Text Extraction from Scene Images Through Local Binary Pattern and Business Features Based Color Image Segmentation

Ranjit Ghoshal, Anandarup Roy, Bibhas Ch. Dhara and Swapan K. Parui

Abstract This article proposes a scheme for automatic extraction of text from scene images. First, we apply a color image segmentation algorithm to a scene image. To improve the color image segmentation performance, we incorporate local binary pattern (LBP) and business features within it. Local Binary Pattern (LBP) operator is a texture descriptor for grayscale images. On the other hand, the business feature describes the variation in intensity. The segmentation procedure separates out certain homogenous connected components from the image. We next inspect these connected components in order to identify possible text components. Here, we define a number of shape based features that distinguish between text and non-text connected components. Our experiments are based on the ICDAR 2011 Born Digital data set. The experimental results are satisfactory.

Keywords Scene image · Color image segmentation · Local binary pattern · Text identification

1 Introduction

Automatic identification and recognition of text in a scene image is useful to the blind and foreigners with language barrier. Moreover, segmentation of such text portions have a fundamental impact on content based image retrieval, document

R. Ghoshal (✉)
St. Thomas' College of Engineering and Technology, Kolkata 700023, India
e-mail: ranjit.ghoshal.stcet@gmail.com

A. Roy · S.K. Parui
CVPR Unit, Indian Statistical Institute, Kolkata, India
e-mail: roy.anandarup@gmail.com

S.K. Parui
e-mail: swapan@isical.ac.in

B.Ch.Dhara
IT Department, Jadavpur University, Kolkata, India
e-mail: bibhas@it.jusl.ac.in

J.K. Mandal et al. (eds.), *Information Systems Design and Intelligent Applications*,
Advances in Intelligent Systems and Computing 340,
DOI 10.1007/978-81-322-2247-7_49

Fig. 1 Sample images (**a**, **b**) and corresponding ground truth images (**c**, **d**) from Born-Digital image data set

processing, intelligent transport systems and robotics. In case of Born-digital images, text is superimposed by a software. Born-digital images are applied in web pages and e-mail as, logo, name or ads. In Fig. 1 two sample images from Born-Digital image dataset are shown. The resolution of the text present in the image and anti-aliasing of text are the major dissimilarities between scenic and born-digital images. There have been several studies on text extraction in the last few years. Wu et al. [1] used a local threshold method to extract texts from gray image blocks containing texts. By considering that texts in images and videos are always colorful, Tsai and Lee [2] developed a threshold method using intensity and saturation features to extract texts in color document images. In recent years, Jung et al. [3] employed a multi-layer perceptron classifier to discriminate between text and non-text pixels. A sliding window scans the whole image and serves as the input to a neural network. High probability areas inside a probability map are considered as candidate text regions. Wavelet transform has also been applied for text identification. In this context Gllavata et al. [4] considered wavelet transform and K-means based texture analysis for text detection. More recently Bhattacharya et al. [5] proposed a scheme based on analysis of connected components (CCs) for extraction of Devanagari and Bangla texts from camera captured natural scene images. In the present article, we first apply Fuzzy c-means (incorporating local binary pattern (LBP) and business features) based clustering algorithm on the color images. With the assumption that text portions are homogeneous in color and lightness and different clusters may contain text portions as different connected components. The next step is the study of these connected components. We define some shape based features that are used to distinguish between text and non-text connected components. We use the ICDAR 2011 Born Digital data set for our experiment.

2 Color Image Segmentation

Color image segmentation is our first step of text extraction. The fuzzy c-means algorithm is used for color image segmentation. Before applying fuzzy c-means we extract some features from the normalized RGB image. Let us consider a pixel p_i of

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

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. Now, we take the business feature [6] for each pixel p_i. The business feature of p_0 is denoted by B_0 that takes into account the variation in intensity and it is computed as:

