A Discriminative Learning Approach for Orientation Detection of Urdu Document Images

Sheikh Faisal Rashid[1], Syed Saqib Bukhari[1], Faisal Shafait[2], and Thomas M. Breuel[1]
[1]Image Understanding and Pattern Recognition (IUPR) Research Group
Technical University of Kaiserslautern, 67663 Kaiserslautern, Germany
s_rashid09@informatik.uni-kl.de, bukhari@informatik.uni-kl.de, and tmb@informatik.uni-kl.de
[2]German Research Center for Artificial Intelligence (DFKI)
D-67663 Kaiserslautern, Germany
faisal.shafait@dfki.de

***Abstract*— Orientation detection is an important preprocessing step for accurate recognition of text from document images. Many existing orientation detection techniques are based on the fact that in Roman script text ascenders occur more likely than descenders, but this approach is not applicable to document of other scripts like Urdu, Arabic, etc. In this paper, we propose a discriminative learning approach for orientation detection of Urdu documents with varying layouts and fonts. The main advantage of our approach is that it can be applied to documents of other scripts easily and accurately. Our approach is based on classification of individual connected component orientation in the document image, and then the orientation of the page image is determined via majority count. A convolutional neural network is trained as discriminative learning model for the labeled Urdu books dataset with four target orientations: 0, 90, 180 and 270 degrees. We demonstrate the effectiveness of our method on dataset of Urdu documents categorized into the layouts of book, novel and poetry. We achieved 100% orientation detection accuracy on a test set of 328 document images.**

I. Introduction

Due to advances in digital image indexing and retrieval tasks [1], there is a growing need in optical character recognition (OCR) of printed/handwritten cursive scripts like Urdu, Arabic and Persian etc. For accurate OCR of document images it is necessary that the document images should be given in right orientation to OCR engine. The process of orientation detection is to determine the actual orientation of document image, which can further be used to transform it into right orientation. Orientation detection methods mainly focus on four target orientations 0, 90, 180 and 270 degrees because usually the document scanning process results in these four orientations. The slight variation besides these four orientation can be handled by skew correction techniques [2]. Most of document image orientation detection work can be categorized into two main categories: 1) landscape and portrait detection and 2) up-down orientation detection. Landscape and portrait can be detected using the global [3] and local [4] projection profiles whereas most of up-down detection techniques are based on the fact that the number of characters ascender are more likely as compared to number of characters descender in Roman script text. H.B. Aradhye [5] proposed a generic method using the character openness to detect the orientation of Roman as well as some other scripts such as Pashto and Hebrew. Lu and Tan [6] introduced the method for script and orientation detection through document vectorization, which encodes document orientation and language information and converts each document image into document vector through exploitation of the density and distribution of the vertical component runs. The performance of their work is reported on a dataset of 492 document images having text lines ranging from 1 to 12 and achieves an accuracy of 98.18% for documents of at least 12 text lines.

Recently, Beusekom et al. [7] proposed a method for combined skew and orientation detection using geometric modeling of Roman script text lines. The method searches for a text line candidates with in a skew range of four orientations top up, top down, top left and top right. The best fit of the model gives the estimate for orientation and skew.

In this paper we propose a discriminative learning approach for the orientation detection of Urdu script using convolutional neural network (CNN) because most of existing orientation detection methods are based on ascender to descender ratio but this information is not available in Urdu script. The idea is to extract connected components from binarized document images and to learn the shape of each connected component in all four orientations. The shape of most of connected components is highly dependent on their orientation and it changes with the change in orientation. Therefore learning the shape of a individual connected component in all four orientations helps in identification of correct orientation of a specific connected component. The orientation of a document image is determined by majority count of connected components for a specific orientation. Convolutional neural network [8] with properties of local receptive fields, weight sharing, and spatial subsampling have been used for various image processing and recognition tasks. A typical architecture of CNN consists of convolutional layers and sub-sampling layers followed by one hidden layer and an output layer. In the proposed method we train a convolutional neural network to learn shapes of connected components (in all four orientations), extracted

978-1-4244-4873-9/09/$25.00 ©2009 IEEE

from Urdu document images and use this shape information for orientation detection.

The rest of the paper is organized as follows: Section 2 gives the description of the proposed method. Section 3 describes the dataset used in this work. Experimental results are discussed in Section 4 followed by conclusion in Section 5.

II. Method Description

This section describes the proposed method for orientation detection of Urdu documents. This section is further divided into three subsections: document image preprocessing, feature vectors generation and convolutional neural network architecture and training criteria.

