A Database of Artificial Urdu Text in Video Images with Semi-Automatic Text Line Labeling Scheme

Imran Siddiqi
Department of Applied Science & Graduate Studies
Bahria University
Islamabad, Pakistan

Ahsen Raza
Department of Computer Software Engineering
National University of Sciences & Technology
Islamabad, Pakistan

imran.siddiqi@bahria.edu.pk, ashen.raza@mcs.edu.pk

Abstract—**This paper describes a novel database of video images containing artificial (superimposed) Urdu text with a semi-automatic text line labeling scheme. The main objective of this study is to provide the community with a standard dataset together with an auto-labeling scheme for algorithmic development and evaluation of textual content based indexing and retrieval systems. We have specifically focused on Urdu text which is increasingly gaining research interest in recent years. The data set comprises 1000 video images collected from 19 different channels of 5 different categories. An attempt is made to capture the maximum possible variation in the text in terms of size, location, appearance and background. The data set is completely labeled by finding the bounding rectangle of each text occurrence facilitating the evaluation of text detection and localization systems. Based on our previous work on text localization, an automatic text labeling scheme is also proposed and the obtained results are compared with manual labeling. Ground truth data, supporting tasks like text recognition and word spotting will be considered in the next version of the data set.**

Keywords-Data Set; Artificial Urdu Text; Text Detection; Text Localization.

I. INTRODUCTION

The availability of data sets is one of the fundamental requirements for development and evaluation in any research domain. Over the recent years, standard databases are becoming increasingly popular in all the scientific research fields. The availability of such data sets not only saves researchers the task of compiling and labeling the database but also provides the possibility of objectively comparing different systems on the same data set. This is further complemented by organization of evaluation campaigns [16, 22] allowing comparison of different techniques under the same experimental conditions as well. Like other research areas, the document analysis and recognition community has also developed a number of standard databases addressing different problem areas. The most popular of these are the databases for handwriting recognition like CEDAR [12], NIST [13], CENPARMI [14], IAM [11] and RIMES [16] for offline while IAM-OnDB [18, 19], UNIPEN [23] and IRONOFF [24] for online recognition. In addition to character and word recognition, some of these data sets have also been used in evaluating tasks like document layout analysis, document segmentation and writer identification/verification.

Another significant research area in the document recognition paradigm is the detection, localization and recognition of artificial and scene text appearing in video images. Scene text recognition finds its applications in autonomous navigation and assistance; ICDAR [1] and KAIST [21] being the two widely used data sets in this domain. Artificial text on the other hand is more useful for applications like semantic indexing and retrieval of video archives. An important component of such keyword based retrieval systems is the detection and localization of textual regions [10]. It has attracted a number of researchers over the last decade [1- 4] and is in fact the subject of our study as well. More specifically, we focus on the artificial Urdu textual content in video images which is relatively a young and unexplored research area as opposed to text in other languages.

Urdu, the national language of Pakistan and a major language of India, has speakers allover the world. Analysis of Urdu documents and recognition/processing of Urdu text is attracting research interest in the recent years [5, 6, 15, 17, 20]. As the research in these areas matures, the need to evaluate the proposed techniques on standard data sets will naturally arise. Contributions have already been made towards the development of handwritten Urdu text data sets [7, 8, 9]. However, despite more than 65 Urdu news, entertainment, sports and religious channels around the world, no attempt has yet been made on the development of an artificial Urdu text data set to the best of authors' knowledge.

In this paper, we present the first version of a collection of video images containing artificial Urdu text. The database is mainly targeted towards the evaluation of artificial text detection and localization systems but may also be extended for Urdu word recognition and word spotting systems. The database comprises a total of 1000 video images captured from 19 different Urdu channels. The ground truth text regions in each image are manually extracted allowing quantitative evaluation of any text localization system. A semi-automatic text labeling is also proposed and compared with manual labeling. The main contributions of this work are:

- A completely labeled artificial Urdu text data set.
- A semi-automatic text labeling scheme.

