J. Appl. Environ. Biol. Sci., 6(3S)96-101, 2016

ISSN: 2090-4274
Journal of Applied Environmental and Biological Sciences
www.textroad.com

Semi-Automated Transcription Generation for Pashto Cursive Script

Riaz Ahmad[1], Muhammad Zeshan Afzal[2], Sheikh Faisal Rashid[3], Saeeda Naz[4,5]

[1]Shaheed Benazir Bhutto University, Sheringal, Paksitan
[2]The Islamia University of Bahawalpur, Pakistan
[3]Department of Computer Sciences and Engineering UET, Lahore, Pakistan
[4]GGPDC No.1 Abbottabad, Higher Education Department, KPK, Pakistan
[5]Department of Information Technology, Hazara University, Mansehra, Pakistan

Received: January7, 2016
Accepted: March 2, 2016

ABSTRACT

Usually, a large amount of transcription data is required for training and benchmarking Optical Character Recognition (OCR) systems for new scripts like Pashto. In case of real image data; mostly the images are acquired through scanning. For supervised training scenarios, it is required to have a ground truth against the corresponding scanned images. Usually, the ground truth is created by transcribing the documents manually, which is an overwhelmingly laborious phase. This work introduces a semi-automated procedure for transcribing Pashto document images using a long short term memory (LSTM) network architecture. The process is applied for the transcription of 1000 images having Pashto ligatures and it improves the transcription performance to around three times of manual method.

KEYWORDS*:* Pashto, Transcription, OCR, Cursive Script, LSTM.

1. INTRODUCTION

In general transcriptions or ground truth data refer the target labels against classes. Thus the availability of transcription or ground truth data, is one of the key requirements for supervised learning. Similarly, transcribed text images in the domain of OCR are equally important for evaluating and benchmarking state of the art techniques. Because, predicted text (output of an OCR) of a text image can easily be checked against its corresponding transcription/ ground truth. Such transcribed datasets are extensively used for benchmarking as well as for training purposes, for example Arabic Printed Text Image (APTI) [1] and IFN[1]/ENIT[2] [2] databases. Further, importance of transcribed images becomes more significant when character based classification is required. As character based classification is only possible if characters in the images are properly labeled. These labels are arranged according to the sequence of their corresponding characters in the image. The optimal way to transcribe these characters, is to use appropriate uni-codes. Because, the uni-codes (labels) can be rendered on text editor directly, and there is no need of intermediate transformation of these transcription. Ultimately, edit distance measures can easily be calculated between predicted text and ground truth text.

This work focuses Pashto language as a test case, and regarding the recognition of Pashto text there is less significant research so far [3]. The dataset being used in this work contains 1000 unique ligatures/ sub-words of Pashto script. An image having some ligatures can be seen in Figure 1.

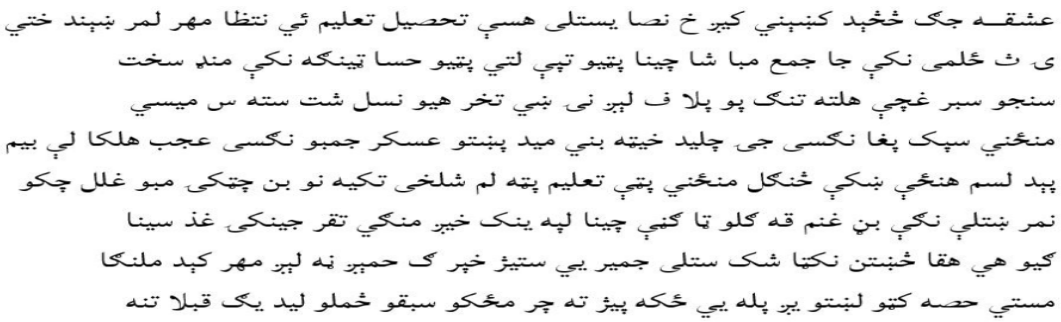

Figure 1. This figure shows an image, which contain Pashto ligatures. These ligatures are used in testset, to evaluate our proposed approach for transcription generation.

