DISCRIMINATIVE LEARNING FOR SCRIPT RECOGNITION

Sheikh Faisal Rashid[1], Faisal Shafait[2], and Thomas M. Breuel[1]

[1]Technical University of Kaiserslautern, Kaiserslautern, Germany
[2]German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany
s_rashid09@informatik.uni-kl.de, faisal.shafait@dfki.de, tmb@informatik.uni-kl.de

ABSTRACT

Document script recognition is one of the important preprocessing steps in a multilingual optical character recognition (MOCR) system. A MOCR system requires prior knowledge of script to accurately recognize multilingual text in a single document. In multilingual documents two scripts can be mixed together within a single text line. Many existing script recognition methods lack the ability to recognize multiple scripts mixed within a single text line. Besides, these methods usually use script dependent features for script recognition thereby limiting their scope to particularly that script. In this paper we propose a discriminative learning approach for multi-script recognition at connected component level by using a convolutional neural network. The convolutional neural network combines feature extraction and script recognition process in one step and discriminative features for script recognition are extracted and learned as convolutional kernels from raw input. This eliminates the need for manually defining discriminative features for particular scripts. Results show above 95% script recognition accuracy at connected component level on datasets of Greek-Latin, Arabic-Latin multi-script documents and Antiqua-Fraktur documents. The proposed method can be easily adapted to different scripts.

Index Terms— Multi-script recognition, MOCR, Discriminative learning, Convolutional neural network

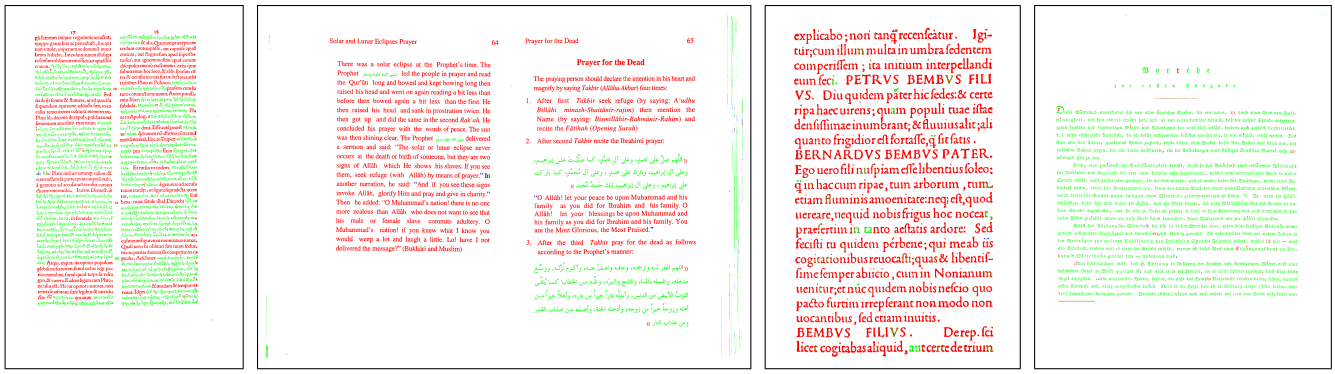
1. INTRODUCTION

Recently, there has been a growing interest in the area of multilingual OCR due to variety of documents available in bilingual or even in trilingual form, for example ancient multi-script documents, multilingual dictionaries, books with line by line or column wise translation, and official documents from some countries like passport forms, examination questions and money order forms etc. Usually a MOCR system combines character level or word level classifiers for different languages or scripts and recognition of a particular character or word is done by its respective classifier. Therefore prior knowledge of a script for each character or word is essential for selection of an appropriate classifier in MOCR system. A brute-force solution would be to train a OCR classifier on more than one language by adding individual characters from all languages into training process. However, this would lead to more classification errors due to increase in number of classes.

In this paper we describe and experimentally test our discriminative learning approach for recognition of multiple scripts present in a single text line. The approach is based on discriminative learning using convolutional neural network and it works at connected component level. The convolutional neural network acts as discriminative learning model, where suitable features for script recognition are automatically extracted and learned. We demonstrate effectiveness of our method on Greek-Latin, Arabic-Latin multi-script documents and Fraktur-Antiqua documents. The preliminary version of this work has also been presented as a short paper in DAS 2010[1].

