Benchmarking recognition results on camera captured word image data sets

Deepak Kumar*
MILE Laboratory
Dept. of EE
Indian Institute of Science
Bangalore, INDIA
deepak@ee.iisc.ernet.in

M N Anil Prasad
MILE Laboratory
Dept. of EE
Indian Institute of Science
Bangalore, INDIA
anilprasadmn@ee.iisc.ernet.in

A G Ramakrishnan
MILE Laboratory
Dept. of EE
Indian Institute of Science
Bangalore, INDIA
ramkiag@ee.iisc.ernet.in

ABSTRACT

We have benchmarked the maximum obtainable recognition accuracy on five publicly available standard word image data sets using semi-automated segmentation and a commercial OCR. These images have been cropped from camera captured scene images, born digital images (BDI) and street view images. Using the Matlab based tool developed by us, we have annotated at the pixel level more than 3600 word images from the five data sets. The word images binarized by the tool, as well as by our own midline analysis and propagation of segmentation (MAPS) algorithm are recognized using the trial version of Nuance Omnipage OCR and these two results are compared with the best reported in the literature. The benchmark word recognition rates obtained on ICDAR 2003, Sign evaluation, Street view, Born-digital and ICDAR 2011 data sets are 83.9%, 89.3%, 79.6%, 88.5% and 86.7%, respectively. The results obtained from MAPS binarized word images without the use of any lexicon are 64.5% and 71.7% fo r ICDAR 2003 and 2011 respectively, and these values are higher than the best reported values in the literature of 61.1% and 41.2%, respectively. MAPS results of 82.8% for BDI 2011 dataset matches the performance of the state of the art method based on power law transform.

Keywords

Word images, Pixel level annotation, MAST-CH toolkit, Cropped word recognition, Benchmarking, Scene images, Born-digital images, sign evaluation.

1. INTRODUCTION

When a camera captured image is presented to an OCR engine, the recognition performance is not necessarily very good. This led to splitting the process of word recognition in camera captured images into two parts, namely localization (or detection) and recognition. In International Conference on Document Analysis and Recognition (ICDAR) 2003, Lucas et. al [4] organized separate competitions for text localization on camera captured images and recognition from the word images cropped by placing a bounding box on the image. They received five entries for text localization and none for word recognition. In the following ICDAR 2005 conference, text localization was the main theme and word recognition was skipped [5].

There are several publicly available data sets for text localization [21]. These are known as IAPR TC11 Reading Systems-Data sets. One may assume that the bounding box information of a word is sufficient for any OCR to recognize. However, we see that ICDAR 2011 Robust reading challenge 2 reports that the best word recognition rate is 41.2% [11]. Figure 1 shows sample word images from this challenge. Even though Mishra et. al [9] report an accuracy of 52%, this performance is only on images from the ICDAR 2003 sample dataset and the authors did not perform any experiments on the actual test set.

Karatzas et. al initiated another robust reading challenge in ICDAR 2011 for born-digital images [10]. Born-digital images are formed by a software by overlaying text on an image. For the competition, these images were collected from web pages and email. Most words present in this data set are oriented horizontally. The reason behind horizontal placement of text may be the simplicity involved in creating the born-digital image using standard softwares.

Low resolution of text and anti-aliasing are the main issues to be tackled in born-digital images, whereas illumination changes and motion blur are difficult problems in the case of camera captured images. These issues indicate the different kinds of complexities involved in processing born-digital and

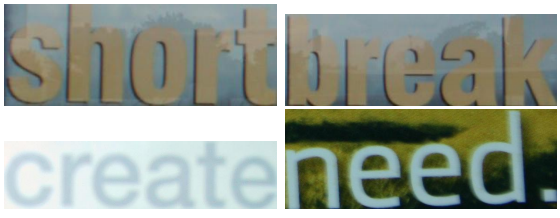

Figure 1: Sample camera captured word images from ICDAR 2011 dataset [11].

*Corresponding author

DAR '12, December 16, 2012, Mumbai, India

camera captured scene images.

An attempt for using top-down approach for word recognition can be observed in the method of sparse belief propagation with lexicon by Weinmann et. al [7]. Similarly, Wang and Belongie use synthetic, custom lexicons on Street view text (SVT) data set [8]. Both Weinmann et. al and Wang et. al use unsegmented character data to train a classifier. When the confidence of the character classifier is less, the top-down approach helps in classification using lexicon. Weinmann et. al use character level image annotation of the training data and textual features to classify the testing dataset. A limitation of this method is that it requires good quality character images with high resolution for training; else the classification may be erroneous.

