Mehran University Research Journal of Engineering & Technology
Vol. 37, No. 2, 429-438, April 2018
p-ISSN: 0254-7821, e-ISSN: 2413-7219
DOI: 10.22581/muet1982.1802.18

Artificial Urdu Text Detection and Localization from Individual Video Frames

SALAHUDDIN UNAR*†,AKHTAR HUSSAIN JALBANI**,MUHAMMAD MOAZZAM JAWAID***, MOHSIN SHAIKH****, AND ASGHAR ALI CHANDIO*****

RECEIVED ON 19.07.2017 ACCEPTED ON 13.11.2017

ABSTRACT

In current era of technology, information acquisition from images and videos become most important task due to the rapid development of data mining and machine learning.The information can be either textual, visual, or combination of these. Text appearing in images or videos is a significant source of information and plays a vital role to perceive it. Developing a unified method to detect the text is hard, as textual properties (i.e. font, size, color, illumination, orientation, etc.) may vary with the complex background. So far, multimedia and computer vision community unable yet to standardize any ideal approach to extract the text smoothly. In this paper, a novel method is proposed to detect and localize artificial Urdu text in individual video frames. Firstly, Sobel and Canny edge detection operators are applied to input frame and are merged with MSER (Maximally Stable Extremal Region) detected regions. Next, geometric constraints are applied to eliminate obvious non-text regions with large and small variations. Further refining of non-text regions is achieved by stroke width transform. SVM (Support Vector Machine) classifier is trained to classify text and non-text objects. Finally, bounding boxes are used to localize the text.Experimental results show that the proposed method is robust and efficient than state-of-the-art methods.

Key Words: Text Detection, Artificial Urdu Text, Video Images, Maximally Stable Extremal Region.

1. INTRODUCTION

Reading the text and localizing it from videos and images became more popular and more challenging task since the last decade [1-3]. Text extraction in video has recently gained much consideration in multimedia understanding systems. The text exists under varying conditions such as fonts, size, color, orientation, and illumination, as shown in Fig. 1. Moreover, the task becomes more challenging when it exists in the multilingual and complex backgrounds. The text can be extracted through some

†Corresponding Author (E-Mail: imulticoder@gmail.com)
* School of Computer Science & Technology, Faculty of Electronic Information & Electrical Engineering, Dalian University of Technology, Dalian 116024, China.
** Department of Information Technology, Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah.
*** Department of Computer System Engineering, Mehran University of Engineering & Technology, Jamshoro.
**** Quaid-e-Awam University College of Engineering, Science & Technology, Larkano.
***** School of Engineering & Information Technology, University of New South Wales, Canberra, Australia.

mechanism and methods for several specific applications like visually impaired people's assistance [4], video indexing and retrieval [5], document analysis [6], content-based image search, automatic translation and so on.Since last decade, many researchers and scientists have been paying more attention to text acquisition from the video images; however, it is still a challenging task due to varying properties of text including unwanted reflection, shadow, and complex backgrounds.

FIG. 1. SAMPLE IMAGES OF ARTIFICIAL URDU TEXT UNDER VARYING CONDITIONS

In last few years, many researchers explored several methods for text detection and localization from videos in order to develop robust video retrieval systems [7-10]. Most of these methods focused on English language and some methods for Chinese and other languages. However, a minimal work can be seen for Urdu text. Urdu is the national language of Pakistan and official language of several cities of India. As reported in Wikipedia, Urdu has 65 million native speakers and 40 million second language speakers worldwide. The language was formed due to the impact of Arabic and Persian languages and is right to left script. Like Arabic and Farsi, Urdu characters can also have distinct shapes as per their position e.g. initial, middle, final or standalone. Urdu has 38 alphabets, and no upper or lower-case characters exist.