1.1. Related work

Previous methods for script identification can be broadly grouped into two approaches: global approaches and local approaches. This categorization is based on feature extraction process employed at global level (a region of text lines) or local level (character, word or single text line) for each individual script. The survey paper of Abirami and Manjula [2] presented a precise overview of some of these methods. Most of the existing methods for script identification work at document level or these methods assume that the document images have different scripts only at a particular place (in a particular column, paragraph or text line) and only few of them consider word level multi-script identification. Hochberg et al. [3] used cluster based templates for the script identification at document level. Later, Spitz [4] presented language identification in Han-based and Latin-based scripts by using vertical position distribution of upward concavities, optical density distribution and most frequently occurring word shapes characteristics. Pal and Chaudhuri [5, 6] worked on separation of text lines from different scripts using projection profiles, water reservoir concepts and existence of head-line(a feature specific to Bangla and Devanagari scripts). Zhou et al. [7] presented the Bangla-English script identification by analyzing connected component profiles and head-line fea-



(a) Greek-Latin script recognition (b) Arabic-Latin script recognition (c) Antiqua script recognition (d) Fraktur script recognition

Fig. 1. Multi-script recognition results in color coding format. For color printed paper: Red color represents Latin and Antiqua scripts, green color represent Greek, Arabic and Fraktur scripts, and blue color represents small components like diacritics or noise. For black & white printed paper: Dark color represents Latin and Antiqua, light color represent Greek, Arabic and Fraktur scripts.

tures. Busch et al. [8] described the use of texture as a tool for determining the script of a document image. Joshi et al. [9] also employed the texture based Gabor filter at multiple scales and then they further used some script dependent features like head line information, statistical features, local features and horizontal profile information. Ma and Doermann [10] performed word level script identification for scanned document images by using Gabor filter and compared the performance of different classifiers. Ramakrishnan and Pati [11] reported word level multi script identification by using Gabor and discrete cosine transform (DCT) features.

2. OUR APPROACH

Our approach works by classifying individual connected components independently and therefore this does not require complex layout analysis for image segmentation into text lines or blocks [12] prior to classification. In this approach we use convolutional neural network (CNN) for learning complex features for recognition of multiple scripts present in a single document. The advantage is that the discriminative features for different scripts are extracted and learned automatically by CNN and there is no need to define script specific features for a particular script. We have already demonstrated this approach for orientation detection in Urdu document images [13].

Document images are binarized with local adaptive thresholding technique [14] before extracting the connected components. We remove small salt and pepper noise and big noise (merged components and borders [15]) by using a heuristic rule, which is based on size of connected components. A connected component is considered as noise if its height and width are less than or equal to 0.3 times or greater than 5 times the median height and median width of all components. Diacritics and punctuations are the small components that occur in both scripts. We also remove these small components from document images using similar heuristic rule i.e. a component is considered as a diacritic or a punctuation if its width and height are equal to or less than 0.7 times the median height and median width of all components. After noise and diacritics removal, remaining connected components are rescaled to 40x40 dimensions and the pixel intensities are normalized between -1 and +1. These rescaled connected components are used as feature vectors to train the CNN for script recognition task. The CNN used in this experiment consists of two convolutional layers with four and eight feature maps followed by two sub-sampling layers. We train the CNN in supervised learning mode with 19600 training and 2000 validation samples from Greek and Latin scripts for 200 epochs with 0.1 learning rate. An on-line error back-propagation algorithm [16] is used for CNN training.

For script recognition, we extract feature vectors (normalized and scaled connected components) as described above and pass these raw features to the CNN for classification. A particular script is recognized based on higher confidence score at output neurons. Script recognition for small connected components like diacritics and punctuations is done by using closest left-right neighboring connected component script information. The script recognition accuracy is further improved by extending the bounding box of each connected component to its left and right by a factor of its height or width (whichever is greater) and then by using the class majority within the neighboring area to relabel the script of that component.

