2011 International Conference on Document Analysis and Recognition

Edge-based Features for Localization of Artificial Urdu Text in Video Images

Akhtar Jamil Imran Siddiqi Fahim Arif Ahsen Raza
Department of Computer Software Engineering
National University of Sciences & Technology
Islamabad, Pakistan
{akhtar.jamil, imran.siddiqi, fahim, ashen.raza}@mcs.edu.pk

***Abstract*—Content-based video indexing and retrieval has become an interesting research area with the tremendous growth in the amount of digital media. In addition to the audio-visual content, text appearing in videos can serve as a powerful tool for semantic indexing and retrieval of videos. This paper proposes a method based on edge-features for horizontally aligned artificial Urdu text detection from video images. The system exploits edge based segmentation to extract textual content from videos. We first find the vertical gradients in the input video image and average the gradient magnitude in a fixed neighborhood of each pixel. The resulting image is binarized and the horizontal run length smoothing algorithm (RLSA) is applied to merge possible text regions. An edge density filter is then applied to eliminate noisy non-text regions. Finally, the candidate regions satisfying certain geometrical constraints are accepted as text regions. The proposed approach evaluated on a data set of 150 video images exhibited promising results.**

Keywords-Urdu text detection; Gradient edge detection; Run length smoothing alogrithm

I. INTRODUCTION

Modern era has seen a tremendous growth in multimedia data in the form of images, audios and videos. Addressing challenges raised by such rapid growth requires efficient techniques for indexing and retrieval of multimedia data-videos being the focus of our interest. In addition to other features, a very powerful attribute for indexing videos is the textual content appearing in them. Exploiting text for indexing and retrieval of videos requires automatic extraction of text from videos and its recognition.

Text appearing in videos is generally classified into two categories: caption text (also known as graphics/artificial text) and scene text. Caption text refers to the text that is artificially embedded in the video at the time of editing, for example names of anchors, score cards, stock exchange data and ticker text etc. Scene text on the other hand refers to the text that naturally appears in the scene during capturing of video, for example text appearing on sign boards and t-shirts. In general artificial text is considered useful for video indexing and retrieval while scene text finds its application in areas like robot navigation, intelligent vehicles and license plate recognition.

Traditionally, the text detection process is divided into three steps: text detection, text localization, and text extraction. In the detection step, text regions and non-text regions are distinguished from each other. In text localization text boundaries are identified from the candidate text regions detected in the previous step. Text extraction refers to removing the background elements so that it contains only text elements. Once the text strings are extracted from the input video image, they could be fed into an Optical Character Recognition System (OCR) for recognition and then subsequent indexing of videos. Our focus in present research is on the extraction of text only and not its recognition.

In this paper we present a method for extraction of artificial Urdu text from video images. More than 65 Urdu channels related to news, entertainment and sports are presently operating around the world. The literature is very rich in text detection for various languages based on Latin or Chinese alphabets. However, the research on detection of Urdu text is still in its infancy and is yet to be fully explored. In the proposed approach we have addressed the problem of caption detection from video images using edge features for horizontally aligned artificial Urdu text.

The rest of the paper is organized as follows. In section II we present related work for text detection in videos/images. Section III provides a detailed description of the proposed approach. We next describe the experimental results and finally present the concluding remarks outlining areas for future research.

II. RELATED WORK

In this section we briefly discuss some of the well-known existing approaches for text detection in videos/images. A comprehensive survey of the techniques proposed till 2004 can be found in [1]. State-of-the art approaches for text detection can be classified into two main categories: Supervised and unsupervised approaches. Supervised approaches employ machine learning methods for detection of textual content. Features extracted from text and non-text regions are used to train a classifier (e.g. support vector machines (SVM) or artificial neural networks (ANN) etc.). Among the well known supervised methods, Support Vector Machines (SVM) have been effectively employed in [2, 3, and 4] where edge and texture based features are used to distinguish text and non-text regions. A combination of SVM and Continuous Adaptive Mean Shift algorithm is investigated in [6]. In [5], the authors employ a local binary patterns (LBP) operator for feature extraction and classify the text and non-text areas using a polynomial neural network (PNN).

The unsupervised approaches for text detection are based on image analysis techniques. They exploit certain statistical and temporal properties of text to identify the textual content

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.226

in an image. The unsupervised methods are generally categorized into region (connected components), edge or texture based methods.

