Impact of Ligature Coverage on Training Practical Urdu OCR Systems

Muhammad Ferjad Naeem*, Noor ul Sehr Zia*, Aqsa Ahmed Awan*, Faisal Shafait*, Adnan ul Hasanⁱ
*School of Electrical Engineering and Computer Science NUST
ⁱ Email: adnan.ulhasan@gmail.com
Email: {14beemnaeem,14besenzia,14mseeaawan,faisal.shafait}@seecs.edu.pk

***Abstract*—A major hurdle in the development of practical Urdu Nastaleeq script OCR is the lack of transcribed data, which is a pre-requisite for training machine learning algorithms. Most of the previous research has focused on UPTI, a publicly available data set with no particular focus on performance on real world images. UPTI contains only 6000 of the most probable 26,000 ligatures of Urdu. We build upon UPTI with a new data set, UPTI 2.0 that covers over 18,000 ligatures of Urdu Nastaleeq, hence covering over 70% of the ligatures that can practically occur. We further train a system on UPTI 2.0 and compare its performance against the only commercial Urdu Nastaleeq OCR system to date. Bidirectional Long Short-Term Memory (BDLSTM) network is employed with Connectionist Temporal Classification (CTC) layer as the recognizer. We show that systems trained on UPTI 2.0 outperform the commercial system.**

***Index Terms*—Urdu OCR, Recurrent neural networks, Ligature coverage, Tesseract, Machine Learning, LSTM, Data set**

I. Introduction

Urdu is the second most common language in the subcontinent. Urdu script is written right to left with numeric information written from left to right. The script is highly cursive and the sentences are written from top right to bottom left. Urdu characters change their shape depending on their position in the word.

Urdu script makes use of 45 alphabets; 5 of which occur in isolation, 10 of them can only be the leading or last alphabet of a word, 2 can only occur in the last position, and 1 can only take middle position, rest 27 alphabets can occur in any position [1]. A ligature is a combination of one or more alphabets to form a word; Urdu words usually have 1 to 8 ligatures per word [2], [3], [4]. Nastaleeq, a script developed in the 14th and 15th century in Iran is the predominant script for Urdu literature. Urdu Nastaleeq is highly context sensitive (Figure 1) [5]. Variation in the nature of script is only part of the challenge; there is significant variation in script depending on how it is published. Historically Urdu literature has been distributed via writings of calligraphers. Calligraphers are skilled individuals trained to write books in traditional Nastaleeq script. Due to the manual nature of the process, there is significant variation in the shapes of ligatures and angles of characters depending on the writer. Modern day Urdu literature in form of books, magazines, newspapers etc. is published in proprietary fonts. There is significant variation between Nastaleeq fonts and systems trained on one font don't generalize well on others. There are several Urdu publishing software available, amongst which, InPage[1] is the industry leader. Literature published using the same software is consistent in style but the proprietary font styles mean that data sets can not be synthesized with these fonts. There is a major short fall of transcribed text for both calligraphy based and published literature. This limits the development of a practical Urdu OCR system.

با	بی	بق	قبا	قب	ب	بری	بتی
8	7	6	5	4	3	2	1
بقول	بکرا	بختاور	بٹ	بیماری	کباب	قبر	بستا
16	15	14	13	12	11	10	9

Fig. 1: Urdu alphabets are highly context sensitive, Figure shows 16 different variations of ب [5].

