Template Matching Based Probabilistic Optical Character Recognition for Urdu Nastaliq Script

Sumayya Ajaz¹, Muhammad Saeed Ahmad², Qaiser Abbas³, Muhammad Ilyas⁴, Mehrun Nisa⁵, Sadia Niazi⁶, Nadia Tabassum⁷, Ijaz Asghar⁸, Abu Sameer⁹

^{1,3,4,9}Department of Computer Science & IT, University of Sargodha, Sargodha, 40100, Pakistan

^{2,5}Department of Computer Science & Information Technology, GSCW University, Bahawalpur, Pakistan

⁶Department of Psychology, University of Sargodha, Sargodha, 40100, Pakistan

⁷Department of Computer Science, Virtual University, Pakistan

⁸Department of English, University of Sargodha, Sargodha, 40100, Pakistan

Email: qaiser.abbas@uos.edu.pk

ABSTRACT

This paper presents a technique for optical recognition of Urdu characters using template matching based on a probabilistic N-Gram language model. Dataset used has the collection of both printed and typed text. This model is able to perform three types of segmentations including line, ligature and character using horizontal projection, connected component labeling, corners and pointers techniques, respectively. A separate stochastic lexicon is built from a collected corpus, which contains the probability values of grams. By using template matching and the N-Gram language model, our study predicts complete segmented words with the promising result, particularly in case of bigrams. It outperforms three out of four existing models with an accuracy rate of 97.33%. Results achieved on our test dataset are encouraging in one perspective but provide direction to work for further improvement in this model.

KEYWORDS: OCR, Template matching, N-Grams, Segmentation, Stochastic, Nastaliq, Urdu

1. INTRODUCTION

Optical character recognition (OCR) is a technique that converts optically written characters, like printed text or handwritten text captured in an image from an imaging scanner or any other imaging device, of any specific language into equivalent computer recognizable formats like Universal Code [27]. Urdu is statutory state language of Pakistan according to the constitution of 1973[1]. Urdu is becoming popular as now many people around the world, like Pakistan, India, Bangladesh, Canada, etc, can speak and understand it [2].

Urdu is cursive in nature just like Arabic. It has 39 characters set, as shown in Fig 1. Almost every character has four different forms which are start, middle, end and isolated as

shown in Fig 2. Due to these different morphological forms of characters, the task for detection becomes more difficult for an OCR system. Besides this, shape and textures-based classification can be used in many other fields [28], [29].

Literature study on Urdu OCRs showed that not very considerable work has been done as compared to other languages like English, Latin, etc. [4]. It's hard because of the complexity in many font styles for typing or writing Urdu characters mainly because of their cursive nature like Naskh, Nastaliq, etc [5]. This motivated us to work on Urdu OCR. To solve this issue of OCR-related less resources and the complex nature of Urdu scripts, we planned to work on the Urdu Nastaliq script. Related work is discussed in Section 2.

Figure 1: Urdu Character set [3].

Figure 2: Different positions of character Meem

In this study, we developed a solution particular to the Nastaliq script. The issues faced in handwritten and printed text of Nastaliq script are minimized with the help of this model. Besides the steps of pre-processing and segmentation, our technique consists of template matching (TM) and recognition through N-Gram (NG) with the support of a stochastic lexicon [13]. Detail is presented in Section III. Results achieved are partially promising and this model will surely help in optical character recognition activities as most of the un-editable digital literature of Urdu is available in Nastaliq script which can be converted into an editable format for further research.

2. RELATED WORK

Since the beginning of Urdu OCR systems, researchers have performed this task using many techniques. Some researchers work for better segmentation techniques and some work for more efficient recognition techniques.

In [6], they presented a segmentation technique, which identified script [26] from a line of text. The script does not mean that it is a character, ligature, or a part of a character. It is just a unit of recognition that can be a character, ligature, or a part of a character. The segmentation into the script is based on the topological relation between the line adjacency graph (LAG) illustration of the text and the baseline of the text.

Elms [7] performed no segmentation, moreover, no ligatures or characters were segmented. Instead, a single connected component was considered as a single segmented character then afterward recognition

was performed using HMM bases recognizer.

Likewise in [8], they didn't segment lines and ligatures but instead they used the whole line as input for recognition. Some features were extracted that were further used for recognition. This same technique was followed in [9, 10].

In [11], they devised an offline technique only for Noori Nastaliq with ligature-centered recognition rather than character-centered recognition that means no segmentation. They used a multi-tier holistic method to identify ligatures from already created ligatures base. At the end with recognized ligatures and some RTS invariant features [24], they performed recognition through the feed-forward neural network.

In [12], the system gets images as input and one by one extracts each line of text. After extraction, a specialized algorithm is used for separating each ligature from every text line. Then different visible features [25] are extracted from each ligature for recognition, which is done by a simple template matching technique.