Connected components based methods [7, 5] generally use a bottom-up (e.g. region growing) or a top-down (e.g. region splitting) approach to group text pixels into clusters. These methods rely heavily on the contrast between text and background and generally fail to retain character shapes in low resolution images.

Edge-based methods [8, 9, and 10] generally segment the text in the image by finding the edges in the image typically followed by some morphological processing. Texture based methods [11, 12] treat text present in images as special type of texture that could serve to distinguish it from non-text regions. These methods perform relatively better in complex backgrounds. However, they produce more false positives when the background contains textures that expose similar properties as the text.

Most of these methods have been developed for text in languages based on the Latin alphabet. Research on Chinese text detection is also quite mature. However, to the best of our knowledge, despite a large number of Urdu news, sports and entertainment channels, work on Urdu text is nonexistent. In the following section we discuss the proposed methodology for detection of artificial Urdu text in video images.

III. PROPOSED SYSTEM

The proposed approach for Urdu text detection mainly rests on an edge based segmentation followed by a pixel density filter and some geometrical constraints. The detected text regions are then binarized to segment the text pixels from background. Figure 1 gives an overview of the proposed scheme, the steps being discussed in detail in the following.

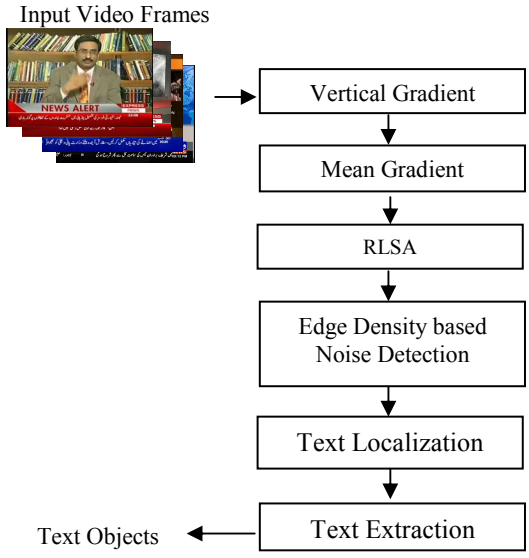


Figure 1. Scheme of the proposed system

The input video image is first converted to grayscale so that only the luminance information in the image is used for text detection. Like many other scripts, vertical strokes are very dominant in Urdu. We therefore compute the vertical gradient using Sobel operator:

$$\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$

Since the characters and partial words in text are likely to appear in groups rather than in isolation, a natural step after edge detection is to enhance the density of (high magnitude) gradient pixels in the text regions. This is generally achieved by sliding a window over the gradient image and performing some operation. The same idea is implemented using accumulated gradients in [10] and gradient difference in [13]. Using a horizontal sliding window of size 1x*s* (*s* being chosen empirically and set to 13 for our experiments), we evaluated both of these operations as well as an average filter where each pixel is replaced by the average value of the gradient in the window. Figure 2 shows a comparison of these approaches. The average filter has less severe effects than accumulation and gradient difference and performed the best in our experiments on Urdu text. All the subsequent steps are therefore performed on the image obtained by averaging the gradient magnitude in the fixed neighborhood (of size *s*) of each pixel. As a result of this process, isolated noisy gradients are suppressed while gradients in the text and 'text-like' regions persist.

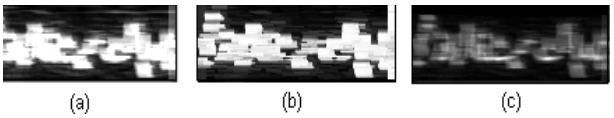


Figure 2. (a)Accumulated gradient (b) Gradient Difference(c) Average Gradinet

The average gradient image is next binarized for which we employ the well-known Otsu's global thresholding algorithm. This step eliminates the gradients with weak intensity giving the likely text regions which appear as isolated connected components in the proximity of one another. To merge these isolated components into words/lines, we next apply the Run Length Smoothing Algorithm (RLSA) in the horizontal direction. As a result, all foreground pixels separated horizontally by less than C pixels (C being the RLSA threshold fixed to 5 for our experiments) are merged into one component.

For images with simple backgrounds, these steps effectively identify the text areas suppressing most of the non-text regions. However, in images with complex backgrounds containing persons, buildings and other similar objects, these objects are also classified as text regions. Since the spatial redundancy of text is likely to be more than that of these false positives, we employ a pixel density filter to suppress these non-text regions.

