HMM-based Multi Oriented Text Recognition in Natural Scene Image

[a]Sangheeta Roy*, [b]Partha Pratim Roy, [c] Palaiahnakote Shivakumara,[d] Georgios Louloudis, [c] Chew Lim Tan,
[b]Umapada Pal
[a]Tata Consultancy Services, Kolkata, India
[b]CVPR Unit, Indian Statistical Institute, Kolkata, India
[c]School of Computing, National University of Singapore, Singapore
[d]Computational Intelligence Laboratory, Demokritos, Greece
email: *roy.sangheeta@tcs.com

Abstract

Recognition of curved text in natural scene image is a challenging task. Due to complex background and unpredictable characteristics of scene text and noise, text characters in strings are often touching that affects the performance of segmentation and recognition. This paper presents a novel approach for curved text recognition using Hidden Markov Models (HMM). From curved text, a path of sliding window is estimated and features extracted from the sliding window are fed to the HMM system for recognition. We evaluate two frame-wise feature extraction algorithms namely Marti-Bunk and local gradient histogram. The proposed approach has been tested on different natural scene benchmark as well as video databases, e.g. ICDAR-2003competition scene images, MSRA-TD500 and NUS. We have achieved word recognition accuracy of about 63.28%*, 58.41% and 53.62%y for horizontal text, non-horizontal text and curved text, respectively.*

Keywords: Scene Text Recognition, Curved Text Recognition, Binarization, Hidden Markov Model.

1. Introduction

Extraction and recognition of text from various types of images, are very effectual in text based application like video and image database retrieval, image annotation, data mining etc. [1, 2]. In the same way, natural scene based text explores automatic detection of street name, location, traffic warning and name of commercial goods [1, 2]. Text appears in complex background, low resolution of natural scene images with different fonts, size, orientation, color and variant alignment. As a result, normal document OCR does not give accurate recognition [3] for natural scene text due to the above mentioned factors. For instance, for the first image shown Fig.1 where we can see curved text with complex background, the publicly available OCR does not recognize the text correctly while the proposed method recognizes the text correctly for the second image shown in Fig.1.Our motivation in this paper is to proposea robust method for curved text recognition in multi-oriented environment.

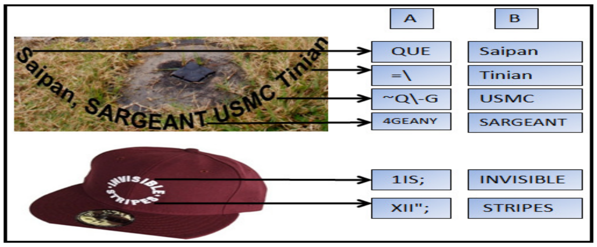


Fig. 1: Text images and their results using (A) available OCR and (B) our approach.

According to literature in natural scene text recognition, we have three classes of methods that are: (i) methods [3] to recognize the segmented text by proposing their own features with classifiers training, (ii) methods [4, 5] to binarize and recognize the text without segmentation of text lines using multiple hypotheses frames works and (iii) methods [6] to enhance the text through binarization to improve recognition rate. Methods of first class work well for specific data and specific languages as they require classifier and number of samples while the methods fall on second class uses several thresholds based on different hypotheses, but it is not clear how different hypotheses are drawn to set thresholds especially. On the other hand, methods of third class propose robust binarization to enhance the text information such that existing OCR gives good accuracy, which does not require classifier and several thresholds setting. Besides, the methods in first and second classes may not work well for the text like curved having complex background and low resolution, where touching between character is quite common and segmenting words and character from curved text lines is non-trivial. Therefore, we propose a method which falls on third class to recognize text irrespective of orientation, contrast and background. For instance, Roy et al. [7] have proposed wavelet gradient fusion method for video text binarization to improve the recognition rate of the video as well as scene text. This method shows that if we improve binarization for the segmented text lines in video and natural scene images, recognition rate can be improved with the available OCRs. However, the method is focused only on horizontal text but not on curved text.