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

where $p_1, \ldots, p_7$ are the eight neighboring pixels of p_0 (shown in Fig. 2). In general, let the business feature for the pixel p_i be B_i. Besides, these three color values and the business feature we consider another feature, i.e., local binary pattern (LBP) [7] value (γ_i) for each pixel p_i. Local Binary Pattern (LBP) operator is a texture descriptor for grayscale images. Texture in a 2D grayscale image is a phenomenon which consists of spatial structure (pattern) and contrast (amount of texture). LBP operator quantizes the pattern of texture. We assume that a text image has different texture regions due to different backgrounds and texts. So, we consider the text segmentation problem as a texture segmentation problem where segmentation process will partition the image into regions based on their texture. LBP is one of the most popular local image descriptors for texture analysis. The LBP operator describes each pixel (c) in an image with a certain binary pattern by calculating the difference of the gray values from a central pixel g_c around its 3 × 3 neighbourhood (W). If the difference of gray values between a neighbouring pixel and the central

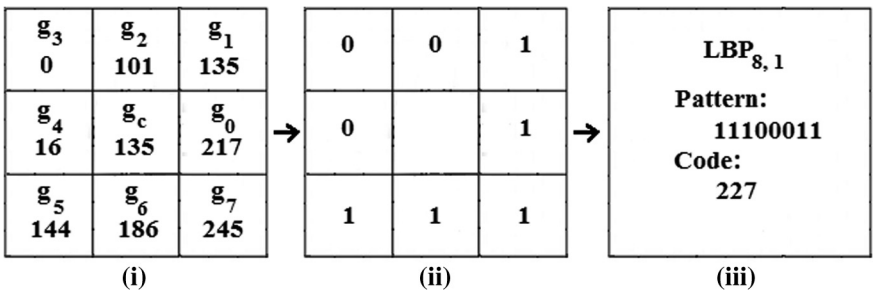

Fig. 3 Example for calculating the local binary pattern (*LBP*) code

pixel is greater than or equal to zero, the value is set to one, otherwise set to zero (see in Fig. 3). Formally, the LBP operator is defined as follows:

$$LBP_{P,W}(c) = \sum_{p=0}^{P} s(g_p - g_c)2^p,$$

where $s(x)$ is 1 if $x \geq 0$ and 0, otherwise and $g_p(p = 0, 1, \ldots, P-1)$ correspond to the gray values of P equally spaced pixels on a 3×3 neighbourhood (W). For each pixel p_i, we consider this as γ_i. Combining, the feature vector ($\mathbf{f}_i$) corresponding to the pixel p_i is: $\mathbf{f}_i = (r_i, g_i, b_i, B_i, \gamma_i)$. These features are sent to the fuzzy c-means clustering algorithm.

3 Connected Component Analysis and Text Extraction Methodology

The color image segmentation produces a number of connected components (CCs) that are spread over several clusters. These connected components include the possible text components. So, we now analyze these connected components to identify text portions. We assume, a single text component is homogeneous in terms of color and lightness. This assumption ensures that a single text component is not broken after clustering. Now, after segmentation, the text parts may make one single cluster. However, more generally, one cluster contains non-text components along with some text components. In order to separate them we proceed as follows. We observe that sufficiently small and large components do not contribute much for text identification. Small connected components generally represent noise and the large connected components are background. So such components are first removed. Moreover, we observe that text like patterns do not generally touch the image boundary. So, we remove all boundary attached connected components using morphological operations. Further, the following criteria are used to extract text portions. All the concerned connected components are subjected to these condition in the same sequence as given. Note that the thresholds may vary if the sequence is altered.

1. *Elongatedness ratio* (ER): The text like components are usually elongated. The elongatedness ratio (ER) is defined as follows:

$$\text{ER} = \frac{\text{total number of boundary pixels}}{\sqrt{(\text{total number of pixels in the CC})}} \tag{2}$$

   Empirically it is found that a component is a text symbol if $2 \leq ER \leq 10$.
2. *Number of complete lobes* (CL): Using Euler number complete lobes can be obtained. Found empirically, a text symbol has less than 15 complete lobes. We simply use lobe, hereafter, since all concerned lobes are complete.

