DOI 10.3233/KES-160344
IOS Press

Recognition of Bangla text from outdoor images using decision tree model

Ranjit Ghoshal[a,*], Anandarup Roy[b], Bibhas Ch. Dhara[c] and Swapan K. Parui[b]
[a]*Department of I. T. St. Thomas' College of Engg. & Tech., Kol, India*
[b]*CVPR Unit, I.S.I. Kol, India*
[c]*Department of I.T., Solt Lake Campus, Jadavpur University, Kol, India*

Abstract. This article proposes a scheme for automatic recognition of Bangla text extracted from outdoor scene images. For extraction, first the headline is obtained, then certain conditions are applied to distinguish between text and non-text. By removing the headline, the Bangla text is partitioned into two zones. Further, an association among the text symbols in these two different zones is observed. For recognition purpose, a decision tree classifier is designed with Multilayer Perceptron (MLP) at leaf nodes. The root node takes into account all possible text symbols. Further nodes highlight distinguishable features and act as a two-class classifiers. Finally, at leaf nodes, a few text symbols remain, that are recognized using MLP classifiers. The association between the two zones makes recognition simpler and efficient. The classifiers are trained using about 7100 samples of 52 classes. Experiments are performed on 250 images (200 scene images and 50 scanned images).

Keywords: Recognition, Bangla text, outdoor image, decision tree, multilayer perceptron

1. Introduction

Automatic recognition of text symbols in a natural scene image is useful to the blind and foreigners having language barrier. Such a recognition method should also include an extraction of text portions from the scene images. Extraction and recognition of texts from outdoor scene images is a challenging problem now-a-days due to the variations in style, color, background complexity etc. The aim of this article is twofold. Here the preliminary task is to extract the possible text symbols from an input scene image. This has been an issue of interest for many years. One of the earliest works was by Jung et al. [1] who employed a multilayer perceptron classifier to discriminate between text and non-text pixels. A good survey of existing methods for detection, localization and extraction of texts embedded in scene images can be found in [2].

Text detection from scene images can roughly be classified into three groups: sliding window based methods [13,14], connected component (CC) based methods [15–17], and hybrid methods [18]. Sliding window based methods are also known as region-based methods. Here, a sliding window is used to search for possible texts in the image and then use machine learning techniques to identify text. These methods are slow as the image has to be processed in multiple scales. Recently, maximally stable extremal regions (MSERs) based methods have become the focus of several recent works [19–21]. More recently, a robust text detection in natural scene images is developed by Yin et al. [23].

Bangla is one of the most popular Indian scripts used by more than 500 and 200 million people respectively in the Indian subcontinent. A unique characteristic of this script is the existence of headlines as shown in Fig. 1. This is the main feature of Bangla text. English text has no such headline. The initial focus of the present work is to exploit the above fact for extraction of Bangla text from images of natural scenes. The only assumption is made here that

[*] Corresponding author: Ranjit Ghoshal, Department of I. T. St. Thomas' College, 4, D. H. Road, Kol-700023, India. E-mail: ranjit.ghoshal@rediffmail.com.

ISSN 1327-2314/17/$35.00

the characters are sufficiently large and/or thick so that using a linear structuring element of a certain fixed length can capture its headlines. Considering Bangla script, Bhattacharya et al. [3] proposed a scheme based on analysis of connected components for extraction of Devanagari and Bangla texts from camera captured outdoor images. For this purpose the headlines were detected using morphology. In this paper, this morphology based scheme is used to obtain possible headlines. Afterwards, the text symbols attached/unattached to the headlines are extracted using certain shape and position based purifications.

The recognition issue also draws interest of many scientists. Pal and Chaudhuri [4] presented a detailed review on recognition of Indian script. However, all the efforts are still done using scanned documents. To the best of authors knowledge no recent works exist that deals with Bangla text separation and recognition from natural scene images. Here a decision tree model is designed for efficient recognition of extracted Bangla texts.

In this recognition system the headline of a Bangla text plays an important role. The headline detection eventually leads us to the zone detection [10]. In Fig. 1 all the three zones separated by headline and baseline are presented.

In this paper, the middle and the lower zones are merged to form the middle-lower zone. This zone has an association with the upper zone as will be observed later. For recognition in middle-lower zone, a decision tree model is used. At each internal node of this tree, one criterion is evaluated that splits the current set of text symbols into two halves. The leaf nodes, finally, encounter only a few text classes and the MLP is used for classification. The association between two zones boost the recognition performance. Here, experiments are performed on a laboratory made dataset consisting of both scene and scanned images. The experimental results establish satisfactory performance of the proposed methods.

In this paper a system for extraction and recognition of Bangla text symbols from scene images is proposed. The proposed system works in three steps. Initially headlines are detected using mathematical morphology [3]. In the next step, some criteria are proposed on the basis of which text symbols are isolated from non-text. These criteria can be applied to scanned images as well, with equal performance. It is shown later. The final step of the system is the proposed decision tree classifier. This decision tree automatically recognizes text symbols based on some pre-defined conditions. To boost the recognition performance, association among text symbols in different zones of a word is considered. These associations finally lead to small groups of symbols, that are easily recognizable. Such a script based boosting has not yet been done, to the best of authors knowledge. Moreover, the proposed system is robust to perspective distortion and image skew at least up to an acceptable level. Also, this system can be extended to scanned documents. In fact, the experiments include scanned documents along with scene images.