The first attempt at an Urdu OCR system was reported by Pal and Sarkar [6]. Their approach was limited to single characters but it opened up the doors for future research. Ten years later, Javed et al [7] reported the first segmentation free Urdu OCR system using Hidden Markov Model (HMM). Sabbour and Shafait [8] used a multi feature k-Nearest Neighbor model and presented a practical approach to unsegmented Urdu OCR. They also presented the popular data set Urdu Printed Text Lines (UPTI). UPTI contains over 10,000 Nastaleeq Text Lines at both line and ligature level in undegraded and degraded form. UPTI was readily taken up as academia's standard for benchmarking Nastaleeq OCR systems. Over the years, there has been a lot of research quoted against UPTI. The first deep learning based approach to Urdu OCR was presented by ul Hasan et al [1]. ul Hasan used Long Short Term Memory networks with Connectionist Temporal Classification layer and achieved an error rate of 5.15%. Hussain and Ali [9] presented a recent approach to segmentation based Urdu Nastaleeq OCR. There have been several other attempts using HMM, kNN and multi dimentional LSTMs. Naz et al [10] reported the current best

[1] http://www.inpage.com/

2379-2140/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDAR.2017.30

error rate of 1.88%.

On the commercial side, CLE Nastaliq OCR [11] is the only available solution for Nastaleeq. CLE Nastaliq OCR is built on a previous release of the open source Tesseract-OCR API [12]. The tesseract engine uses polygon approximation [12] to detect characters and an adaptive classifier performs recognition. The adaptive classifier can be trained without the need of extensive language data. The ability to train with little training data means that language models are not an integral part of its working.

A major hurdle in the development of a machine learning based practical OCR system for Urdu is the lack of transcribed data. A practical Urdu OCR system must be capable of recognizing majority of the ligatures in Urdu language. Manually transcribing a data set covering the most probable ligatures is both time consuming and costly. A viable solution is to synthesize text for training a machine learning system as state-of-the-art OCR systems [13] have shown that LSTM-based OCR systems are fairly robust in dealing with small degradations and differences between scanned and synthesized text. The research from academia on Urdu OCR [1], [14], [15] has focused on improving results on synthetic data, in particular UPTI [8], a publicly available Nastaleeq data set. However, UPTI is not a good representation of the dynamic nature of the Urdu Language and in particular, Nastaleeq script. UPTI has only 6000 unique ligatures of Urdu Nastaleeq; while a humble estimate puts the unique ligatures in Urdu to be above 26,000 [16]. We test systems trained on UPTI on real world scanned images and highlight a major limitation of UPTI, it generalizes poorly on scanned images. Our proposal to improve accuracy is to increase the ligature coverage in UPTI data set. We extend upon UPTI with a new data set, UPTI 2.0, that contains over 18,000 unique ligatures of Urdu; hence covering the most probable ligatures an OCR system can observe in its lifetime. We further study the impact of ligature coverage on the performance of an OCR system and compare the results against the commercial OCR system. We show that UPTI 2.0 models can perform at par with the commercial OCR system.

This paper is further divided in four sections. Section II discusses the training model and the new data sets. Section III includes experiments to test effect of ligature coverage. Section IV discusses the results and Section V concludes the paper.

II. Methodology

We now discuss the methodology for our experiments. We also introduce our new data sets that make the comparison possible.

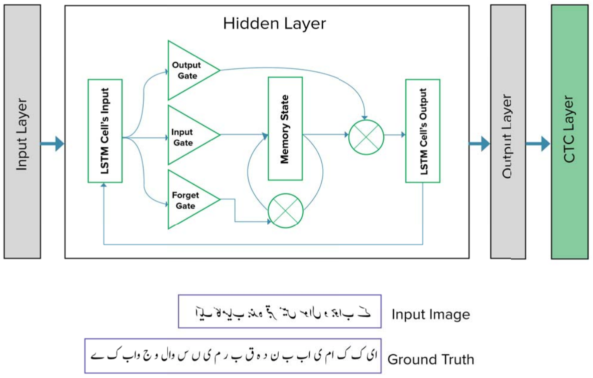


Fig. 2: Bidirectional Long Short-Term Memory Network is used with Connectionist Temporal Classification layer in the classifier. The text images are flipped horizontally while the ground truth is not changed allowing for Left to Right recognition of Urdu Text by the Network.

