbers (SVHN) dataset. We test the potential of PCANet based feature extraction technique for this problem, which has proven to be successful in various image classification problems. We apply support vector machines (SVM) based classifier to recognize digits. To the best of author's knowledge, no other study reports similar analysis.**

I. Introduction

Optical Character Recognition (OCR) has been at the core of pattern recognition research. Current OCR systems have achieved sufficient accuracy in recognizing document images, nevertheless more work is required to achieve good performance in natural images [1]. The performance of the state of the art techniques lag behind human performance in complex natural scene images. The problem becomes more complex in this context mainly due to low resolution, non-contrasting backgrounds, large illumination differences, orientation problems, motion-blur and de-focused images. Other problems include great variety in text colors, font styles and wide range of backgrounds, which sometimes make impossible even for human to recognize text in such images. In this paper, our focus is on a sub-problem, recognizing digits in natural scene images.

We study the problem of digit classification in context of street view house numbers (SVHN). The ability to recognize multi-digit house numbers is an essential requirement of modern-day map applications. Being able to recognize the house numbers in a particular street can enable an application to precisely pinpoint an address with greater accuracy. Netzer et. al. [2] introduced a data-set suitable for such an application, i.e. house numbers acquired from Google street view images. They provided the data-set in a format similar to MNIST handwritten digits data-set [3] (32x32 images), but sometimes contains multiple digits, however, it considers only the digit near center of the image. The complexity of the problem can be understood looking at some images from the data-set (see Figure 1 [2]).

In this paper, we test the potential of PCANet [4] based feature extraction techniques and provide suitable architecture or tuned settings for PCANet.

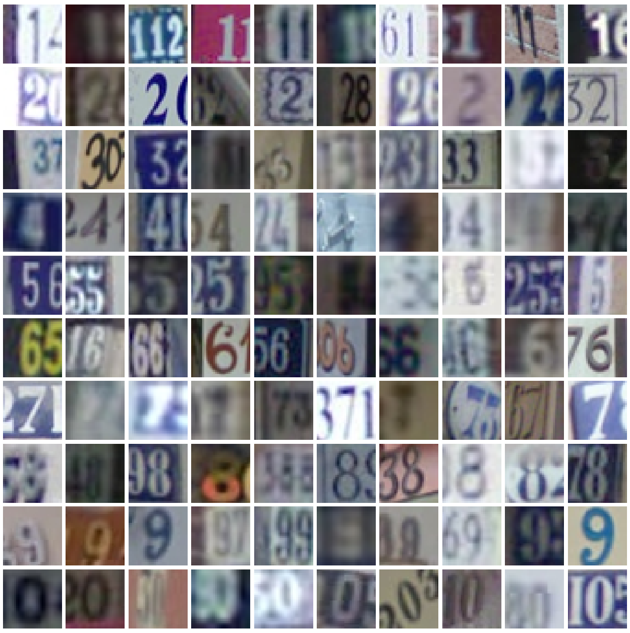

Fig. 1. 32x32 cropped samples from the SVHN dataset

Rest of the paper is organized as follows: overview of related research is presented in section II, section III discusses methodology of the study, presenting a brief introduction of PCANet and PCANet architecture for SVHN recognition task, results and discussion is given in section IV and section V presents conclusion of the study.

II. Related Research

Researchers proposed PCA based recognition approach for handwritten digits recognition in [5], similarly DCT features based classification approach for handwritten digits classification task has been proposed in [6]. In context of natural images text recognition, template-matching [7] and multiple hand-crafted features [8] based approach has been applied previously. Several other techniques proposed for this task apply implicit feature learning schemes, e.g. convolutional networks. Convolutional networks learn features from pixels and [9] showed that they outperformed hand-designed features. Authors in [1] applied ConvNets for SVHN recognition task and presented similar conclusions.

Convolutional neural networks (CNNs) [3],[10] in many ways are similar to other neural networks, however, affine transformations at each layer can be carried out as a discrete convolution which makes the CNNs efficient [11]. Several factors have accounted for the recent success of CNNs, i.e.

algorithmic advances, availability of large size training sets and increase in computational resources.

Netzer et. al. [9] compared performance of various hand-designed features (being used in state of the art OCR systems). They depicted that weighted direction code histogram (WDCH) binary images based features [12] result in 63.33 % accuracy. They also applied histogram of gradients based features [13] to classify SVHN dataset and reported 85% recognition rate. Stacked sparse auto-encoders based approach described in [14] has also bee applied and resulted in 89.7% accuracy. They also concluded that recognition accuracy improves with increase in training samples for all the techniques compared.