2. Color image segmentation

Here, color image segmentation is the first step of text extraction. Prior to segmentation, let us describe the features that are extracted from the normalized RGB image. Consider a pixel p_i of the image. Then p_i can be described by the tuple $(r_i,\ g_i,\ b_i)$ i.e. the normalized R, G and B values. Besides these color values, the business feature [5] for each pixel p_i is taken. The business feature of p_0 is denoted by B_0 that takes into account the direction of variation in intensity and it is computed as:

$$\begin{aligned} B_0 = \frac{1}{12}(&|p_1 - p_2| + |p_8 - p_1| + |p_0 - p_3| \qquad (1)\\ &+ |p_7 - p_0| + |p_5 - p_4| + |p_6 - p_5|\\ &+ |p_3 - p_2| + |p_4 - p_3| + |p_0 - p_1|\\ &+ |p_5 - p_0| + |p_7 - p_8| + |p_6 - p_7|). \end{aligned}$$

Where $p_1, \ldots, p_7$ are the eight neighboring pixels of p_0 (see Fig. 2).

Hence, the feature vector $\vec{f}$ for a pixel p_0 is defined as: $\vec{f} = \{r_0, g_0, b_0, B_0\}$. After collecting all these features the Gaussian Mixture Model (GMM) [6] is applies to obtain an unsupervised clustering of the feature vectors for an image. The GMM describes the distribution of the feature vectors using a convex combination of Gaussian distributions. The expectation maximization (EM) algorithm is used to estimate the GMM parameters. In order to estimate the number of mixture components, several passes are performed of EM with increasing clusters. Further the *integrated classification likelihood* (ICL) [7] criterion is applied. The optimum number of components is at the first local minimum of the ICL criterion. Finally, the individual components of the mixture model correspond to the clusters in the feature set.

After EM, a single spectral cluster when mapped to an image, may contain several spatial connected

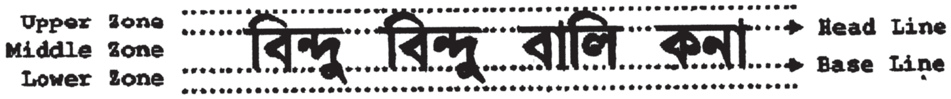


Fig. 1. Illustration the zones of Bangla Script.

components. It is observed that between two such adjacent components from two different clusters, there is a thin boundary consisting of pixels with spectral values that are different from both these clusters. In fact, a pixel on such a boundary really comes from any one the two adjacent components. These thin boundaries do not correspond to any particular physical entities in the original image. Hence, these boundary pixels may be merged with any of the adjacent clusters without harm. It is observed that if there is a thin boundary between two adjacent components, its thickness varies from one pixel to three pixels. So, the 5×5 (or larger) neighborhood around a pixel of such a boundary will necessarily contain pixels of at least two clusters other than the boundary. Let there be K clusters $C_1, C_2, \ldots, C_K$. Let p_i be a pixel belonging to cluster C_j. It is examined that the $n \times n$ neighborhood around p_i, where n ($\geqslant 5$) is the size of the window. Further, a set $H = \{h_1, h_2, \ldots, h_K\}$ is constructed where h_l denotes the number of pixels surrounding p_i and belonging to C_l. So, at least two elements of H, namely, h_l and h_n ($l \neq n \neq j$) become positive. If there does not exist at least two such surrounding clusters C_l and C_n, it may be ignored the current pixel p_i and proceed to the next. However, even if it is found a pixel with two or more different surrounding clusters, it may not be a boundary pixel. A thin object may have a pixel having a few pixels of two or more different clusters around it. Hence, a condition is imposed that, if it is found that h_l and h_n ($l \neq n \neq j$) both greater than $\lfloor h_j/2 \rfloor$, then p_i is treated as a boundary pixel. Such a pixel is to be merged with one of the two surrounding components corresponding to h_l and h_n. Next the posterior probabilities of p_i for these two clusters are found. Now p_i is assigned to the cluster corresponding to the higher posterior probability.

3. Bangla text symbol extraction

Most of the Bangla characters are connected by the headline. Hence, in order to extract the Bangla text portions, the headline that joins them should be detect.

p_2	p_1	p_8
p_3	p_0	p_7
p_4	p_5	p_6

Fig. 2. Eight neighbors of a pixel p_0.

