Text Detection in Natural Scene Images

U.Elakkiya
Assistant Professor, Department of Information Technology, Sri Ramakrishna Institute of Technology,Coimbatore

M.Safa
Assistant Professor, Department of Information Technology, SRM University, Chennai

Abstract — The text detection and localization in scenery images are important for content based analysis. Here the problem is due to complex background, the non-uniform illumination, variations of text, font and line orientation. A hybrid approach is used to detect and localize the text.The text region detector is designed to estimate the text confidence map based on the text components, segmented by local binarization method. A conditional Random Field (CRF) model should consider the unary and binary component relationships and it is presented to label the components as "text "or "non-text". Finally, text components are grouped into text lines with energy minimized method.

Key Terms — text localization, text detection, Conditional Random field (CRF).

I. INTRODUCTION

In the field of computer vision, pattern recognition is the greatest amount of interests shown in content retrieval from images and videos. The content are in kind of objects, color, texture, shape as well as their relationships. The information provided by an image is useful for the content-based image retrieval, as well as for indexing and classification purpose. Since, text data can be embedded in an image or video in different font , styles, size, colors, and against complex background, the problem of extracting the text region becomes a challenging one[4]. The text understanding system encompasses three main steps: text detection, localization and text extraction from background and text recognition. Text extraction from background is dealing mainly with lighting and complex backgrounds. It is a step to prepare data for Optical Character Recognition. Text is an object that has to be extracted properly for better recognition. The final step is text recognition, to convert the characters into ASCII values to understand the text. The aim is that, segmenting the text from background, and isolate the text pixels from the background [2].

With the wide usage of digital image capturing devices such as digital cameras, mobile phones and PDAs, content-based image analysis techniques are receiving intensive attention in recent years. Among all the contents in images, text information has inspired great interests, as it can be easily understood by both human and computer. Though many efforts have been devoted, it remains challenging due to variation of text size, font, and degraded images with noises. The existing methods can be categorized into two classes:region based method and connected component(CC) method.

In Region-based method [3,5], the text regions have distinct characteristics from non-text regions such as structure and texture. It consists of two stages: text detection and text localization. For text detection, the features of local regions are extracted if they contain text. Then specific grouping or clustering approaches are used to localize text regions accurately. In CC-based methods [3, 5] the text can be seen as set of separate connected components, each has distinct intensity and color distributions. These methods consist of three stages: 1) CC extraction to segment CC from images, 2) CC analysis is to determine whether they are text components by heuristic rules or classifiers 3) to group text components into text regions (e.g. lines, words).

In existing methods have some problems difficult to be solved. For Region-based methods [1], the speed is relatively slow and the performances are sensitive to text alignment orientation. In CC-based methods, text components are hard to be segmented without prior information of text position. Designing fast and reliable is also difficult because there are too many text-like components in images. Performance of region based method is sensitive to the text orientation. Most of these methods can localize the text containing many characters in horizontal alignment.

In the proposed system, to overcome these difficulties by using hybrid approaches are robustly used to localize text in natural scene images. A text region detector is designed to generate a text confidence map, based on which components are segmented with local binarization. Then the Conditional Random Field (CRF) model considering of both unary component and neighboring component relationship is presented for component analysis. Finally, the energy minimization based approach is used to group to text components into text lines [2].

The region based methods are based on observations that the text regions have distinct characteristics from non-text regions by distribution of gradient strength and texture properties. This method consists of two stages:1) text detection - is used to estimate the text confidence in local image regions through classification 2) text localization - is used to cluster the local text regions into text blocks, and the text verification method is used to remove the non-text regions for further process. An earlier method, proposed by Wu *et al*.[8] used a set of Gaussian derivative filters to extract texture features from the image region. With the corresponding filter responses, all the image pixels are assigned to one of three classes ("text", "non-text", and "complex background"), then k-means clustering and morphological operators are grouped the text pixels into text regions. Unlike region-based methods are based on observations that text can be seen as a set of connected components, each of which has distinct geometric features, and neighbor components have spatial and geometric relationships. These methods normally consist of three stages: 1) CC extraction to segment text components from images; 2) CC analysis to filter out non-text components using heuristic rules or classifiers; 3) post-processing is to group text components into text blocks(e.g., words and lines).