Authors in [9] estimated human level performance for SVHN recognition task as a function of image height in pixels. Their results suggested that 100% human performance can be achieved with image height ≥ 76 pixels, however, even human performance is around 90 ± 1.5% in the task being studied, i.e. 32 x 32 images.

In this paper, we consider a simple PCA filters based deep learning network, PCANet,optimize its parameters for SVHN recognition task . To the best of author's knowledge, similar analysis has not been performed or reported in literature for the task being studied.

III. Methodology

We first compute the features from training images using PCANet and train a linear SVM classifier to perform classification on test images. This section describes the features extraction and classification techniques applied.

A. Features Extraction

Features extraction process is very important step in pattern recognition problems. Considering the complexity of the problem, rather than applying traditional OCR approach, i.e. localization, segmentation, binarization, feature extraction and recognition, we adopt a different approach. We apply a powerful feature extraction technique, PCANet, which is capable of extracting class-level features all the way from input images, which makes the approach simple from an algorithm design perspective.

PCANet or PCA network is a recent convolutional neural network that compounds the advantages of deep learning and PCA [4]. PCANet is different from other convolutional networks in the sense that it learns the filter banks by applying PCA on the input images rather than attempting to find optimal filters. It is advantageous as compared to other convolutional networks because it does not demand large amounts of dataset for good performance and takes lesser time to compute features.

Idea behind PCANet is simple as it is illustrated in Figure 2. A typical two layer PCANet is initialized by applying PCA to overlapping patches of input image. It's first layer filters are composed of the selected principle components and image patches projections to these principle components formulate the output of units in first layer. Similar methodology is repeated at the second layer in cascaded fashion. Further, it applies binary quantization and hashing for two-stage cascaded filters to concatenate and perform binary to decimal conversion. Final output of PCANet is determined by computing block wise histograms from quantized images and further applying spatial pyramid pooling technique.

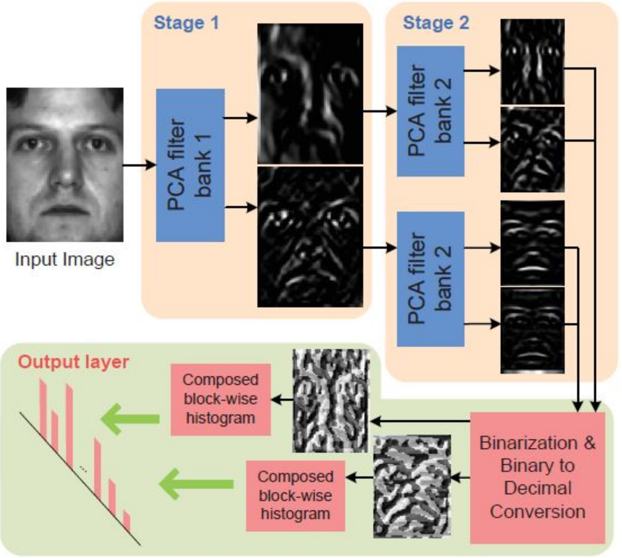
Stage 1
Stage 2
PCA filter bank 1
PCA filter bank 2
PCA filter bank 2
Input Image
Output layer
Composed block-wise histogram
Composed block-wise histogram
Binarization & Binary to Decimal Conversion

Fig. 2. Illustration of PCANet feature extraction technique

PCANet has a number of parameters that need to be tuned for each application, these parameters are: the number of layers N, filter size at each stage k, number of filters at each layer L, histogram block size, and the overlap ratio. In order to fine tune PCANet parameters for SVHN classification task, we performed three fold cross validation on training data.

Firstly, we fixed all other parameter as suggested in original paper and changed number of layers to three. As suggested in original paper, performance improvement could not be achieved but training and testing time was much increased. Further, we considered different values for overlap ratio and found that original value of 0.5 is suitable for this task. Taking inspiration from common setting of Gabor filters, we decided number of filters L to be 8.

Optimal value for filter size k is found to be 7. Different experiments has been performed to find the optimal value for histogram block size and 7 is determined to be final value. Spatial pyramid pooling is another tool in PCANet that can be applied at the final stage. Although this tool did not affect performance for problems addressed in the original paper, but due to the complexity of the problem and specially cluttered background and noise considerations, it improved the performance of technique overall by almost 1% and feature learning time is slightly increased.

We also considered different tweaks at pre-processing stage to test how does it affect the performance of PCANet. Firstly we transformed the color space of images from RGB to YUV and IHLS colos space, in both cases performance deteriorated. Then, we also considered applying RGB to grayscale transformation and applying PCANet and it does not affect the performance. Considering it an interesting finding, we computed DCT coefficients of grayscale images and then ap-

TABLE I. CLASSIFICATION RESULTS

Total Samples	26032
Correctly Recognized	23662
Accuracy	90.90%

plied PCANet, which resulted in degradation in performance. We also looked at the effect of standardizing images before applying PCANet and we determined that it could not improve the performance.