3.1. Morphology based headline detection

Here, mathematical morphology is applied to obtain all horizontal lines inside an image. Later, applying the following conditions the possible headlines are detected. At first sufficiently small and large connected components (CC) are removed. For each remaining CCs, its skeleton image is obtained by morphological operation. Let us denote the skeleton image by A. On this skeleton image, morphological opening operation is performed to extract straight lines. It is evident that opening of an image A with a linear structuring element B can effectively identify the horizontal line segments present A. However, a suitable choice of the length of B is crucial for processing at the latter stages and this length is fixed to 21 pixels for the present dataset. This length of the structuring element is selected based on the exhaustive experimental study. To select the length of the structuring element, experiments are performed for different sizes of structuring elements within the range $13 \leqslant B \leqslant 27$. It is observed that 21 pixels length structuring elements gives the best performance. The effect of skeletonisation removes pixels at the boundaries of the image but does not allow components to break apart. After skeletonisation, only the required pixels remain and thus a less number of lines detected. However, if no such lines are detected, it is needed to apply morphology on the original CC. Also, it may encounter a component small enough that could not find any horizontal line. Such components remain as it is.

Now, all the headlines from all the components are accumulated. Let a line be denoted by L_i. Let H_u and H_l be the heights of the portions of the corresponding CC lie at the upper and at the lower half of the line L_i. Bangla characters mostly lie at the lower portion

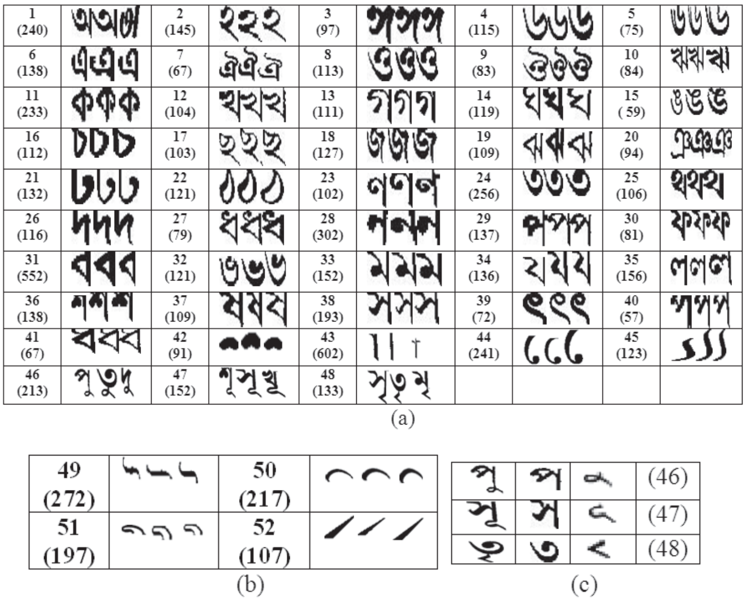

1 (240)
2 (145)
3 (97)
4 (115)
5 (75)
6 (138)
7 (67)
8 (113)
9 (83)
10 (84)
11 (233)
12 (104)
13 (111)
14 (119)
15 (59)
16 (112)
17 (103)
18 (127)
19 (109)
20 (94)
21 (132)
22 (121)
23 (102)
24 (256)
25 (106)
26 (116)
27 (79)
28 (302)
29 (137)
30 (81)
31 (552)
32 (121)
33 (152)
34 (136)
35 (156)
36 (138)
37 (109)
38 (193)
39 (72)
40 (57)
41 (67)
42 (91)
43 (602)
44 (241)
45 (123)
46 (213)
47 (152)
48 (133)
(a)
49 (272)
50 (217)
51 (197)
52 (107)
(b)
(46)
(47)
(48)
(c)

Fig. 3. Text components in (a) middle-lower and (b) upper zone. (c) First column shows composite texts and corresponding lower zone symbols are in third column. Serial numbers indicate class numbers.

of the headline, only a small portion may reside at the upper of the headline. So, for a headline L_i it should have $H_u < H_l$. Thus at the next step all the lines L_i for which $H_u < H_l$ are considered. These lines mostly represent the headlines.

3.2. Separation of text portions

First proceed with headline attached components. The following criteria are used to identify text portions. All the concerned CCs are subjected to these criteria in the same sequence as given. After conducting an exhaustive experimental study, the thresholds are selected. Note, the thresholds may differ if the sequence is altered.

1. *Boundary attached components*: Generally text like patterns are not attached with boundary of the image. So, first all the boundary attached CCs are removed using morphological reconstruction.
2. *Elongatedness ratio* (ER): This was used by Roy et al. [8]. Empirically it is found that a component with ER value greater than 5 is a text symbol.
3. *Number of complete lobes*: Using Euler number [9] complete lobes can be obtained. Found empirically, a text symbol has less than 9 complete lobes.
4. *Aspect ratio*: It is found by experiments, that the aspect ratio (height/width) of a non-text become less than 0.3 or greater than 2.0.
5. *Object to background pixels ratio* (r): Due to the elongated nature of Bangla script, only a few object pixels fall inside text bounding box. On the other hand, elongated non-texts are usually straight lines, so, contribute enough object pixels. It is observed that $0.3 \leqslant r \leqslant 3$ could identify text symbols.

All the headline attached text symbols are therefore separated.