 ISBN: 978-1-61208-195-3

The rest of the paper is organized as follows. In the next section, we discuss data acquisition followed by some characteristics and statistics of the collected data. We then present the manual ground truth labeling followed by the proposed automatic labeling methodology. The results of automatic labeling are then compared with manual labeling. Finally, we give the concluding remarks and discuss some possible future enhancements to the present database.

II. DATA ACQUISITION

Videos from 19 different Urdu channels were captured using Pinnacle Studio Movie board. All videos were recorded at a resolution of 720x576 and stored in 'avi' format. In an attempt to have natural and unconstrained content, multiple videos from the same channel were recorded at different times on a given day. Individual images from videos were then extracted in such a way that there is no repetition of textual content in different images and the maximum variation of text positions, sizes, colors and backgrounds is captured. All images are stored in 'png' format.

The major part of textual content in each image is in Urdu. In some cases however, the images also contain some occurrences of text in other languages (for example, English, Pashto, etc.). These occurrences are inevitable in some of the Urdu channels we have considered. All such occurrences were separately identified and recorded as well.

In the next section, we discuss in detail the different aspects of the data set as well some useful statistics.

III. CHARACTERISTICS AND STATISTICS

The data set comprises a total of 1000 video images extracted from 19 different channels which are grouped into 5 different categories. These categories are chosen to be news, entertainment, sports, business and religious channels. The number of images in each of the categories is summarized in Figure 1. Naturally, the number of news channels and consequently, the number of images in this category is more than any other category due to a large number of Urdu news channels operating around the world. These images are also rich in textual content due the presence of a continuous ticker text. A more detailed distribution of words in images can be found in Figure 2.

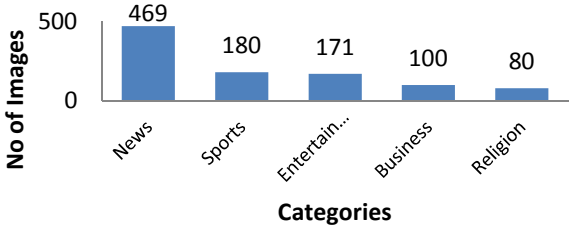


Figure 1. Distribution of images in categories

There are a total of 23833 Urdu words in the collected images. As already discussed, some of the images contain occurrences of English text as well, which make up a total of 5324 English words. A small number (120) of Pashto words also exist in the collected images. In addition to words, the images contain 3339 numerals as well. Table 1 gives an idea of the number of words per image in the data set and some other detailed statistics of the database. On the average, each image contains about 25 Urdu words, 4 words in another language and 4 numerals.

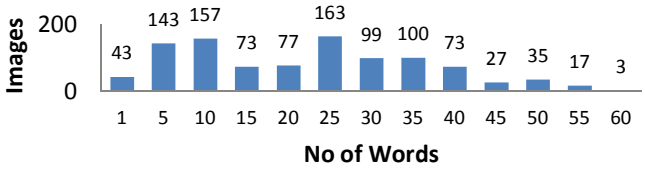


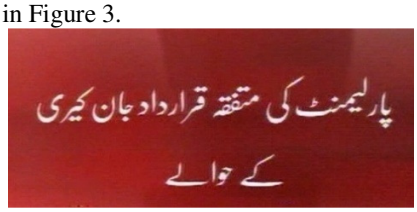
Figure 2. Distribution of words in images.

IV. NAMING & MANUAL LABELING

Once the images are collected, each category as well as each channel is assigned a three digit code. Each image is also given a three digit identification number. These codes are then used to name the images using the convention:

CategoryCode_ChannelCode_ImageID

Some of the images along with example names are illustrated in Figure 3.



(a) 002_003_004

(b)004_016_010

Figure 3. Example images containing artificial Urdu text

For quantitative evaluation of any system using these images, the ground truth data must be labeled. This, naturally, is time consuming, expensive and error prone task [1]. In the first version of the database, we have targeted text localization systems which require the coordinates of all text regions as ground truth data. Labeling of text regions in images is carried out manually using simple software (Figure 4) that allows opening an image and drawing rectangles over the textual content. The x and y-coordinates and width, height of each rectangle are stored in a data file. We have also proposed an automatic labeling scheme the results of which are compared with manual labeling as will be discussed in the subsequent sections.