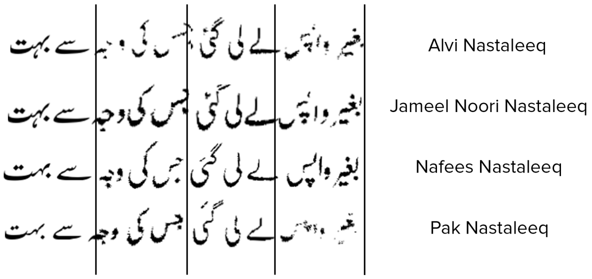

Alvi Nastaleeq
Jameel Noori Nastaleeq
Nafees Nastaleeq
Pak Nastaleeq

Fig. 3: UPTI 2.0 is available in four publicly available fonts. There are 100,000 unique text lines for the train set and 20,000 text lines for the test set. Each text line is available in four fonts and each font is available in four degradation levels. The figure shows a text line in four different fonts from top to bottom. Each line is further divided in four sections showing different degradations. From left to right we showcase non degraded, low, medium and high degraded variants.

A. LSTM Networks

Recurrent Neural Networks are an excellent model to recognize complex speech and text sequences due to their context aware property; However, they suffer from vanishing and exploding gradient problems. Long Short-Term memory networks [17] overcome these limitations by using a forget gate. LSTM have been successfully used for tasks such as unconstrained handwriting recognition [18], Fraktur script recognition [19] and speech recognition [20]. Connectionist Temporal Classification (CTC) [21] layer makes it possible to train Recurrent Neural Networks without the need for manually aligning the ground truth with the input. Bidirectional LSTM [19] have been reported to give excellent results on Urdu Nastaleeq script [1]. We use Bidirectional Long Short-Term memory networks with Connectionist temporal classification layer (CTC) [21] to train our system (Figure 2). Each image is resized to a height of 48 before feeding it to the network. ul Hasan et al [1] flipped the ground truth and passed the image without changes. We horizontally flip the images and no modification is done on the ground truth before passing to the network. We show that LSTM can process Urdu as a left to right language. A bigger learning rate would give faster convergence at the cost of possible oscillations around the global minima, a smaller learning rate would ensure that the system converges to the global minima at the cost of slower training. We use an exponentially decaying learning rate of 0.001 to get the best of both. We use Steepest Gradient Descent as used by [1], [19] with a momentum of 0.9. The size of the hidden layer is kept at 256 LSTM cells each in both forward and backward layers. We trained our network till it achieved an error rate of 6% on UPTI to recreate the results achieved by [1] and perform a sanity check. The error metric used to evaluate each system is Edit Distance [17].

1) Tensor LSTM (TLSTM) Trainer: We implemented our models using the popular Machine Learning API from Google, TensorFlow [22]. This model will be released as part of our OCR Library TLSTM. TLSTM supports Long Short Term Memory Cell [17], [19] based single and multi layer networks with Connectionist Temporal Classification (CTC) [21] layer. TLSTM can be used to train models for both Left to Right and Right to Left languages. Being built on TensorFlow, TLSTM supports GPU acceleration for OCR systems. GPU acceleration allows for faster training at a fraction of time compared to traditional RNN libraries. This allows for rapid prototyping and hypothesis testing. Our results show that TLSTM achieved an error rate of 6% on UPTI in 14 hours of training on a Nvidia Titan X GPU compared to 50+ hours required for RNNLib [23] on CPU as used in [1].

B. Data set

As discussed earlier, much of the previous Urdu Nastaleeq research has been quoted on the popular UPTI data set. We now introduce three new data sets focused on ligature coverage and real world scanned images.