Wang and Belongie used Amazon's Mechanical Turk for annotation of SVT images. Bounding boxes were placed around the words spotted. The placement of these bounding boxes was not defined rigorously. The resulting irregularities in the word bounding boxes add additional complexity to the segmentation task, which can be inferred from the low f-score reported in the literature.

2. SEMI-AUTOMATED BINARIZATION

Benchmarking is not a good idea, if ground truth is not explicitly defined rigorously. The five data sets have different definitions for bounding box and contain human errors while inserting bounding boxes for cropping the words. All the databases provide text level annotation as ground truth. We have created pixel level annotation for the word images from five different, publicly available, standard data sets. To our knowledge, annotation at the pixel level and on several data sets has not been carried out until now. Small subsets have been annotated and utilized for experiments [9, 13, 14] from different data sets. We have annotated 3606 word images at the pixel level. Annotation has been carried out by a semi-automated process with assured quality. The huge task of thoroughly cross-checking the pixel level annotation has also been accomplished to reduce human errors to a minimum.

We split the task of recognition of the words from the word images into segmentation and recognition tasks. We require sophisticated algorithms to segment an image. In document imaging community, conventional research primarily focused on digitization of scanned documents. In the section on annotation, we discuss known algorithms for segmentation of word images, all of which have been implemented to facilitate the semi-automated annotation process. These algorithms contributed in improving the speed of pixel level annotation.

Annotated pixel level word images can be used to train and test any classifier. However, several good optical character recognition (OCR) engines are already available for Roman script [18, 19, 24, 25]. Hence, we use the trial version of Omnipage Professional 16 OCR [24] for recognition of characters in the binarized images to create the benchmark recognition result. Hence, our interest in the annotation algorithm and annotating data sets is meant to mainly cater to research in improved segmentation of characters.

If a single data set is used in the experiments, it may lead to a data specific approach. So, to confirm that our approach to semi-automated binarization is data set independent, we cover five different data sets for annotation. These data sets are from ICDAR 2003 competition [4], ICDAR 2011 competition [10, 11], Street view [8], born-digital and Sign evaluation data sets [7].

3. MAST-CH ANNOTATION TOOL

A multi-script annotation toolkit for scene text (MAST) was developed by MILE lab in 2011 [12]. MAST is designed to annotate scene images with multiple word images with reasonably good resolution. Using seed points input by the user, the tool uses region growing and annotates at the pixel level, with a bounding box and text annotation. It can currently handle ten different Indic scripts and has the capability for adding plug-ins with suitable layout for new scripts.

Some of the data sets used contain low resolution word images, which cannot be handled by MAST. Further, for all the word image data sets studied in this work, the text in the word image is already available as ground truth. Hence, MAST-CH, an enhanced version of character annotation tool kit recently developed by us, does not have provision to generate text level annotation for different scripts. It handles a single word image at a time and annotates characters at the pixel level using multiple automated segmentation methods, followed by user selection of the best segmented output.

We have added new functionalities based on feedback from MILE lab project staff, who helped annotate the various data sets. A GUI of the tool kit with the buttons and a single window for image is shown in Figure 2. 'LOAD' button enables us to load images from a particular directory. 'NEXT' and 'PREV' buttons provide options to the user for skipping and going back and forth during annotation, which help in rapid annotation of clean word images. Heavily degraded word images, which require more time to annotate, can thus be skipped. Those skipped images may or may not be tagged later. 'SAVE' button saves an annotated word image in .bmp format, also containing information on component order and in .tiff format, containing colour map for individual components segmented. GUI also displays whether the currently loaded word image has already been tagged or not. 'VIEW MASK' button overlays the obtained segmentation mask on the original word image for verification of the correctn ess of segmentation.

3.1 Enhancing speed of annotation

MAST segments words by region growing on the seeds placed by the user. Region growing fails with very low resolution characters. Thus, to reduce the manual task and also to speed up segmentation time, the seed growing option has been removed in MAST-CH, and replaced by the use of known segmentation algorithms.

For segmentation, we have provided a drop down button giving 'BINARIZE' and 'INVERT' options. A user can invoke the suitable option based on the relative colors of the foreground and background. Using multiple approaches, we create 16 different segmentation outputs. First, the original RGB image is converted to HSV and CIE $L^*a^*b^*$ formats. Then, each of the images is split into the three individual planes and Otsu's threshold [1] is applied individually on the resulting nine planes, which are essentially gray level images. In addition, we form three clusters using the RGB information directly and obtain the three permutations of the clusters formed (each of the 3 clusters and the union of the other two clusters at a time). Finally, we apply Kittler's robust automatic threshold selection algorithm on the intensity image of the word [2].