3. *Aspect ratio* (AR): The aspect ratio of a text component is defined as: $AR = \min\{(height/width), (width/height)\}$ and for a text component its value lies in a compact range. Non-text components generally have irregular shapes, hence their aspect ratio falls outside the range. We empirically found that the AR value of a text component satisfies $0.1 \leq AR \leq 1$.
4. *Object to background pixels ratio* (OBR): Due to the elongated nature of text, only a few object pixels fall inside text bounding box. On the other hand, elongated non-texts are usually straight lines, and contribute enough object pixels. We observed $0.02 \leq OBR \leq 0.1$ or $0.95 \leq OBR \leq 1.0$ for text symbols.
5. *Axial ratio* (AXR): Axial ratio (AXR) of any shape is the ratio of the lengths of the two axes. Here, we calculate the *Major Axis Length* and *Minor Axis Length* of a connected component. The AXR is defined as:

$$AXR = \frac{\text{Major Axis Length}}{\text{Minor Axis Length}} \tag{3}$$

   It is empirically found that the AXR value of a text component is less than 7.
6. *Thickness* (TH): Thickness (TH) is defined by Ghoshal et al. [8] of a connected component is calculated as: let h_i and v_i be the horizontal and the vertical run lengths of a pixel p_i at the ith position of a component CC_j. We next compute the minimum of h_i and v_i and further constitute a set $\text{MIN}_j = \{m_i$ such that $m_i = \min(h_i, v_i), \forall i\}$. Thus MIN_j denotes the set of all the minimum run lengths considering all the pixels of CC_j. The thickness TH_j of the component CC_j is defined as the element whose frequency is maximum in the set MIN_j. We have observed $5 \leq TH \leq 20$ can successfully identify text symbols.

Since, our concerned Born Digital dataset provides ground truth only for the text components. In Fig. 1c and d, we have shown two ground truth images corresponding to the input images of Fig. 1a and b. In fact, it is rather unrealistic to assume any prior shape of the non-text components. Thus, from extensive experiments using the ground truth of text-components, we determined the threshold values.

4 Results and Discussion

The experimental results are based on ICDAR 2011 Born Digital data set [9]. This data set contains 420 training images and 102 test images. Born-digital images are inherently low-resolution in order to be transmitted online efficiently. Therefore, extracting text from born-digital images is an interesting research work. First, we apply Fuzzy c-means algorithm for segmenting the images into a set of homogenous connected components. We observe that our segmentation could preserve the text like components. Now, let us first consider the "molex" image (Fig. 4a) as an example. Figure 4b shows the result after performing color image segmentation. Individual clusters are shown in Fig. 4c–g. Figure 4h represents the result obtained after applying

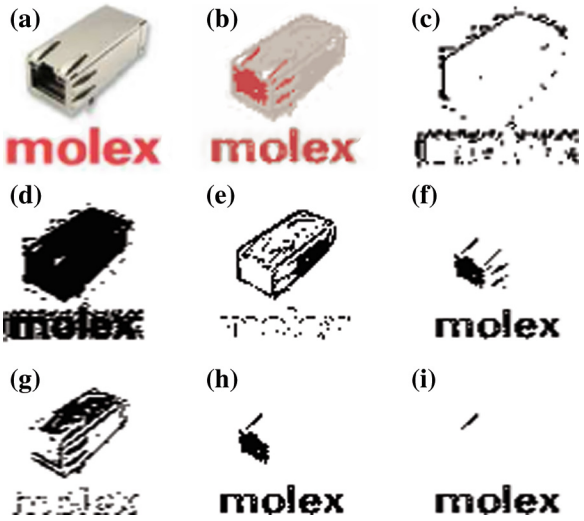

Fig. 4 Steps of the proposed text extraction method: **a** input image. **b** Segmented image. **c–g** Individual clusters. **h** Result obtained after applying criteria 1–5. **i** Result obtained after applying thickness feature (criterion 6)

a series of criteria (i.e., sufficiently small and large components, boundary attached components, elongatedness ratio, number of complete lobes, aspect ratio, object to background pixels ratio and axial ratio). Here, the feature "to remove boundary attached connected component" does not occur because no boundary attached connected components are present in the individual clusters. Finally, after applying the feature "thickness" we obtain the result shown in Fig. 4i. Here, we notice that all the text components are successfully identified along with one non-text component. Finally, after applying our set of features sequentially, we can separate out the text components. In Table 1, some of the original images (first and third rows) and the extracted text components from them (second and forth rows) are shown. We observe from Table 1 that often some non-text components are included and some text components are still missing. However, during segmentation, some of the text components are not separated from the non-text portions. Such text portions are essentially not included in the extracted text components. Thus, to evaluate our text extraction technique, we consider the final text extraction images of training set with the ground truth and compute the precision, recall and F-measure. The performance evaluation is based on true positive (TP), false positive (FP) and false negative (FN) pixels in order to calculate recall and precision metrics.