1) UPTI 2.0: UPTI 2.0 builds upon UPTI [8] and extends its text corpus. UPTI 2.0 covers 18,000 unique ligatures. UPTI 2.0 was formed by collecting samples from Books, News articles and the Web. In total, 120,000 text lines were collected covering over 18,000 ligatures of Urdu Language. These text lines were then rendered using Pango [24] to form the data set. UPTI 2.0 comes in four publicly available fonts namely Alvi Nastaleeq, Jameel Noori Nastaleeq, Nafees Nastaleeq and Pak Nastaleeq. Each font variant is further available in four degradation levels to represent real world wear and tear in form of elasticity, smudge, fading and other variations. The degradation levels are listed as no degradation, low, medium and high. Hence each text lines is available in 16 variants. Figure 3 shows the different variants of UPTI 2.0. In total there are 1,920,000 text line images available, making UPTI 2.0 the biggest and most comprehensive data set for Urdu Nastaleeq.

TABLE I: UPTI performance on FZKR
The Commercial System performs better than all the five models.

System	Accuracy%
Alvi Nastaleeq	66.98
Jameel Noori Nastaleeq	70.15
Nafees Nastaleeq	52.95
Pak Nasataleeq	50.64
Multi Font	72.42
CLE Nastaliq OCR	**77.20**

2) FZKR: FZKR data set consists of text lines extracted from scanned images of InPage published books. These books were published using InPage and later transcribed to form the data set. There are 5,900 text line images available with their ground truth. Much of the modern day printed Urdu Literature is published in Inpage and an OCR system's performance on this data set is a good representation of its performance on real world scanned images.

3) URTI: URTI or Urdu Real World Text Images builds upon the data set from Shafait et al [25] and provides text line images and their ground truth. The data set consists of scanned text lines from Urdu Magazines, Newspaper, Poetry and Novels published in both printed and calligraphy form. There are 971 text lines from magazines, 233 text lines from books, 282 lines from Poetry and 694 text lines from Novels. This data set is aimed towards testing performance of OCR systems on unconstrained images. URTI will allow researchers to develop systems that bridge gap between traditional and modern day publishing.

III. Experiment Configuration

We perform multiple sets of experiments focused on ligature coverage, font variation and their impact on recognition of real world data. We further benchmark our network against the Commercial System CLE Nastaliq OCR [11].

A. Performance of UPTI on FZKR

We train five different variants of UPTI. Four of the variants are single font models and the fifth model is a multi font model containing the previous four fonts. The ratio of undegraded to degraded lines (from the low degradations set) is 0.7 to 0.3. Training sets of 9000 images of UPTI are used for each system with the multi font model containing each font in equal proportion. The networks are trained to convergence against a validation set of 1000 images. The five models are then tested on 1000 lines from FZKR data set and bench marked against the Commercial System. The results are shown in Table I.