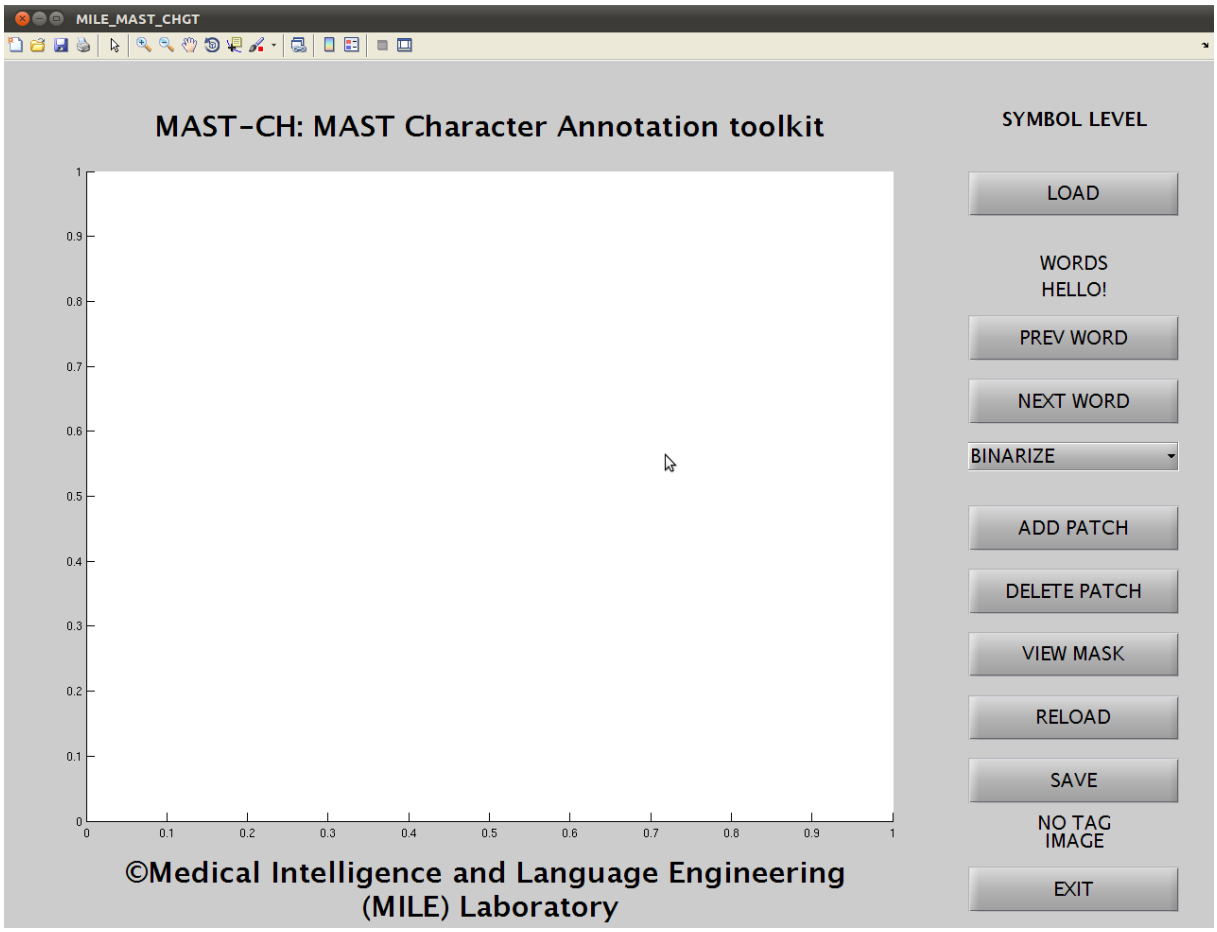

Figure 2: Graphical user interface of the MAST-CH annotation tool developed on MATLAB platform.

We display all of these segmentation results in another window and provide a manual keyboard input for the user to select one of the results. Figure 3 shows the sixteen different segmentation results displayed to the user for a sample image from SVT dataset. The screen shot also shows a text box with provision for a user entry through the keyboard to choose the best segmented output for either fully automated annotation or for subsequent manual correction. As seen in Fig. 3, the optimal choice is '2', since it has minimal requirement for manual editing. Once the user makes a choice, a mask is generated and overlaid on the original image. This semi-automated technique has improved the speed of segmentation and reduced the fatigue of the annotators. If the mask generated has distinct or well separated characters, then the user can save the annotated result by clicking the 'SAVE' button. If none of the segmentation results are satisfactory, the user can choose '0' and thus no mask will be ge nerated.

