# Text Detection in Color Images

J.Sushma
Lecturer
Dept. of Electronics & Communication Engg
DMS SVH College of Engineering
Machilipatnam-521001,
Andhra Pradesh, India
sushmajala@yahoo.co.in

M.Padmaja
Assistant Professor
VR Siddhartha Engineering College,
Vijayawada,-520007
Andhra Pradesh, India
padmaja19m@gmail.com

*Abstract*— **Content-based multimedia database indexing and retrieval tasks require automatic extraction of descriptive features that are relevant to the subject materials i.e., images, video etc. The typical low-level features that are extracted in images and video include measures of color, texture, or shape. Although these features can easily be obtained, they do not give a precise idea of the image content. Extracting more descriptive features and higher level entities, such as text and human faces is important. Text embedded in images and video, especially captions provide brief and important content information, such as the name of players or speakers, the title, location, date of an event, etc. Besides, text-based search has been successfully applied in many applications, while the robustness and computation cost of feature matching algorithms based on other high-level features is not efficient enough to be applied to large databases. The objective of this paper is to compare two basic approaches of text extraction in natural (non-document) images namely; edge-based and connected-component based. These algorithms are implemented and evaluated using a set of images of natural scenes that vary along the dimensions of lighting, scale and orientation. Accuracy, precision and recall rates for each approach are analyzed to determine the success and limitations of each approach. Recommendations for improvements are given based on the results.**

***Text Extraction, Detection, Localization, Dilation, Variance, Enhancement***

## I. INTRODUCTION

Image is defined as a two dimensional function of f(x, y), where x and y are spatial coordinates and f indicates the intensity or the gray level of the image at that point. Text that appears in images contains important and useful information. The content can be in the form of objects, color, texture, shape as well as the relationships between them. The text data can be embedded in an image or video in different font styles, sizes, orientations, colors, and against a complex background. The problem of extracting the candidate text region becomes a challenging one.

Text extraction is the ability to extract (pull out) text from a document or image. A TE system receives an input in the form of a still image or a sequence of images. The images can be in gray scale or color and the text in the images. The text data is particularly interesting, because text can be used to easily and clearly describe the contents of an image. Since the text data can be embedded in an image or video in different font styles, sizes, orientations, colors, and against a complex background. Different approaches for the extraction of text regions from images have been proposed based on basic properties of text. The text has some common distinctive characteristics in terms of frequency and orientation information, and also spatial cohesion. Spatial cohesion refers to the fact that text characters of the same string appear close to each other and are of similar height, orientation and spacing. Two of the main methods commonly used to determine spatial cohesion are based on edge based and connected component features of text characters.

Edge based method is focused in the search of those areas that have a high contrast between text and background. In this way, edges from letters are identified and merged. Once these regions are recognized, spatial cohesion features are applied in order to discard false positives. Connected component method used a bottom-up approach by iteratively merge sets of connected pixels using a homogeneity criterion leading to the creation of flat-zones or Connected Components (CC). At the end of the iterative procedure all the flat-zones are identified. Also in this case spatial cohesion features are applied. Both edge-based and CC-based methods could be included under the same group, region-based methods.

## II. PRINCIPLE OF TEXT

Text extraction is the ability to extract (pull out) text from a document or image. A TE system receives an input in the form of a still image or a sequence of images. The images can be in gray scale or color and the text in the images.

978-1-4244-4711-4/09/$25.00 ©2009 IEEE

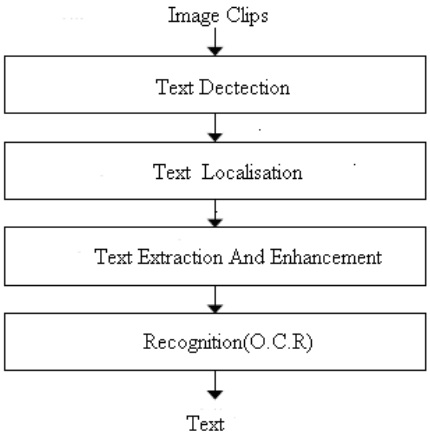


Figure1: Architecture of Text Extraction

Text detection refers to the determination of the presence of text in a given frame (normally text detection is used for a sequence of images).Text localization is the process of determining the location of text in the image and generating bounding boxes around the text. Text extraction is the stage where the text components are segmented from the background.

Enhancement of the extracted text components is required because the text region usually has low-resolution and is prone to noise. Thereafter, the extracted text images can be transformed into plain text using OCR technology.

Edge based method is focused in the search of those areas that have a high contrast between text and background. In this way, edges from letters are identified and merged. Once these regions are recognized, spatial cohesion features are applied in order to discard false positives.

