REVIEW: ROBUST APPROACH TO DETECT AND LOCALIZE TEXT FROM NATURAL SCENE IMAGES

Khushbu C. Saner [1], Mrs. Neeta. A. Deshpande[2]

[1]Student M.E.Computer, Matoshri College of Engineering & Research Center, Nashik, India, khushbusaner@gmail.com

[2]Associate professor, Matoshri College of Engineering & Research Center, Nashik, India, deshpande_neeta@yahoo.com

ABSTRACT

In natural scene large amount of information is hidden which can be automatically recognized and processed. To get this automatic segmentation, detection and recognition of visual text entities in natural scene images should be done. For extracting the text from natural scene images, robust approaches like combination of algorithm such as by combining the edge and connected component based are adopted. Efficient outcome is the result of combining. The connected component based algorithm is more robust to scale and lighting condition as compared to edge based. In literature unary component properties and binary contextual component of conditional random field model are preferred to detect non-text components. In region based method, the speed is relatively slow and performance is sensitive to text alignment orientation and for connected component based methods cannot segment text component accurately. Combination of these two models can achieve a far more robust algorithm, which would be invariant to scale, lighting and orientation changes. This robust approach will fail to segment text. The paper emphasizes a strong focus on four methods of detecting and localizing the text already proposed in the literature. A brief review along with comparative analysis and future scope of the algorithm are focused in the paper

Keywords: Connected component, region based, text detection, text localization, text extraction

ISSN 2229-5518

1 INTRODUCTION

With the increasing use 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, since it can be easily understood by both human and computer, and finds wide applications such as license plate reading, sign detection and translation, mobile text recognition, content-based web image search.

For detecting text from natural images have many method. Here we consider only region based and connected component based method. According to Yi-Feng Pan, Xinwen Hou, and Cheng-Lin Liu [1] the existing methods of text detection and localization can be roughly categorized into two groups: region-based and connected component (CC)-based.

A region-based method consists of following stages:

1) Text detection to estimate text existing confidence in local image regions by classification,
2) Text localization to cluster local text regions into text blocks,
3) Text verification to remove non-text regions for further processing.

And a connected component based method consist of following stage

1) CC extraction to segment candidate text components from images;
2) CC analysis to filter out non-text components using heuristic rules or classifiers;
3) Post-processing to group text components into text blocks (e.g., words and lines).

These paper methods are categorized according to which technique will be used for detection and localization. That is CC based method and region based method. These methods are complementary to each other. If we merge this technique we get the robust output. Remaining section arrange as follow 2 Techniques/Method for text detection and localization 3 Conclusion.

2 TECHNIQUES FOR TEXT DETECTION AND LOCALIZATION

Following subsection gives overview of already implemented text detection and localization method.

2.1 Comprehensive method

Michael R. Lyu and Min Cai [2] implemented comprehensive method to detect text from multilingual video. For efficient video text detection, localization, and extraction method, this emphasizes the multilingual capability over the whole processing. This method is also robust to various background complexities and text appearances. The text detection is carried out by edge detection, local thresholding, and hysteresis edge recovery. The coarse-to-fine localization scheme is then performed to identify text regions accurately. The text extraction consists of adaptive thresholding, dam point labeling, and inward filling.

Steps for text recognition in video

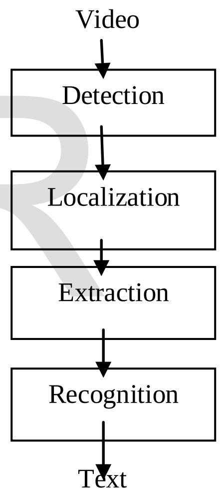


Fig 1: Steps for text recognition

Multilingual text characteristic classified as follow

a) Language Independent Characteristics
   It contains contrast, color, orientation and stationary location.
b) Language Dependent
   Such as stroke density, font size, aspect ratio and stroke statistic.

2.1.1 Working:

Main problem in multilingual text detection is large font size range. For this region researcher added sequential multiresolution paradigm. They use Sobel detector [6] to detect edges for two reasons: a) it is isotropic, so that the strokes of all directions are equally treated and b) it generates

ISSN 2229-5518

double edges, which make text areas much denser than non text areas.