Recently, a notable progress can be seen on Urdu text. The researchers have investigated robust methods for Urdu OCR (Optical Character Recognition) [11-13], Urdu handwritten text recognition [14], Urdu document analysis [15]. However, extraction of Urdu text from video images is still rare explored research area as compared to other languages (e.g. English, Chinese). As there are more than 105 Urdu TV channels of sports, movies, music, news, religious and education worldwide, there is great need to extract Urdu text more robustly. Considering this, we have proposed a new approach which efficiently detects and localizes artificial Urdu text from video frames.

In this paper, we propose a framework which robustly detects and localize Urdu text. The framework is robust and efficient than the state-of-the-art methods available for Urdu language. The contributions of this paper are:

- A novel framework based on MSER and SWT (Stroke Width Transform) is proposed to detect artificial Urdu text.
- The proposed method is robust for complex background images with minimum computational complexities.

The rest of the paper is organized as follow: In Section 2, literature work available for Urdu and other languages is described. In section 3, the proposed methodology is presented. Experimental results and performance evaluation are given in Section 4. Section 5 concludes the results and future direction of proposed work.

2. RELATED WORK

The text in video/image can be categorized as artificial text and scene text. Artificial text can be seen in the form of captions, subtitles, annotations, and is embedded during editing of video. The algorithms to detect such text are mostly developed for indexing and retrieving purpose. However, scene text naturally and coincidently appears in different objects (e.g. signboards, walls, buildings) of an image and its algorithm mostly focuses on real-time applications.

The methods to detect the text can be divided into two types: region-based methods and CC (Connected Components) based methods. In region-based method, a learning-based method is usually employed to discriminate text regions from non-text regions by using some textural features (e.g. LBP (Local Binary Pattern), HOG (Histogram of Oriented Gradients), DCT (Discrete Cosine Transform), Gabor filter, etc.) to train a classifier such as AdaBoost, SVM or ANN (Artificial Neural Network). The CC-based method, generally creates distinguished CCs by using some image properties such as stroke, edge, width and color, and then some geometric constraints are employed to eliminate false positives. Recently, CC-based methods have gained more attention due to the efficient results and less computation time.

Khare et. al. [14] proposed a method to detect multilingual and arbitrarily-oriented text based on moments and gradient directions. Sobel and Canny edge are applied to input frame for finding automatic windows. A deviation based iterative procedure with K-means clustering is applied between consecutive frames. The gradient in inner and outer directions of pixels of edge components has been used to eliminate non-text regions. The extraction of full-text lines is proposed via Sobel edge images. Jamil et. al. [16] proposed a method to detect horizontal Urdu

text in video frames. Vertical gradient and average gradient magnitude are applied in individual input frame for binarization purpose. Then RLSA (Run Length Smoothing Algorithm) is employed to merge positive text regions, and edge density filter based on geometric constraints is employed to remove non-text regions.

Raza et. al. [15] introduced a method to detect multilingual artificial text which computes the wavelet transform to detect potential text regions. Gabor filtersare used for validation of text and non-text regions.To further remove non-text regions, FFT (Fast Fourier Transform) is engaged. The detected candidate text regions are validated using GLCM (Gray-Level Co-Occurrence Matrix) based features. Finally, the text is extracted using binarization threshold. The proposed method has greater computational complexity.

Jamil et. al. [3] proposed a hybrid approach to detect and identify multilingual script (i.e. English, Chinese, Arabic, Hindi, Urdu) in video frames. This work is an extension of their previous work presented in [16-17]. The proposed system is a generic framework and combines supervised and unsupervised approaches. For text detection, Sobel mask is applied to compute gradient value. Then ANN is employed to validate potential text candidates. Script identification is finally achieved using texture features based on LBP.