In [4], they performed the image preprocessing procedure and then selected some unique features. In the end, they performed recognition using multiple classifiers. Some of the classifiers they used are named neural networks, HMM, fuzzy logic-based model, and conditional random fields.

In [14], they developed a technique (DT-OCR) that performed some preprocessing using median filters and global thresholding. Then the features were extracted using different methods named as Hu moments, Zernike moments, and PCA. Lastly, the recognition was done using a decision tree classification algorithm. They claimed the best precision using Hu moment feature extraction technique as 92.06%. This model has similarities with our approach and a comparative study has been which is represented in Section 4 of the results.

Recently, an OCR study is presented with Linear Discriminant Analysis (LDA) for the Pashto language which also uses Nastaliq script [21]. Zernike moments are used for recognition with 10-fold cross-validation and achieved 63.71% of accuracy from the dataset of handwritten images. A comparative study is also made and presented in Section 4 of the results.

In another investigation [22], a line segmentation algorithm (LSA) is used for the handwritten and printed text of Urdu because the accuracy of an OCR model is directly

proportional to segmentation. An approach of counting pixels is used before removing the skewness of the text. Finally, they came up with 96.7% and 98.3% of line accuracy. Their comparative analysis is depicted in Section 4.

As it is advocated in most studies that accuracy is directly proportional to segmentation, so in [23], a segmentation-free (SF) OCR study is investigated. In this study, they captured ligatures from the Urdu Nastaliq text and then these ligatures were clustered and experimented with neural network. They secured 72.31% and 73.13% of accuracy on the whole dataset and average of clusters, respectively. Their comparative results are displayed in Section 4.

3. METHODOLOGY

The methodology of our work TMP-OCR (Template Matching based Probabilistic OCR) is partially extracted from [15]. They worked on the word segmentation problem and guessed the proper sequence of segmentation using ligature's unigram, bigram and trigram. But we have used the character's unigram, bigram and trigram to guess the more suitable next character. The design diagram of our system is given in Fig 3 which is discussed next.

3.1. Data Collection

Data collection is always the most critical part of any research work as poorly collected data can lead to wrong conclusions. We needed to collect three types of data:

Urdu raw corpus of printed textbooks is collected to create a stochastic lexicon. To see what is a stochastic lexicon, follow the text in [16]. The large set of both handwritten (UHaT Dataset) and printed text images¹ are collected to build and test our Urdu Optical Character Recognition (UOCR) system.

A list of Urdu Unicode characters is prepared from which bigrams and trigrams of characters are made. A stochastic lexicon from the Urdu corpus is built with the help of this list of unigram, bigrams and trigrams [17].

3.2. Pre-Processing

First of all, the input image is converted into a grey scale image. Then that greyscale image is binarized using Sauvola thresholding [18] with a 12x12 window with a threshold of 0.1. A sample image is displayed in Fig 4 after the step of pre-processing.

3.3. Segmentation

The segmentation process is further divided into different steps which are as follow:

Figure 3: TMP-OCR Design Diagram

Figure 4: Image after pre-processing

3.3.1. Line Segmentation:

The first step in segmentation is always line segmentation. Many researchers have used the horizontal projection technique to segment lines as it is the easiest way to segment a line. The horizontal projection of image in Fig 4 is given in Fig 5 and the output image is displayed in Fig 6 which clearly identifies the lines of the text in that image.

Ligature Segmentation: After line segmentation, we need to extract ligatures. For this purpose, we used connected component labeling (CCL), which is the most used

Figure 5: Horizontal Projection of Image in Fig. 4

Figure 6: After line segmentation

technique for ligature detection [11]. After using this technique, each diacritic is detected as a separate ligature. To understand and connect each diacritic with its main body we sketched a baseline using CCL as portrayed in Fig 7. Each main body must intersect the baseline, and, in this way, diacritics can be separated from the main body. In the end, each diacritic is connected with its main body as the image is displayed in Fig 8.

Figure 7: Baseline

Figure 8: Image after Ligature Segmentation

Character Segmentation: For character segmentation, at first a ligature is thinned using Jang Chin thinning algorithm [19], [20] and it gave us quite decent results and the output image of the word 'nahIN' is given in Fig 9.

Figure 9: Thinned Ligature

After thinning, the characters are segmented using corner points and junctions [21]. After separating ligatures, diacritics are associated using the left assignment rule for Urdu. Output image of Fig 9 is given in Fig 10 after segmentation of characters.

Figure 10: Segmented characters.

3.4. Recognition

As we have discussed earlier that we have used statistical features so for that purpose, the Probabilities of characters are calculated from the collected corpus containing 10145 words.