However, in Bangla, some text symbols do not meet the headline. Now consider such components. These components, though not connected, must be close to one/more of already detected text symbols. Then, if the area of the bounding box is increased enough, the possible text symbols may lie inside it. With this view, the width of the bounding box is increased by its height and the height by an empirical threshold. The components inside this modified bounding box are subjected to the previous criteria to extract the text symbols.

3.3. Headline removal and zone detection

Now, it is needed to remove the headlines to isolate a text symbol. Sometimes a major part of a symbol may be joined with the headline. If the headline is removed totally, these symbols become broken. Instead, all the pixels over the headlines are examined.

Let such a pixel p_i (the i^{th} pixel) belongs to a component C_j. Let v_i be the vertical run length of p_i at the i^{th} position considering only the corresponding component C_j. Experimentally the thickness of the headline is considered to be H (say). Now, if $v_i \leqslant H$, it can be concluded that p_i is a pixel solely from the headline and thus removed. After removing the headline, several small components (i.e. consisting a few pixels) may exist in the image. Since, the actual Bangla characters are thick enough, so the components having small thickness may be removed. This procedures works as follows. Let h_i and v_i be the horizontal and the vertical run lengths of a pixel p_i at the i^{th} position of a component C_j. Next the minimum of h_i and v_i is computed and further constitute a set $MIN_j = \{m_i \text{ s.t. } m_i = \min(h_i, v_i),\ \forall i\}$. Thus MIN_j denotes the set of all the minimum run lengths considering all the pixels of C_j. The thickness T_j of the component C_j is defined to be the maximum frequent element of the set MIN_j. A predefined tolerance threshold (T_H say) is fixed for component thickness. If $T_j < T_H$ then the corresponding component C_j could be removed.

In previous studies (e.g. [10]), the headline was used to partition a Bangla word into three zones as in Fig. 1. However, to obtain the lower zone, one may need to identify the "*Base Line*" (Fig. 1). This leads to additional computations. Here, the detection of lower zone is dropped and proceed with the upper and the middle-lower zones (the latter being the portion below the headline). The symbols in different zones are shown in Figs 3(a), (b) with class numbers (serial numbers). The numbers in brackets indicate the total number of samples of the corresponding class in the dataset. The three classes: 46, 47 and 48 have lower zones text connected with middle zone text symbols. Let us term these as "composite" symbols. Only three lower zone components are considered, ূ, ু and ৃ . Figure 3(c) describes how a lower zone symbol is connected with a middle zone symbol.

4. Association between upper and middle-lower zones

The two zones produced by removing the headline play an important role during recognition. The upper zone contains some specific text symbols. So recognition in the upper zone has a high accuracy. On the other hand, middle-lower zone consists of a large symbol set, so vulnerable for recognition. However, once the upper zone symbol is identified, an assumption could be made about middle zone symbols. Thus, the initial set may be reduced. This assumption is based on prior knowledge about Bangla language. In Bangla, there are a few certain pre-defined middle zone symbols for a particular upper zone symbol. Moreover, upper zone symbols exists frequently in Bangla text. So, this very association greatly improves the accuracy of the proposed recognizer. Of course, if all the texts are present inside the middle-lower zone, this association could not be applied. Considering the upper zone text symbols from Fig. 3(b) the corresponding possible middle-lower zone text classes are enlisted in Table 1.

Table 1
Association between upper and middle zone text symbols

Upper zone text class	49	50	51	52
Middle zone text class	2–5, 21, 22, 43, 44	43	43	2, 11, 13, 16, 18, 19, 21–33, 35–38, 40, 41

5. Bangla text recognition

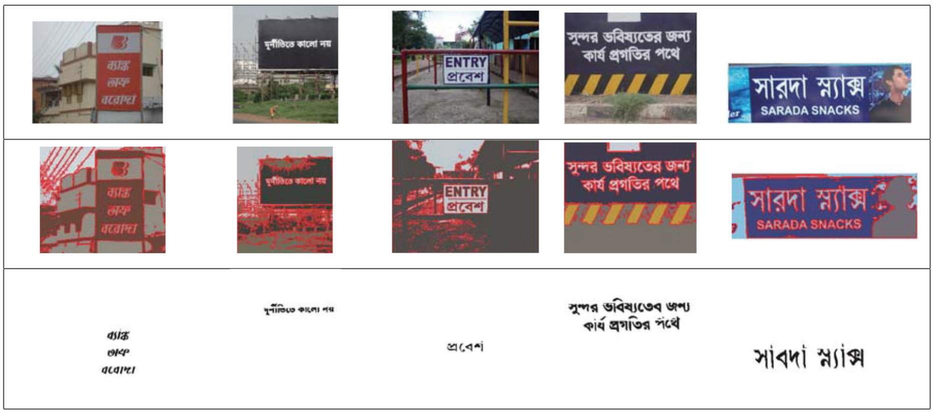
The middle-lower zone text symbols may be partitioned into two. One of them consists of all the composite text symbols. The study by Parui et al. [11] described a method to filter out and recognize lower zone symbols from a composite symbol. After its application, only the middle zone symbols remain. The upper zone symbols are subjected to MLP classifier directly, while a decision tree is applied to the middle zone symbols to form small groups. MLPs are then applied to each such group. For MLP, the wavelet feature described by Bhowmik et al. [12] is used.