The extraction result can also be improved by enhancing the video quality. Considering this, some authors investigated methods to improve the video image quality. Shivakumara et. al. [18] proposed a method for text detection in mobile captured videos. Inspired from [19], the authors employed fractals to enhance the low contrast video text and K-means clustering algorithm toenhancethe image. The cluster with the maximum values is considered to be the text components. The fractal expansion is explored further in gradient domain to eliminate misclassified non-text components. To classify non-text components, optical flow properties are proposed. Finally, likely text candidates grouped into text lines by direction-guided boundary growing method. Roy et. al. [20] presented a fractional calculus based method to improve the quality of video frames acquired by Laplacian operation. However, the pixels are broken in proposed enhanced image, and it seems difficult to recognize the text.

3. PROPOSED METHOD

In this section, we describe a novel framework to detect and localize artificial Urdu text in video frames. The general workflow of our proposed methodology is shown in Fig. 2. Firstly, Sobel and Canny edge filters are applied to input frame and are merged with MSER detected regions in next step. Then some geometric constraints are employed to remove obvious non-text regions. Further, thetrue text regions are validated by stroke width and the SVM classifier. Finally, the detected text lines are formed into groups using bounding boxes. The detail of the proposed framework is as follows:

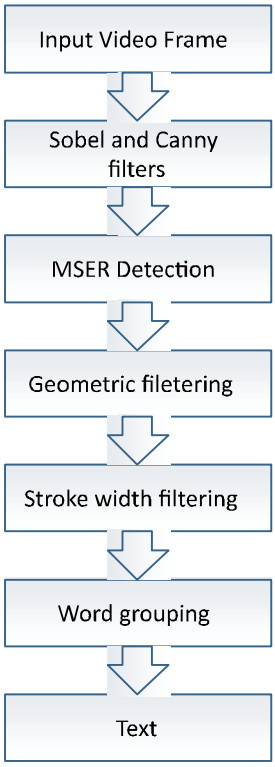


FIG. 2. GENERAL WORKFLOW OF PROPOSED METHOD

3.1 Text Detection

To detect textual regions in an individual frame, first,we used Sobel and Canny edge detectors to find potential edges, as shown in Fig. 3(b-c). Then MSER [21] feature detector is employed as it extracts maximum features due to high contrast and regular color intensities. The detected regions are shown in Fig. 3(d). The pixel area is setto $120 < \text{Area} < 400$ and the threshold is set to 3. The Sobel and Canny filters are merged with MSER to cope with blurred frames.

3.2 Localization and Validation

The obtained text regions are localized to text and non-text objects and then validated using geometric constraints to filter out non-text objects.Binarization of input frame is enhanced via Otsu's method. To filter out obvious non-text objects, we used simple geometric constraints such as width, height, aspect ratio. The objects having maximum and minimum variations are eliminated first.Then we set the aspect ratio of objects between 0.2 as Urdu text can have connecting characters.We used different constraints from ICDAR [22] which are observed best features by [23], and are given as follows:

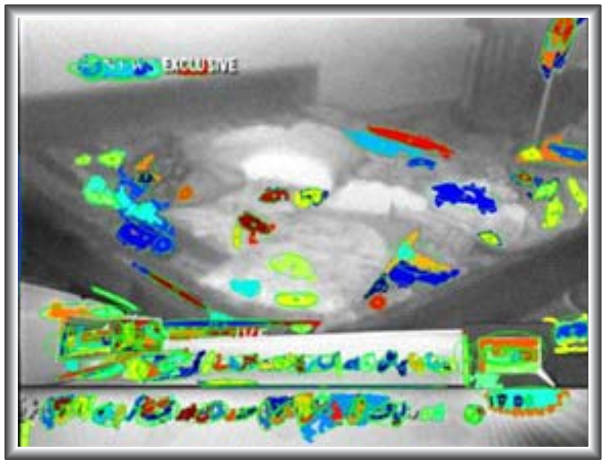
$$\text{Aspect\_Ratio} = \frac{\max(w,h)}{\min(w,h)} \tag{1}$$

$$\text{Solidity} = \frac{\text{area}}{\text{convex\_area}} \tag{2}$$

(a) ORIGINAL IMAGE

(b) SOBEL IMAGE

(c) CANNY IMAGE



(d) MSER REGIONS

FIG. 3. DETECTED TEXT REGIONS USING SOBEL, CANNY, AND MSER

$$\text{Ocuupy\_Ratio} = \frac{\text{area}}{\text{h} \times \text{w}} \tag{3}$$

$$\text{Compactness} = \frac{\text{area}}{\text{perimeter}^2} \tag{4}$$

$$\text{Strok\_Width\_Size} = \frac{\text{Stroke\_Width}}{\max(\text{h}, \text{w})} \tag{5}$$