Table 1. Script Recognition Accuracy for Greek and Latin Scripts

	Training set		Validation set		Test set		
	Nos. of samples	CNN accuracy (%)	Nos. of samples	CNN accuracy (%)	Nos. of samples	CNN accuracy (%)	Accuracy after left-right neighbor rule (%)
Greek	9800	99.41	1000	96.40	11302	95.16	97.65
Latin	9800	98.92	1000	97.80	10828	97.58	99.15
Average	19600	99.16	2000	97.10	22130	96.37	**98.40**

Table 2. Script Recognition Accuracy for Arabic and Latin Scripts

	Training set		Validation set		Test set		
	Nos. of samples	CNN accuracy (%)	Nos. of samples	CNN accuracy (%)	Nos. of samples	CNN accuracy (%)	Accuracy after left-right neighbor rule (%)
Arabic	24000	97.03	2000	98.95	6037	97.80	99.30
Latin	24000	99.31	2000	97.70	1221	90.60	91.92
Average	48000	98.17	4000	98.33	7258	94.20	**95.61**

3. EXPERIMENTAL RESULTS

We evaluate performance of the described multi-script recognition approach on Greek-Latin, Arabic-Latin and Fraktur-Antiqua document datasets. For Greek-Latin script recognition, we have nineteen ancient documents that contain Greek and Latin scripts mixed with each other. This dataset is manually processed to generate ground truth for training and testing the approach. We use twelve documents for training the CNN and seven documents for testing. The evaluation results for Greek-Latin script recognition are given in Table 1. For Arabic-Latin script recognition, we train the CNN on seventy scanned pages from an Arabic story book and on a subset of UW-III dataset[17] documents for Arabic and Latin scripts respectively. However, we test our approach on five scanned pages from a different book that contain Arabic and Latin scripts mixed within text lines. We get slightly less recognition accuracy on Arabic-Latin test dataset, as shown in Table 2, because the test dataset is entirely different from training dataset. But these Arabic-Latin script recognition results also reflect generalization capability of our approach on even unseen document images. For Antiqua-Fraktur recognition, we train CNN on three ancient Fraktur document images and training samples for Antiqua are taken from subset of Greek-Latin document images because these documents are in Antiqua typeface. The testing is performed on two Fraktur document images and two available Antiqua document images. Table 3 shows the evaluation results on Antiqua and Farktur recognition task. The script recognition accuracy is improved by incorporating class majority in left-right neighboring area of every connected component. We also represent the script recognition output as a color coded image that shows output of CNN in terms of color intensities. These color intensities reflect confidence of each script present in the document. This color coded output can be further analyzed by some statistical techniques to improve the script recognition accuracy.

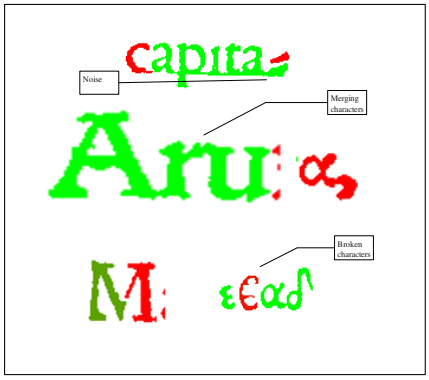


Fig. 2. Errors due to noise, merged and broken characters

Figure 1 shows output of our approach in color coded format for all scripts after applying left-right neighbor class majority rule.

4. DISCUSSION

In this paper we present a multi-script recognition approach by using discriminative learning at connected component level. We use a convolutional neural network as discriminative learning model to extract and learn suitable features for multi-script recognition task. One observation was that due to appearance based discriminative property of different scripts, a script can be recognized based on the shape of its individual characters. Convolutional neural networks with properties of local receptive fields, weight sharing, and spatial sub-sampling layers have ability to learn discriminative features from the raw image [18]. Therefore the use of CNN at connected component level to learn discriminative shape based features is an efficient approach for multi-script recognition as shown by results given in Table 1, 2 and 3. The

Table 3. Script Recognition Accuracy for Antiqua and Fraktur

	Training set		Validation set		Test set		
	Nos. of samples	CNN accuracy (%)	Nos. of samples	CNN accuracy (%)	Nos. of samples	CNN accuracy (%)	Accuracy after left-right neighbor rule (%)
Antiqua	10600	98.23	2000	97.90	1194	87.0	97.40
Fraktur	10600	98.07	2000	98.60	4130	92.27	95.81
Average	21200	98.15	4000	98.25	5324	89.64	**96.61**

approach also has generalization capabilities because it gives good results on different target documents that are not part of training process. The datasets used in these experiments have noise in terms of touching or broken characters and smudge (e.g. ink spots or spread) as shown in Figure 2. These noisy components are removed before training the CNN, therefore CNN does not give good recognition accuracy for these components. It is observed that most of the recognition errors are due to noise and we may obtain better results on clean datasets. Another observation is that CNN is sensitive to character shapes in terms of slight variations e.g, thick or thin writing strokes and this problem can be overcome by adding more training samples that contain all these variations.