The design of a tree classifier has three components: (a) a tree skeleton or hierarchical ordering of the class labels, (b) choice of features at each non- terminal node, and (c) the decision rule at each non-terminal node. In this study the tree is a binary tree where the number of descendants from a non-terminal node is two. While traversing the tree, only one feature is tested at each non-terminal node. To decide the feature at a particular non-terminal node, the occurrence statistics of the characters are considered. If the set of patterns at a non-terminal node can be sub-divided into two sub-groups by examining a feature so that the sum of occurrence probabilities of one group is roughly equal to that of the other group, the resulting binary tree is optimum in time complexity, assuming that the

Table 2
The groups obtained from decision tree

Group	A	B	C	D	E	F	G	H	I	J
Class	2–5, 8, 9, 18, 24, 32, 42, 44, 45	20	22	17	15, 39	7, 30	21, 26	37	10, 19	43
Group	K	L	M	N	O	P	Q	R	S	–
Class	6	13, 23, 35, 36 38, 40	16	27, 31, 41	11	29	12, 14, 25, 34	1	28, 33	–

Table 3
Some images (first row), the corresponding segmentation results (second row) and extracted text components (third row) after headline removal. Often, some non-text components are included in text class



time required to test a feature is constant. However, a set of features to design such an optimal tree is not always possible. A semi-optimal tree is generated out of the available features. For a given non-terminal node, a feature is selected that best separates the group of patterns in the above sense.

5.1. Decision tree classifier for middle zone symbols

The decision tree evaluates one criterion at each internal node that deterministically splits the current set of text symbols into two subsets. Proceeding this way, each leaf node finally contains a small group of symbols. For each such group, an MLP classifier is used. Before describing the tree, let us concentrate on the leaf level subsets (groups) resulting from of the decision tree. Table 2 suggests, most of such groups are small. So efficient classifiers can be designed. In practice, the association method can be used (Section 4) to make the groups even smaller. For example, when class 49 is present at the upper zone, only the classes 2, 3, 4, 5 and 44 are candidates from Group A.

Let us now present the decision tree model in Fig. 4. Each criterion is given a number and described below.

1. This criterion checks if any lobe exists in the text symbol.
2. This criterion checks the existence of vertical line (VL) in the text symbol. The VL is found by applying morphological operations with linear structuring element having 90% height of the concerned component image.
3. This criterion compares if the lobe width is less than a threshold $T_1 = 7$.
4. Eliminating the upper zone, the text symbol ঠ results only a lobe. By applying a threshold $T_2 = 0.88$ this criterion separates out ঠ .
5. This criterion finds text symbols (e.g. Group L) with rightmost VL.
6. Some text symbols (e.g. Group N) has a wide lobe comparable to the width of the component. This criterion decides if the lobe width is only a threshold ($T_3 = 1.35$) less than the component width.
7. Using a threshold $T_4 = 18$ indicating component height, the lobe may be vertically at the lower or upper halves. Two separate groups can be formed.
8. The contour of the VL is traced, and check if exactly one other component is connected with

Table 4
Recognition (in %) of the middle zone text symbols

Text class	Some symbol at upper zone	No symbol at upper zone	Text class	Some symbol at upper zone	No symbol at upper zone	Text class	Some symbol at upper zone	No symbol at upper zone
1	–	99.6	16	100	98.0	31	83.9	83.9
2	94.7	92.6	17	–	95.8	32	91.6	88.6
3	94.7	–	18	93.8	87.7	33	86.0	83.1
4	89.0	85.6	19	93.3	84.0	34	–	82.9
5	93.4	–	20	–	98.7	35	92.5	91.4
6	–	94.2	21	93.8	93.0	36	92.5	92.3
7	–	94.7	22	99.0	98.3	37	91.5	90.9
8	–	90.9	23	92.5	90.9	38	92.3	91.9
9	–	93.4	24	91.3	87.9	39	–	94.7
10	–	87.1	25	92.1	78.3	40	93.3	92.1
11	95.3	93.1	26	93.5	92.4	41	78.0	78.0
12	–	80.3	27	83.0	83.0	42	–	94.1
13	86.2	86.2	28	85.7	84.8	43	95.6	94.3
14	–	82.0	29	94.5	93.2	44	94.3	93.6
15	–	95.0	30	95.0	94.3	45	–	90.6