However, the work presented in [8] works for multi-oriented text in graphical documents. Here, different segmentations cut are chosen and finally a dynamic programming based approach is used for the segmentation and recognition of touching characters. In the same way, the method proposed in [9] recognizes isolated characters in multi-scale and multi-orientation documents. However, these methods are developed for scanned images but not for scene text in natural scene images. Therefore, these methods cannot be applied on scene text directly due to complex background. There are few methods [10, 11, 12] which extract curved text from natural scene images. However, the scope of these methods is limited to text detection and extraction but not binarization and recognition.

In this paper, we present a novel method of multi oriented text recognition in scene images using HMM and performs segmentation free recognition in curved text image. The input is given by the text line portions of a given image. The output consists of the recognition of the text of each text line irrespective of the orientation of the text lines. Here, we propose an efficient method of sliding window from oriented text to feed HMM for text string recognition. We evaluate two frame-wise feature extraction algorithms namely:- LGH (Local Gradient Histogram) and MB (Marti-Bunke). Our contribution in this paper is to recognize multi-oriented text using curvy-linear sliding window based HMM. To the best of our knowledge, there is no work on curvy-linear text recognition in scene image.

2. Proposed Approach

We note that there are several sophisticated methods for text line detection and extraction in natural scene images irrespective of contrast, text type, orientation, background variation etc. and we use method [10] using gradient directional features to obtain text candidates and boundary growing for extracting curved text lines. Therefore, the output of text detection method is the input to the proposed method in this work for recognition.

The proposed approach is arranged into four sub-sections. In Section 2.1, we applied a robust binarization method of text extraction. In Section 2.2, we propose a method to find the path of sliding window using text-pixel analysis. Feature selections and HMM based training are discussed, respectively, in Section 2.3 and Section 2.4.

2.1 Binarization of Text Line

We propose to use Wavelet-Gradient-Fusion [7] method to convert text line image into binary image. Here, this approach does fusing of horizontal, vertical and diagonal information obtained by the wavelet and the gradient on text line images to enhance the text information. An unsupervised clustering algorithm is applied on row-wise and column-wise pixels separately to extract possible text information. The union operation on row-wise and column-wise clusters provides the text candidates information. With the help of Canny of the input image, the method identifies the disconnections based on mutual nearest neighbour criteria on end points and it compares the disconnected area with the text candidates to restore the missing information. Next, the method uses connected component analysis to merge some subcomponents based on nearest neighbour criteria. The foreground (text) and background (non-text) regions are separated based on the observation that the color values at edge pixel of the components are larger than the color values of the pixel inside the component. The sample results are shown in Fig.2 where binarization method works well for different fonts and orientation of texts. Interestingly, one can notice from binarization results that there is a touching in “curved” text. This makes recognition more complex and challenging.

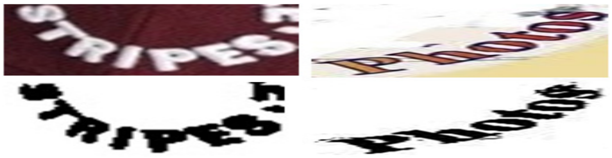


Fig. 2: Binarization result of text line image from image

2.2 Path of Sliding Window

From the binary text line images shown in Fig. 2, the foreground pixels of text lines are chosen and fed to a curved fitting algorithm. For this purpose, we use a general polynomial function *f(x)* as defined below

$$f(x) = a_n x^n + \cdots + a_2 x^2 + a_1 x^1 + a_0 \quad \ldots (1)$$

Where, $n > 0$ and $a_0, a_1, .., a_n$ are real numbers. Eq.1 provides the pathway of the sliding window in our approach (See Fig.3). In our present system, we chose the value of n=2 according to the experimental results. For feature analysis, the fixed width sliding window is placed at the left most position of the curved line and is moved to the next positions in steps. Since text line is the input for this work, the method finds starting position of text line easily. The height (H) of the text line is estimated from the average of local height computed at each points in f(x) bounded in the text region as $H = \frac{1}{n}\sum_{t=1}^{n} \mathrm{h(t)}$. *h(t)* is the local height of text region at a given step in curved f(t) and it is computed as the distance between the foreground pixels in both extreme sides. H is used to normalize the feature in each sliding window position.

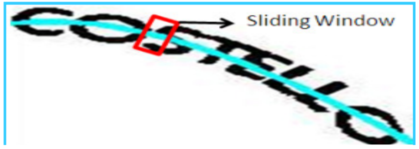


Fig. 3: The path of sliding window shown.