'RELOAD' button is used to load a saved mask and the corresponding original image. This is useful to examine the already annotated images. To minimize human errors, we have cross-checked the annotated word images thrice by different annotators.

3.2 Use of polygons to refine segmentation

A degraded image may not get segmented properly. This may be due to illumination changes, occlusion or low resolution of characters. For such cases, we have enabled manual editing of the segmentation by providing polygonal masks. These masks can be used to add parts of characters which are merged to the background or delete parts of the background that get added to a character. 'ADD PATCH' button provides the option for adding pixels in the polygonal format to the annotated mask. 'DELETE PATCH' button facilitates deletion of the background segmented as characters or splitting merged characters. When add or delete option is selected, we can place a single polygon at a given time. Mask will be modified based on the operation performed and the shape and position of the polygon. The annotation tool then asks whether the same operation needs to be continued. If the user chooses 'yes', then the user can place another polygon to modify the annotated characters. If the choice is 'no', then the tool exits this edit loop.

Figure 4 illustrates the application of polygonal masks to refine the best segmentation result chosen by the user. Two sample images are shown, with the automated segmentation outputs and the manually refined outputs, obtained after editing by invoking polygonal masks. In the case of the top word image in Fig. 4, a part of the background in the top right corner is marked by the tool as foreground (also seen in Fig. 3). This patch of wrong binarization is selected using a polygonal mask and removed in a single step. In the