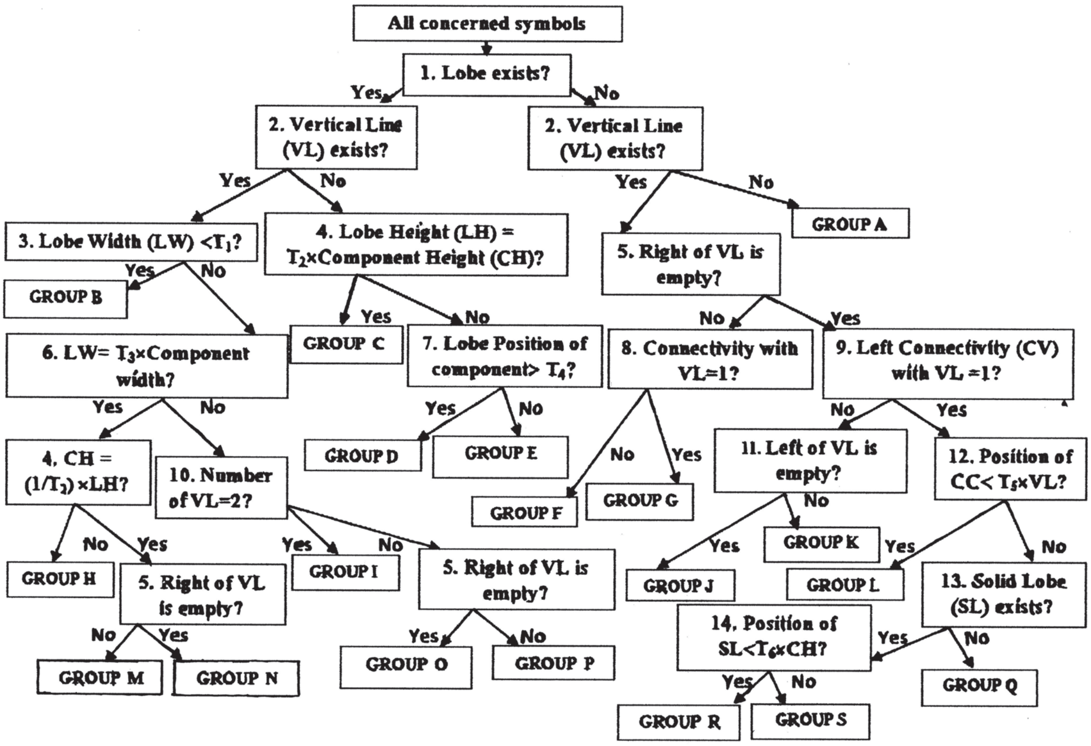


Fig. 4. The decision tree for middle zone symbol recognition.

the VL. To cope perturbations, such a component must have at least a 20 pixel run, connected with the VL. Further the term "connectivity" is used to denote number of components connected with the VL.

9. Certain groups can be formed if left connectivity is one.
10. The number of VLs in this criterion are checked. The second VL is, however, smaller than the first. So, a shorter (80% of component height) structuring element is used.
11. Group J is only a VL. So this is sufficient criterion to separate.
12. Only one component (CC) is connected with VL. The boundary contour is traced on which the first pixel meet of this component. This criterion

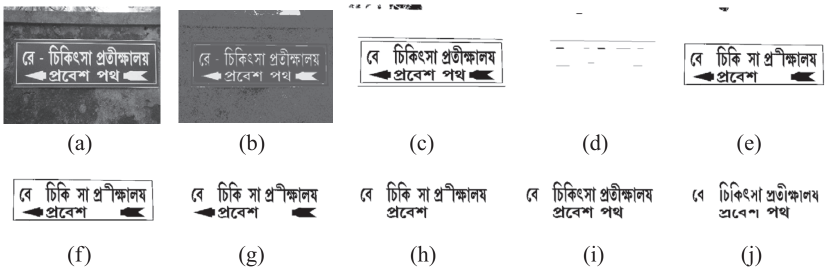


Fig. 5. (a) Input "X-Ray" image, (b) segmented image, (c) after removing too small and large components, (d) the headlines and (e) attached components, (f) after removing boundary attached components, (g) after testing aspect ratio, (h) after testing ratio r, (i) headline unattached components, and (j) after removing the headlines.

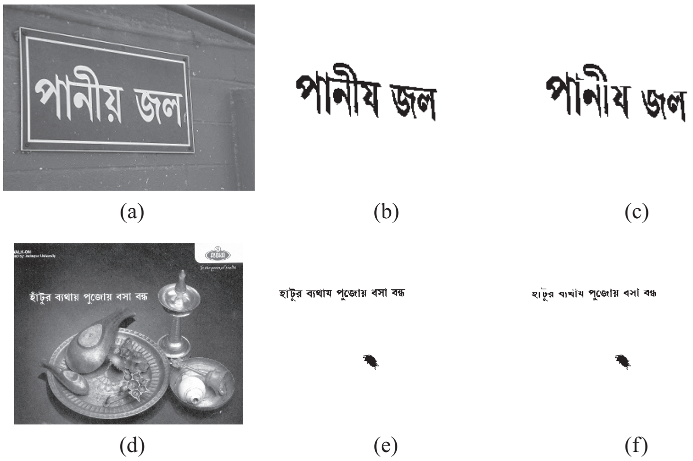


Fig. 6. (a) Input "Water" image, (b) its text portion, (c) after headline removal and (d) Input "Festival" image, (e) its text portion, and (f) after headline removal.

checks if the position of this pixel is only a threshold ($T_5 = 0.5$) high from the bottom of the VL.