 ISBN: 978-1-61208-195-3

TABLE I SOME STATISCTICS OF THE DATABASE

Category	Number of Images	Average Urdu words/ image	Second language(s)	Average second language words/ image	Average numeral(s) /image	Total Urdu words	Total second language words	Total numerals
News	469	17	Pashto	5	2	7860	120	1000
			English				2430	
Sports	180	30	English	6	3.5	5450	1070	615
Entertainment	181	26	English	5	3.5	4650	960	635
Business	100	31	English	5	6	3076	532	632
Religion	70	40	English	3	6.5	2797	212	457
Overall	**1000**	**23.8**	**-**	**5.3**	**3.3**	**23833**	**5324**	**3339**

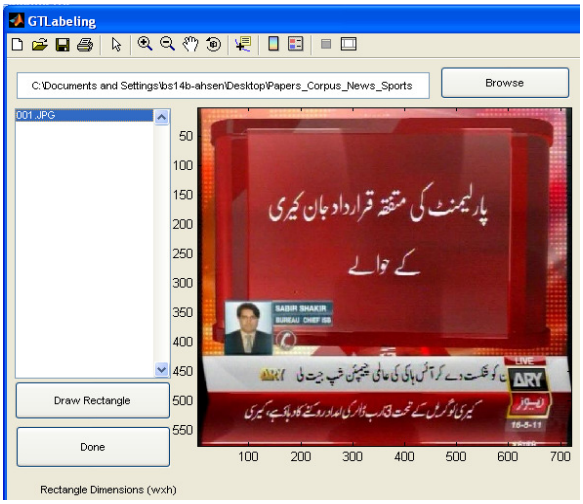


Figure 4. The user interface of the ground truth labeling software

An important factor in labeling is when to start a new rectangle. This will consequently affect the evaluation performance of tested systems depending upon the metric used. We investigated the ICDAR labeling methodology [1] but it cannot be replicated for Urdu text due to different characteristics of the script, for example, non uniform alignment within the same line of text (Figure 5). We therefore devised a labeling methodology that is based on the following heuristics:

- A single rectangle is drawn over a line of text that belongs to the same semantic unit (for example a sentence), when the words in the line are horizontally algined, are of the same size and, do not have a significant inter-word distance (Figure 5a).
- If different blocks of text in the same line have non-uniform size/alignment, they are split into different rectangles so that minimum background becomes part of the rectangles (Figure 5b and Figure 5c).
- In case of overlapping words, separate (overlapping) rectangles are drawn for each of the words (Figure 5d).

Naturally, these heuristics are subjective and the definitions of terms like 'alignment' and 'size difference' may vary from one individual to another. This however is an inherent limitation with such manual labeling. A solution could be to shift from block level to pixel level where each individual pixel is identified as being text or non-text. This, however, is an extremely time consuming job and is not considered in the present version of the database.

Figure 5. Sample labeled images showing the labeling methodology

The ground truth data for each image is stored in the accompanying data file. The data for the entire set of images is also stored in a single file. A part of the the ground truth data file is illustrated in Figure 6. Each line in the data file contains the image name, the language of textual content and the coordinates of the bounding rectangle. The complete data set along with ground truth data is publically available for download [35].

```
#--- Groud Truth Data---#
#
# Urdu Artificial Text Data Set
#
# format: 001_003_100 Urdu 11 435 372 43
#
#      001   -> Category Code
#      003   -> Channel Code
#      100   -> Image ID
#      Urdu  -> Language of textual content
#      11    -> x-coordinate of bounding rectangle
#      435   -> y-coordinate of bounding rectangle
#      372   -> width of bounding rectangle
#      43    -> height of bounding rectnagle
#

001_003_100 Urdu 11 435 372 43
001_003_100 Urdu 412 441 54 23
001_003_100 Urdu 477 436 122 39
001_003_100 Urdu 120 490 401 49
```

Figure 6. A snapshot of ground truth data file.