Where h and w are height and width of thetext region, respectively. The area refers to the area of convex hull and solidity is a scalar which specifies the amount of the pixels in the convex hull. We set the Euler number <-2 that determines the number of objects in region minus the number of the holes in the objects. Based on stated geometric properties, many non-text objects are removed as shown in Fig. 4(a).

3.3 Segmentation and Extraction

Extraction step eliminates background pixels with the foreground. We use stroke width variation [24] for further extraction of true text regions. It is the length of a straight line from a text pixel to another pixel towards its gradient direction.Stroke width measures the width of curves and lines which can make a character. Text regions can have less stroke width variation, while non-text regions can have more variations.Skeleton image of remaining text regions is obtained by computing distance transform from each pixel to its nearest boundary pixel.We set the threshold rate to 0.3 and apply the procedure to each region filtered from previous step. Stroke width distance further segments non-text objects, as shown in Fig. 4(b).

3.4 Character Classification

The adjacent characters are grouped together to form a straight line. For this purpose, we train an SVM classifier with three different features including HOG, MDF (Mean Difference Feature) and SD (Standard Deviation) to reject false positives. To merge individual characters into wordsa [x,y,w,h]format is followed.Bounding boxesare imposed by finding neighboring textual regions. The bounding boxes having one textual regionare eliminated, as shown in Fig. 5(a) and true text regions are localized, as shown in Fig. 5(b).

4. EXPERIMENTAL RESULTS

In this section, we will briefly present the experimental results and performance evaluation. All experiments are implemented and executed on a computer with 6 GB Ram and 3.10 GHz CPU Intel Core-i3-2100.

(a) NON-TEXT OBJECTS REMOVED BY GEOMETRIC PROPERTIES

(b) NON-TEXT OBJECTS REMOVED BY STROKE WIDTH

FIG. 4. VALIDATION OF TEXT AND NON-TEXT OBJECTS USING GEOMETRIC CONSTRAINTS AND STROKE WIDTH

4.1 Dataset

We evaluated the proposed approach on publicly accessible Artificial Urdu Text Dataset [25] and compared the results with state-of-the-art methods available for Urdu text.The dataset consists of 1000 individual video images which are captured from different Urdu TV channels (e.g. News, Sports, Business, Entertainment, and Religion). All images have a uniform dimension of 720x576 pixels and have "png"file format.

4.2 Experimental Setup

To evaluate the performance, we used area based precision p and recall r measures, which are universally accepted, and are defined as:

$$p' = \frac{\sum_{r_e \in E} m(r_e, T)}{|E|} \tag{6}$$

$$r' = \frac{\sum_{r_t \in T} m(r_t, T)}{|T|} \tag{7}$$

$$f = \frac{1}{\frac{\alpha}{p'} + \frac{(1-\alpha)}{r'}} \tag{8}$$

where E is the estimated words, T is the ground truth targets. The frequency measure f is used to combine precision and recall. The relative weights of precision and recall are controlled bya. All the performance measures are computed for each image and then an average result is set for the performance of proposed approach.

The proposed approach achieved an overall precision rate of 83%, recall 88% and f-score 85%. We compared the result with state-of-the-art methods available for Urdu text. It can be noted that the proposed method outperformed the available methods. The obtained results are shown in Table 1. Fig. 6(a-b) demonstrates the true detected text and complex text examples.