2.3 Feature Extraction

For feature extraction, we propose a sliding window of fixed width to extract a sequence of frames from the curved text. Before the feature extraction stage, the frames are normalized to a pre-defined height as described in the previous section. Two different types of features were used in our approach and described below.

Marti-Bunke feature: In this approach[13], the sliding window has a width of 1 pixel, moving from left to right and at each position 9 geometrical features are extracted. Three global features capture the fraction of black pixels, the center of gravity, and the second order moment. The remaining six local features consist of the position of the upper and lower contour, the gradient of the upper and lower contour, the number of black-white transitions, and the fraction of black pixels between the upper and lower contours.

LGH feature: The second feature extraction approach is based on the calculation of the local gradient histogram (LGH) [14]. Here, a sliding window traverses the image and each window is sub-divided into 4 × 4 (4 rows and 4 columns) regular cells. From all pixels in each cell a histogram of gradient orientations is calculated.We consider 8 orientations thus the final feature vector which is the concatenation of the 16 histograms results in a feature vector containing 128features.

2.4 Hidden Markov Models

The text recognition system is performed using Hidden Markov Models (HMMs). The basic models considered in this approach are character models. It contains a fixed number of hidden states (S_1,S_2,...,S_N) arranged in a left-to-right linear topology. These states emit observable feature vectors $O\epsilon\ IR^m$ with output probability distributions given by a Gaussian Mixture Model (GMM), starting from the second state (first and last states correspond to input and output states, respectively) in Bakis topology. For a model λ, if O is an observation sequence O= (O_1,O_2,..,O_T) assumed to have been generated by a state sequence Q= (Q_1, Q_2.,Q_T), of length T, we calculate the observation's probability or likelihood as follows:

$$P(O,Q\mid\lambda\ ) \\ = \sum_{Q} \pi_{q1} b_{q1}(O_1) \prod_{T} a_{qT-1\ qT}\ b_{q_T}(O_T) \qquad (2)$$

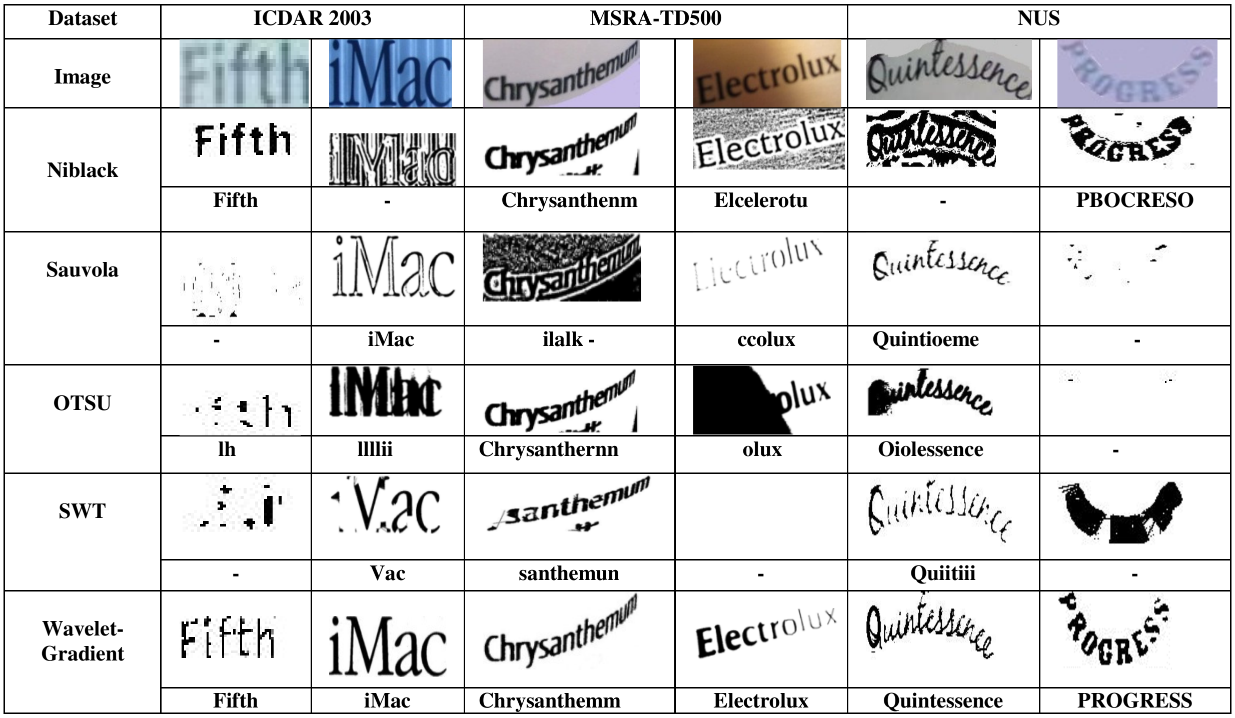
where$\pi_1$is initial probability of state 1, a_{ij} is transition probability from state i to state j and $b_i(O)$is output probability of state *i.* In training phase, the transcriptions of the text line image together with the feature vector sequences are used in order to train each character model. Using the transcription of each text line, character models are concatenated in order to produce the text line model. The text line model re-estimates the initial output probability distributions of $b_i(O)$ with the Baum-Welch algorithm until the likelihood of the training set is maximized. The recognition is performed using the Viterbi algorithm. It finds the character sequence having the best likelihood using the text line feature vector sequence. For our HMMs implementation, we used the popular HTK toolkit [15].