Algorithm for sequential multiresolution paradigm.

1) Sampled frames from video segment per second
2) Convert into 256 –level gray scale image
3) Using Sobel detector to detect edge from image
4) Calculate edge strength using level to highlight edge and corner
5) Text detection
6) Text localization
7) Text region scale up with f(l) to generate working image with desired resolution
8) Get original co-ordinate of text region

Above algorithm gives single pass or single level of detecting original co-ordinate of text region. To get the final text region, the algorithm has to execute at each level and then merge the all result of each pass. Before the final text region in a frame are fed to the text extraction algorithm, a multiframe verification process take place to eliminate those regions that are transitory or have already been detected in the previous frames. Therefore, the text extraction only processes the first-emerging text regions to produce the binary OCR-ready text images. Now we will see the how the text detection, text localization and multiframe verification will be performed.

2.1.1.1 Text Detection Algorithm:

Here they used two concentric squares to scan the edge map scaled for the current level, called the EMP. The lager square called 'window' and smaller one called 'kernel'. The height of kernel is h and that of window is 3h.

Assume horizontal orientation.

1) Scan the edge map kernel by kernel
2) Selection criteria of h are related to the minimum (8) and maximum (24) font to be handled in a level. Check kernel < 8 and window < 24 then set h=10.
3) To calculate local threshold for kernel, analyze the back ground complexity and the edge strength histogram inside the window. The local threshold for the kernel T_{kernel} is determine as follow[2]

$$T_{kernel} = \begin{cases} S_{max} & \sum_{i=1..3h} P_i < \varepsilon \\ T_{low} & MAX\_NUM\_BL\_ROWS \geq MIN\_SPC \\ T_{high} & MAX\_NUM\_BL\_ROWS < MIN\_SPC \end{cases} \quad (1)$$

Set $T_{kernel} = S_{max}$ to suppress all edges pixel in the kernel

T_{low} and T_{high} are determined from the edge strength histogram of the window. And rang of edge strength = 64 bins for quick computation.

4) To get some text edges back use text like area recovery filter
5) Text like area recovery occurred in two step
   a) Text like edge labeling
   b) Hysteresis edge labeling

a) Text like edge labeling:

Result of local threshold is store in new edge map, called EMP'

i) All edge pixel in the EMP' are labeled as "NON-TEXT"
ii) Employ a 10 x 4 rectangle to scan the EMP' row by roe stepping 5 x 2.
iii) If edge density in current rectangle is > ζ, labeled as "TEXT".

   ζ= 100 Chinese text region, 100 English region and 100 non text region on different background.
iv) Suppressed all edge pixel with the "NON TEXT" label.
v) For each "TEXT" edge pixel, a hysteresis thresholding mask is applied to bring back some lower-contrast edge pixels in its neighborhood.
vi) Derived mask using stroke-blurred model.

b) Hysteresis edge labeling:

After text like edge labeling they get the strong edge pixel is surrounded by some lower strength edge pixels.

i) 5 X 5 masks are derived to create a hysteresis threshold surface around the strong edge pixel.

ii) Let H(i , j)($-2 \leq i , j \leq 2$) = mask

iii) If H= centered at a "TEXT" then use Hysteresis edge labeling

6) Non text edges are removed to highlight text area.
7) Get the EMP' contains highlight text area.

From this it is clear that text like area recovery enhances the edge density of text area and removes spares edges Here text detection completed and can perform text localization on EMP'.

2.1.1.2 Text Localization:

They proposed coarse-to-fine localization [5] method to handle both multilingual texts and complex layouts. For this they used both horizontal and vertical projection method

Whole EMP' is the only region to be processed.

R_0= First region in the processing array.

Algorithm:

First Check if R_0not vertically divisible then is used as a whole for the vertical projection

Else all sub region of R_0 enter the vertical projection. If an input region of vertical projection is horizontally divisible then add all its sub regions are added to the processing array

1) Select the first region from the processing array if the processing array is empty then terminate.
2) Perform horizontal projection inside the region
3) Check if it is divisible then each sub region perform step 4

   Else the region perform step 4
4) Perform vertical projection
5) Check if it divisible then each sub region added to the processing array

   Else check the valid aspect ratio for the indivisible region then add to the resulting text

   Else discard the false region.
6) Go to step 1)

2.1.1.3 Multiframe Verification:

The main aim of multiframe verification is to eliminate those text regions that are transitory or have already been detected in the previous frames.