II. PROPOSED WORK

In the proposed, hybrid approach is used to detect and localize the text in natural scene images. Since local region detection can detect scene texts even in noisy images, a text region detector is used to estimate the probabilities of text position and scale which helps to segment the candidate text components with efficient local binarization algorithm. To combine unary and binary contextual component relationships, a conditional random field (CRF) model with supervised parameter is proposed. Finally, text components are grouped into text lines or words robustly with an energy minimized method. Experimental results on English and multilingual image datasets show that the proposed approach yields higher precision and recall performance compared with state-of-the-art methods. The proposed approach reports higher performance on the ICDAR 2005 text locating competition dataset. It extends in several aspects, i.e., more technical details of the system are given, In more classifiers and learning criteria are evaluated for training the CRF model, the text line grouping method is extended to word level and further evaluation on a multilingual image dataset is added[1].

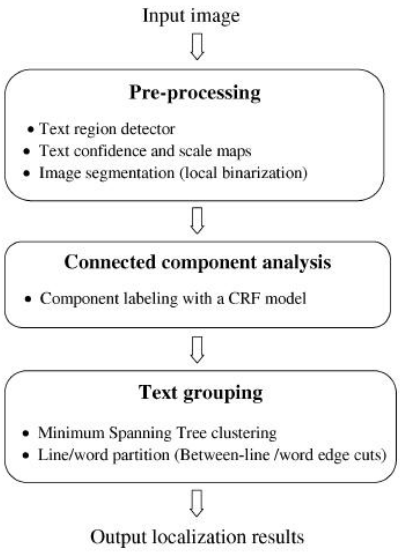

Fig. 2.1 Flowchart of Proposed System

III. IMPLEMENTATION

A. Pre-processing

To extract and utilize the local text region information, a text region detector is designed to estimate the text confidence and the scale based on the text components and that can be segmented and analyzed accurately.

1) Text-Region Detector:

The original colored image is converted into a gray level image, on which the image pyramid with

scale step 1.2 is built with nearest interpolation and to capture text information of different scales. A text region detector is designed by integrating a widely used feature descriptor: HOG[11] and a boosted cascade classifier. In detail, within
each 16 × 16 window sliding in a layer of the image pyramid, define several sub-regions by horizontally and vertically partitioning the window[10].

2) Text confidence and scale maps:

To measure the confidence that one region contains text, we translate the Wald boost [5] output, no matter accepted or rejected into posterior probability based on boosted classifier.

$$P_t(y_i s_i) = \frac{P_t(S_i y_i) P_t(y_i)}{\sum_{y_i} P_t(S_i y_i) P_t(y_i)}$$

$$= \frac{P_t(S_i y_i) P_{t-1}(y_i | accepted)}{\sum_{y_i} P_t(S_i y_i) P_{t-1}(y_i | accepted)}, \qquad (1)$$

By the probability calibration, the text confidence map for each layer of image pyramid is obtained. All the pixel confidence and the scale values are at different pyramid layers are projected back to the original image for calculating the final text confidence and scale maps.

3) Image Segmentation:

To segment candidate CCs from the gray-level image, Nib lack's local binarization algorithm [7] is adopted due to its high efficiency and non-sensitivity for image degrading. The formula to binarize each pixel x is defined as

$$b(x) = \begin{cases} 0, & if\ gray(x) < \mu_r(x) - k.\sigma_r(x); \\ 255, & if\ gray(x) > \mu_r(x) + k.\sigma_r(x)\ ; \\ 100, & otherwise \end{cases} \qquad (2)$$

Where $\mu_r(x)$ and $\sigma_r(x)$ are the intensity mean and standard deviation (STD) of the pixels within a -radius window
centered on the pixel x and the smoothing term k is empirically set to 0.4. After local binarization, assume that each local region, the gray-level values of foreground pixels are higher or lower than the average intensity,

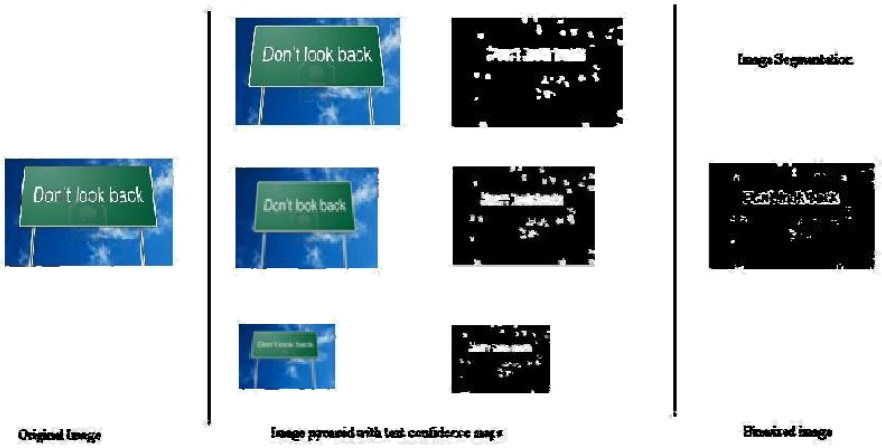

Fig.3.1 Example for Pre-Processing Stage

Candidate text components and those values 100 are not considered further. On the text confidence map, the red

color represents high probability of text and the blue color represents low probability. On the text scale map, each pixel with the text confidence greater than 0.9 is drawn a circle whose radius indicating the estimated text scale[7].