V. AUTOMATIC TEXT LABELING

In this section we present the proposed text localization scheme for automatic labeling of text regions in the database. Naturally, such automated techniques do have inherent problems with the accuracy and precision of the labeled regions and human assistance is required to correct these labeling errors. In our case, the automatically labeled regions are also compared with the manually labeled regions as will be discussed shortly.

The labeling scheme is primarily based on a series of image processing operations. The complete flow of the proposed scheme is shown in Figure 7. It is to be noted that the labeling algorithm operates on a single image and does not use any temporal features (e.g., redundancy of textual content in the video). This allows localization of textual occurrences on individual frames as well where the complete video is not available.

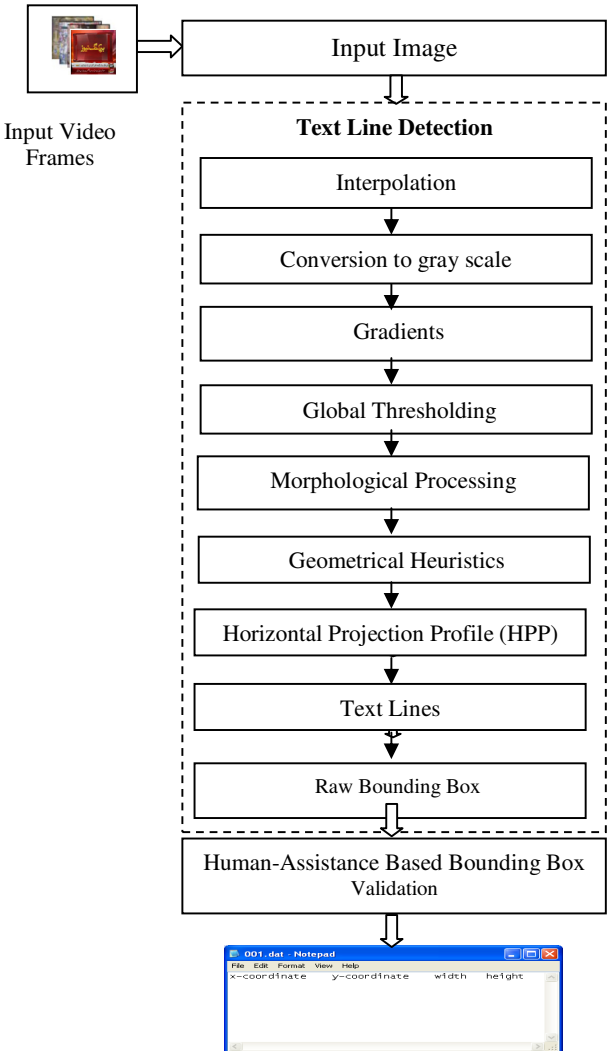


Figure 7. A general flow of text line labeling scheme.

In comparison to English, Urdu lexicon detection is much more challenging with main difficulties being the different shapes of alphabets depending upon their position within the words, the high frequency of diacritics, non-uniform inter and intra word distances and the occurrence of strokes in all directions. These factors make the detection of Urdu text more difficult as many non-text regions may also posses these text-like characteristics.

Our automatic labeling methodology is inspired from [2] with modifications for Urdu text and is mainly based on our previous work on text localization the details of which can be found in [31]. Similar methods have been used for detection of Farsi and Arabic texts [32-34] as well, which are quite similar in nature to Urdu text.

For localization of textual Urdu content in an image, as a first step, the image is converted to gray scale so that further processing is independent of the color information of image. Then, we conditionally resize the image using bi-cubic interpolation to an experimentally known size of 720x576. This gives a smooth estimate of the gray level at any desired point in the image [25]. Since the text is supposed to be readable on the screen, there is a high contrast between the textual content and its background which can be exploited to extract the boundaries of the text regions. We evaluated a number of standard edge detection filters, and finally, chose the Sobel filter for boundary detection as it preserves most of the edges and gives a strong boundary lining for isolated words.