1) Check whether a text region is "new" or "old" first depends on the location comparison between the current one and those on the previous frame.
2) If two text regions on consecutive frames have same locations, we then use the signature comparison to determine whether they are the same text or not.
3) Signature distance metric is designed to tolerate possible location offsets and edge density changes of the same text over time.

2.1.1.4 Text Extraction:

Here text extraction means to convert the grayscale image in an accepted text region into the OCR-ready binary image, in that all pixels of characters are in black and others are in white.

The main problems of text extraction come from three aspects:

1) The unknown color polarity
2) Various stroke widths;
3) The complex background.

As the multiframe verification is done, then bitmap integration over time [10] is often used to remove the moving background of a text region. To handle the problem of text and background has similar colors. To solve this problem, they developed a statistic color-polarity classification approach [26], discussed below

Color –polarity classification approach:

1) Uses the Otsu thresholding method for generating a binary text image
2) To classify the color polarity explores the statistic relationships between the number of black edges and that of white edges. Accuracy reaches 98.5%.
3) If the color is dark do the bitmap integration process, if the text color is dark,
4) After that we get an image of a text region with text in light color

To handle larger font size, they normalize the height of text image to 24 pixel keeping its aspect ratio. They had proposed the language in-dependent text extraction method that consist of three steps

1) Adaptive thresholding
2) Dam point labeling
3) Inward filling

2.2 Contour based robust algorithm:

Yangxing LIU and Takeshi [3] implemented a contour-based robust algorithm for text detection in color images. Here they present a new approach to accurately detect text in color images possibly with a complex background. This algorithm is based on the combination of connected component and texture feature analysis of unknown text region contours. First, describe color image edge detection algorithm to extract all possible text edge pixels. Then connected component analysis is performed on these edge pixels to detect the external contour and possible internal contours of potential text regions. The gradient and geometrical characteristics of each region contour are carefully examined to construct candidate text regions and classify part non-text regions. Then each candidate text region is verified with texture features derived from wavelet domain. Finally, the Expectation maximization algorithm is introduced to binarize each text region to prepare data for recognition.

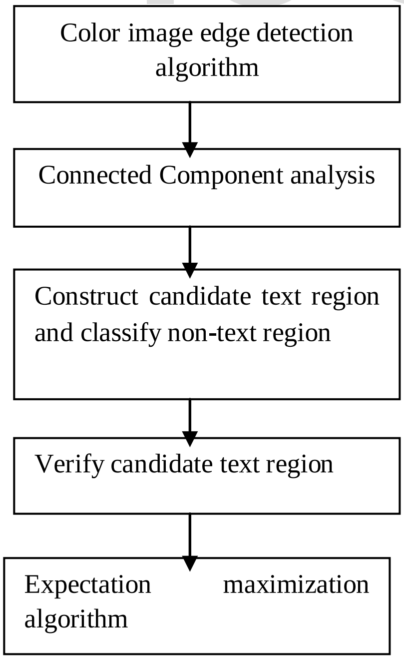


Fig 2: Flow of contour based robust algorithm

2.2.1 Working:

This algorithm is mainly based on following properties of text

1) Contour gradient
2) Structural information
3) Texture property

Now we explain this algorithm in detail. This algorithm is mainly used for detecting the text from color image. Hence they include these properties of text.

2.2.1.1 Text detection steps:

1) Edge Detection
   i) Edge detector
   ii) Edge Extraction algorithm
2) Generation of the candidate of text regions
   i) Region contour Extraction
   ii) Selection of candidate text region
3) Text Region Verification
4) Text Region Binarization

2.2.1.2 Explanation

1) Edge Detection:
   i) To detect edge from image we required gradient magnitude and gradient direction at any given point.
   ii) To calculate gradient magnitude and gradient direction we have to calculate larger eigenvalue and its corresponding eigenvector direction. Then calculate gradient direction angle using eigenvector.
   iii) Apply median filter to input image c to reduce the noise of input images while preserving sharp edges.
   iv) Compute matrix for each pixel on the image from this we get eigenvalues and eigenvector
   v) The following two condition for acceptance of pixel
      vi) Eigenvalue(pixel) > that of neighboring pixels ,closest to eigenvector direction
      vii) Pixel gradient magnitude > adaptive thresholding
   viii) Make character contour continuous perform conditional dilation on edge collection
   Also gradient magnitude of center pixel must be exceeding to adaptive threshold and gradient direction must be less than predefined angular tolerance.

ix) Then all character and non-character edge pixels which has high local color contrast are remained in image edge map.

x) Join the connected edge pixel to form region contour. By CC analysis to generate text region candidate.