(a) WORD GROUPING

(b) DETECTED TEXT

FIG. 5. BOUNDING BOXES ON DETECTED AND LOCALIZED TEXT.

TABLE 1. PERFORMANCE COMPARISON OF PROPOSED METHOD

Methods	Precision	Recall	F-Score
Our Method	0.83	0.88	0.85
Jamil et. al. [3]	0.58	0.84	0.69
Siddiqi and Raza [25]	0.71	0.80	0.75
Raza et. al. [26]	0.75	0.86	0.80
Jamil et. al. [15]	0.80	0.89	0.84

(a) TRUEDETECTED TEXT

(b) FALSE DETECTED TEXT DUE TO COMPLEX BACKGROUND AND FONTS

FIG. 6. TEXT DETECTION EXAMPLES

5. CONCLUSION

In this paper, we have investigated a robust approach to detect and localize artificial Urdu text in individual video frames. The framework provides efficient method to detect and localize Urdu text. The detection of textual regions in images is achieved with Sobel and Canny operators and the results are then merged with MSER detected regions. Simple geometric constraints are applied in next step to eliminate obvious non-text regions having very large and very small region area.We employed stroke width to obtain distance map in order to further remove non-text objects.Text and non-text objects are classified using SVM classification. Finally bounding boxes are used to localize the text. We performed the evaluation on Artificial Urdu Text Dataset and achieved better results compared to state-of-the-art methods.In future, we will further investigate and apply unsupervised learning technique to cope with blurred and multi-oriented text and implement Urdu OCR for recognition of detected text.

ACKNOWLEDGEMENT

Authors are extremely thankful to anonymous reviewers for their valuable comments and suggestions that helped us to improve the quality of the script.

REFERENCES

[1] Wei, Y., Zhang, Z., Shen, W., Zeng, D., Fang, M., and Zhou, S., "Text Detection in Scene Images Based on Exhaustive Segmentation", Signal Processing Image Communication, Volume 50, pp. 1-8, June, 2017.

[2] Khan, N., and Puri, S., "A Study on Text Detection Techniques of Printed Documents", IEEE Proceedings of International Conference on Wireless Communication Signal Processing Networking, pp. 2478-2482, 2016.

[3] Jamil, A., Batool, A., Malik, Z., Mirza, A., and Siddiqi, I., "Multilingual Artificial Text Extraction and Script Identification from Video Images", International Journal of Advanced Computer Science Application, Volume 7, No. 4, pp. 529-539, 2016.

[4] Joan, S.P.F., and Valli, S., "An Enhanced Text Detection Technique for the Visually Impaired to Read Text", Information Systetem Frontiers, pp. 1-18, September, 2016.

[5] Saravanan, D., "Effective Video Data Retrieval Using Image Key Frame Selection", Proceedings of 1st International Conference on Computer Intellegent Informatics, pp. 145-155, 2017.

[6] Wang, Y., Jmathew, J., Saber, E., Larson, D., Bauer, P., Kerby, G., and Wagner, J., "Scanned Document Enhancement Based on Fast Text Detection", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1961-1965, 2016.

[7] Mittal, A., Roy, P.P., Singh, P., and Raman, B., Rotation and Script Independent Text Detection from Video Frames Using Sub Pixel Iapping", Journal of Visual Communication and Image Representation, Volume 46, pp. 187-198, 2017.

[8] Sun, Y., and Yu, J., "Robust Scene Text Detection for Multi-Script Languages Using Deep Learning", 23rd International Conference on Multimedia Modeling, Volume 10133, pp. 209–220, 2017.

[9] Jaderberg, M., Simonyan, K., Vedaldi, A., and Zisserman, A., "Reading Text in the Wild with Convolutional Neural Networks",International Journal Computer Vision, Volume 116, No. 1, pp. 1-20, 2016.

[10] Yu, C., Song, Y., and Zhang, Y., "Scene Text Localization Using Edge Analysis and Feature Pool", Neurocomputing, Volume 175, Part-A, pp. 652-661, 2016.