13. Some components has solid lobes (e.g. Group R, S). To separate those components this is the correct criterion to be checked.
14. With $T_6 = 7$ the position of the solid lobe is determined.

Here, all the thresholds are selected based on the experiments of the training samples.

6. Results and discussions

The concerned laboratory made dataset consists of 200 images captured by a digital still camera (14.1MP) along with 50 scanned document images.

First, the text extraction results are presented. Consider the "X-Ray" image (Fig. 5(a)) as an example. The headline (Fig. 5(d)) attached components are shown in Fig. 5(e). Comparing this with Fig. 5(a), it is noticed that all but three text symbols are present. Next, after eliminating boundary attached components the blob at the top of the image (Fig. 5(f)) can be removed. The ER and counting the number of lobes have no effect on this image. Further, the box like non-text, surrounding the text portion are filtered out by testing the aspect ratio (Fig. 5(g)) and the arrow like components using ratio r (Fig. 5(h)). Afterwards, Fig. 5(i) gives the results of the identified headline unattached text symbols. It may be noted that the text symbols absent in Fig. 5(e) are now identified successfully. Finally, the symbols after removing the headlines are shown in Fig. 5(j). Note that there are only ⌒ and ৎ text symbols in the upper zone. Moreover, this particular image has compound characters প্র and ক্ষ . Such characters are beyond of the present study. In the next example, it is highlighted an important aspect of the dataset. In Fig. 6(a), the "Water" image is shown. This image has perspective distortion. Though any perspective correction procedure is not

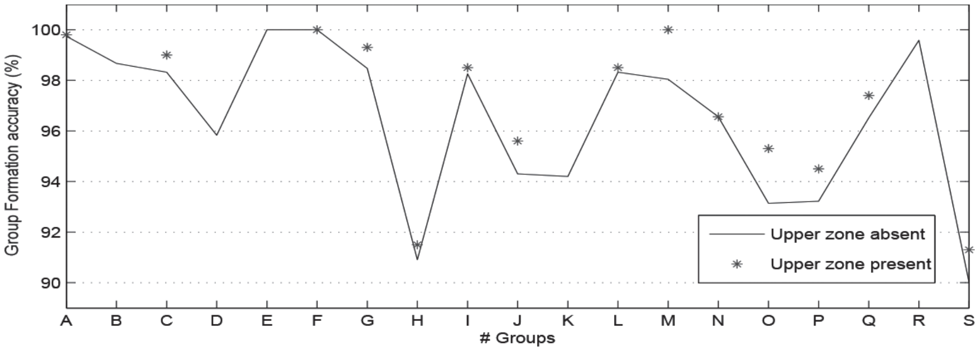


Fig. 7. Group formation accuracy depending presence/absence of upper zone.

applied, the proposed algorithm extracts text symbols successfully (Fig. 6(b)). The headline removal also works satisfactorily as presented in Fig. 6(c). The next example concerns a scanned image. In Fig. 6(d) the scanned "Festival" image is presented. The corresponding text portions are displayed in Fig. 6(e) and the headline is removed in Fig. 6(f). The notable point here is, though the headline is removed perfectly, a small blob remain on the image. Nevertheless, the procedure could cope scanned images with perfection. A few other sample images and the corresponding results are presented in Table 3. The first three images are natural scene with embedded Bangla text, whereas the last two are banners containing Bangla text. Finally, the precision and recall values of the text extraction algorithm based on 250 images, are respectively 70.7% and 73.3%.

Now the work of Parui et al. [11] is applied to all middle-lower zone text symbols. Besides, the upper zone symbols are recognized separately, using MLP. The lower zone classes 46, 47 and 48 give recognition accuracies 93.63%, 91.65% and 92.46% respectively. On the other hand, for the upper zone classes 49, 50, 51 and 52 the recognition accuracies from MLP are 94.36%, 92.68%, 93.74% and 94.52% respectively. In Fig. 7, it is described how accurately a group is formed depending upon the upper zone is absent/present. Note that the presence of upper zone makes most groups to form more accurately. These groups are produced by the decision tree, so Fig. 7 measures the reliability of the decision tree. Finally, the recognition accuracy of all the middle zone symbols is presented in Table 4. This results include the error produced by all previous operations. Note that accuracy is higher if some upper zone symbol is present. The association reduces the size of the groups even to a singleton. Since the MLP in upper zone performs well, the final recognition becomes satisfactory.

According to the author's knowledge only one work [24] exist that deals with recognition of Bangla text from scene images. Now, this proposed method is compared with the existing method. The proposed method (92.0%) outperforms the existing method (85.93%) in terms of average recognition accuracy.

7. Summary and future scope

This article proposes a recognition scheme of Bangla text symbols embedded in outdoor natural scene images. The proposed method introduces a decision tree classifier based on structural and topological features. This tree produces small symbol groups, well recognized by MLP. A script based boosting is proposed to improve recognition accuracy. The proposed scheme is extendable to printed scanned documents. Future studies may aim at the use of probabilistic scoring at the internal nodes instead of a deterministic decision.

Acknowledgement

This work is funded by the CSIR SRF grant (File No: 9/202 (0034) 2K11-EMRI).