B. Classification

Once the features has been computed, next task is to train a classifier and test the performance of suggested approach. Considering the massive size of feature vectors and low number of classes, i.e. 10, we decided to employed linear SVM classifier. We used $C = 1$ as a mis-classification penalty cost term to avoid over-fitting.

IV. RESULTS

This section presents the results achieved with suggested approach. We used the train and test splits provided in original data-set, extra data provided for training is not used. Hence, 73257 digits have been used for training and 26032 digit samples have been used to test the proposed approach. Classification results are depicted in Table I.

In Table II, accuracies comparison of our approach to different state-of-the-art approaches has been presented. It is evident from the comparison that suggested approach outperforms the standard hand-crafted approaches in this task, also the proposed approach performance is almost comparable to human performance. Average testing time per sample is 400 milli seconds (ms).

TABLE II. ACCURACIES COMPARISON OF DIFFERENT TECHNIQUES TO OUR APPROACH

ALGORITHM	ACCURACY on SVHN-TEST
HOG	85.0%
WDCH	63.3%
STACKED SPARSE AUTO-ENCODERS	89.7%
HUMAN PERFORMANCE	90 ± 1.5 %
OUR APPROACH	90.9%

V. CONCLUSION

A sub-problem of text recognition in natural images, digits recognition, has been studied in this paper. The problem being studies is more hard as compared to standard OCR task for document recognition. Main reasons for its complexity are background variations, out of focus images, large variations in text color and style, large illumination differences and non-contrasting backgrounds. The severity of the problem can be realized from the fact human performance is a function of image dimensions. We have suggested an approach based on PCANet and optimized its parameters for this task using 3-Fold cross validation and also tried some pre-processing tricks. SVHN data-set representing house numbers from street-view data has been used to evaluate performance of the proposed approach. Proposed technique outperforms many state-of-the-art approach and its performance is approach human performance. For future work, we suggest to train a PCANet with extended training data (including extra training data provided in SVHN dataset) and less number of training samples. This analysis would be interesting because it can answer the question whether performance of PCANet improves with increase in training data as other techniques and how much data is necessary to achieve reasonable performance with PCANet.

REFERENCES

[1] Sermanet, P., Chintala, S., & LeCun, Y. (2012, November). Convolutional neural networks applied to house numbers digit classification. In Pattern Recognition (ICPR), 2012 21st International Conference on (pp. 3288-3291). IEEE.

[2] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, Andrew Y. Ng Reading Digits in Natural Images with Unsupervised Feature Learning NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011.

[3] Y. LeCun, L. Bottou, Y. Bengio & P. Haffner, "Gradient-based learning applied to document recognition, Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, 1998.

[4] Chan, T.; Jia, K.; Gao, S.; Lu, J.; Zeng, Z.; Ma, Y., "PCANet: A Simple Deep Learning Baseline for Image Classification?," in Image Processing, IEEE Transactions on , vol.PP, no.99, pp.1-1, doi: 10.1109/TIP.2015.2475625.

[5] Li, R., & Zhang, S. (2011). Handwritten Digit Recognition Based on Principal Component Analysis and Support Vector Machines. In Advances in Computer Science, Environment, Ecoinformatics, and Education (pp. 595-599). Springer Berlin Heidelberg.

[6] S. S. Ali & M. U. Ghani, Handwritten Digit Recognition using DCT and HMMs, in 12th International Conference on Frontiers, Islamabad, Pakistan., 2014.

[7] T. Yamaguchi, Y. Nakano, M.Maruyama, H.Miyao, & T. Hananoi. "Digit classification on signboards for telephone number recognition. In ICDAR, pages 359363, 2003.

[8] T. E. de Campos, B. R. Babu, & M. Varma. "Character recognition in natural images. In Proceedings of the International Conference on Computer Vision Theory and Applications, Lisbon, Portugal, February 2009.

[9] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, & A. Y. Ng. "Reading digits in natural images with unsupervised feature learning. In NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2011.

[10] Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 36, 193202.

[11] Goodfellow, I. J., Bulatov, Y., Ibarz, J., Arnoud, S., & Shet, V. (2013). Multi-digit number recognition from street view imagery using deep convolutional neural networks. arXiv preprint arXiv:1312.6082.

[12] F. Kimura, T. Wakabayashi, S. Tsuruoka, & Y. Miyake. Improvement of handwritten japanese character-recognition using weighted direction code histogram. Pattern Recognition, 30(8):13291337, Aug. 1997.

[13] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In CVPR, pages I: 886893, 2005.

[14] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research, 11:33713408, 2010.