B. Connected Component Analysis

For connected components analysis (CCA), propose a conditional random field (CRF) model to assign candidate components as one of the two classes (either text or non-text) by considering both unary component properties and binary contextual component relationships.

1) Conditional Random Field:

CRF[3] is a probabilistic graphical model which has been widely used in many areas such as natural language processing, computer vision, and handwriting recognition and document analysis. Given a set of observation variables X=($x_{1,\ldots}$, xn) with state variables (labels) Y= (y1… yn). Let G
(V, E) be a graph such that Y is indexed by the vertices V of G. then (X, Y) is a conditional random field when the probability of Y conditioned on X obeys the Markovian property

$$p(_{yi}|X,_{yj},j\neq i)=p(_{yi}|X,_{yj},j\in Ni), \qquad (3)$$

where N_i is neighborhood set of x_i. Based on the Hammersley-Clifford Theorem, the conditional probability can be written in the form

$$p(y|X)=\frac{1}{z}\prod_{c\in C}\varphi(Xc,Yc), \qquad (4)$$

where c is the set of all cliques in the graph, ψ represents the real-valued potential function and Z, defined as $Z=\sum_Y \prod_{c\in C}\varphi(X_cY_c)$, is a normalization factor

CC labeling with CRF:

Based on the definition of CRF, formulate CC analysis into CC labeling problem: given an image containing candidate text component set $X = (x_1, x_2 ...)$, on which a 2D undirected graph is constructed, the objective of the CRF is to find the best component label $Y^* = (y^*_1, y^*_2 ...)$ is to minimize the total graph energy[3].

2) Energy Function:

CCA methods only consider individual component properties, and thus, are prone to misclassifying components when the image is cluttered and noisy, The CRF model is used to explore contextual component relationships as well as unary component properties. Unlike the methods that use third-order cliques, only second-order cliques is considered due to its high efficiency, and this method performs sufficiently because of the effective pre-processing stage and supervised learning of CRF parameters. The total energy functions is defined as

$$E(X, Y, C, \Lambda) = \sum_i E_u(x_i, y_i, \lambda_u) + \frac{\lambda_c}{n_i}\sum_{j\in N_i} E_b(x_i, x_j, y_i, y_j, \lambda_b) \qquad (5)$$

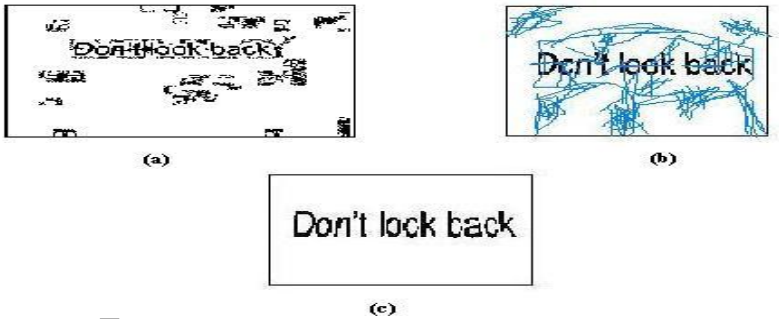

Fig. 3.2 Example for CCA stage. (a) Components passing through unary classification thresholds. (b) Component neighborhood graph. (c) Text components after CRF labeling.

Base Classifiers:

Different types of classifiers has been used for approximating the energy functions of CRFs, such as generalized linear model, Gaussian mixture model, Boosting and SVM. Two types of artificial neural network (ANN) classifiers are chosen as base classifiers: single-layer perceptron (SLP) and multi-layer perceptron (MLP).

Unary Component Features:

To characterize the single component's geometric and textural properties, six types of unary component features are used such as Normalized width and height, Aspect ratio, Occupation ratio, Confidence, Compactness, Contour gradient.

Binary component Features:

To characterize the spatial relationship and geometric and textural similarity between two neighboring component and, six types of binary component features are used as Shape difference, Spatial distance, Overlap ratio, Ordering confidence, Scale ratio, Gray-level difference.