5. REFERENCES

[1] S.F Rashid, F. Shafait, and T.M. Breuel, "Connected component level multiscript identification from ancient document images," in *9th IAPR Workshop on Document Analysis Systems, (short paper)*, Boston, MA, USA, June 2010.

[2] S. Abirami and D. Manjula, "A survey of script identification techniques for multi-script document images," *International Journal of Recent Trends in Engineering*, vol. 1, no. 2, pp. 255–257, May 2009.

[3] J. Hochberg, P. Kelly, T. Thomas, and L. Kerns, "Automatic script identification from document images using cluster-based templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 176–181, February 1997.

[4] A.L. Spitz, "Determination of the script and language content of document images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 3, pp. 235–245, March 1997.

[5] U. Pal and B.B. Chaudhuri, "Automatic identification of English, Chinese, Arabic, Devnagari and Bangla script line," in *ICDAR '01: Proceedings of the Sixth International Conference on Document Analysis and Recognition*, Washington, DC, USA, 2001, pp. 790–794.

[6] U. Pal and B.B. Chaudhuri, "Script line separation from Indian multi-script documents," in *ICDAR '99: Proceedings of the Fifth International Conference on Document Analysis and Recognition*, Washington, DC, USA, 1999, pp. 406–409.

[7] L. Zhou, Y. Lu, and C.L. Tan, "Bangla/English script identification based on analysis of connected component profiles," in *7th IAPR Workshop on Document Analysis Systems*, Nelson, New Zealand, Feb 2006, vol. 3872 of *Lecture Notes in Computer Science*, pp. 243–254.

[8] A. Busch, W.W. Boles, and S. Sridharan, "Texture for script identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1720–1732, November 2005.

[9] G. D. Joshi, S. Garg, and J. Sivaswamy, "Script identification from Indian documents," in *7th IAPR Workshop on Document Analysis Systems*, Nelson, New Zealand, Feb 2006, vol. 3872 of *Lecture Notes in Computer Science*, pp. 255–267.

[10] H. Ma and D. Doermann, "Word level script identification for scanned document images," in *Proc. SPIE Document Recognition and Retrieval XI*, San Jose, CA, USA, December 2003, pp. 124–135.

[11] P.B. Pati and A.G. Ramakrishnan, "Word level multi-script identification," *Pattern Recognition Letters*, vol. 29, no. 9, pp. 1218–1229, July 2008.

[12] F. Shafait, D. Keysers, and T.M. Breuel, "Performance evaluation and benchmarking of six page segmentation algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 941–954, June 2008.

[13] S.F. Rashid, S.S. Bukhari, F. Shafait, and T.M. Breuel, "A discriminative learning approach for orientation detection of urdu document images," in *13th IEEE Int. Multi-topic Conference, INMIC09*, Islamabad, Pakistan, Dec 2009, IEEE.

[14] F. Shafait, D. Keysers, and T. M. Breuel, "Efficient implementation of local adaptive thresholding techniques using integral images," in *Proc. SPIE Document Recognition and Retrieval XV*, San Jose, CA, USA, January 2008, pp. 101–106.

[15] F. Shafait, J.V. Beusekom, D. Keysers, and T.M. Breuel, "Document cleanup using page frame detection," *International Journal on Document Analysis and Recognition*, vol. 11, no. 2, pp. 81–96, Nov 2008.

[16] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification (2nd Edition)*, Wiley-Interscience, 2 edition, November 2000.

[17] I. Guyon, R.M. Haralick, J.J. Hull, and I.T. Phillips, *Data sets for OCR and document image understanding research*, pp. 779–799, World Scientic Singapore, 1997.

[18] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, November 1998.