Corresponding Author: Saeeda Naz, GGPDC No.1 Abbottabad, Higher Education, Department, KPK, Pakistan & Department of Information Technology, Hazara University, Mansehra, Pakistan

Figure 2. Shows the corresponding predicted text (output of our proposed semi-automatic ground-truth generator) for the image shown in the Figure 1.

Initially, each ligature/image was transcribed with an integer id. Such transcriptions are only applicable to ligature based classification, but they fail when character based classification is required. Therefore, for character based classification, proper transcriptions for each corresponding character are required. Such transcription can be achieved by typing each label manually against the corresponding character in the image. However, writing transcriptions not only require professional typists but also need more time. This tedious task presents bottle necks for benchmarking of different OCR systems for script like Pashto.

This paper presents a semi-automatic ground-truth generator; which can be used for generating ground truth data for images having Pashto text. The proposed semi-automatic ground-truth generator is based on OCRopus [4]. In which we adapt a bidirectional LSTM (BLSTM) architecture. Another, the proposed approach can easily be adaptable to any other scripts, because OCRopus has been already using for many different script since 2008.

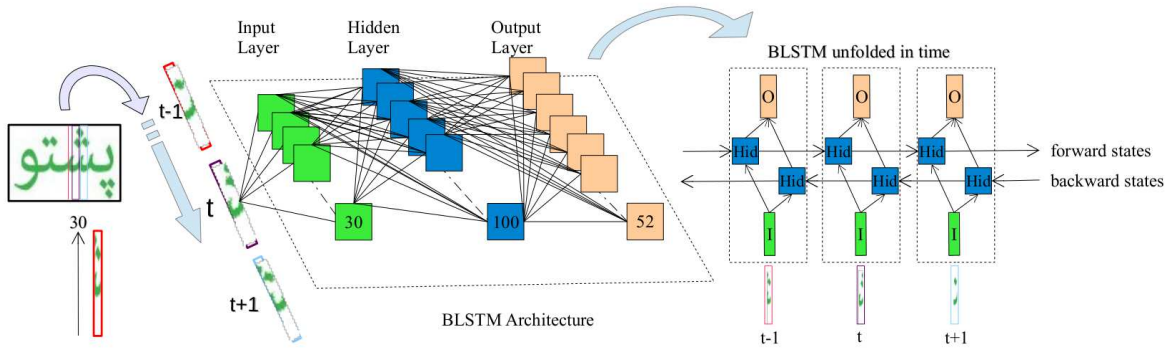

Figure 3. The proposed BLSTM architecture, with 30 units of input layer, 100 units of hidden layer and 52 units of output layer. The bi-directionality is shown, when network is unfolded in time.

The proposed system is trained on 100,000 words of Pashto. We have tested our proposed model on images having ligatures. The intention behind the checking of ligatures is based on two factors. First, in cursive script, salience of text shape is mainly dependent on ligatures instead of characters, because characters

usually lost their identify due to cursive nature. These are the ligatures which retain their shapes in almost all cases. Second, the ligatures being used in this research are the most frequent ligatures and having sufficient variation of constituent characters.

No.	Unicode (Hex)	Pashto Characters	No.	Unicode (Hex)	Pashto Characters
1	u+0627	ا	23	u+069a	ښ
2	u+0628	ب	24	u+0635	ص
3	u+067e	پ	25	u+0636	ض
4	u+062a	ت	26	u+0637	ط
5	u+067c	ټ	27	u+0638	ظ
6	u+062b	ث	28	u+0639	ع
7	u+062c	ج	29	u+063a	غ
8	u+0686	چ	30	u+06a9	ک
9	u+062d	ح	31	u+06ab	ګ
10	u+062e	خ	32	u+0644	ل
11	u+0681	څ	33	u+0645	م
12	u+0685	ځ	34	u+0648	و
13	u+062f	د	35	u+0646	ن
14	u+0689	ډ	36	u+06bc	ڼ
15	u+0630	ذ	37	u+0647	ه
16	u+0631	ر	38	u+06cc	ى
17	u+0693	ړ	39	u+06d0	ې
18	u+0632	ز	40	u+064a	ي
19	u+0696	ږ	41	u+0641	ف
20	u+0698	ژ	42	u+0642	ق
21	u+0633	□	43	u+06cd	ۍ
22	u+0634	ش	44	u+0626	ئ