- Stochastic Lexicon: In our work, the

stochastic lexicon is a lexicon containing the probabilities of characters using N-grams. Different probabilities are calculated as follows:

- Character Unigram: Unigram probabilities for characters are simply the count of each individual character divided by the total number of occurrences of all characters in a corpus. This can be mathematically written as in Equation 1:

$$P(c_i) = \frac{C(c_i)}{\text{Total number characters}(N)} \quad (1)$$

Where $P(c_i)$ is the probability of i^{th} character.

3.4.1. Character Bigram:

Bigram probabilities of characters are the count of one character that appeared with its preceding character by the total number of occurrences of the preceding character. This can

$$P(c_i|c_{i-1}) = \frac{C(c_{i-1}c_i)}{C(c_{i-1})} \quad (2)$$

3.4.1. Character Bigram:

After character segmentation in Fig 3, if the first character of a ligature is detected, then that character is finally attempted to be identified accurately through template matching with the original input image. This template matching rescales each detected character with the original image using three reference points x, y, and z accordingly. In this way, it will keep on detecting the characters of ligatures. However, when it fails, the stochastic lexicon displayed in Fig 3 is scanned to get the next best probable character. In this way, the whole ligature is recognized.

4. RESULTS

In our experiment, the results achieved using unigram and trigram with template matching technique were not very hopeful as they were 57.33% and 67.13% respectively. But the outcome using bigram with template matching was very satisfying as it was 97.33%, which is quite comparative with some related studies discussed in Section 2. We noticed that the results were totally dependent on segmentation. If the ligatures are properly segmented into the correct number of proper characters, then the results can be better in the

case of unigram and trigrams. Further, our model's results for more frequent ligatures were better than others.

In a comparative study, the results of our model are compared with other existing models [14], [21], [22], and [23] only, and the rest of the related studies presented in Section 2 are lower in results and also performed a decade earlier, that's why they are not presented in the following Figure 11.

Figure 11: Comparison with existing studies

Our template matching-based probabilistic model TMP-OCR is a little behind the line segmentation-based LSA-OCR. We performed line segmentation but then those lines were further segmented into ligatures from which then characters were predicted finally. The main reason that TMP-OCR is lagging behind the LSA-OCR is the preprocessing step. In our methodology, preprocessing is traditionally a conversion to greyscale and then binarization while in LSA-OCR, they performed an adaptive thresholding approach. They used black and white, grey and RGB scales. They also performed Otsu's thresholding, filtration, and binarization. There are other issues which are discussed next.

5. DISCUSSION AND ISSUES

We designed a model with a quite simple approach and found little behind than the accuracy of LSA-OCR. The reason of this less accuracy has the issues including filtering, erosion, dilation, thinning, scaling, thresholding, skewing, etc. Following is the picture of the issues observed during experimentation. The address of these issues can make the accuracy better.

- Filtering, erosion, dilation techniques during the preprocessing of images can increase the accuracy remarkably.
- During normalization of dataset, thinning is the major process which helps

in finalizing the morphology of the word or character.

- Black and White scale provides clues in tricky situations.
- Thresholding techniques separate the background and foreground of images.
- Images should be skewed for better prediction. etc.

These are mainly the issues due to which we are lagging the state of art model LSA-OCR, however, the simplicity of our approach is no doubt a niche among the studies of Urdu OCRs, which outperformed the three existing state-of-the-art models except LSA-OCR.

6. CONCLUSION

In the future, a larger size corpus can be used for creating TMP-OCR because the calculated probabilities directly affect the recognition rate. A more error-free and robust segmentation unit is needed to be designed because the segmentation affects our recognition system more than the probabilities itself. An overly segmented image could ask you to select more characters in a ligature which is then subsequently resulted in wrong understanding. Same is the case with under segmentation. Hence, the segmentation unit needs improvement. Further, issues discussed in Section 5 should be addressed in future study.

REFERENCES

- [1] J. Pooja, K. Taneja, H. Taneja. "Which OCR toolset is good and why: A comparative study." Kuwait Journal of Science, vol. 48, issue 2, 2021.
- [2] L. M. Paul, G. F. Simons, C. D. Fennig. "Ethnologue: languages of the world, Dallas, Texas: SIL International." Online version: <http://www.ethnologue.com>, vol. 12, issue 12, 2010.
- [3] M. Sabahat, S. Zaman, M. W. Anwar. "Printed Urdu Nastalique script recognition using analytical approach." 13th International Conference on Frontiers of Information Technology (FIT). IEEE, pp. 334-340, 2015.
- [4] U. Pal, A. Sarkar. "Recognition of printed Urdu script." Proceedings of Seventh International Conference on Document Analysis and Recognition,