2) Generation of the candidate of text regions

Here we need group of connected edge pixel from different region contour through CC analysis. After that they extract CCs directly from contour. This method is based on the principal that character is fully determined by its contour. Region contour extraction algorithm work as 8-neighborhood-connectivity algorithm.

Region contour extraction algorithm:

a) Scan the image from left to right and top to bottom

b) Initialize the class label of each pixel = 0

c) For edge pixel examine its four neighbors which already encounter in the scan.

d) Then label edge pixel using different cases

i) If only one neighbor has class number > 0 and gradient direction difference between neighbor and edge pixel < angular tolerance then assign this class number to edge pixel.

Otherwise

ii) If more than one neighbor has class number > 0 and gradient direction difference among these neighbor and edge pixel < angular tolerance then assign one of the label to edge pixel and make note of equivalence

Otherwise

iii) Assign new class number to edge pixel

e) After finishes scanning of whole image equivalent label pair sorted into equivalence classes and assign unique label to each class.

f) Then in second scanning of w the hole image ,each label is replaced by the label assign to its equivalence classes

g) From above we can link edge pixel of a possible character into a closed contour and get many CCs from edge map.

h) For creating bounding rectangles we can use leftmost, rightmost, topmost and bottommost coordinates of each component. For selecting part of the components as candidate text regions and remove false alarms.

i) For developing the criteria for filter out the non-text region, following features of text observed for counter pixel

- Average gradient magnitude

  Here if average gradient magnitude of text > non-text, then calculate average value of each CCs region

- Gradient direction variance

Within a text region, the variance must be greater than PI (180 degree)

- Number of contour pixels

  Text block should have more edge pixels than some non-text blocks.

Here most of non-text region contour will be eliminated and candidate text region will be generated.

3) Text Region Verification

From above section it will be clear that it would be add false alarms. They consider every region of interest separately to remove false alarms. They select Harr wavelet as the basis for our texture characterization.

- Convert input image into gray level image
- Consider two Haar wavelet bases of one-dimensional.
- Calculate tensor product transform
- Image is processed with discrete wavelet transform
- Compute wavelet moment features to capture each candidate text region texture property.
- Checked the second and third wavelet moment of each candidate text region to verify whether it is a text region or not. To

filter out false alarms and verify true text regions.

- Remove the whole region problem by considering spatial relations between the possible internal and external contour of a character.

Using this they remove false alarm and hole region

4) Text Region Binarization

Here binarize each text region separately and feed the result to OCR. Necessity of performing color clustering is to avoiding variance of pixel color. They also used EM algorithm for adjusting the parameters of the Gaussian probability model to maximize the likelihood of data in dataset. The iteration stops when the difference between two successive iterations becomes negligible. In this work they use the filling algorithm to select two different colors of two pixels inside and outside in text contour to carefully initialize for EM algorithm, thus we can binarize this text region via measuring the Euclidian distance between each pixel color and two mean color vectors.

2.3 Automatic Detection and Recognition of Signs from Natural Scenes

In this approach [4] they merge 1) multiresolution and multiscale edge detection 2) adaptive searching, 3) color analysis,4) affine rectification in a hierarchical framework for sign detection with different priority at each phase to handle the text in different sizes, orientations, color distributions and backgrounds. To recover deformation of the text regions caused by an inappropriate camera view angle they use affine rectification. Rather using binary information for OCR, they extract features from an intensity image directly. To effectively handle lighting variations they propose a local intensity normalization method, a Gabor transform is used to obtain local features, lastly a linear discriminate analysis (LDA) method is used for feature selection. They have applied the approach in developing a Chinese sign system, which can automatically detect and recognize Chinese signs as input from a camera, and translate the recognized text into English. The procedure can significantly improve text detection rate and optical character recognition (OCR) accuracy.