In order to separate the text boundaries from the background, we next binarize the gradient image. This is one of the most critical steps as the subsequent steps are very sensitive to the binarization threshold. We experimenetd with a number of local [26, 27, 28] as well as global [29] thresholding algorithms and finally employed Otsu's thresholding [29] to binarize the gradient image.

As a result of binarization, most of the background is suppressed and the likely text boundries appear as connected components in the proximity of one another. These isolated components need to be merged together into words and ultimately text lines. This is implemented using the standard morphological operations of dilation and erosion. Dilation is carried out to merge all horizontally aligned components together which effectively is the merging of loosley connected characters into words. Dilation is followed by erosion which eliminates the falsely merged components. As a final step we employ the traditionally used geometrical constraints on the identified textual regions 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.

Since the labeling is done at line level, the text lines are extracted from the identified textual regions using the well-known horizontal projection profiles [30]. Naturally the localized text lines are not always accurate/precise and need human intervention for validation.

The auto-labeled image is presented to the user with rectangles on potential text regions and the possibility the resize, move, add or delete the text rectangles. Once

 ISBN: 978-1-61208-195-3

validated, the coordinates of each text rectangle in the image are saved to a file. This semi-automatic labeling greatly reduces the effort involved as compared to a total manual labeling. The results of the proposed labeling are also very promising as discussed later in the paper. The detected textual regions can also be used for content based image retrieval (CBIR) applications [2]. The steps involved in labeling are illustrated on an example image in Figure 8. These steps are similar to those as in [2, 31] with adjustment of parameters for Urdu text.

Figure 8. Various steps of proposed labeling scheme. (a) Original image.(b) Grayscale image (c) Sobel filter (d) Binarized gradients (e) Morphological processing (f) Geometrical constraints (g) Detected text lines (h) Manual validation/correction (i) Ground truth data saved to file.

VI. Evaluation Metrics

The performance of a text localization system is traditionally quantified using the precision and recall. The problem however is that the text rectangles detected by a given system will not have a 1-1 correspondence with the text rectangles in the ground truth data. In addition, the size of these rectangles will also vary. To handle these issues, area based definitions of precision and recall are generally used. If G represents the ground truth text area in an image and D the detected area, we have:

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

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

More sophisticated metrics have also been proposed in the literature. For example, the ICDAR scene text database [1] defines a metric that searches for the true match of a detected rectangle in the set of ground truth rectangles. Wolf and Jolion [20] improved the ICDAR measure by introducing a novel performance measure that takes into account one-to-one, one-to-many (splits) and many-to-one (merges) scenarios as well.

All these metrics are equally applicable in case of Urdu text as well. Since the only information required in these metrics is the coordinates of the bounding rectangles, they can be easily calculated on the developed data set.

VII. Experimental Results

To evaluate the effectiveness of the proposed automatic labeling, we compared the results of auto-labeling (without human intervention) with manual labeling. On the data set of 1000 images, we achieved an overall precision of 71% and recall of 80% as summarized in Table II. We also studied how the performance of the localizer varies with the size of the image. These results are presented in Figure 9 and indicate that the performance is not very sensitive to the resolution of the image. The errors in terms of false positives, false negatives and misplaced rectangles can then be corrected by human intervention which naturally is much efficient as opposed to a complete manual labeling. Some results of the labeling scheme on a variety of backgrounds are presented in Figure 10.