Table 1. The Shapes Of Pashto Characters And Unicode Information

The performance evaluation of the proposed semi-automatic ground-truth generator has been done through an experiment. The experiment is divided into two phases. In first phase, total *4* images of Pashto text are given to *10* different typists. They are asked to type the corresponding text in an editor of their own choice. In this way total time consumed by each typist is recorded. In second phase, the same images along with predicted text (output of propose ground-truth generator) are given to the same *10* typists. They are asked to rectify only the errors in the produced predicted text. Similarly, total time consumed by each typist for completion of second phase is recorded. Thus, the average time taken for completion of phase two is 68% less than average time taken by phase one, which signify that our proposed semi-automated ground-truth generator is more efficient in generating transcriptions compare to manual work. An example of the predicted text from proposed semi-automatic ground-truth generator is shown in Figure 2. In addition, the Unicode (Hexadecimal) of *44* Pashto characters along with their shapes are given in Table 1.

The remaining paper is organised as; Related work is summarized in Section 2. A detail description of the proposed technique is available in Section 3. Section 4 explains how we carried out the experiments. In Section 5, we discuss the results, and finally in Section 5 there is conclusion and future work.

2. RELATED WORK.

Mostly related work regarding ground-truthing is based on segment and transcribed methodology. In this methodology each image is first segmented into text and non text, usually this step has been performed manually. Then, binarization is done automatically, and after binarization line segmentation is performed automatically. However, line segmentation need to be corrected manually. Similarly, transcription alignment is done automatically but it also need to be corrected manually. Furthermore, line normalization, feature extraction, word segmentation are done automatically. But word segmentation once again may need manual correction. An approach based on mentioned methodology was developed for degraded documents by [5]. Similarly another approach was also reported for generating ground-truth from German manuscripts using IAM Historical Handwriting Database (IAM-HistDB) [6]. Additionally, the importance of policies for ground-truthing was explored in [7], with finding that tighter bounding boxes can improve classification up to 45%.

However, adaptation of the above methodology for cursive script is difficult due to complex nature of cursive script. Particularly, the segmentation of words into its constituent characters is one of the most pattern

recognition challenges. Therefore, in this research we also avoid the segmentation based approach, and exploit the learning capability of neural networks along with labeling alignment by Connectionist Temporal Classification (CTC).

3. PROPOSED MODEL

It is already mentioned, that the proposed model is based on LSTM. LSTM has been dominantly used in sequence learning, because it has proven robustness in the domain of sequence learning [8][10]. Therefore, the proposed approach is based on LSTM.

We adapt the BLSTM architecture with optimal parameters. Hidden layer size is 100, which is also reported with optimal performance for Urdu language [8]. Overall architecture of BLSTM network consist of three layers, such as input layer, hidden layer and output layer. The size of the input layer is 30, follow by hidden layer with 100 LSTM units and size of the output layer is *52* i.e. total number of classes/ unique characters). The output layer is a CTC layer, which align the most probable label for target sequences. Learning rate is set to $1e^{-4}$. Architecture of the proposed model can be seen in Figure 3.

The system is trained on 100,000 words of Pashto. These words are collected from different web resources. These web resources are as follows:

- http://www.khabarial.com/
- http://www.khpalapashtu.com/
- http://www.kitabtoon.com/
- http://www.larawbar.net/
- http://www.tolafghan.com/
- http://afghanpost.com/
- http://www.sabawoon.com/
- http://rashad.benawa.com/

Each word is rendered into image with Karor font-shape and their corresponding transcription are saved in a text file.

During training each image is given a normalized height of *30* pixels. Each normalized image is then fed to LSTM along with its ground truth. After, *100k* iteration a model along with network error is saved. When, the network error is converged to minimum, then the training process is stopped. After this, a series of network models associated with network errors are available. Network model having minimum network error is chosen for testing.