The (foreground) pixel density is estimated locally for each region by using a rectangular sliding window that scans the whole image from left to right and top to bottom with a

step equal to one fourth of the size of the image window. Naturally, the local pixel density in text regions is likely to be high as opposed to noisy non-text regions. The window is identified as text region if the pixel density is greater than a predefined threshold (0.35). Windows not satisfying this density criterion are likely to be non-text regions and are suppressed.

Once the text regions are identified, we localize the words/lines by performing a connected component analysis on the image and representing each component by a rectangular bounding box.

As a final step we employ the traditionally used geometrical constraints on the localized rectangles to eliminate the ones that do not satisfy the geometrical properties of text. These constraints are based on the aspect ratio and minimum height and width of the detected rectangles giving a set of rectangles which are likely to contain textual content.

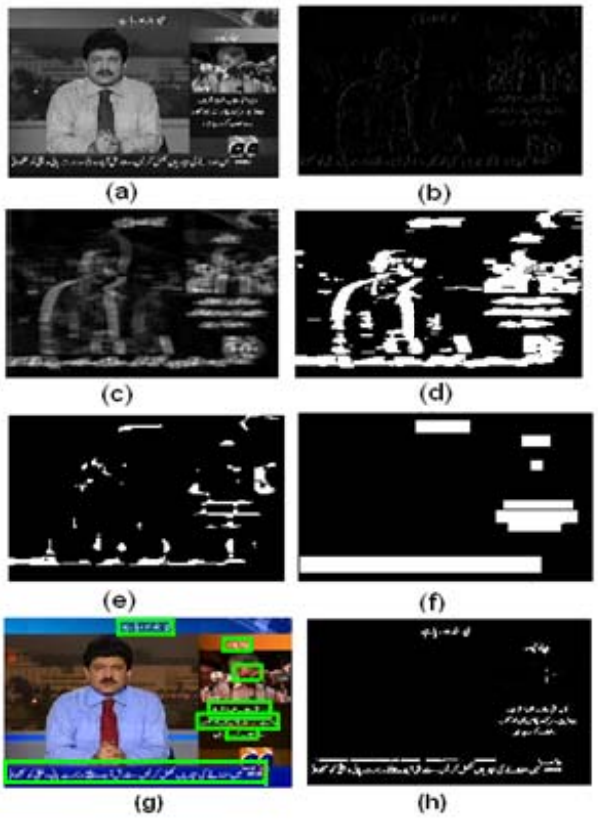


Figure 3. (a) Grayscale image (b) Vertical Gradients (c) Mean gradient image (d) RLSA based Image (e) Noise removal with text pixel density filter (f) Candidate text regions (g) Text detection result on the input video image (h) Extracted text from input image.

Once the text regions in the image are localized, we need to extract the text pixels from the detected rectangles removing the background. This process virtually translates to a two-class segmentation problem or binarization. We evaluated a number of binariztion algorithms including Otsu's global thresholding [14] as well as a number of local binarization algorithms by Sauvola [15], Wolf [10] and Feng [16], all being different variants of the Niblack algorithm [18].

Ideally, these binarization methods should be rated as a function of their performance on an OCR. However, since a mature Urdu OCR is non-existent and recognition of text is not a part of our present research, these methods are subjectively evaluated, Wolf's algorithm [10] turning out to be the most suitable for the images under study. Figure 3 shows the results of each of the steps from original gray scale image to text extraction.

IV. Experimental Results

The application of text detection and localization to Urdu text is its infancy only so naturally benchmark data sets of Urdu video images are not available. We therefore collected a total of 150 Urdu video images from a variety of sources including news channels, sports videos, talk shows, serials and movies.

Figure 4. Text localization and extraction results for video images taken from different sources. (a) Movie credits (b) News tickers (c) Oil prices

Although a number of sophisticated evaluation metrics [19, 20] have been proposed, we have evaluated our present study on the most commonly used area based metric to compute the precision and recall (and their harmonic mean F-measure). If G represents the ground truth text area in an image and D the detected area, the area-based precision and recall are defined as:

$$Recall = \frac{(G \cap D)}{G}$$

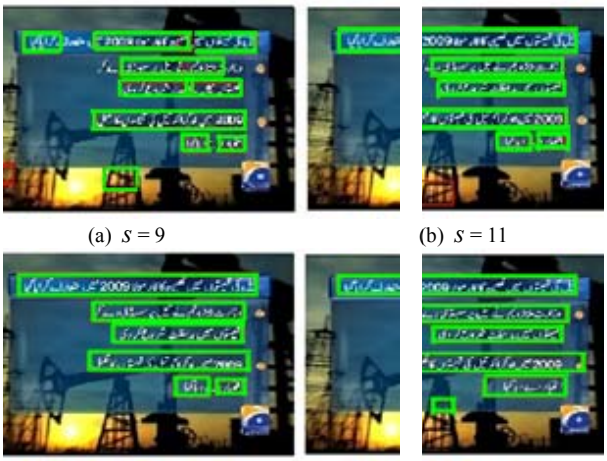
$$Recall = \frac{(G \cap D)}{D}$$

The idea can be extended to multiple images by simply summing up area of intersection and dividing by the total ground truth area (in N images) for recall and the total detected area of precision. Using the proposed method we achieve an overall recall of 81% and a precision of 77% as summarized in Table 1. Text extraction results on few images from different video sources are illustrated in Figure4.

TABLE I. PERFORMANCE OF THE PROPOSED METHOD

Data Set	Precision	Recall	F-measure
150 Images	0.77	0.81	0.79

We also carry out an analysis of the sensitivity of the system performance to different parameters. We first study the effect of the size of averaging window (s) used to average the gradient magnitude. The performance does not show drastic changes with respect the window size as can be seen from Figure 5 where detection results for different sizes of averaging window are illustrated. The overall performance on 150 images is summarized in Figure 6 where a window size of 13 achieved the best results on the data set under study.



(a) $s = 9$ (b) $s = 11$

(c) $s = 13$ (b) $s = 15$

Figure 5. Text detection results on different sizes of averaging window

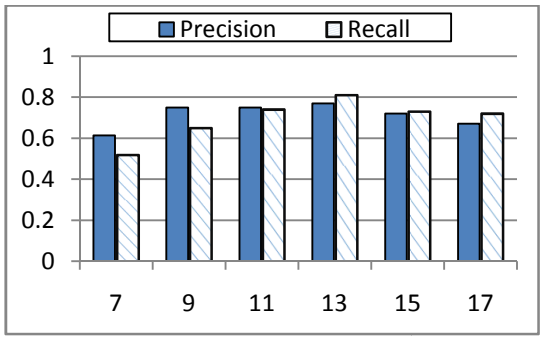


Figure 6. Effect of size of averaging window (s) on precision and recall

We also studied the effect of the size of pixel density filter on performance of the system. Since the filter has a binary output (classifies each window as text or non-text), the performance is relatively more sensitive to its size as compared to that of averaging filter. Figure 7 illustrates as subset of the window sizes evaluated where a size of 10x10 achieved the highest precision and recall.

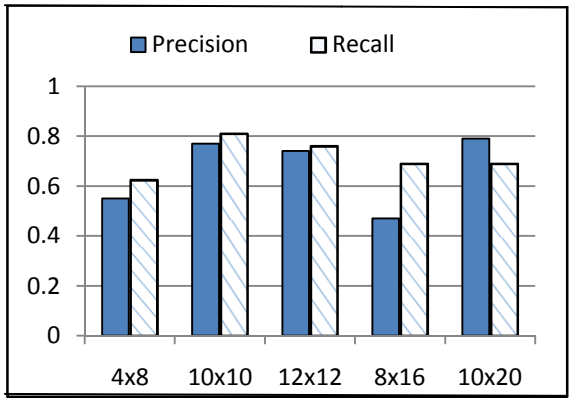


Figure 7. Effect of size of pixel density filter on precision and recall

In general, the proposed method performs well on the variety of images evaluated. It has however certain limitations. Since the method relies heavily on edge features, complex backgrounds containing elements exhibiting 'text-like' properties are also detected as text regions. Isolated words are also hard to detect as they might be filtered out by the pixel density filter. Figure 8 shows some examples of false negatives and false positives.

Figure 8. Examples of false negatives and false positives.

V. CONCLUSION AND DISCUSSION

In this paper, we addressed the problem of extracting Urdu textual content from video images. The proposed system is based on a gradient based approach where the average gradient in the neighborhood of each pixel is computed and horizontally aligned gradients are merged together. A pixel density based filter is then used to distinguish text and non-text regions followed by the application of some geometrical constraints. The localized text regions are finally binarized and are ready to be used in the subsequent steps of indexing and retrieval. The proposed scheme exhibited reasonably good detection rates on a wide variety of Urdu video images. The present system is designed for and evaluated on individual video frames only

not exploiting the temporal redundancy of text. Text appearing in videos is likely to persist for some time and exploiting this redundancy could serve to enhance the precision and recall of the system. We also plan to extend the current system from text extraction to a complete indexing and retrieval application which could be based either on text recognition or word-spotting. Both of these are still unexplored areas on Urdu text.