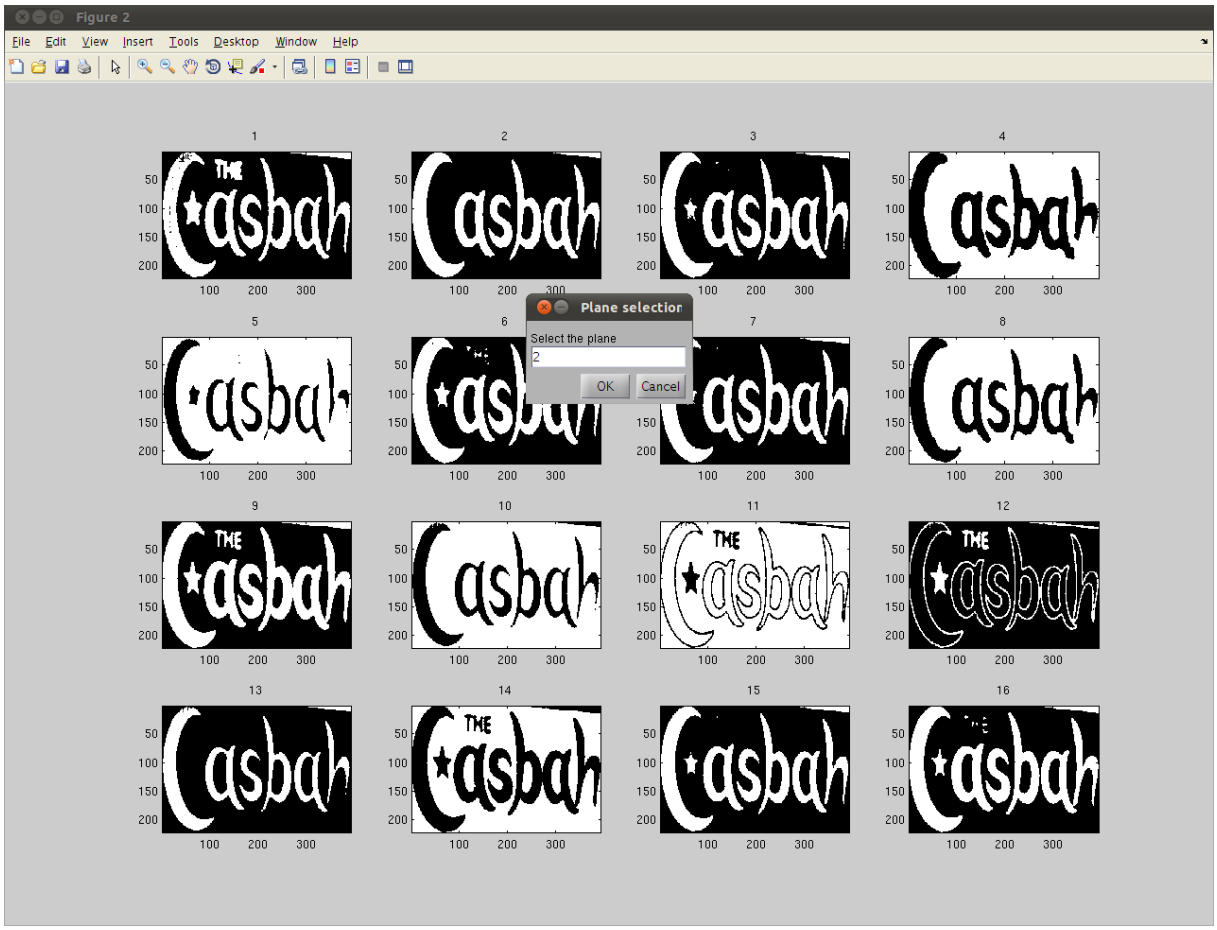

Figure 3: Sixteen distinct automated segmentation results for a sample image from SVT dataset. User selects the best result using a keyboard input (shown in the middle).

Figure 4: Use of polygons to refine the best automated segmentation results. (a) Two sample images from SVT dataset. (b) Best automated segmentation result chosen by the segmenter. (c) Segmented images after refinement by deletion and/or inclusion of appropriate regions defined using polygons.

bottom word image, multiple polygonal masks are used to remove and add patches as required to arrive at the proper segmentation result.

4. POSTPROCESSING FOR ENHANCED FOREGROUND-BACKGROUND DISCRIMINATION

Normally, any scanned document image contains top, left, right and bottom margins, which help the OCR software to unambiguously distinguish foreground (text) from the background gray values. However, as shown in Fig. 5(a), when we binarize a scene or a born-digital word image, margins do not exist since we have segmented at the word boundary.

Figure 5: Binarized image (a) before and (b) after postprecossing by background padding.

In such cases, where characters touch the boundary of the image, we observe difficulty in recognition with the standard OCR engines. To overcome this issue, we ensure margins in all the directions, by adding zero rows at the top and the bottom of the image, equal to half the original number of rows in the word image. Similarly, we pad zero columns on both the left and right sides of the word image. We refer to these images as *postprocessed* binarized images [see Fig. 5(b)]. Postprocessed binarized images are sent to the OCR for recognition. Recognition rates on binarized images are reported in the next section.

5. BENCHMARK RESULTS ON DIFFERENT DATA SETS

We have considered five word image data sets for experimentation. Only the test images from ICDAR, SVT, PAMI and Born-digital data sets have been annotated using the annotation tool described in Sec. 3. Images with visually distinguishable boundaries between characters and background are tagged. Others have been ignored, since if a human being cannot tag the text, we cannot expect an algorithm to either segment or recognize it.

These data sets cover different types of degradations. Each word in the data set has been appropriately tagged in such a way that there is minimal visual distortion with respect to the respective original image. In all the data sets, we have considered only the testing set for the segmentation and recognition experiments. We may possibly be able to improve character segmentation using word images from the training set, but we have not attempted it in this current exercise.

Table 1 compares the word recognition rates (WRR) and, where appropriate, edit distance measures (EDM) also, of images segmented by MAST-CH and MAPS algorithm with the best results in the literature and baseline - for the five publicly available word image data sets. In the following subsections, we separately discuss the recognition results for each of the five different data sets experimented upon. The reported benchmark results and those of MAPS algorithm [17] have been obtained using Nuance Omnipage OCR on the pixel level segmented word images. Definitely, the numbers will slightly vary if we use any other standard OCR and hence the benchmark results we report here indicate a rough level of recognition that can be achieved, rather than the exact maximum value attainable in current circumstances. The baseline results reported have been obtained by giving the raw, coloured word images to Omnipage OCR without any preprocessing or binarization. Further, some of the results co mpared from the literature have been obtained using custom lexicons, synthetically created from the ground truth. Since availability of such lexicons is impossible in a real life recognition scenario, the results reported on MAPS algorithm and MAST-CH segmented images have been obtained without the use of any such custom lexicons.

5.1 ICDAR 2003 data set

Robust reading competition was first conducted in ICDAR 2003 [4]. There were five entries for text localization and none for word recognition. Mishra et. al [9] express the importance of binarization for word images and show 52% recognition on only the sample dataset, but not on the test set. This result explains that an equal importance should be given to word recognition. ICDAR 2003 Test data set consists of 1110 word images, all of which have been segmented by the authors. The word recognition rates are tabulated in Table 1.