Next sub section will show you how text detection will be done in this method.

2.3.1 Text detection:

It is difficult for automatically detecting sign from system and to detect text from these images because of the image will contains affine deformations, secularity, highlights and shadows.

We have to check image condition such as lighting, distance, orientation and view angle with respect to effect on text properties such as size, color, intensity, affine, highlight. View angle is used to adjust different focus lens of camera. To focus out these changes in image, they use hierarchical detection. The hierarchical detection contains edge detection, color analysis and affine rectification algorithm. For edge detection they combine multiresolution and multiscale for detecting text in different size. They use Gaussian mixture model (GMM) to represent background and foreground. And for performing color segmentation in selected color spaces. The affine rectification is used to recover deformation of the text regions caused by an inappropriate camera view angle. They perform text detection again in rectified regions within the image to refine detection results.

Following steps for text detection

1) Detection of candidate text
2) Color modeling
3) Layout Analysis and Affine Parameter Estimation
4) Affine rectification

1) Detection of candidate text:

They use edge-based features in the coarse detection phase because for text detection image intensity is the main source. Other hand, the gradient of the intensity (edge) is less sensitive to lighting changes. Keeping this in mind they have made some assumption about text

a) High contrast to its background in both color and intensity images.
b) Every character is composed of one or several connected regions.

c) The characters in the same context have almost the same size in most of the cases, especially for Chinese.

d) The characters in the same context have almost the same foreground and background patterns.

To calculate the edge set they use the multiscale Laplacian of Gaussian (LOG) edge detector. The properties of edge set and edge path are connected with each other. They added more criteria for edge candidate filtering and pyramid structure to handle various variations in implementation.

2) Color modeling

They use a GMM to characterize color distributions of the foreground and background of a sign. For each character, they use the following model:

For each character = (1-deviation) Color distribution of background + deviation (color distribution of foreground).

Deviation must be in between 0 to 1.

Some observations are

a) Character with more strokes has a higher percentage in its foreground than that of character with fewer strokes

b) A character with bold font has less variance in its foreground than the same character without bold

c) The background is the opposite: a bold font character has a larger variance in its background than a normal font character.

Basically signs are designed for human perception. There exist multiple lighting sources and shadows, contrasts of foreground and background might change significantly across the entire sign. Hence, they model the distribution of each character separately rather than the entire sign as a whole.

3) Layout Analysis and Affine Parameter Estimation

The aim of layout analysis is to line up characters in an optimal way, so that characters that belong to the same context will be grouped together. A text layout has two cluster features: intrinsic and extrinsic features. Both the intrinsic and extrinsic features can give hints for layout analysis. The extrinsic features can also provide some information on the affine transform that occurred while snapping the image. The text on sign will be present in row based or column based structure. They have been considered that text on the sign might be in a line. If the line is tilted to a certain degree will cause problems in layout analysis.

4) Affine Rectification

It is possible to reconstruct a front view of the sign if we know the normal of the sign plane under the camera coordinate system. A proper interpolation should be applied because it is not a one-to-one mapping for the digitalized image. We use a B-spline interpolation in our current implementation. To avoid additional blur in the interpolation, all edge pixels are interpolated only along the edge direction while the other points are done by surface interpolations.

They used the intensity based OCR instead of binary OCR. If we use binary features for OCR, we cannot guarantee effectively removing noises before the binarization processing. The intensity based OCR uses a gray scale image as its input. First convert a color image into a gray scale image before further process it. Gray scale sign images come in two forms: bright characters on a dark background or dark characters on a bright background. Then resolve these two into a single case by inverting the foreground and background if the text is darker than its background. To determine whether the text is darker or brighter, they apply a method similar to the one described in color modeling, but only in gray scale (single component).They used a localized intensity normalization method before feature extraction to reduce intensity distribution changes. And for feature extraction they used a Gabor wavelet for its superior mathematic properties Gabor wavelet is a sinusoidal plane wave with a particular frequency and orientation, modulated by a Gaussian envelope. It can characterize a spatial frequency structure in the image while preserving information about spatial relations, and is therefore suitable for extracting orientation-dependent frequency contents from patterns.