A. Document Image Preprocessing

Scanned document images were preprocessed before extracting raw feature vectors for convolutional neural network training. Preprocessing steps are described below.

1) Binariztion: Scanner has been used for digitizing document images in gray scale format. Complete experimental setup in this paper is based on binarized images. Therefore, binarization of scanned images is considered as a first document image preprocessing step. Different state-of-the-art binarization methods can be classified into two groups: (i) global binarization (like Otsu [9]) and (ii) local binarization (like Sauvola [10]). Global binariation estimate single threshold for a complete image, whereas local binarization calculate threshold for each pixel individually based on the neighborhod information. In general, local binarization works better than global binarization under different types of document image degradations, like non-uniform shading, blurring, etc. But local binarization methods are slower than global binarization methods. Shafait's local binarization method [11] overcomes this problem by using integral images [12] for local threshold computation. We used Shafait's binarization in this paper.

2) Noise removal: Binarized document images contain marginal noise [13] (like borders), which have been removed by using heuristic rules: for example, a connected component is considered as marginal noise if its height is greater than 5 times of median height or width is greater than 5 times of median width. Apart from this most of the Urdu characters consist of small connected components like dots and diacritics. These dots and diacritics have similar shapes in all possible orientations. Since our method is based on learning the variations in the shape due to change in orientation and these small connected components can produce a negative effect on training. Therefore, small connected components are also considered as noise and have been removed by using heuristic rules: for example, a connected component is considered as small noisy component if its height is smaller than 85% of median height and width is smaller than 85% of median width. Figure 1 shows original grayscale image and images after binarization and noise removal.

3) Feature Vectors Generation: The extracted connected components are rescaled to 40x40 dimension and are used as raw feature vector for CNN training and evaluation. In this rescaling we downscale or upscale a component depending on size of its width or height to the size of 40x40 matrix while keeping the aspect ratio of the connected component intact. This step is very important because our method uses the shape information for orientation learning and change in aspect ratio after rescaling will cause shape degradation. This shape degradation may effect in recognition accuracy. Feature vector files are generated for all connected components extracted from training, validation and testing dataset with their class labels. Connected component class labels are selected among 0, 1, 2 or 3 for 0, 90, 180, or 270 degree(s) orientation respectively.

B. CNN Architecture and Training Criteria

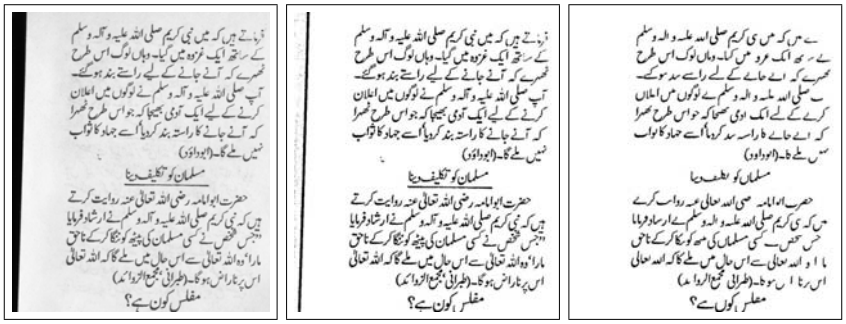
In the proposed method we exploit the properties of convolutional neural network as discriminative learning model for orientation detection of scanned Urdu document images. The overall architecture used in this method is depicted in Figure 2.

A convolutional neural network is a kind of multilayer neural network with a built in capability of feature extraction from raw input data. The general working of CNN consists of extraction of simple features at high resolution and convert them into more complex features at a coarse resolution. The coarser resolution is obtained by using a sub sampling layer with a subsampling factor of two. In the proposed architecture, the first convolutional layer, C1, has 10 (5x5) convolutional kernels corresponding to 10 high resolution features. The second convolutional layers has 20 (5x5) convolutional kernels corresponding to 20 complex features. Each convolutional layer is followed by subsampling layers S1 and S2 respectively with a subsampling factor of two. The output of S2 is then feed forwarded to fully connected hidden layer with 100 hidden units. The output layer consists of 4 units corresponding to 4 orientation classes. The first 4 layers of this neural network can be viewed as trainable feature extraction layers connected to a trainable classifier in the form of two fully connected layers. The number of hidden units controls the capacity and the generalization of the overall classifier.

For training the convolutional neural network we use feature vectors extracted from 60% of book dataset (complete description of dataset is given in Section III) under different orientations. We train the network for 200 epochs with 0.1 learning rate. An online error backpropagation algorithm [14] is used to train the CNN.