Table 1 shows a large gap in recognition rate between the postprocessed and non-processed MAST-CH segmented images. This is because the low resolution text images are not recognized properly without proper margins formed by the background. The result reported by Mishra et. al [16] is based on only 829 images, a subset of ICDAR 2003 test dataset. Hence, the reported result is scaled to the total number of images in the test set for comparison in Table 1. Further, the net performance of 61.1% has been obtained using a synthetic custom lexicon derived from the ground truth of the test set, whereas the performance of 64.5% by MAPS algorithm and the benchmark result of 83.9% have been obtained without the use of any such lexicon.

5.2 PAMI 2009 data set

The next two benchmarked data sets were originally used to demonstrate the ability of top-down approach for character and/or word recognition. A lexicon provides information from the top layer to the middle layer during the classification stage. Using N-gram statistics derived from the limited lexicon formed from the test data sets themselves, the authors show improvement in word recognition rate.

Sign evaluation dataset was prepared by Weinmann et. al [7]. This dataset of 215 word images consists of horizontally aligned characters only, except for one or two. The degradation in the images is also minimal. Hence the recognition rates reported by different methods are all close. In Table 1, the benchmark recognition rate of 89.3% on this dataset is the highest among all the data sets. As compared to the best result of 86.1% obtained using the limited lexicon, MAPS segmentation achieves a reasonable performance of 80%, without the use of the limited lexicon.

5.3 SVT 2010 data set

Wang and Belongie introduced street view text (SVT) data set obtained as part of Google Street View project [8]. The SVT test data set consists of 647 labeled words from businesses around where the images were obtained. It also provides a synthetic lexicon created out of the ground truth, in which each word image has an associated custom dictionary of 50 words including the actual word. Apart from the other degradations, these word images undergo motion blur. Severe motion blur and low resolution created difficulty for the authors in annotating a few word images (around 2%).

This data set consists only of name and location information of businesses. The bounding box tagged by Amazon's Mechanical Turk is not perfect. A rough bounding box is placed around the spotted word, which was listed by the Google search engine. This imprecise bounding box itself provides another layer of difficulty for locating the presence of text within the annotated bounding box. The erroneous tagging of bounding boxes has led to lower recognition rate using both open source and proprietary OCR engines.

Mishra et. al show a high recognition rate of 73.3% by top-down approach [16]. However, this result is based on the custom synthetic lexicon, built from the test word ground truth. With a limited lexicon, one can hit upon the proper word more easily than with full lexicon, as discussed by Weinmann et. al [7]. Mishra et. al use bigram probabilities

Table 1: Comparison of word recognition rates (WRR) and edit distance measures (EDM) of images segmented by MAST-CH and MAPS algorithm with the best results in the literature and baseline - for the five publicly available word image data sets. The baseline results have been obtained by running Omnipage on the word images without binarization. Results reported with MAPS algorithm and benchmark values neither use the training dataset nor any custom lexicon. One of the best reported results [7] uses a limited lexicon and two others [16] use a synthetic custom lexicon, both derived from the ground truths of the respective test sets.

Data sets —>	ICDAR2003	PAMI 2009	SVT 2010	BDI 2011		ICDAR 2011	
Performance measure —>	WRR	WRR	WRR	WRR	EDM	WRR	EDM
Benchmark results (MAST-CH+BkGnd padding)	**83.9**	**89.3**	**79.6**	**88.5**	**51.3**	**86.7**	**60.1**
MAST-CH segmentation	68.0	69.8	74.8	83.1	99.2	—	—
MAPS algorithm [17]	64.5	80.0	30.1	82.8	120.7	71.7	199.7
Best result in the literature	61.1 [16]	86.1 [7]	73.3 [16]	82.9 [15]	108.7 [15]	41.2 [11]	176.2 [11]
Baseline (Omnipage)	41.0	50.2	27.7	63.0	282.2	31.4	494.9

extracted from this custom lexicon to improve recognitiion. The recognition rates are tabulated in Table 1.

In order to study the actual recognition rates obtainable in real life situations, we did not employ any such custom lexicons. The poor result of 30.1% obtained by MAPS algorithm clearly shows that it is not suited for the SVT dataset; the bounding boxes are much bigger than the enclosed words in many cases, and hence, the key assumption made by the MAPS algorithm that the midline contains both foreground and background information is not satisfied. If bounding boxes had been properly annotated, it might have resulted in a higher recognition rate. Further, the images also contain motion blur, which cannot be adequately handled by the segmentation approaches used for the midline.

5.4 Born-digital 2011 data set

Karatzas et. al [10] initiated a new robust reading competition in 2011. This competition was based on email attached and web images. These images are known as born-digital images (BDI), since they are created using software such as Adobe Photoshop [19], Gimp [20] and Microsoft Paint [23].