3. Experimental Results and Discussions

We test our approach on 3 different datasets namely: ICDAR 2003 [16], MSRA-TD500 [17]and NUS [10], where text appears in different orientations like horizontal, non-horizontal, curved etc. ICDAR 2003 data is popular data for scene text detection and recognition as it contains varieties of images. However, most of text lines in ICDAR data are in horizontal direction. Therefore, we use MSRA-TD500 dataset that contains both horizontal and non-horizontal text images with variety of background complexity as in ICDAR. However, the orientation of the text is limited to non-horizontal straight and there was no curved text. To test effectiveness of the proposed method in terms of orientation and contrast variation, we use NUS data that contains curved text and their resolution was low because of video data. The numbers of the words in the datasets used for our experiment are 852, 310 and 300, respectively.

We use a sample data set from these datasets as validation set to fix our HMM parameters such as the number of states, the number of Gaussian distributions and width of sliding window. In Fig 5, a detailed analysis of performance on datasets is observed in terms of state and Gaussian number. We choose 5 states for each character models and 8 GMMs for each state. Fig. 5 shows that the best performance is obtained with LGH feature. From the experiment, it is noted that in LGH, sliding window width of 6 pixels with overlapping ratio 50% provides better result.

The text recognition accuracy is performed using a 5-fold cross validation from ICDAR datasets, i.e. the dataset is divided into 5 parts and among them 4 parts are considered as training set and the other part is used for testing. We combine the data for other 2 datasets as number of data in each dataset is less. For training and testing, we use 5-fold cross validation for this merged dataset. The average of these results is reported in Table I. Here, the character recognition accuracy is 70.27%. Wenote that the character recognition rate of the proposed approach is higher than the recognition rate 67% reported for ICDAR-2003 competition data [3, 18]. Recognition scores at the word level have also been given in Table I. The recognition of word accuracy in ICDAR dataset is achieved up to 63.28%.



Dataset	ICDAR 2003		MSRA-TD500		NUS	
Niblack	Fifth	-	Chrysanthenm	Elcelerotu	-	PBOCRESO
Sauvola	-	iMac	ialk -	ccolux	Quintioeme	-
OTSU	lh	llllii	Chrysanthernn	olux	Oiolessence	-
SWT	-	Vac	santhemun	-	Quiitiii	-
Wavelet-Gradient	Fifth	iMac	Chrysanthemm	Electrolux	Quintessence	PROGRESS

Fig. 4: Comparison of recognition results obtained by the proposed HMM based method using different binarization methods. The recognized texts using HMM are shown below the respective binarized images.

Mishra et al. [18] reported the word recognition rate52.04% in ICDAR-2003 competition data. It is clear from Table I that the proposed method can recognize the non-horizontal and curved video data too with good accuracy.

