Scale and Rotation Invariant OCR for Pashto Cursive Script using MDLSTM Network

Riaz Ahmad*, Muhammad Zeshan Afzal*, Sheikh Faisal Rashid†, Marcus Liwicki‡, Thomas Breuel‡
*riaz@iupr.com, afzal@iupr.com, rashid@iupr.com, TU-Kaiserslautern, Germany
†marcus.liwicki@dfki.de, DFKI Kaiserslautern, Germany
‡tmb@iupr.com, TU-Kaiserslautern, Germany

Abstract—Optical Character Recognition (OCR) of cursive scripts like Pashto and Urdu is difficult due to the presence of complex ligatures and connected writing styles. In this paper, we evaluate and compare different approaches for the recognition of such complex ligatures. The approaches include Hidden Markov Model (HMM), Long Short Term Memory (LSTM) network and Scale Invariant Feature Transform (SIFT). Current state of the art in cursive script assumes constant scale without any rotation, while real world data contain rotation and scale variations. This research aims to evaluate the performance of sequence classifiers like HMM and LSTM and compare their performance with descriptor based classifier like SIFT. In addition, we also assess the performance of these methods against the scale and rotation variations in cursive script ligatures. Moreover, we introduce a database of 480,000 images containing 1000 unique ligatures or sub-words of Pashto. In this database, each ligature has 40 scale and 12 rotation variations. The evaluation results show a significantly improved performance of LSTM over HMM and traditional feature extraction technique such as SIFT.

Keywords. Pashto, OCR, Cursive Script, HMM, LSTM, SIFT

I. INTRODUCTION

OCR of Pashto script is important due to its rich character set. Pashto contains 44 regular characters, including all characters from Arabic and Persian and 36 characters out of 38 characters from Urdu. The additional characters of Pashto script can be seen in Figure 1. This makes Pashto character set a generic one and a successful OCR for Pashto script will be applicable to Arabic and Persian as well. There has been a little published research on OCR of Pashto script but interest is growing due to the need of digitization of such scripts.

Recognition of Pashto is different from OCR of other scripts like Latin, Chinese, Japanese and Korean (CJK). Because in Latin and CJK, shape of individual characters may have very little variations for printed text and discriminative features of the characters almost remain the same. However, OCR of cursive script languages like Pashto is particularly difficult due to complex word formation rules and context sensitivity. In context sensitivity, shape of individual character changes when the character connects to other character inside a word [1], [2]. In addition to context sensitivity, the concept of “joiner” and “non-joiner” characters produce “space-omission” and “space-insertion” complexities [3]. Furthermore, the way Pashto is written makes it more difficult and less reliable to extract baselines, unlike Latin or CJK scripts, where characters line up precisely on baselines.

The objective of current research is to evaluate existing state-of-the-art OCR methodologies like LSTM based recurrent neural networks, HMM and descriptive based classifiers like SIFT for the recognition of Pashto ligatures. Pashto recognition is a complex problem due to the presence of different character level variations in shape, font, scale and rotation in real document images. These such variations can be seen in Figure 2. In Pashto, words can be decomposed into ligatures or sub-words. Examples of such Pashto ligatures or
The proposed work is based on our previous work[6]. However, the focus of current work is to evaluate state-of-the-art OCR methodologies like HMM and RNNs for the recognition of Pashto ligatures and compare their performance with SIFT based classifier. Additionally, the dataset has also been extended to include more scale and rotation variations.

B. Pashto Image Database

The dataset used in this research was first created by Mehreen Wahab [13]. At that time the dataset had 4000 Pashto sub-word images with 1000 unique shapes and four scale variations (12pt, 14pt, 16pt and 18pt) in font size for each sub-word. The sub-words are extracted from Pashto novel “Da Jwand ao Da Seray” means “This life and these faces”. In this work, the original dataset is extended synthetically to...
480,000 images by introducing forty scales and twelve rotation variations to the unique 1000 Pashto words. The dataset has been split into training and testing sets by randomly selecting eight rotations for training and four rotations for testing for each scale. The testing set is then further divided into validation and test set by randomly selecting two rotations for each scale. In this way, our training dataset contains 320,000 images, while testing and validation sets contain 80,000 images respectively.

II. PROPOSED METHODOLOGY

Despite the fact that scale normalization and slant detection techniques have been frequently used in pre-processing stage of OCR, why is it interesting to build a scale and rotation invariant OCR? The answer to this question is based on two logical arguments.

- A holistic classifier should learn complete shape of the object irrespective of its scale and rotation.
- If a classifier like multidimensional LSTM can learn from temporal sequences in all directions then feeding an input in a predefined or certain orientation and scale becomes useless.