VI. ACKNOWLEDGMENT

This research work is funded by the Higher Education Commission (HEC) of Pakistan.

REFERENCES

[1] K. Jung, K.I. Kim and A.K. Jain."Text information extraction in mages and video: a survey". Pattern Recognition, 37, 2004.

[2] Marios Anthimopoulos, Basilis Gatos, Ioannis Pratikakis, A Hybrid System for Text Detection in Video Frames, The Eighth IAPR Workshop on Document Analysis Systems, 2008.

[3] Rongrong Wang, Wanjun Jin, Lide Wu. A Novel Video Caption Detection Approach Using Multi Frame Integration. In Proceedings of the 17th International Conference on Pattern Recognition(ICPR'04), pp. 1051-4651,2004.

[4] Guangyi Miao, Qingming Huang, Shuqiang Jiang, Wen Gao. Coarse-to-fine video text detection, 2008.

[5] Jun Ye, Lin-Lin Huang, XiaoLi Hao, Neural newtowork based text detection in videos using local binary Patterns, pattern recognition, 2009.

[6] K. L Kim, K. Jung and J. H. Kim. Texture-Based Approach for Text Detection in Images using Support Vector Machines and Continuous Adaptive Mean Shift Algorithm . IEEE Transactions on Pattern Analysis and MachineIntelligence, Vol. 25, No. 12, December 2003.

[7] Fan, W.; Sun, J.; Katsuyama, Y.; Hotta, Y. & Naoi, S. Text Detection in Images Based on Grayscale Decomposition and Stroke Extraction Proc. Chinese Conf. Pattern Recognition CCPR 2009, 2009.

[8] Palaiahnakote Shivakumara, Trung Quy Phan, and Chew Lim Tan, 'a laplacian approach to multi-oriented text detection in video , IEEE transactions on pattern analysis and machine intelligence, vol. 33, 2009.

[9] Teo Boon Chen D. Ghosh S. Ranganath, video-text extraction and recognition, 2004.

[10] C. Wolf, J-M. Jolion and F. Chassaing, Text Localization, Enhancement and Binarization in Multimedia Documents. In Proceedings of the 16th International Conference on Pattern Recognition ICPR'02, Quebec, Canada, pp. 1037-1040, 2002.

[11] C. Zhu, W. Wang and Q. Ning, Text Detection in Images Using Texture Feature from Strokes. LNCS - Advances in Multimedia Information Processing, pp. 295-30, 2006.

[12] V.Wu, R.Manamatha, and E.Riseman, "Textfinder: an automatic system to detect and recognized text in images", *IEEE Trans. On PAMI*, Vol.20, 1999.

[13] P. Shivakumara, T.Q. Phan and C. L. Tan. A Gradient Difference based Technique for Video Text Detection . 10th International Conference on ocument Analysis and Recognition, 2009.

[14] N.Otsu. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man and Cybernetics, 1979.

[15] J.Sauvola, T,Seppanen, S.Haapakoski, M.Pietikainen, "Adaptive Document Binarization", 4th Int. Conf. On Document Analysis and Recognition, Ulm, Germany, pp.147-152 (1997).

[16] Meng-Ling Feng and Yap-Peng Tan, "Contrast adaptive binarization of low quality document images", IEICE Electron. Express, Vol. 1, No. 16, pp.501-506, (2004).

[17] Wei Fan, W.; Sun, J.; Katsuyama, Y.; Hotta, Y. & Naoi, S. Text Detection in Images Based on Grayscale Decomposition and Stroke Extraction Proc. Chinese Conf. Pattern Recognition CCPR 2009, 2009.

[18] W. Niblack. An introduction to digital image processing, pp. 115-116. Prentice-Hall, Englewood Cliffs (NJ), 1986.

[19] Wolf and J.-M. Jolion. Object count/area graphs for the evaluation of object detection and segmentation algorithm. International Journal on Document Analysis and Recognition (IJDAR), 2006.

[20] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young. Icdar 2003 robust reading competitions: Entries, results and future directions. *International Conference on Document* Analysis and Recognition (ICDAR), 2003.