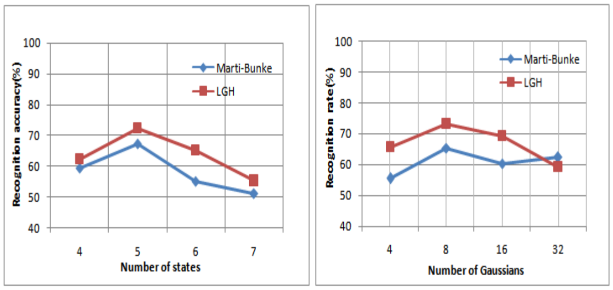


Fig 5: Recognition rates against the number of Gaussians and the number of states.

It is interesting to observe the comparative recognition accuracy of the proposed HMM-based method on multi oriented binarization texts, generated by different existing binarization methods. For comparative study, we use a freely available OCR [19] and we compare our HMM-based method with this freely available OCR. Five baseline methods of binarization (Niblack[20], Sauvola [21], Otsu [22], SWT[23], Wavelet-gradient[7]) are used and the recognition results of our HMM-basedmethod and thefreely available OCR are compared on these five different binarized images. The comparative results obtained from the three datasets are shown in Table II, III and IV. It is noticed from the recognition results that the freely OCR engine does not produce good recognition accuracy where we consider input images in different orientation with low contrast, complex background, distorted text and different fonts. The reason for the poor result lies in the use of OCR as it cannot handle text present in non-horizontal orientation. This illustrates the advantage of using thecurvy-linear sliding window based proposed HMM approach. From the tables it can be seen that HMM-based method gives better results than the other method for all the datasets as well as for all the binarization methods.

TABLE I: CHARACTER AND WORD RECOGNITION PERFORMANCE BY PROPOSED METHOD (%)

Database	Accuracy (%)	
	Character	Word
ICDAR	70.27	63.28
MSRA-TD500	68.58	58.41
NUS	64.7	53.62

The samples of qualitative results obtained from the proposed method are shown in Fig 4. The recognized texts obtained using proposed HMM-based method are shown below the respective binarized images in the figure. From the Table it can be seen that our approach can recognize the text even if they are touching in multi-orientation. Fig.6 shows some wrong recognition results of characters obtained from the proposed method. Analysing the recognition result, we find classification errors and noted that those errors are mainly caused by ambiguous characters, such as {L, I}, {O, D}, {h, n}, {e, c} etc. Therefore, there is a scope for further improvements by making feature robust.

TABLE II: COMPARISON OF WORD RECOGNITION RESULT (%) IN HORIZONTAL DATA (ICDAR DATASET)

	ICDAR	
Methods	Proposed HMM-based method	OCR[19]
Niblack [20]	54.18	40.72
Sauvola [21]	48.34	35.15
OTSU [22]	58.12	39.34
SWT [23]	57.13	53.16
Wavelet-Gradient [7]	63.28	60.68

TABLE III: COMPARISON OF WORD RECOGNITION RESULT (%) IN NON-HORIZONTAL DATA (MSRA-TD500 DATASET)

	MSRA-TD500	
Methods	Proposed HMM-based method	OCR[19]
Niblack [20]	48.15	25.24
Sauvola [21]	42.59	29.13
OTSU [22]	52.27	31.17
SWT [23]	45.12	30.78
Wavelet-Gradient [7]	58.41	33.34

TABLE IV: COMPARISON OF WORD RECOGNITION RESULT (%) IN CURVED DATA (NUS DATASET)

	NUS	
Methods	Proposed HMM-based method	OCR [19]
Niblack [20]	44.16	11.23
Sauvola [21]	39.23	18.13
OTSU [22]	46.37	15.56
SWT [23]	44.12	11.78
Wavelet-Gradient [7]	53.62	19.78

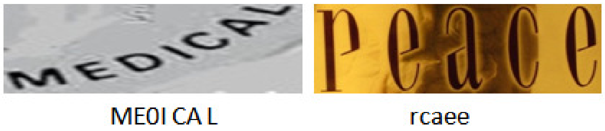


Fig. 6: Some text images and their wrongly classified character recognition results shown below the respective images.