In these images, the text has been placed by a user in an interactive fashion through a software. Hence, the fonts available in the system are only used to create the text pixels. Thus, when the recognizer uses a similar font, the recognition rate can be higher with born-digital images. This is evident from the fact that the baseline recognition result itself is 63%. This data set has better word boundary definition than others. A background margin of four pixels exists around the text bounding box to provide the context of the image.

BDI 2011 data set consists of 918 test word images. Only one participant competed in the ICDAR 2011 word recognition competition. Abbyy Fine Reader (applied on the image without binarization) was used as the baseline method by the competition organizer to compare with the performance of submitted algorithms. Since only TH-OCR algorithm [6] competed in the competition and could not beat the baseline method, it was mentioned as a honorary entry.

The resolution of born-digital word images are low compared to scene word images. Further, they are affected by anti-aliasing. Due to these two factors, characters in the word images merge when a global threshold is applied. Kumar and Ramakrishnan [15] applied power-law transform to remove the affect of anti-aliasing. By varying the γ value in the power-law transform, they showed that the merged characters can be split and obtained the best recognition result of 82.9%. The performance of MAPS algorithm is very close, with a value of 82.8%. As per our MAST-CH segmentation and background padding, the best obtainable recognition rate is about 88.5%.

A new metric, called edit distance measure, was introduced in this competition. This metric gives equal weights for addition and deletion of characters from the word. The calculated distance is normalized to that word. The normalized weights for all the words are added to form the edit distance measure reported for the complete data set. Table 1 compares the total edit distance and word recognition rate for the born-digital image data set.

5.5 ICDAR 2011 data set

This data set given for ICDAR 2011 Robust reading challenge Task 2 is almost a subset of ICDAR 2003 dataset, where repeated words have been removed and a few additions have been made. All the images removed are from scene images and are not considered either in the testing or training sets of ICDAR 2011 competition. The test set consists of 716 word images. The different recognition rates are shown in Table 1. We can observe that the recognition rates of both MAPS algorithm and benchmark experiments are higher than their ICDAR 2003 values.

Here, we could not assess the recognition rate of non-processed MAST-CH segmented images. Polarity reversal of some images by the OCR itself has led to erroneous outputs, the reason being that the bounding boxes specified are tight. Word images in the test dataset do not have any background pixels around the word boundary, unlike the Born-digital 2011 data.

6. DISCUSSION

We observe that the recognition rate on MAST-CH (semi-automated) segmented images is better than the rest for all the data sets. The recognition rates of MAPS method [17], proposed by the authors, are comparable or better than the best in the literature, except for SVT 2010 data set. In street view dataset, the recognition rate is often poor due to skew or curvy nature of words. Figure 6 shows sample images from SVT 2010 database with different degradations.

Figure 6: Street view images with different degradations. (a) Improper bounding box. (b) Low resolution. (c) Curved and degraded text. (d) High degree of motion blur.

We undertook the huge task of annotation in order to show that good segmentation of word images is the key to recognize characters/words well. Further, the benchmark result obtained for each data set is an indirect indicator of the quality or degradation of the images in the data set, and can be used to evaluate the effectiveness of any algorithm with respect to this possibly maximal rate. Even though top-down approach is useful in improving the recognition rate on specific data sets using custom lexicons synthetically created for each test image, it is felt by the authors that this approach is not practical in most real world situations, where one cannot expect such custom lexicons to be available. Further, Weinmann et. al [7] showed that the recognition performance reduces with full lexicon.

Around 85% word recognition has been achieved with the semi-automatically segmented images. Thus, if we focus on proper segmentation of characters in spite of all possible degradations, we can improve the recognition rate. Here, all the word images were segmented in such a way that individual components in the segmented image can be properly recognized or classified by a classifier or an OCR engine.

7. CONCLUSION

We have made both the MAST-CH annotation tool and the annotated data sets publicly available for download from our MILE website[1]. Any researcher can download the binarized images and use them as ground truth for their segmentation algorithms. Any issues in using the annotation tool or errors observed in the annotation of the images may kindly be reported to the authors. The recognition rate varies across OCR engines and also with their versions. The tabulated results emphasize the need for good segmentation, since it is the major part of the bottom-up approach. The good segmentation performance of MAPS algorithm can be indirectly seen from the achieved word recognition rates for the ICDAR, PAMI and BDI data sets.