TABLE II Performance of the Proposed Method

Data Set	Precision	Recall	F-measure
1000 Images	0.71	0.80	0.75

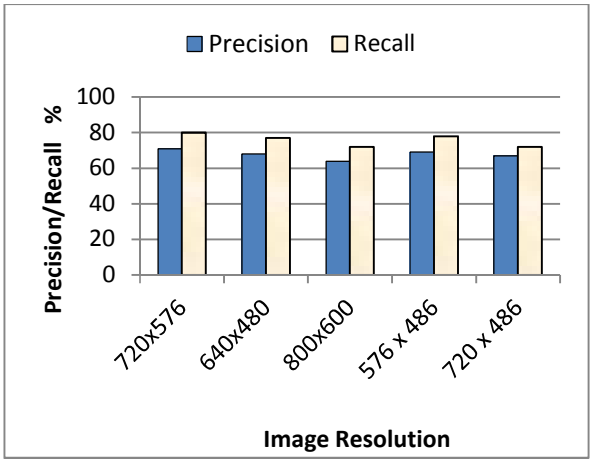


Figure 9. Performance of the proposed method on different image resolutions

VIII. Conclusion and Perspectives

In this paper, we presented the first version of a novel data set for Urdu artificial text along with a semi-auto text labeling scheme. This first version of the database is specifically targeted towards the evaluation of Urdu text localization systems. The ground truth data for each textual occurrence is saved to a file and can easily be used for evaluating such systems using any of the standard metrics. The present ground truth data is based on manual labeling

 ISBN: 978-1-61208-195-3

but we intend to use the proposed labeling scheme with human assistance to generate the ground truth data in the next version of the dataset.

We also plan to include the actual transcription of text in the next version, which will also allow the evaluation of Urdu text recognition and word spotting systems. The size of the data is also likely to double in the next version with additional channels in each of the categories. A similar dataset with text in languages based on the Latin alphabet is also under development finally leading to a huge collection of video images with unconstrained multilingual text. The authors expect that these data sets will prove to be useful for the document recognition community.

Figure 10. Auto-text labeling results (without manual validation) on a variety of images present in the database.

REFERENCES

[1] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 robust reading Competitions". Proc. 7th International Conference on Document Analysis and Recognition (ICDAR'03), 2003.

[2] C. Wolf and J. Jolion, "Extraction and recognition of artificial text in multimedia documents", Pattern Anal. Appl., 2004, pp.309-326.

[3] K. C. Kim, H. R. Byun, Y. J. Song, Y. M. Choi, S. Y. Chi, K. K. Kim, and Y. K. Chung, "Scene text extraction in natural scene images using hierarchical feature combining and verification," Proc. 17th International Conference on Pattern Recognition, vol. 2, pp. 679–682, 2004.

[4] M. Abdel-Mottaleb, N. Dimitrova, R. Desai, and J. Martino, "CONIVAS: Content-based image and video access system" ,Proc. of ACM Multimedia, pp 427-428, Boston 1996.

[5] M. Humayoun, "Urdu morphology, orthography and lexicon extraction". Masters Thesis, Department of Computing Science, Chalmers University of Technology, 2006.

[6] N. Durrani and S. Hussain, "Urdu word segmentation", Proc. 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, June 2010.

[7] M. W. Sagheer, C. L. He, N. Nobile, and C. Y. Suen "A new large Urdu database for off-line handwriting recognition", Proc. ICIAP 2009, LNCS 5716, pp. 538–546, 2009.

[8] D. Becker and K. Riaz, "A study in Urdu corpus construction", Proc. of the 3rd Workshop on Asian Language Resources and International Standardization,Taipei, Taiwan. 2002.

[9] M. Ijaz and S. Hussain, "Corpus based Urdu lexicon development". Proc. Conference on Language and Technology, Peshawar, Pakistan, 2007.

[10] C. Lui, C.Wang, and R. Dai. "Text detection in images based on unsupervised classification of edge based features", Proc. International Conference on Document Analysis and Recognition (ICDAR 2005), pp 610–614, 2005.

[11] U.-V. Marti and H. Bunke. "A full english sentence database for off-line handwriting recognition", Proc. International Conference on Document Analysis and Recognition (ICDAR'99)", pp 705–708, 1999.

[12] J. Hull, "A database for handwritten text recognition research"; IEEE Trans. on Pattern Analysis and Machine In- telligence, 16(5):550–554, May 1994.