III. Urdu Dataset Description

Urdu belongs to cursive scripts and is very different from Roman script. Some of characteristics of Urdu language are that Urdu is written from right to left, characters have connections between each other to form a word and some characters have special symbols called diacritics above or below the character. As mentioned above Urdu characters



(a) Original grayscale image (b) Binarized image (c) Noise removed image

Fig. 1. Sample images after binarization and noise removal

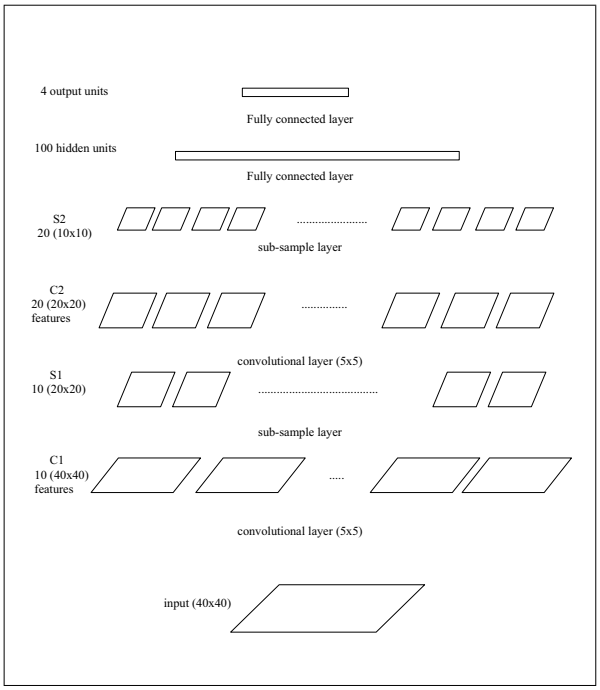


Fig. 2. Architecture of Convolutional Neural Network for orientation detection

cannot be categorized into ascenders or descenders as a discriminative feature like in Roman script for orientation detection. We use Urdu dataset[1] of scanned images from different Urdu publishing sources. The complete dataset is categorized into five subcategories book, novel, poetry, magazines and newspapers (based on its publishing source), out of which book, novel and poetry has been used in this work. The original dataset with these three categories consists of 59 scanned images and each scanned image consists of 2 document pages. This dataset is available in 0 degree orientation, referred as correct orientation, and dataset for other orientations is generated by rotating the original dataset images into other orientations like 90, 180 and 270 degrees. After these rotations we have a dataset of 236 images and since each image has 2 document pages therefore we have 472 document pages in total. We use only 112 document pages from book dataset for training and 16 document pages from book dataset for validation of CNN. Rest of dataset is used for evaluation of the method in two experiments. Figure 3 shows the images of all three type of datasets in all four orientations.

[1]The subset of dataset is available for public access at http://www.dfki.uni-kl.de/∼shafait/urdu-documents.zip and complete dataset can be obtained by personal contact.