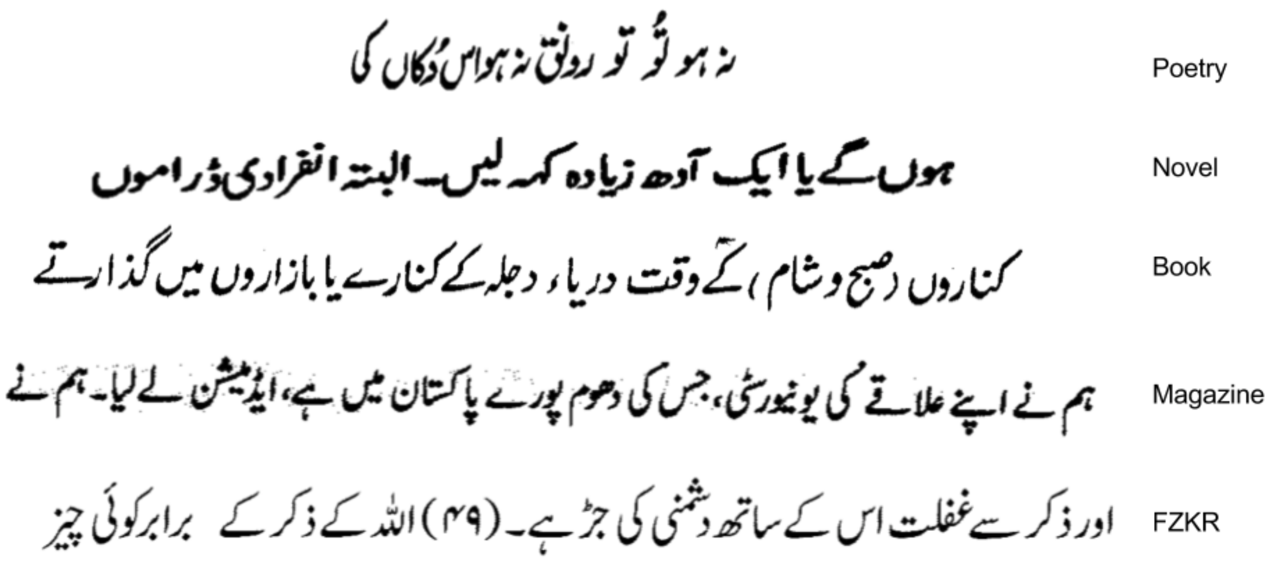

نہ ہو تُو تو رونق نہ ہواس دُکاں کی
Poetry
ہوں گے یا ایک آدھ زیادہ کہہ لیں۔ البتہ انفرادی ڈراموں
Novel
کناروں (صبح وشام) کے وقت دریا، دجلہ کے کنارے یا بازاروں میں گذارتے
Book
ہم نے اپنے علاقے کی یونیورسٹی، جس کی دھوم پورے پاکستان میں ہے، ایڈمیشن لے لیا۔ ہم نے
Magazine
اور ذکر سے غفلت اس کے ساتھ دشمنی کی جڑ ہے۔ (۴۹) اللہ کے ذکر کے برابر کوئی چیز
FZKR

Fig. 4: FZKR and URTI present an accurate account of real world data. Figure shows the four subsets of URTI in form of Poetry, Novel, Book and Magazine as well as a sample from FZKR data set. Poetry and Books are still written in calligraphy style for visual pleasure. Periodicals such as Newspapers, Novels and Magazines have little spacing between words to maximize coverage on paper. Literature such as FZKR is written in consistent easy-to-read format.

TABLE II: UPTI 2.0 performance on FZKR
The Multi Font model performs better than the Commercial OCR system.

System	Accuracy%
Alvi Nastaleeq	68.55
Jameel Noori Nastaleeq	74.81
Nafees Nastaleeq	55.69
Pak Nasataleeq	53.40
Multi Font	**78.33**
CLE Nastaliq OCR	77.20

B. *Performance of UPTI 2.0 on FZKR*

We now train five different models of UPTI 2.0. Similar to previous experiment, four of the models are single font models and the fifth model is a multifont model. The ratio of undegraded to degraded lines(from the low degradation set) is 0.7 to 0.3. A training set of 80,000 images, covering 18,000 ligatures, is used for each system. The multi font model has equal proportion of each font. The models are trained to convergence against a validation set of 10,000 images. The 5 models are again tested on 1000 lines from the FZKR data set and bench marked against the Commercial System. The results are shown in Table II

C. *Performance on URTI*

The best performing Multi Font Model from UPTI 2.0 is now bench marked against the Commercial System on the URTI data set. The performance is measured on the four subsets of URTI and the results are shown in table III.

TABLE III: URTI Test Results

Subset	Accuracy%	
	UPTI 2.0 MultiFont	CLE Nastaliq OCR
Magazine	49.31	49.7
Book	41.9	49.88
Poetry	40.6	35.45
Novel	42.3	61.35

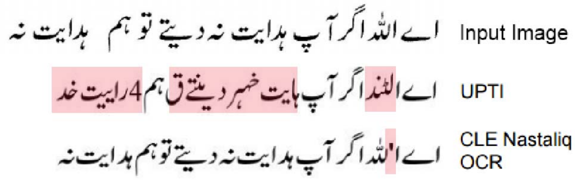

اے اللہ اگر آپ ہدایت نہ دیتے تو ہم ہدایت نہ
Input Image
اے اللند اگر آپ ہایت خہر دیتے ق ہم 4رایت خد
UPTI
اے اللہ اگر آپ ہدایت نہ دیتے توہم ہدایت نہ
CLE Nastaliq OCR

Fig. 5: UPTI Multi Font model failed on complex ligatures while the Commercial System recognized them fine.