2.4 Hybrid Approach

Yi-Feng Pan, Xinwen Hou, and Cheng-Lin Liu [1] invented hybrid approach. In that they used both methods i.e. region based and connected component. As earlier saw that

ISSN 2229-5518

both method having complementary advantage and disadvantages. Hence by merging this method we get the better output

Algorithm:

1) Convert the color image into gray level
2) Design text region detector to detect text regions each layer of the image pyramid
3) To calculate the text confidence and scale project the text confidence and scale information back to the original image,
4) To generate candidate text components scale-adaptive local binarization is then applied
5) By using a CRF model combining unary component properties (including the text confidence) and binary contextual component relationships is used to filter out non-text components.
6) Then neighboring text components are linked with a learning-based minimum spanning tree (MST) algorithm and between-line/word edges are cut off with an energy minimization model to group text components into text lines or words.

Here they give input image which converted to gray scale then apply tree region detector to get the text confidence using Wald boost classifier. Then using posterior probability calculates the text confidence and scale map. Apply text confidence and scale map to get the final text confidence and scale map. Using scale map image segmentation will be performing. For labeling the component with CRF model we use text confidence. Then apply the minimum spanning tree for clustering component. Then apply line/word partition to get the localization of text.

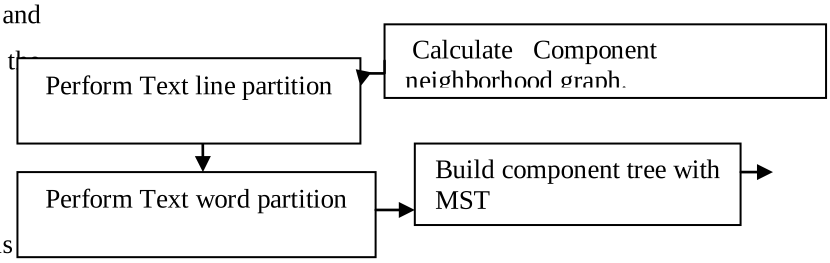
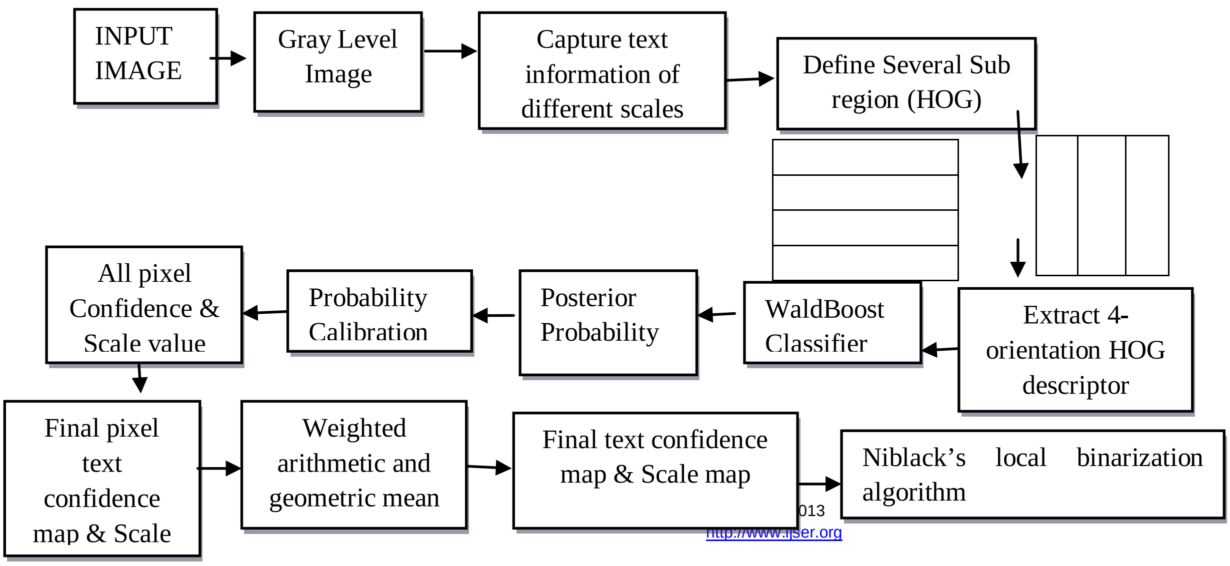


Fig 3: Flow of Hybrid Approach



Components pass through unary classification thresholds.