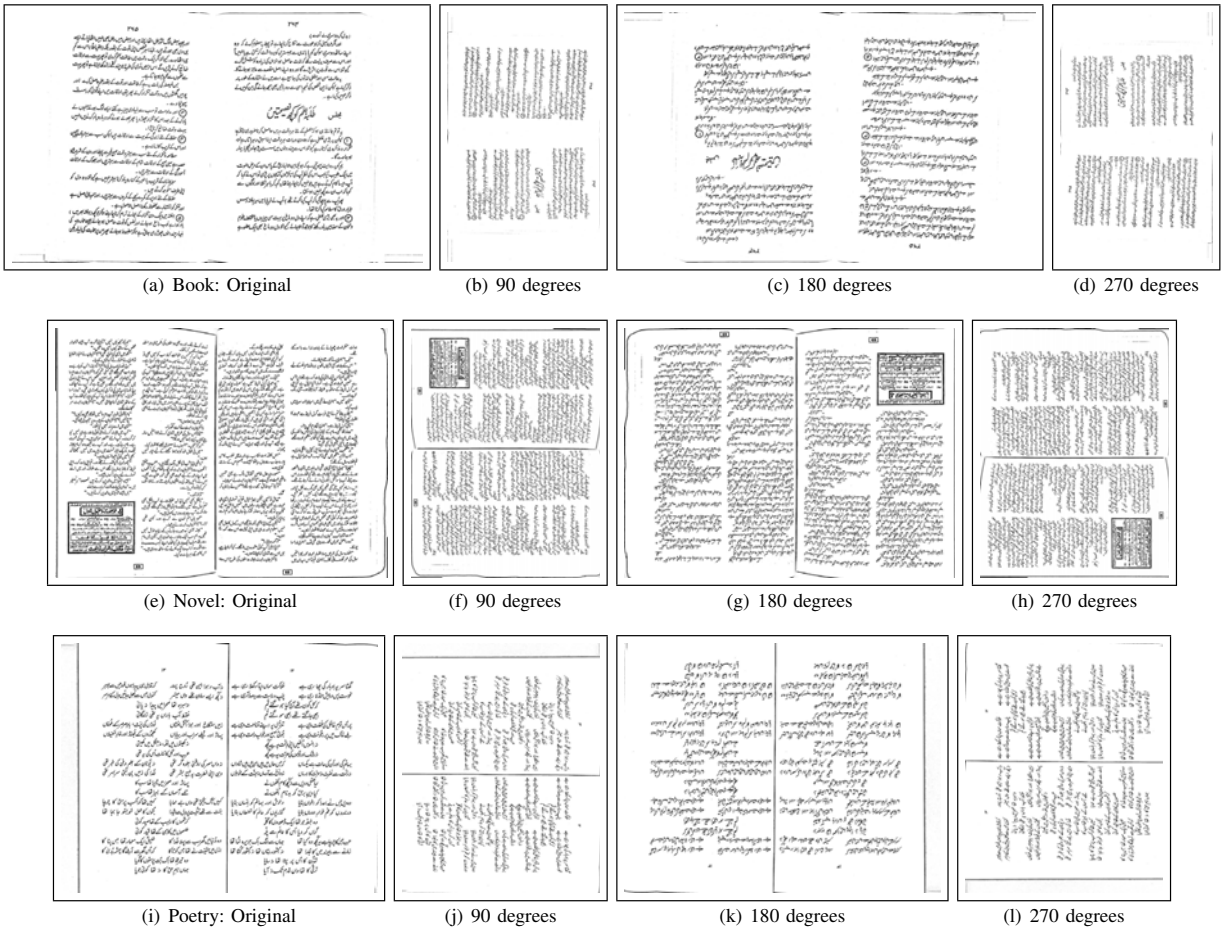
IV. Experimental Results and Evaluation Criteria

The dataset selected for evaluation has two distinctive features, layout and font or text-printing technology. In Urdu publishing system, usually book, poetry etc. are written by 'Katibs' (persons who have skill to write in different calligraphy styles and fonts) and have variability in shape of similar ligatures. However, novel, magazine and newspaper etc. are printed using printing stamps of different ligatures. The layout of each type of document is also different, for example poetry has unique layout of writing words over other words etc. We evaluate our proposed method on two different criteria: 1) connected component level orientation detection accuracy 2) page level orientation detection accuracy.

We perform two experiments for evaluation of our method for orientation detection. The description of each experiment is given below.

A. Experiment 1

In first experiment we use book dataset for validation and evaluation of our method. As mentioned above 60% of book dataset is used for training the neural network therefore we use remaining 40% of book dataset for validation (20%) and testing (20%) purposes. We also measure orientation recognition accuracy for training dataset. In this experiment we achieved overall 88% accuracy for training dataset, 87% accuracy for validation dataset and 87.96% accuracy for testing dataset at connected component level. Accuracy for testing dataset for all orientations is given in Table I. The results obtained in this experiment shows that we achieve around 88% accuracy for complete book dataset at component level and 100% accuracy at page level, because page



(a) Book: Original (b) 90 degrees (c) 180 degrees (d) 270 degrees

(e) Novel: Original (f) 90 degrees (g) 180 degrees (h) 270 degrees

(i) Poetry: Original (j) 90 degrees (k) 180 degrees (l) 270 degrees

Fig. 3. Sample images from all three categories of Urdu dataset with different orientations

level orientation is determined by majority count among orientations determined for all connected components. The 12% error at connected component level may be due to the reason that some of connected components loose shape information during rescaling process or their shape has minor variations in all or some of possible orientations.

B. Experiment 2

In second experiment we evaluate our method for novel and poetry datasets. The results obtained for all possible orientations in this experiment are given in Table II. These results show the robustness and generalization capabilities of our method because our training dataset consists of only book document images but our method is capable to categorize novel and poetry document images into correct orientations. We obtained less connected component level accuracy on novel dataset as compare to accuracy achieved for book and poetry datasets because novel dataset has variations in layout and printing. These variations cause variability in the shapes of similar Urdu ligatures and characters. Since our neural network is trained for one type of printed shapes therefore it gives less accuracy for other type of printing style. We can achieve high accuracy at component level for novel or other printed style category by providing more training samples containing all the variations in layout, printing style and font to the CNN. However we achieved 100% accuracy at page level for these different categories of dataset.