Acknowledgment

We express our heartfelt thanks to Shanti Devaraj, Shanti S and Saraswathi S for careful annotation of the word images and committed interaction with the authors. Improvement in UI of the annotation tool was possible only from their feedback. The annotation work has minimal errors and the credit goes to them. Without them, our dream to annotate all the images at the pixel level could not have been accomplished. The authors also thank Technology Development for Indian Languages (TDIL), Department of Information Technology, Government of India for partially funding this research work. Finally, we thank the reviewers for their valuable comments, which helped us to significantly improve the presentation of our work.

[1]http://mile.ee.iisc.ernet.in/mile/download.html

8. REFERENCES

[1] N. Otsu, "A thresholding selection method from gray-level histogram", *IEEE Trans. on Systems, Man and Cybernetics*, vol. 9, pp. 62–66, March 1979.

[2] J. Kitler, J. Illingworth and J. Foglein, "Threshold selection based on a simple image statistic," *CVGIP*, vol. 30, pp. 125–147, 1985.

[3] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Second Edition, Prentice Hall, 2002.

[4] S. M. Lucas et. al, "ICDAR 2003 Robust reading competitions: entries, results and future directions", *IJDAR*, vol. 7, no. 2, pp. 105–122, June 2005.

[5] S. M. Lucas et. al, "ICDAR 2005 text locating competition results," Proc. VIIIth *ICDAR*, pp. 80–85, 2005.

[6] H. Liu, X. Ding, "Handwritten character recognition using gradient feature and quadratic classifier with multiple discrimination schemes", Proc. 8th *ICDAR*, pp.19-25, 2005.

[7] J. J. Weinmann, E. Learned-Miller and A. R. Hanson, "Scene text recognition using similarity and a lexicon with sparse belief propagation", IEEE Trans. PAMI, vol. 31, no. 10, pp. 1733–1746, 2009.

[8] K. Wang and S. Belongie, "Word spotting in the wild", Proc. 11th *ECCV*, pp. 591–604, 2010. http://vision.ucsd.edu/∼kai/svt/

[9] A. Mishra, K. Alahari and C. V. Jawahar, "An MRF model for binarization of natural scene text", Proc. 11th *ICDAR*, pp. 11–16, 2011.

[10] D. Karatzas, S. Robles Mestre, J. Mas, F. Nourbakhsh and P. Pratim Roy, "ICDAR 2011 Robust reading competition - challenge 1: reading text in born-digital images (web and email)", Proc. 11th *ICDAR*, pp. 1485–1490, 2011. http://www.cv.uab.es/icdar2011competition/

[11] A. Shahab, F. Shafait and A. Dengel, "ICDAR 2011 Robust reading competition - challenge 2: reading text in scene images", Proc. 11th *ICDAR*, pp. 1491–1496, 2011.

[12] T. Kasar, D. Kumar, M. N. Anil Prasad, D. Girish and A. G. Ramakrishnan, "MAST: Multi-script annotation for scene images toolkit", Proc. *Joint workshop on MOCR and AND*, pp. 1–8, Beijing, China, September 2011.

[13] T. Kasar and A. G. Ramakrishnan,"Multi-script and multi-oriented text localization from scene images", Proc. 4th *International workshop on Camera-Based Document Analysis and Recognition (CBDAR)*, Lecture Notes in Computer Science, Vol. 7139, pp. 1–14, 2012.

[14] A. Shahab, F. Shafait, A. Dengel, and S. Uchida, "How salient is scene text?", Proc. *International Workshop on DAS*, pp. 317–321, 2012.

[15] D. Kumar and A. G. Ramkrishnan, "Power-law transformation for enhanced recognition of born-digital word images", Proc. XIth *International Conference on Signal Processing and Communications (SPCOM)*, pp. 1-5, July 2012.

[16] A. Mishra, K. Alahari and C. V. Jawahar, "Top-down and bottom-up cues for scene text recognition", *IEEE conference on CVPR*, 2012.

[17] D. Kumar, M. N. Anil Prasad and A. G. Ramakrishnan, "MAPS: Midline analysis and propagation of segmentation", Proc. 8th *ICVGIP*, 2012.

[18] Abbyy Fine reader. http://www.abbyy.com/

[19] Adobe Reader. http://www.adobe.com/products/acrobatpro/scanning-ocr-to-pdf.html

[20] GIMP. http://www.gimp.org/

[21] IAPR TC11 Reading Systems-Data sets List, http://www.iapr-tc11/mediawiki/index.php/Datasets

[22] MATLAB. http://www.mathworks.com/products/matlab/

[23] Microsoft Paint. http://www.microsoft.com/

[24] Nuance Omnipage reader. http://www.nuance.com/

[25] Tesseract OCR engine. http://code.google.com/p/tesseract-ocr/