IV. Results and Discussion

From the first set of experiments, we analyzed the results of UPTI. The single font models performed poorly with Jameel Noori Nastaleeq having the best accuracy figure. This is attributed to the fact that Jameel Noori Nastaleeq font takes its inspiration from the InPage fonts. The multi font model scored the best among the UPTI systems. After careful analysis of the results, we found UPTI models failing on complex ligatures (Figure 5). This was the inspiration behind creating UPTI 2.0. The Commercial System performed better than all UPTI

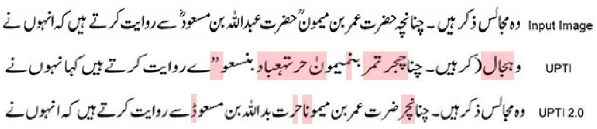
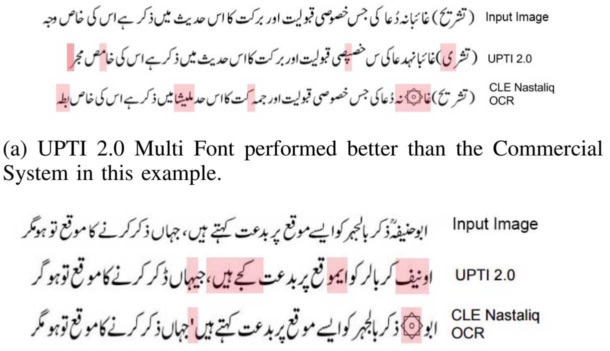

Input Image
UPTI
UPTI 2.0

Fig. 6: UPTI 2.0 Multi Font model performed better on complex ligatures than UPTI Multi Font.


Input Image
UPTI 2.0
CLE Nastaliq OCR
Input Image
UPTI 2.0
CLE Nastaliq OCR

(a) UPTI 2.0 Multi Font performed better than the Commercial System in this example.

(b) Commercial System performed better than UPTI 2.0 Multi Font in this example.

Fig. 7: Performance of UPTI 2.0 Multi Font and Commercial System varies depending on the input image.

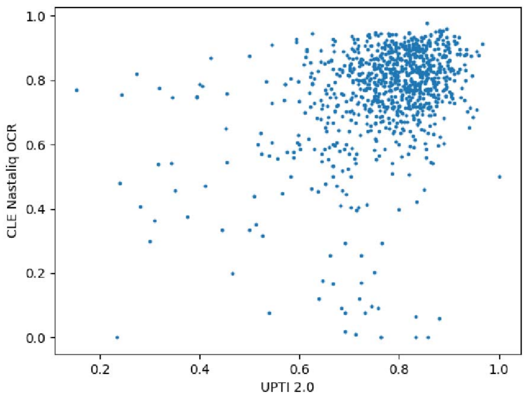


Fig. 8: Scatter Plot shows UPTI 2.0 Multi Font and the Commercial System can complement each other. The correlation coefficient of the two systems is 0.32 which indicates potential for combining their outputs.

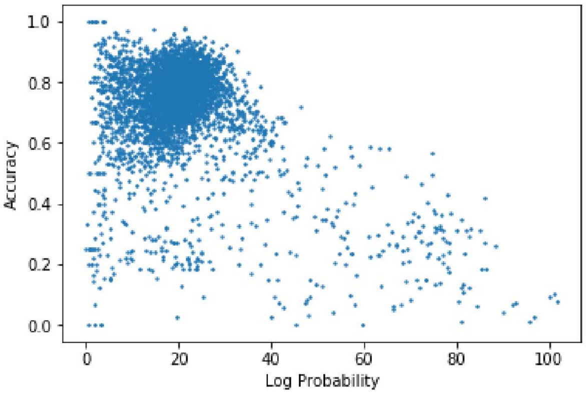


Fig. 9: There is very low correlation (0.04) between the log probability of CTC output sequence and the accuracy. This indicates that log probability is not a suitable metric of accuracy.

models.