[13] R. Wilkinson, J. Geist, S. Janet, P. Grother, C. Burges, R. Creecy, B. Hammond, J. Hull, N. Larsen, T. Vogl, and C. Wilson, "The first census optical character recognition systems", Proc. Conf. #NISTIR 4912, The U.S. Bureau of Cen- sus and the National Institute of Standards and Technology, Gaithersburg, MD, 1992.

[14] C. Suen, C. Nadal, R. Legault, T. Mai, and L. Lam, "Computer recognition of unconstrained handwritten numerals", Proc. of the IEEE, 7(80):1162–1180, 1992.

[15] M. W. Sagheer, N. Nobile, C. L. Hev, and C. Y. Suen, "A novel handwritten Urdu word spotting based on connected components" Proc. International Conference on Pattern Recognition(ICPR) ,2010.

[16] E. Augustin, M. Carré, E. Grosicki, J. M. Brodin, E. Geoffrois, and F. Preteux, "Rimes evaluation campaign for handwritten mail processing", Proc. of the Workshop on Frontiers in Handwriting Recognition, pp. 231–235, 2006.

[17] U. Pal and A. Sarkar, "Recognition of printed Urdu script", Proc. of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03), 2003.

[18] M. Liwicki and H. Bunke, "IAM-OnDB - An on-line English sentence database acquired from handwritten text on a whiteboard", Proc. of the Eighth International Conference on Document Analysis and Recognition, 2005.

[19] E. Indermühle, M. Liwicki, and H. Bunke, "IAMonDo database: an online handwritten document database with non-uniform contents", Proc. of the 9th IAPR International Workshop on Document Analysis Systems, 2010.

[20] C. Wolf and J.-M. Jolion, "Object count/area graphs for the evaluation of object detection and seg-mentation algorithm", International Journal on Document Analysis and Recognition (IJDAR), 8(4) 280–296, 2006.

[21] http://ai.kaist.ac.kr/home/DB/SceneText, 12-02-2012.

[22] D. Pallett, "A look at NIST's benchmark ASR tests: past, present, and future", Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, 2003.

[23] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, and S. Janet, "Unipen project of on-line data exchange and recognizer benchmarks", Proc. of the 12th International Conference on Pattern Recognition, 1994.

[24] C. Viard-Gaudin, P. M. Lallican, P. Binter, and S. Knerr, "The ireste on/off (ironoff) dual handwriting database", Proc. of the Fifth International Conference on Document Analysis and Recognition, 1999.

[25] C. R. Gonzalez and R. E. Woods, Digital Image Processing. (2nd edition), 2001.

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

 ISBN: 978-1-61208-195-3

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

[28] C. Wolf, J-M. Jolion, and F. Chassaing, "Text localization, enhancement and binarization in multimedia documents", Proc. of the 16th International Conference on Pattern Recognition ICPR'02, Quebec, Canada, pp. 1037-1040, 2002.

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

[30] M. Ben Halima, H. Karray, and A. M. Alimi, "A comprehensive method for Arabic video text detection, localization, extraction and recognition", PCM 2010, Part II, LNCS 6298, pp. 648–659, 2010.

[31] A. Jamil, I. Siddiqi, F. Arif, and A. Raza, "Edge-based features for localization of artifiical Urdu text in video images", Proc. of the 11th Int'l Conference on Document Analysis and Recognition (ICDAR 2011), China, pp 1120-1224, 2011.

[32] M. Moradi, S. Mozaffari, and A.A. Orouj, "Farsi/Arabic text extraction from video images by corner detection", Proc. 6th Iranian Machine Vision and Image Processing Conference, Oct 2010.

[33] M. Ben Halima, H. Karray, and A.M. Alimi, "Arabic text recognition in video sequences", Proc. International Conference on Informatics, Cybernetics, and Computer Applications, July 2010.

[34] M. Ben Halima, H. Karray, and A.M. Alimi, "AVOCR: Arabic video OCR", Proc. 5th International Symposium on I/V Communications and Mobile Network (ISVC), Sept 2010.

[35] https://sites.google.com/site/artificialtextdataset/, 12-02-2012.

 ISBN: 978-1-61208-195-3