GET TEXT LOCALIZATION

TABLE I: SUMMARY OF REVIEW

Sr. No	Author	Title	Method	Description	Fault
1	M. Lyu, J. Q. Song, and M. Cai,	Comprehensive method for multilingual video text detection, localization, and extraction,	Region-based	Detect candidate text edges of various scales with a Sobel operator. A local thresholding procedure used to used to filter out non-text edges, and the text regions are then grouped into text lines by recursive profile projection analysis	Only focuses on Chinese and English language. 1)it cannot detect motion texts due to the assumption of stationary text 2)non horizontally aligned texts cannot be localized
2	Y. X. Liu, S. Goto, and T. Ikenaga,	“A contour-based robust algorithm for text detection in color images	Connected Component Based	Extracts candidate CCs based on edge contour features and removes non-text components by wavelet feature analysis. Within each text component region, a GMM is used for binarization by fitting the gray-level distributions of the foreground and background pixel clusters.	Minimizes classification error – not number of false negatives.
3	X. L. Chen, J. Yang, J. Zhang, and A.Waibel,	“Automatic detection and recognition of signs from natural scenes,	Region based	Proposed a hierarchical sign text detection framework by combining multi-scale Laplacian of Gaussian (LOG) edge detection, adaptive search, color analysis and affine normalization.	Noise level increases
4	Yi-Feng pan Xinwen Hou,and Cheng-Lin Liu	A Hybrid approach to detect and localize texts in natural scene Images	Region based and Connected component	The image first pass through the pr-processing steps from that we get the location and size of text. Then it will passes to the CCA stage to get the connected component. then these component are cluster using minimum spanning tree and group in to line or word.	Approach fails on some hard-to-segment texts Speed need to be accelerated further

3 CONCLUSION

In this paper we revised the existing method for detection and localize the text. Table I show summary of the related work.

From this summary we clarify that the best approach is the hybrid approach. In remaining three approaches use the either region based or connected component. In region based method the text will be detected and localize be texture analysis. Hence detection and localization accurately even in noise. In CC-based method directly segment candidate text components by edge detection or color clustering. Having lower computation cost and the located text components can be directly used for recognition. In region-based methods, 1) speed is relatively slow,2) the performance is sensitive to text alignment orientation. For CC-based methods1) cannot segment text components accurately without prior knowledge of text position and scale 2) designing fast and reliable connected component analyzer is difficult since there are many non-text components which are easily confused with texts when analyzed individually..And in hybrid approach [1] these two methods are get merged to get the better output. From above it is clear that they are complementary. The hybrid approach also having some disadvantages 1) approach fails on some hard-to-segment texts 2) Speed need to be accelerated further.

REFERENCES

[1] Yi-Feng Pan, Xinwen Hou, and Cheng-Lin Liu,A Hybrid Approach to Detect and Localize Texts in Natural Scene Images, IEEE Trans on Image Processing, VOL. 20, NO. 3, March 2011

[2] M. Lyu, J. Q. Song, and M. Cai, A comprehensive method for multilingual video text detection, localization, and extraction, IEEE Trans. Circuits Syst. Video Technol., vol. 15, no. 2, pp. 243 255, 2005.

[3] Y. X. Liu, S. Goto, and T. Ikenaga, A contour-based robust algorithm for text detection in color images, IEICE Trans. Inf. Syst., vol. E89-D, no. 3, pp. 1221 1230, 2006.

[4] X. L. Chen, J. Yang, J. Zhang, and A.Waibel, "Automatic detection and recognition of signs from natural scenes," IEEE Trans. Image Process., vol. 13, no. 1, pp. 87–99, Jan. 2004.

[5] M. Cai, J. Song, and M. R. Lyu, "A new approach for video text detection," in Proc. Int. Conf. Image Process., Rochester, NY, Sep. 2002, pp.117–120.

[6] I. Sobel, "An isotropic 3_3 image gradient operator," in Machine Vision for Three-Dimensional Scenes, H. Freeman, Ed. New York: Academic, 1990, pp. 376–379.