- [5] P. Jerin, S. Siripragada, V. P. Namboodiri, C. V. Jawahar. "Revisiting Low Resource Status of Indian Languages in Machine Translation." In 8th ACM IKDD CODS and 26th COMAD. pp. 178-187, 2021.
- [6] M. Hong, S. Nazir, Z. Shuo, W. Guan, Analyzing the Classification Techniques for Bulk of Cursive Languages Data: An Overview. Scientific Programming, 2021.
- [7] A. M. Elgammal, M. A. Ismail. "A graph-based segmentation and feature extraction framework for Arabic text recognition." Proceedings of Sixth International Conference on Document Analysis and Recognition. IEEE. pp. 622-626, 2001.
- [8] A. J. Elms, "A connected character recogniser using level building of HMMs." Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5). Vol. 2. IEEE. pp. 439-441, 1994.
- [9] L. Zhidong, et al. "Robust language-independent OCR system." 27th AIPR Workshop: Advances in Computer-Assisted Recognition. Vol. 3584. International Society for Optics and Photonics, pp. 96-104, 1999.
- [10] P. Mario, V. Maergner. "HMM based approach for handwritten Arabic word recognition using the IFN/ENIT-database." Proceedings of Seventh International Conference on Document Analysis and Recognition. Vol. 3. IEEE Computer Society. pp. 890-890, 2003.
- [11] S. A. Husain, S. H. Amin. "A multi-tier holistic approach for Urdu Nastaliq recognition." the Proceedings of International Multi Topic Conference. Karachi, Pakistan, 2002.
- [12] Z. A. Shah, "Ligature based optical character recognition of Urdu-Nastaleeq font." International Multi Topic Conference. Abstracts. INMIC IEEE, pp. 25-25, 2002.
- [13] Abbas, Qaiser. "A Stochastic Prediction Interface for Urdu." Intelligent Systems and Applications, vol. 7, issue 1. pp. 94-100. 2014.
- [14] K. Khan, R.U. Khan, A. Alkhalifah, N. Ahmad. "Urdu text classification using decision trees." 12th International Conference on High-capacity Optical Networks and Enabling/Emerging Technologies (HONET). IEEE, 2015.
- [15] M. Akram, S. Hussain. "Word segmentation for urdu OCR system." Proceedings of the Eighth Workshop on Asian Language Resources. pp. 88-94, 2010.
- [16] Davis, Mark, et al. "Unicode 3.1." Unicode Standard Annex 27. Available: <http://www.unicode.org/reports/tr27/tr27-4.html>, 2001.
- [17] S. Jaakko, M. Pietikäinen. "Adaptive document image binarization." Pattern recognition, vol. 33. issue 2. pp. 225-236. 2000.
- [18] S. T. Javed, S. Hussain. "Improving Nastalique specific pre-recognition process for Urdu OCR." 13th International Multitopic Conference. IEEE, pp. 1-6, 2009.
- [19] B. K. Jang, T. C. Roland. "One-pass parallel thinning: analysis, properties, and quantitative evaluation." IEEE Computer Architecture Letters 14(11): pp. 1129-1140, 1992.
- [20] M. Ahmed. "Urdu optical character recognition system MS thesis." Diss. National University of Computer & Emerging Sciences, 2007.
- [21] S. Jehangir, S. Khan, S. Khan, S. Nazir, A. Hussain. "Zernike Moments Based Handwritten Pashto Character Recognition Using Linear Discriminant Analysis." Mehran University Research Journal of Engineering and Technology 40(1): pp. 152-159, 2021.

- [22] M. Saud, et al. "An Efficient Skewed Line Segmentation Technique for Cursive Script OCR." Scientific Programming, 2020.
- [23] M. J. Rafeeq, Z. Rehman, A. Khan, I. A. Khan, W. Jadoon. "Ligature categorization based Nastaliq Urdu recognition using deep neural networks." Computational and Mathematical Organization Theory, vol. 25, issue 2, pp.184-195. 2019.
- [24] T. Zia, Q. Abbas, M. P. Akhtar. "Evaluation of Feature Selection Approaches for Urdu Text Categorization." International Journal of Intelligent Systems & Applications, vol. 7, issue 6, 2015.
- [25] T. Zia, M. P. Akhter, Q. Abbas. "Comparative study of feature selection approaches for Urdu text categorization." Malaysian Journal of Computer Science, vol. 28, issue 2, pp.93-109, 2015.
- [26] Q. Abbas, M. S. Ahmed, S. Niazi. "Language Identifier For Languages Of Pakistan Including Arabic And Persian." International Journal of Computational Linguistics (IJCL) 1(03), pp.27-35. 2010.
- [27] Naeem, A., et al. "Shape and Texture Based Classification Of Citrus Using Principal Component Analysis", International Journal of Agricultural Extension, vol. 9, issue 1, pp.15-24. 2021.
- [28] Muhammad, W., et al, "Classifying Urdu Verbs Using Rule Based Approach", LGU Research Journal of Computer Science and Information Technology, vol. 5, issue 1, pp. 57-64, 2021