4. Conclusion

We have proposed a novel methodology for recognizing curved text line using HMM. Here, we took pain in investigating the application of HMM in the datasets by varying the number of states and GMMs. The novel approach is based on estimating curved line of foreground pixel and extracting feature vector placing sliding window on image by following curved line. We demonstrated that by using simple feature we are able to achieve higher recognition rate about 63.28% which is better that the state of the art method. This result validates that proposed system is promising for real world application in addition to robust and invariant to size and orientation of the text present in the scene and video image.

References:

1. D. Doermann, J. Liang and H. Li, “Progress in Camera-Based Document Image Analysis”, In Proc. ICDAR, 2003, pp. 2106-2116.
2. J. Zang and R. Kasturi, “Extraction of Text Objects in Video Documents: Recent Progress”, In Proc. DAS, 2008, pp. 5-17.
3. L. Neuman and J. Matas, “A Method for Text Localization and Recognition in Real World Images”, In Proc. ACCV, 2010, pp. 770-783.
4. J. M. Odobez and D. Chen, “Robust Video Text Segmentation and Recognition with Multiple Hypotheses”, In Proc. ICIP, 2002, pp 433-436.
5. R. Huang, S. Oba, P. Shivakumara and S. Uchida, “Scene Character Detection and Recognition Based on Multiple Hypotheses Framework”, In Proc. ICPR, 2012, pp. 717-720.
6. S. Jetley, S. Behlhe, V. K. Koppula and A. Nagi, “Two-Stage Hybrid Binarization around Fringe Map based Text Line Segmentation for Document Images”, In Proc. ICPR, 2012, pp. 343-346.
7. S. Roy, P. Shivakumara, P. P. Roy and C. L. Tan, “Wavelet-Gradient-Fusion for Video Text Binarization” In Proc. ICPR, 2012, pp. 3300-3303.
8. P. P. Roy, U. Pal, J. Lladós and M. Delalandre. "Multi-Oriented Touching Text Character Segmentation in Graphical Documents using Dynamic Programming". PR, vol. 45, pp. 1972-1983, 2012.
9. U. Pal, P. P. Roy, N. Tripathy and J. Lladós. "Multi-Oriented Bangla and Devnagari Text Recognition". PR, vol. 43,pp. 4124-4136, 2010.
10. N. Sharma, P. Shivakumara, U. Pal, M. Bluemenstein and C. L. Tan, “A New Method for Arbitrarily-Oriented Text Detection in Video”, In Proc. DAS, 2012, pp. 74-78.
11. C. Yao, X. Bai, W. Liu, Y. Ma and Z. Tu, ‘’Detecting Texts of Arbitrary Orientations in Natural Images’’, CVPR 2012.
12. P. Shivakumara, T. Q. Phan and C. L. Tan, “A Laplacian Approach to Multi-Oriented Text Detection in Video”, IEEE Trans. on PAMI, 2011, pp. 412-419.
13. U. V. Marti and H. Bunke, “Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system”, IJPRAI, 2001, pp. 65–90.
14. A. Rodrı́guez-Serrano and F. Perronnin, “Handwritten word-spotting using hidden Markov models and universal vocabularies”, Pattern Recognition, 2009, pp. 2106-2116.
15. S. J. Young et al., “The HTK Hidden Markov Model Toolkit Book”, Entropic Cambridge Research Laboratory, 1995.
16. S. Lucas et al.,ICDAR 2003 Robust Reading Competitions. In Proc. ICDAR, 2003: pp. 682-687
17. C. Yao1, X. Bai1, W. Liu, Y. Ma and Z. Tu. "Detecting Texts of Arbitrary Orientations in Natural Images", In Proc. CVPR, 2012, pp. 1083-1090.
18. A. Mishra, K. Alahari, C. V. Jawahar: “An MRF Model for Binarization of Natural Scene Text”. In Proc. ICDAR, 2011: 11-16.
19. Tesseract. http://code.google.com/p/tesseract-ocr/
20. W. Niblack, “An Introduction to Digital Image Processing”, Prentice Hall, Englewood Cliffs, 1986.
21. J. Sauvola et al., “Adaptive Document Binarization”, In Proc. ICDAR, 1997, pp. 147-152.
22. N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," IEEE Trans. on Systems, Man, and Cybernetics, Vol. 9, 1979, pp. 62-66.
23. B. Epshtein, E. Ofek, Y.Wexler: Detecting text in natural scenes with stroke width transform. CVPR 2010: 2963-297