V. CONCLUSION

In this paper we present a new approach for orientation detection of Urdu document images. We use a convolutional neural network as discriminative learning model to learn the orientation of Urdu documents varying in layout and printing techniques. A CNN is trained from Urdu book dataset taking connected components from each page image in all four (0, 90, 180 and 270 degree(s)) orientations as raw feature vector. Orientation of a specific page image is determined by extracting connected components from the page image and scaled at 40x40 dimension to form a set of raw feature vectors. These raw feature vectors are passed to CNN for classifying orientation of each connected component. The orientation of the page image is determined by majority count of orientations among connected components. The page level accuracy is dependent on numbers of connected components

TABLE I

ORIENTATION DETECTION ACCURACY FOR BOOK DATASET

Dataset	Document Orientation	Connected Component Level Accuracy (%)	Overall Connected Component Accuracy (%)	Page Level Accuracy (%)
Book	0	80.54	87.96	100
	90	85.04		
	180	93.8		
	270	92.44		

TABLE II

ORIENTATION DETECTION ACCURACY FOR NOVEL AND POETRY DATASETS

Dataset	Document Orientation	Connected Component Level Accuracy (%)	Overall Connected Component Accuracy (%)	Page Level Accuracy (%)
Novel	0	63.64	76.29	100
	90	74.49		
	180	81.17		
	270	85.84		
Poetry	0	74.2	84.63	100
	90	83.42		
	180	93.13		
	270	87.76		

extracted from page and hence may be decreased for pages having very less numbers of Urdu words or characters. The proposed approach has been evaluated on a subset of publicly available dataset [15] of Urdu document images from book, novel and poetry categories. We obtained 100% page level accuracy and 83% connected component level accuracy. Due to discriminative learning behavior of our proposed approach, the approach can be applied to detect the orientation of other scripts accurately. Our method does not require text line extraction because it does not require ascenders to descenders ratio for orientation detection as required by most of previous available techniques.

REFERENCES

[1] T. M. Breuel. The OCRopus open source OCR system. In *Proc. SPIE Document Recognition and Retrieval XV*, pages 0F1–0F15, San Jose, CA, USA, Jan. 2008.

[2] B. V. Dhandra, V. S. Malemath, H. Mallikarjun, and R. Hegadi. Skew detection in binary image documents based on image dilation and region labeling approach. In *18th International Conference on Pattern Recognition*, volume 2, pages 954–957, 2006.

[3] T. Akiyama and N. Hagita. Automated entry system for printed documents. *Pattern Recognition*, 23(11):1141–1154, 1990.

[4] D. X. Le, G. Thoma, and H. Weschler. Automated page orientation and skew angle detection for binary document images. *Pattern Recognition*, 27(10):1325–1344, 1994.

[5] H. B. Aradhye. A generic method for determining the up/down orientation of text in roman and non-roman scripts. *Pattern Recognition*, 38(11):2114–2131, 2005.

[6] S. Lu and C. L. Tan. Automatic document orientation detection and categorization through document vectorization. In *Proceedings of the 14th annual ACM International Conference on Multimedia*, pages 113–116, 2006.

[7] J. Beusekom, F. Shafait, and T. M. Breuel. Resolution independent skew and orientation detection for document images. In *Proceedings of SPIE Document Recognition and Retrieval XVI*, Jan. 2009.

[8] Y. LeCun and Y. Bengio. *Convolutional Networks for Images, Speech and Time Series*, pages 255–258. MIT Press, 1995.

[9] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions Systems, Man and Cybernetics*, 9(1):62–66, 1979.

[10] J. Sauvola and M. Pietikainen. Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236, 2000.

[11] 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*, pages 101–106, San Jose, CA, USA, Jan. 2008.

[12] P. Viola and M. J. Jones. Robust real-time face detection. *Int. Journal of Computer Vision*, 57(2):137–154, 2004.

[13] K. C. Fan, Y. K. Wang, and T. R. Lay. Marginal noise removal of document images. *Pattern Recognition*, 35(11):2593–2611, 2002.

[14] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324, 1998.

[15] F. Shafait, Adnan ul Hasan, D. Keysers, and T. M. Breuel. Layout analysis of urdu document images. In *10th IEEE International Multitopic Conference (INMIC 2006), Islamabad, Pakistan.*, Dec 2006.