Jointly supervised Training:

An advantage of CRF, compared with other graphical models such as MRF, is that the model parameters can be estimated by supervised training. If the base classifier (for evaluating energy functions) can be trained by gradient descent, its parameters can be estimated jointly with the CRF. In CRF criterion to learn the parameters is the conditional log-likelihood (CLL), to maximize the conditional probability p (Y|X). It defined as

$$\mathbf{L_{cll}(\Lambda)= -log[p(Y\ X)]=E(X, Y, C, \Lambda) + log(Z).} \qquad \mathbf{(6)}$$

Two alternative criteria are used such as saddle point approximation and LOG-likelihood of Margin.

Saddle point Approximation:

SPA is a measurement to approximate the probability distribution with the Maximum a posteriori (MAP) estimation which has been used to overcome the computational intractable problem.

LOG-Likelihood of Margin (LOGM):

Another criterion is LOGM recently proposed for learning prototype-based classifiers, and can be generalized to the learning of structural models. The LOGM criterion is the logarithm of smoothed classification error, differing from the minimum classification error (MCE) criterion only by the logarithm, which makes the loss to be convex with respect to the hypothesis margin. The LOGM criterion is different from the SPA in that it approximates the normalization factor Z by summing the energy of the truth label (Y^c) and the minimal energy among the incorrect labels (Y^t).

The negative log-likelihood (l_{LOGM}) can be computed by

$$\mathbf{l_{LOGM}= -log\ [1-\sigma\ [E(X, Yc, C, \Lambda)-E(X, Y^r, C, \Lambda)]]} \qquad \mathbf{(7)}$$

C. Text Grouping

To group text components into text lines, we presented a learning based method by building neighboring components into minimum spanning tree(MST) and cutting off inter-line edge with an energy minimization model.

1) MST Building:

Based on the observations, cluster the text components into tree with MST based on a learned distance metric, which is defined between two components as a linear combination of some features as

$$\mathbf{Dist_{metric}(x_i, x_j) =W_{dist} .F_{ij}} \qquad \mathbf{(8)}$$

where Fij is the vector of between-component features and W_{dist} is the vector of combining weights, which is learned by the mean squared error on a dataset of component pairs labeled in two classes: within line/word and extra line/word. With the learned metric, the MST can be efficiently constructed with the Kruskal algorithm[6]. Define a linear distance metric whose parameters are learned with the perceptron algorithm to estimate the similarity measurement between two components, where features are defined as Table 3.1

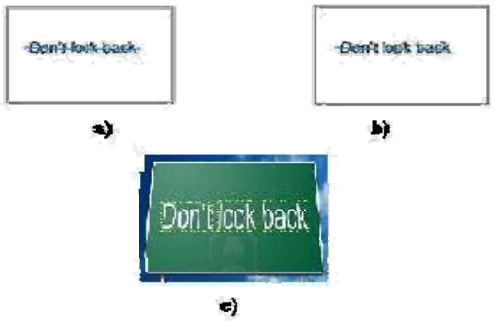

Fig. 3.3 Example of the text grouping stage. (a) Building component tree with MST. (b) Textword partition. (c) Text localization results.

Distance metric feature	Text line feature
centroid distance (horizontal and vertical)	line regression error
	line height
box boundary distance (horizontal and vertical)	line number
	cut edge score
shape difference	inter-line distance (horizontal and vertical)
color difference	

Table 3.1 Distance metric and text line features.

2) Edge Cut:

With the initial component tree built with the MST algorithm, between line/word edges need to cut to partition the tree into subtrees, each of which corresponds to a text unit (line or word)[12].

Line Partition:

To formulate the edge cutting in the tree as a learning-based energy minimization problem. In the component tree, each edge is assigned one of two labels:―linked‖ and ―cut‖, and each edge is subtree corresponding to a text line is separated by cutting the ―cut‖ edges. The objective is t_i find the optimal edge labels such that the total energy of the separated subtrees is minimal. The total text line energy is defined as

$$E(L) = \sum_{i=1}^{N} W_{line} \cdot F_i \qquad (9)$$

Where N is the number of subtrees(text lines), F1 is the feature vector of a text line, and Wtime is the vector of combibing weights.

Assume that the initial component tree is split into N candidate lines L= (l_1, …. ,l_N) by cutting edges E= (e_1,e_2,……e_{N-1}), the text line features are defined as follows:

―**Line regression error**. Each text line is approximately straight, the sum of line regression errors (*error*)is used to measure this regularity.

―**Cut score**. The value of learned distance metric ($dist_{metric}$) is a good measurement of the similarity between two components. The sum of logarithm of the N-1 cut edge distances..