- A pixel is classified as TP if it is ON in both Ground Truth (GT) and binarization result images.
- A pixel is classified as FP if it is ON only in the binarization result image.
- A pixel is classified as FN if it is ON only in the GT image.

The recall metric shows the ratio of the number of pixels, which our method truly classified as foreground, to the number of all pixels classified as foreground from the ground truth image. Precision metric is the ratio of the number of pixels, which our method truly classifies as foreground, to the number of all pixels which

Table 1 Some images (first row, third row) and the corresponding extracted text components (second row, fourth row)

classified as foreground. Setting C_{TP} as the number of TP pixels, C_{FP} as the number of FP pixels and C_{FN} as the number of FN pixels, recall (RC) and precision (PR) metrics are given as follows:

$$RC = \frac{C_{TP}}{(C_{FN} + C_{TP})}, \quad PR = \frac{C_{TP}}{(C_{FP} + C_{TP})} \tag{4}$$

Recall and Precision metric have values between zero and one. These metrics are closer to one for better results. The overall metric that is used for evaluation is the F-measure (FM) which is calculated as follows:

$$FM = (2 \times RC \times PR)/(RC + PR) \times 100\,\% \tag{5}$$

The recall, precision and F-measure values of our algorithm obtained on the basis of the training set of ICDAR 2011 Born Digital data set images are respectively 66.23, 64.53 and 65.37 %. On the other hand, for the test set, we apply the evaluation criteria of Clavelli et al. [10]. They proposed a number of measures to assess the segmentation quality of each of the text-parts defined in the ground truth. Here, we obtain the recall, precision and F-measure values respectively 63.31, 61.73 and 62.51 %.

5 Summary and Future Scope

This article provides an automatic identification of text entities embedded in scene images. It is based on color image segmentation followed by extraction of several connected component based features that lead towards identification of text

components. The proposed technique is not very sensitive to image color, text font, skewness and perspective effects. The results obtained on ICDAR 2011 Born Digital data set are quite satisfactory. It can be extended to extract texts present in scanned document images also. In future, we plan to study the use of machine learning tools to improve the performance.

References

1. Wu, V., Manmatha, R., Riseman, E.M.: Textfinder: an automatic system to detect and recognize text in images. IEEE Trans. Pattern Anal. Mach. Intell. **21**(11), 1224–1229 (1999)
2. Tsai, C., Lee, H.: Binarization of color document images via luminance and saturation color features. IEEE Trans. Image Process. **11**(4), 434–451 (2002)
3. Jung, K., Kim, I.K., Kurata, T., Kourogi, M., Han, H.J.: Text scanner with text detection technology on image sequences. In: Proceedings of International Conference on Pattern Recognition, vol. 3, pp. 473–476 (2002)
4. Gllavata, J., Ewerth, R., Freisleben, B.: Text detection in images based on unsupervised classification of high frequency wavelet coefficients. In: Proceedings of International Conference on Pattern Recognition, vol.1, pp. 425–428 (2004)
5. Bhattacharya, U., Parui, S.K., Mondal, S.: Devanagari and Bangla text extraction from natural scene images. In: Proceedings of the International Conference on Document Analysis and Recognition, pp. 171–175 (2009)
6. Mandal, A.K., Pal, S., De, A.K., Mitra, S.: Novel approach to identify good tracer clouds from a sequence of satellite images. IEEE Trans. Geosci. Remote Sensing. **43**(4), 813–818 (2005)
7. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 1224–1229 (2002)
8. Ghoshal, R., Roy, A., Bhowmik, T.K., Parui, S.K.: Decision tree based recognition of Bangla text from outdoor scene images. In: Proceedings of the 18th International Conference on Neural Information Processing, pp. 538–546 (2011)
9. Karatzas, D., Robles Mestre, S., Mas, J., Nourbakhsh, F., Roy, P.P.: ICDAR 2011 robust reading competition-challenge 1: reading text in born-digital images (web and email). In: Proceedings of 11th International Conference of Document Analysis and Recognition (2011)
10. Clavelli, A., Karatzas, D., Lladós, J.: A framework for the assessment of text extraction algorithms on complex colour images. In: Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, DAS'10. ACM, pp. 19–26 (2010)