[11] Choudhary, p., and Nain, n., "A Four-Tier Annotated Urdu Handwritten Text Image Dataset for Multidisciplinary Research on Urdu Script", ACM Transactions on Asian and Low Resource Language Information Processing, Volume 15, No. 4, 2016.

[12] Din, I. Malik, Z., Siddiqi, I., and Khalid, S., "Line and Ligature Segmentation in Printed Urdu Document Images Line and Ligature Segmentation in Printed Urdu Document Images", Journal of Applied Environmental Biological Science, March, 2016.

[13] Ahmad, I., Wang, X., Li, R., and Rasheed, S., "Offline Urdu Nastaleeq Optical Character Recognition Based on Stacked Denoising Autoencoder", Communication Theoretical System, pp. 146-157, 2016.

[14] Khare, V., Shivakumara, P., Paramesran, R., and Blumenstein, M., "Arbitrarily-Oriented Multi-Lingual Text Detection in Video", Multimedia Tools Application, pp. 1-31, 2016.

[15] Raza, A., Siddiqi, I., Djeddi, C., and Ennaji, A., "Multilingual Artificial Text Detection Using a Cascade of Transforms", Proceddings of International Conference on Document Analysis and Recognition, pp. 309-313, 2013.

[16] Jamil, A., Siddiqi, I., Arif, F., and Raza, A., "Edge-Based Features for Localization of Artificial Urdu Text in Video Images", Proceedings International Conference on Document Analysis and Recognition, pp. 1120–1124, 2011.

[17] Jamil, A., and Abidi, A., "A Hybrid Approach for Artificial Urdu Text Detection in Video Images", Proceedings of International Conference on Pattern Recognition, pp. 1944-1947, 2012.

[18] Shivakumara, P., Wu, L., Lu, T., Tan, C.L., Blumenstein, M., and Anami, B.S., "Fractals Based Multi-Oriented Text Detection System for Recognition in Mobile Video Images", Pattern Recognition, Volume 68, pp. 158-174, 2017.

[19] Xu, H., Zhai, G., and Yang, X., "Single Image Super-Resolution with Detail Enhancement Based on Local Fractal Analysis of Gradient", IEEE Transactions on Circuits and Systems for Video Technology. Volume 23, No. 10, pp. 1740–1754, October, 2013.

[20] Roy, S., Shivakumara, P., Jalab, H.A., Ibrahim, R.W., Pal, U., and Lu, T., "Fractional Poisson Enhancement Model for Text Detection and Recognition in Video Frames", Pattern Recognition, Volume 52, pp. 433-447, 2016.

[21] Bouaziz, B., Zlitni, T., and Mahdi, W., "AViTExt: Automatic Video Text Extraction; A New Approach for Video Content Indexing Application", 3rd International Conference on Information and Communication Technologies: From Theory to Applications, pp. 1–5, 2008.

[22] Lucas, S.M., Panaretos, A., Sosa, L., Tang, A., Wong, S., and Young, R., "Robust Reading Competitions", Proceedings of International Conference on Document Analysis and Recognition, pp. 682-687, 2003.

[23] Bergasa, L.M., and Yebes, J.J., "Text Location in Complex Images", Proceedings International Conference of the Association for Pattern Recognition, pp. 617–620, 2012.

[24] Li, Y., and Lu, H., "Scene Text Detection via Stroke Width", Proceedings International Conference of the Association for Pattern Recognition, pp. 681-684, 2012.

[25] Siddiqi, I., and Raza, A., "A Database of Artificial Urdu Text in Video Images with Semi-Automatic Text Line Labeling Scheme", 4th International Workshop on Quality of Multimedia Experience, pp. 75-81, 2012.

[26] Raza, A., Abidi, A., and Siddiqi, I., "Multilingual Artificial Text Detection and Extraction from Still Images", Proceedings of Document Recognition and Retrieval, 2013.