―**Line Height.** The line height is defined as the sum of the candidate text line distance ($dist_{base}$) between the top and bottom component centroids (xc_{top},xc_{bottom}) along the orthogonal orientation of the base line normalized by the line height.

―**Spatial distance.** The Euclidean distance (*dist*) between the centroids of two components linked with the cut

edge reflects the distance between text lines. The sum of N-1 distance is taken.

—**Bounding Box distance.** The spatial distance between the bounding boxes of two components linked with the cut edge is defined as another feature for characterizing the between-line spatial relationship.

The distance is defined as $dist_{box}(e_i) = \sqrt{bw_i^2 + bh_i^2}$, where bw_i and bh_i denote the distance between the closest left-right and top-bottom borders between two boxes.

—**Line number.** To avoid over-splitting the component tree, text line number N is used is used as regularization.

The text lines are tabulated in Table 3.2

3) *Word Partition:*

The difference in the word-level features, which as defined as:1) word number; 2) components centroid distances of cut edges; 3) component bounding box distances of the cut edges; 4)bounding box distances between words separated by cut edges; 5) ratio between the component centroid distance of the cut edge within separate words.

Finally text words are corresponding to partitioned subtrees can extract and the ones containing too small components are removed as noises

Feature	Definition
Line regression error	$LRE = \sum_{i=1}^{N} error_i$
Cut score	$CS = \sum_{i=1}^{N-1} \log[dist_{metric}(e_i)]$
Line height	$LH = \sum_{i=1}^{N} dist_{base}(xc_{top}, xc_{bottom})$
Spatial distance	$SD = \sum_{i=1}^{N-1} dist(e_i)$
Bounding box distance	$BBD = \sum_{i=1}^{N-1} dist_{box}(e_i)$
Line number	$LN = N$

Table 3.2.Text line features for N cut Edges

IV.CONCLUSION

To localize scene texts by integrating the region information into a robust CC-based method. The binary contextual component relationships, the unary component properties, are integrated in a CRF model, whose parameters are jointly optimized by supervised learning. The aspects in this methods are region-based is very helpful for text component relationships and the CRF model differentiates the text components from the non-text components is better than classifiers, learning-based energy minimization method can group text components into text lines..

REFERENCES

[1] Yi-Feng Pan, Xinwen Hou, and Cheng-Lin Liu ‗A Hybrid Approach to Detect and Localize Texts in Natural Scene Images', *IEEE Transactions on image processing*, vol. 20, no. 3, pp. 800-813,Mar 2011

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

[3] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling Sequence data. In *Proc. Eighteenth Int.l Conf. Machine Learning (ICML'01)*, pages 282–289, San Francisco, USA,2001.

[4] J. Liang, D. Doermann, and H. P. Li. Camera-based analysis of text and documents: a survey. *Int'l J. Document Analysis and Recognition*, 7(2-3):84–104, 2005.

[5] J. Sochman and J. Matas. Wald boost – learning for time constrained sequential detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'05)*, pages 150–156, San Diego, USA, 2005.

[6] R. Sedgewick. *Algorithms in C, Part 5: Graph Algorithms*,Third Edition. Addison-Wesley Professional, 2001.

[7] W. Niblack. *An Introduction to Digital Image Processing*. Strandberg Publishing Company, Birkeroed, Denmark,1985

[8] V.Wu, R. Manmatha, and E. M. Riseman, ―Finding text in images,‖ in *Proc. 2nd ACM Int. Conf. Digital Libraries (DL'97)*, New York, NY, 1997, pp. 3–12.

[9] H. Takahashi, ―Region graph based text extraction from outdoor images,‖ in *Proc. 3rd Int. Conf. Information Technology and Applications (ICITA'05)*, Sydney, Australia, 2005, pp. 680–685.

[10] Zhu K. H., Qi F. H, Jiang R. J., Xu L., Kimachi M., Wu Y., and Aizawa T., ―Using Adaboost to detect and segment characters from

natural scenes,‖ in Proc. 1st Conf. Caramera Based Document Analysis and Recognition (CBDAR‘05), Seoul, South Korea, 2005, pp. 52–59.

[11] Dalal N. and Triggs B (2005) ‗Histograms of oriented gradients for human detection‘ in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR‘05), San Diego, CA, pp. 886–893

[12] D.-Q. Zhang and S.-F. Chang, ―Learning to detect scene text using ahigher-order MRF with belief propagation,‖ in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops RW'04),*Washington, DC, 2004, pp. 101–108.