Furthermore, the way Pashto is written makes it more difficult and less reliable to extract baselines like in Latin or CJK scripts where characters line up precisely on a baseline. Due to these problems normalization does not guarantee same recognition results neither at same scales nor at different scales. This concept is clearly explored in this research by doing counter experiment with HMM. As the holistic approach based classifiers should learn from the temporal sequences in all direction, but HMM mostly learns from sequences in one direction. Therefore, HMMs are evaluated using only normalized images with no rotation, but still they give worst results compare to LSTM and SIFT. This signifies importance of scale and rotation invariant OCR for Pashto cursive script.

In this work, three different Pashto ligature recognition systems has been presented using three different recognition methodologies. The first system is based on MDLSTM and recurrent neural networks. The system is evaluated against scale and rotation variations present in the Pashto ligatures dataset. Second recognition system is developed using SIFT based classifier. This system uses the same training and testing sets that have been used for the evaluation of LSTM based system. The third system uses ligature level hidden Markov models for the recognition of Pashto ligatures that have only scale variations. Next sections explain the proposed recognition systems in detail.

A. LSTM-Based Ligature Recognition of Pashto Script

A multi dimensional LSTM network is trained for the recognition of Pashto ligatures with different scale and rotation variations. The proposed network learns from the raw pixel values taken from an input block (image patch) from all four directions (i.e., left to right, right to left, top to bottom and bottom to top). The LSTM network is organized in a hierarchical manner as shown in Figure 5. The network uses image patches(input blocks) and preprocess the images for suitable feature extraction. The extracted features are then fed to higher layers for further processing. Overall network topology consist of the parameters such as input block size, hidden block size, sub-sample size and LSTM layer size. The size of LSTM layer decides the number of units in the layer. Sub-sampling size describes the number of feed-forward \( \text{tanh} \) units in the sub-sampling layers, which are inserted between every pair of hidden layers. Input block is the size of the patches in which the input image is divided for further processing and the hidden block size specifies the patch size at each hidden layer. The block size i.e 3x3 for each hidden layer is kept same for our experiments. The actual detail of the parameters is as follows:

The processing of each image is carried out by dividing it into small patches of size 3x3. After the input image has been divided into patches, these patches are further passed to the LSTM layers for processing. The remaining network layers
are described as follows: There are 3 hidden layers consisting of LSTM cells. The size of the each layer is 4, 20 and 100 respectively. The last hidden layer has the maximum size in order to provide as much feature as possible to the output layer for better classification. The hidden layers are fully connected. These three layers are further separated by two sub-sampling layers. These sub-sampling layers have size of 12 and 20 respectively. The sub-sampling layers are feed-forward \textit{tanh} layers. They reduce the number of weights significantly. Proposed network is trained on 320,000 images, while testing and validation are also done parallel with the training. The test and validation sets each contain 80,000 images. The network carried out 153 epochs for its complete training and each epoch has taken almost 7 hours and 44 minutes approximately. The task of learning and testing is set to classification rather than sequence learning. The learning rate was set to $1e^{-5}$, and the momentum was set to 0.9. RNNLib\textsuperscript{1} is used for the development of LSTM based Patho OCR system.

B. SIFT-Based Ligature Matching of Pashto Script

SIFT has been considered state-of-the-art feature extraction methodology for scale and rotation invariant object recognition. Therefore, in this work SIFT based ligature matching techniques is proposed for scale and rotation invariant Pashto ligature dataset. In this system, SIFT descriptor is extracted from each individual ligature image, and then the descriptor is appended to a huge descriptor. The huge descriptor is populated by indexing the image descriptor with the image label. Testing has been done by matching all possible pairs of SIFT features extracted from test image against the feature vectors from huge descriptor. Two SIFT features are said to be similar, if nearest neighbor of one has angle less than dist-ratio times 2nd. The value of dist-ratio is 0.6. Thus, an image descriptor from huge descriptor which provides maximum score for test image is selected as target image.

SIFT based recognition is significantly slower and more memory intensive than other recognition techniques. The descriptors required 22.5GB memory and matching of just one image required more than 3 minutes. In order to speed up recognition, the matching algorithm is performed on 10 parallel units. SIFT based system is developed using SIFT-Python library\textsuperscript{2}.