In testing, the chosen network model is tested against 1000 images. Each image has a unique ligature of Pashto language, which is basically available in Karor font, further detail about creation of these ligatures can be found in [9]. The predicted text is saved in a text file with same file name as image file. At this stage, we are unaware of the exact performance of our proposed LSTM model. This is the stage where the need of proper transcription can be feel.

4. EVALUATION OF PROPOSED MODEL

Evaluation of the proposed tool is made by comparing two statistics obtained by two different phases. However, it is worth mentioning to understand the nature of the testset. We have created four images, in each image around *250* ligatures are pasted from our basic dataset. Each image has *14* lines and each line has *18* ligatures. These four pages are considered for manual as well as automated transcription generation. Therefore, in first phase the ground truth are generated manually and in the second phase the ground truths are generated automatically for the four pages. These two phases are explained in the following paragraphs.

In first phase four images are given to ten typists. These typists are hired in local community, and they are professional in composing and typing such documents. They are told to type the contents of each image, provided an editor of their own choice. They are further instructed to use stop watch, to record the time. In this way all the typists have completed phase one. The detail report of average time and individual time consumed by each typist can be seen in Table 2. The typists are referred with TW-Number, where Number is just id i.e. ranging from one to ten, and refer a particular typist. Each row in the table shows that how much time is taken for one particular typist to finish phase one. The average time taken to compose these four pages in phase one is *29* minuets and *33* seconds.

Typist	Time (mm:ss)
TW-1	29:33
TW-2	32:17
TW-3	27:16
TW-4	26:35
TW-5	28:31
TW-6	30:51
TW-7	28:34
TW-8	33:32
TW-9	31:21
TW-10	30:11
Total average time taken by Task1: 29:33 minutes	

Table 2. Each row represent total time for completion of writing four pages by each typist.

In second phase, the same four images are given to the same ten typists along with the predicted text. And this time they are instructed to make corrections only in the provided text corresponding to the given images. They are also instructed to use stop watch to record time. In this way, phase two is also completed. The detail report of time taken by each typist for completion of this phase is shown in Table 3. The average time taken to make the correction in predicted text is 9 minuets and 38 seconds.

5. RESULTS.

The statistics of two different scenario yield that our proposed semi-automatic ground-truth generator is more efficient in creating transcription for text images of Pashto language compare to manual method. The results in term of time consumed can be seen in Figure 3. Which shows that proposed semi-automatic ground-truth generator is three time faster than manual method. After testing and rectification of predicted text, we finally achieved the transcriptions for *1000* ligatures. The transcription/ ground-truth are evaluated against the predicted text, and we found that the proposed semi-automatic ground-truth generator has recognised *2555* characters out of *3160*, which gives around 80.85% character recognition rate. In other words, due to this accuracy around 80% key punching for characters typing is already retained, however to search and rectify a character in ligatures where miss-classification is occurred required overhead time of 12%. Compare to other approaches [5][6], our proposed approach also need manual correction and transcription alignment. However, it has an advantage over other approaches that user is not bothered in doing text, line and word segmentation manually. Although, the proposed approach is checked on limited ligatures and the time consumed for rectification is around *20* minutes less than the time required to write complete transcription (i.e. *30* minutes), a more prominent difference may be observed when large data is under consideration.