The next set of experiments were performed on models trained on UPTI 2.0. The single font models from UPTI 2.0 performed better than their UPTI counter parts. Jameel Noori Nastaleeq was again the best performing single font model. The accuracy of single font Jameel Noori Nastaleeq model highlights the importance of having font representation in the training data. After careful analysis we found UPTI 2.0 models performing better than UPTI on complex ligatures (Figure 6). Indic scripts are highly context sensitive and there has been no previous work to our knowledge that studies the effect of ligature coverage on the generalization of an OCR system. UPTI 2.0 allows for better language representation of Urdu by covering a greater number of ligatures and the performance difference proves the hypothesis discussed from Section I.

The MultiFont UPTI 2.0 model performed slightly better than the Commercial OCR system. The performance difference between UPTI 2.0 and the Commercial System varied with one outperforming the other and vice versa depending on the image (Figure 7).

We further studied the performance difference between UPTI 2.0 MultiFont model and the Commercial System. The scatter plot of their accuracy on each line is shown in Figure 8. There is a possibility of combining the two systems to form a more accurate Nastaleeq OCR engine.

The third set of experiments on URTI highlighted the shortcomings of current OCR systems. Both systems failed on unconstrained images. There is a big gap between OCR systems for modern published Urdu literature and traditionally published, printed and calligraphed text. Further research needs to be performed to develop systems for unconstrained Urdu text. A huge portion of Urdu literature exists in old calligraphy based books and it needs to be digitized for preservation.

V. Conclusion

We presented a new data set for Urdu Nastaleeq and showed that ligature coverage has an impact on the accuracy of Indic script OCR systems. Future systems can be trained on UPTI 2.0 for better generalization on real world scanned images. LSTM and Deep Learning based models generalize better on real world images than traditional models. CLE Nastaliq OCR was engineered with InPage published scanned images in mind while our models were trained on synthetic text. However, both systems generalized to the same point.

There are several instances of labelling images without ground truth for training OCR systems in recent literature. Ahmed and Fink [26] proposed a method of using unlabelled images for training by running a trained model on the data and filtering the output based on confidence; this is an iterative approach to training on unlabelled images. We tested this method on BDLSTM with CTC using the log probability of the output sequence from the CTC. We found that there is no correlation between log probability and accuracy (Figure 9). Future research needs to be done to find better parameters for this unsupervised approach on LSTM based OCR systems. Moreover, new data sets need to be developed for traditional Urdu text to train OCR systems and preserve historic literature.

References

[1] A. Ul-Hasan, S. B. Ahmed, F. Rashid, F. Shafait, and T. M. Breuel, "Offline printed Urdu Nastaleeq script recognition with Bidirectional LSTM networks," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 1061–1065.

[2] G. S. Lehal, "Choice of recognizable units for Urdu OCR," in *Proceeding of the workshop on document analysis and recognition*. ACM, 2012, pp. 79–85.

[3] G. S. Lehal and A. Rana, "Recognition of nastalique urdu ligatures," in *Proceedings of the 4th International Workshop on Multilingual OCR*. ACM, 2013, p. 7.

[4] G. S. Lehal, "Ligature segmentation for urdu ocr," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 1130–1134.

[5] A. Wali and S. Hussain, "Context sensitive Shape-substitution in Nastaliq writing system: analysis and formulation," in *Innovations and Advanced Techniques in Computer and Information Sciences and Engineering*. Springer, 2007, pp. 53–58.