C. HMM-Based Ligature Recognition of Pashto Script

The third system is developed using ligature based HMMs. In this work, only the scale variations are considered and rotation variations are dropped from training and testing datasets. This reduced the training set to 26,549 images and the test set to 6,755 images. The purpose of dropping rotation variations is the variant nature of HMM against rotation. In this system, each ligature, along with its scale variants, is modeled using a single right to left, multi-state HMM. For HMM, input images are first cropped and then rescaled to the height of 30 pixels. The ligature based HMMs are trained by using pixel based raw features extracted from each input image [14]. Best recognition performance is achieved using 10 states HMM with 512 Gaussian mixture. The HMM topology used in this work can be seen in Figure 6. HTK toolkit\textsuperscript{3} is used for the development of proposed HMM based Pashto OCR.

III. RESULTS AND DISCUSSION

Quantitative comparison of the three approaches shows that LSTM-based method (recognition rate of 98.9%) significantly outperform both SIFT (94.3%) and HMM-based (89.9%) methods. Results are depicted in Figure 7. A careful interpretation of these results is required. In principle, all three methods could likely be improved by different choices of parameters, different training schedules, and different model structures. However, in all three cases, we chose reasonable and common defaults that have been used previously in other OCR applications. The three systems also do not solve quite the same problem: the HMM-based recognizer was given inputs that were normalized for scale and rotation, and it therefore did not have to generalize across such variation in input, meaning that the HMM-based recognizer had an easier problem to solve than either the LSTM or SIFT recognizer. The lower performance of HMM-based methods compared to

\begin{figure}[ht]
\centering
\includegraphics[width=\textwidth]{Fig6.png}
\caption{Fig. 6. Shows 10 state HMM topology having self loops and one state skip. The 10 state topology provides the best recognition accuracy so far.}
\end{figure}

\begin{figure}[ht]
\centering
\includegraphics[width=\textwidth]{Fig7.png}
\caption{Fig. 7. Recognition performance of LSTM, HMM and classification based on SIFT. LSTM not only learns the script dependent patterns but also learns the scales and rotation variations very well.}
\end{figure}

\end{document}
LSTM-based methods observed here is consistent with similar observations for other scripts like in [15] and [5].

Based on analysis of different error cases, we reached the following tentative conclusions. SIFT has been found good in cases, where the classification is required among the inter-classes. On the other hand SIFT has miss classified images where classification is required among intra-classes. An example of miss classification is shown in Figure 8, and it is clear that when base sub-word is same and other additional parts of characters are not the same then SIFT makes miss classifications. It is also observed that SIFT is more sensitive to image resolution rather than the shape. It means that SIFT can generate different number of features for same shape with different resolutions. Further, when dealing with huge dataset, the size of SIFT descriptor becomes another issue. In our case, it is not practical to wait for just one shape (connected component) to be classified in more than 3 minutes. Though, there are other ways to speed up this process by implementing SURF features or Visual Bag of Words etc. But certainly we will compromise in some part of recognition rate.

A very important aspect about the use of SIFT features is the selection of problem domain. For example, the problem domain in this research is to recognize Pashto cursive script. Our intention was to check different state of the art approaches against the scaling and rotation variations in the Pashto cursive script. SIFT features are considered to be the robust features against scale and rotation. However, it fails to give good results compare to LSTM. Alternatively LSTM is a good choice to handle these issues, though the training will take longer time than SIFT, but it is only once.

IV. CONCLUSION AND FUTURE WORK

Our results suggest that LSTM is currently the best method for recognizing Pashto script: even simple implementations yield comparatively low error rates with little parameter tuning or effort. Future work with LSTM-based Pashto recognition includes exploring more complex LSTM architectures, as well as feature extraction. SIFT-based methods have higher error rates. One potential advantage of SIFT-based methods is that it is fairly easy to understand how they work and to explain why they fail when they fail. Furthermore, transformations and invariance under transformation are explicitly engineered into SIFT-based methods, giving potentially better control over the degree of invariance and integration with other computer vision techniques for such approaches. We believe that SIFT-based methods can potentially be improved by match refinement strategies. HMM-based methods perform worst of the three methods, even though they were actually given already normalized data. The benchmarks in this paper, as well as previous results from the literature don’t let HMM-based methods appear to be very promising for high performance OCR compared to LSTM. Nevertheless, we plan on understanding the behavior of HMM-based recognizers and subclasses of inputs where they outperform LSTM methods.

Since our overarching goal is to make Pashto writing more generally available, our next steps are now to integrate the LSTM-based recognizers not only into a scan-based OCR but also into camera-based OCR system for Pashto script.

ACKNOWLEDGMENT

This study is supported by Shaheed Benazir Bhutto University, Sheringal and Higher Education Commission (HEC) Pakistan.

REFERENCES


Fig. 8. Miss-classification due to shape similarity. Here sub-words are same (intra-class) but additional parts of characters are not same. SIFT has misclassified certain number of such images.