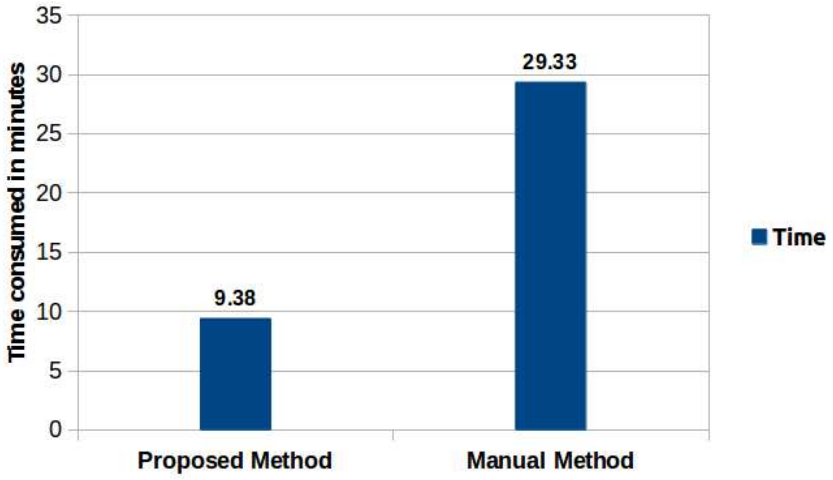

Figure 3. Shows that the consumed time of manual method is three time of our proposed semi-automatic approach for same quantity of transcriptions generation.

Typist	Time (mm:ss)
TW-1	09:13
TW-2	11:23
TW-3	08:45
TW-4	08:11
TW-5	08:44
TW-6	10:17
TW-7	09:37
TW-8	11:43
TW-9	10:51
TW-10	10:26
Total average time taken by Task2: 09:38 (mm:ss)	

Table 3. Total time taken for each typist to rectify the predicted text against the given images.

5. CONCLUSION AND FUTURE WORK.

On the bases of results, we can conclude that our proposed semi-automatic ground-truth generator is three time more efficient for transcription creation compare to manual method. Transcribed datasets can easily be evaluated against different state of the art approaches, and this is very important for new language like Pashto. We believe that, our proposed approach will play an important role in dataset extension by generating proper transcriptions as well as in the development of mature OCR for Pashto script. This is one of the mile stones toward robust OCR system for Pashto script.

Our next steps are now to create a standard image database of Pashto text. In which we will use the proposed approach for transcript generation. The database will contain scan-based as well as camera-captured images.

REFERENCES

1. Slimane, F., Ingold, R., Kanoun, S., Alimi, A. M., Hen-nebert , J. (2009), A New Arabic Printed Text Image Database and Evaluation Protocols. In proc. of 10th IEEE International Conference on Document Analysis and Recognition (ICDAR), Barcelona (Spain), 946-950.
2. Pechwitz, M., Maddouri, S. S., Margner, V., Ellouze, N., Amiri, H. (2002)., IFN/ENIT-Database Of Handwritten Arabic Words, 7th Colloque International Francophone sur l'Ecrit et le Document (CIFED), Hammamet, Tunis, 2, 127-136.
3. Naz, S., Hayat, K., Razzak, M. I., Anwar, M . W., Madani, S. A., Khan, S. U. (2013), The optical character recognition of Urdu-like cursive scripts, Pattern Recognition, 47(3), 1229-1248.
4. Breuel. T. M. (2008), The OCRopus open source OCR system, Proc. SPIE 6815, Document Recognition and Retrieval XV, 68150F.
5. Bal, G., Agam, G., Frieder, O. (2008), Inter-active degraded document enhancement and ground truth generation, Document Recognition and Retrieval, 6815.
6. Fischer, A., Indermuhle, E., Bunke, H., Viehhauser, G., Stolz, M. (2010), Ground Truth Creation for Handwriting Recognition in Historical Documents, Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, 3-10.
7. Moll, M. A., Baird, H. S. and Chang An, C. (2008) Truthing for Pixel-Accurate Segmentation, In Document Analysis Systems, the Eight IAPR Int. Workshop, 379-385.
8. Ul-Hasan, A., Ahmed, S. B., Rashid, S. F., Shafait , F., Breuel. T. M. (2013), Offline Printed Urdu Nastaleeq Script Recognition with Bidirectional LSTM Networks, 12th International Conference on Document Analysis and Recognition, 1061-1065.
9. Wahab, M., Amin, S. H., Ahmed, F. (2009), Shape analysis of Pashto Script and Creation of Image Database for OCR , International Conference on Emerging Technologies (ICET), IEEE, Islamabad, Pakistan, 287 - 290.
10. Graves, A. (2012), Offline Arabic Handwriting Recognition with Multidimensional Neural Networks, Book Chapter, Guide to OCR for Arabic Scripts, Springer, 297-313.