[6] U. Pal and A. Sarkar, "Recognition of printed Urdu Script." in *ICDAR*, vol. 2003, 2003, pp. 1183–1187.

[7] S. T. Javed, S. Hussain, A. Maqbool, S. Asloob, S. Jamil, and H. Moin, "Segmentation free Nastalique Urdu OCR," *World Academy of Science, Engineering and Technology*, vol. 46, pp. 456–461, 2010.

[8] N. Sabbour and F. Shafait, "A Segmentation-free approach to Arabic and Urdu OCR." in *DRR*, 2013.

[9] S. Hussain, S. Ali *et al.*, "Nastalique segmentation-based approach for urdu ocr," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 18, no. 4, pp. 357–374, 2015.

[10] S. Naz, A. I. Umar, R. Ahmad, I. Siddiqi, S. B. Ahmed, M. I. Razzak, and F. Shafait, "Urdu Nastaliq recognition using Convolutional–Recursive Deep Learning," *Neurocomputing*, 2017.

[11] S. Hussain, A. Niazi, U. Anjum, F. Irfan *et al.*, "Adapting Tesseract for complex scripts: An example for Urdu Nastalique," in *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*. IEEE, 2014, pp. 191–195.

[12] R. Smith, "An overview of the Tesseract OCR Engine," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 2. IEEE, 2007, pp. 629–633.

[13] T. M. Breuel, A. Ul-Hasan, M. Al Azawi, and F. Shafait, "High performance OCR for printed English and Fraktur using LSTM networks," in *ICDAR*, Washington D.C. USA, aug 2013.

[14] S. Naz, K. Hayat, M. Razzak, M. Anwar, S. Madani, and S. Khan, "The Optical Character Recognition of Urdu-like cursive scripts," *Pattern Recognition*, vol. 47, no. 3, pp. 1229–1248, 2013.

[15] S. Naz, A. I. Umar, R. Ahmad, S. B. Ahmed, S. H. Shirazi, and M. Razzak, "Urdu Nastaĺiq Text Recognition system based on Multidimensional Recurrent Neural Network and Statistical features," *Neural computing and applications*, vol. 26, no. 8, 2015.

[16] I. U. Khattak, I. Siddiqi, S. Khalid, and C. Djeddi, "Recognition of Urdu Ligatures-a holistic approach," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*. IEEE, 2015, pp. 71–75.

[17] E. S. Ristad and P. N. Yianilos, "Learning string-edit distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 5, pp. 522–532, 1998.

[18] T. M. Breuel, A. Ul-Hasan, M. A. Al-Azawi, and F. Shafait, "High-performance OCR for printed English and Fraktur using LSTM networks," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 683–687.

[19] M. Liwicki, A. Graves, H. Bunke, and J. Schmidhuber, "A novel approach to On-line handwriting recognition based on Bidirectional Long Short-Term Memory networks," in *Proc. 9th Int. Conf. on Document Analysis and Recognition*, vol. 1, 2007, pp. 367–371.

[20] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with Recurrent Neural Networks." in *ICML*, vol. 14, 2014, pp. 1764–1772.

[21] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling unsegmented sequence data with Recurrent Neural Networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[22] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale Machine Learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[23] A. Graves, "Rnnlib: A Recurrent Neural Network Library for sequence learning problems," *[OL][2015–07-10]*, 2013.

[24] O. Taylor, "Pango, an Open-source Unicode text layout engine," in *Proceedings of 25th Internationalization and Unicode Conference*, 2004.

[25] F. Shafait, D. Keysers, T. M. Breuel *et al.*, "Layout analysis of Urdu document images," in *Multitopic Conference, 2006. INMIC'06. IEEE*. IEEE, 2006, pp. 293–298.

[26] I. Ahmad and G. A. Fink, "Training an Arabic handwriting recognizer without a handwritten training data set," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*. IEEE, 2015, pp. 476–480.