References

[1] K. Jung, I.K. Kim, T. Kurata, M. Kourogi and H.J. Han, Text scanner with text detection technology on image sequences, *Proc. of Int. Conf. on Pattern Recognition* **3** (2002), 473–476.

[2] J. Liang, D. Doermann and H. Li, Camera based analysis of text and documents: A survey. *Int. Journ. on Doc. Anal. and Recog. (IJDAR)* **7** (2005), 84–104.

[3] U. Bhattacharya, S.K. Parui and S. Mondal, Devanagari and bangla text extraction from natural scene images, *Proc. of the Int. Conf. on Document Analysis and Recognition* (2009), 171–175.

[4] U. Pal and B.B. Chaudhuri, Indian script character recognition: A survey, *Pattern Recognition* **37** (2004), 1887–1899.

[5] A.K. Mandal, S. Pal, A.K. De and S. Mitra, Novel approach to identify good tracer clouds from a sequence of satellite images, *IEEE T. Geoscience and Remote Sensing* **43**(4) (2005), 813–818.

[6] M.A.T. Figueiredo and A.K. Jain, Unsupervised learning of finite mixture models. *IEEE Trans. on PAMI* **24**(3) (2002), 381–396.

[7] C. Biernacki, G. Celeux and G. Govaert, Assessing a mixture model for clustering with the integrated completed likelihood, *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(7) (2000), 719–725.

[8] A. Roy, S.K. Parui, A. Paul and U. Roy, A color based image segmentation and its application to text segmentation, *Proc. of Ind. Conf. on Computer Vision, Graphics & Image Processing*, (2008), 313–319.

[9] S. Di Zenzo, L. Cinque and S. Levialdi, Run-Based Algorithms for Binary Image Analysis and Processing, *IEEE Trans. Pattern Anal. Mach. Intell.* **18**(1) (1996), 83–89.

[10] B.B. Chaudhuri and U. Pal, A complete printed bangla ocr system, *Pattern Recognition* **31** (1998), 531–549.

[11] S.K. Parui, U. Bhattacharya, A. Datta and B. Shaw, A database of handwritten bangla vowel modifiers and a scheme for their detection and recognition, *Proc. of Workshop on Computer Vision Graphics and Image Processing*, (2006), 204–209.

[12] T.K. Bhowmik, P. Ghanty, A. Roy and S.K. Parui, Svm-based hierarchical architectures for handwritten bangla character recognition, *Int. Journ. on Doc. Anal. and Recog. (IJDAR)* **12** (2009), 83–96.

[13] X. Chen and A. Yuille, Detecting and reading text in natural scenes, *Proceedings of IEEE Conference of Computer Vision and Pattern Recognition (CVPR)* **2** (2004), 366–373.

[14] J.-J. Lee, P.-H. Lee, S.-W. Lee, A. Yuille and C. Koch, Adaboost for text detection in natural scene, *Proceedings of International Conference of Document Analysis and Recognition (ICDAR)* (2011), 429–434.

[15] B. Epshtein, E. Ofek and Y. Wexler, Detecting text in natural scenes with stroke width transform. *Proceedings of IEEE Conference of Computer Vision and Pattern Recognition (CVPR)* (2010), 2963–2970.

[16] C. Yi and Y. Tian, Text string detection from natural scenes by structure-based partition and grouping, *IEEE Trans. on Image Processing* **20**(9) (2011), 2594–2605.

[17] C. Yi and Y. Tian, Localizing text in scene images by boundary clustering, stroke segmentation, and string fragment classification, *IEEE Trans. on Image Processing* **21**(9) (2012), 4256–4268.

[18] Y.-F. Pan, X. Hou and C.-L. Liu, A hybrid approach to detect and localize texts in natural scene images, *IEEE Trans. on Image Processing* **20**(3) (2011), 800–813.

[19] A. Shahab, F. Shafait and A. Dengel, ICDAR robust reading competition challenge 2: Reading text in scene images, *Proceedings of the International Conference of Document Analysis and Recognition* (2011), 1491–1496.

[20] A. Shahab, F. Shafait and A. Dengel, A head-mounted device for recognizing text in natural scenes, *Proceedings of the Fourth International Workshop CBDAR*, Beijing, China 2011, pp. 29–41.

[21] L. Neumann and J. Matas, Real-time scene text localization and recognition, *Proceedings of the IEEE Conference of Computer Vision and Pattern Recognition (CVPR)* (2012), 3538–3545.

[22] K. Jung, K. Kim and A. Jain, Text information extraction in images and video: A survey, *Pattern Recognition* **37**(5) (2004), 977–997.

[23] X.-C. Yin, S. Yin, K. Huang and H.-W. Hao, Robust text detection in natural scene images, *IEEE Trans. Pattern Anal. Mach. Intell* **36**(5) (2014), 970–983.

[24] R. Ghoshal, A. Roy and S.K. Parui, Recognition of Bangla text from Scene Images through Perspective Correction. *Proc. of International Conference on Image Information Processing (ICIIP)*. (2011).