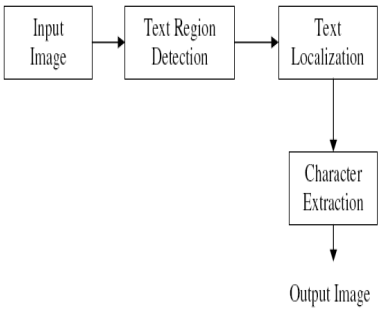


Figure 2: Block Diagram for Edge Based Text Extraction.

The basic steps of the edge-based text extraction algorithm are given below, and diagrammed in Figure2. The details are explained in the following sections.

1. Create a Gaussian pyramid by convolving the input image with a Gaussian kernel and successively down-sample each direction by half. (Levels: 4)
2. Create directional kernels to detect edges at 0, 45, 90 and 135 orientations.
3. Convolve each image in the Gaussian pyramid with each orientation filter.
4. Combine the results of step 3 to create the Feature Map.
5. Dilate the resultant image using a sufficiently large structuring element (7x 7) to cluster candidate text regions together.
6. Create final output image with text in white pixels against a plain black background. The procedure for extracting a text region from an image can be broadly classified into three basic steps:

(1)Detection of the text region in the image,

(2)Localization of the region, and

(3)Creation of the extracted output character image.

### *A. Detection*

Given an input image, the region with a possibility of text in the image is detected. A Gaussian pyramid is created by successively filtering the input image with a Gaussian kernel of size 3x3 and down sampling the image in each direction by half. Down sampling refers to the process whereby an image is resized to a lower resolution from its original resolution. Each level in the pyramid corresponds to the input image at a different resolution. These images are next convolved with directional filters at different orientation kernels for edge detection in the horizontal ($0^o$), vertical ($90^o$) and diagonal ($45^o$, $135^o$) directions.

After convolving the image with the orientation kernels, a feature map is created. A weighting factor is associated with each pixel to classify it as a candidate or non- candidate for text region. A pixel is a candidate for text if it is highlighted in all of the edge maps created by the directional filters. Thus, the feature map is a combination of all maps at different scales and orientations with the highest weighted pixels present in the resultant map.

### *B. Localization*

The process of localization involves further enhancing the text regions by eliminating non-text regions .One of the properties of text is that usually all characters appear close to each other in the image, thus forming a cluster. By using a morphological dilation operation, these possible text pixels can be clustered together, eliminating pixels that are far from the candidate text regions. Dilation is an operation which expands or enhances the region of interest, using a structural element of the required shape and/or size. The process of dilation is carried out using a very large structuring element in

order to enhance the regions which lie close to each other. In this algorithm, a structuring element of size [7x7] has been used. Figure below shows the result before and after dilation.

The resultant image after dilation may consist of some non-text regions or noise which needs to be eliminated. An area based filtering is carried out to eliminate noise blobs present in the image. Only those regions in the final image are retained which have an area greater than or equal to 1/20 of the maximum area region.

Color histograms have been widely used for object recognition. Though in practice these histograms often vary slowly under changes in viewpoint, it is clear that the color histogram generated from an image surface is intimately tied up with the geometry of that surface, and the viewing position. A method is developed to create color histogram based on the color gradients and it is invariant under any mapping of the surface which is *locally* affine, and thus a very wide class of viewpoint changes or deformations [5].

### *C. Character Extraction*

The common OCR systems available require the input image to be such that the characters can be easily parsed and recognized. The text and background should be monochrome and background-to-text contrast should be high. Thus this process generates an output image with white text against a black background.

## III. ALGORITHM FOR CONNECTED COMPONENT BASED TEXT REGION DETECTION

The input image is preprocessed to facilitate easier detection of text regions. As proposed in, the image is converted to the YUV color space (luminance + chrominance), and only the luminance(Y) channel is used for further processing. The conversion is done using the MATLAB function 'rgb2ycbcr' which takes the input RGB image and converts it into the corresponding YUV image. The individual channels can be extracted from this new image. The Y channel refers Connected component based method used a bottom-up approach by iteratively merge sets of connected pixels using a homogeneity criterion leading to the creation of flat-zones or Connected Components (CC). At the end of the iterative procedure all the flat-zones are identified. Also in this case spatial cohesion features are applied. Both edge-based and CC-based methods could be included under the same group, region-based methods.

The basic steps of the connected-component text extraction algorithm are given below, and shown in Figure. The details are discussed in the following sections.

1. Convert the input image to YUV color space. The luminance(Y) value is used for further processing. The output is a gray image.
2. Convert the gray image to an edge image.
3. Compute the horizontal and vertical projection profiles of candidate text regions using a histogram with an appropriate threshold value.
4. Use geometric properties of text such as width to height ratio of characters to eliminate possible non-text regions.
5. Binarize the edge image enhancing only the text regions against a plain black background.
6. Create the Gap Image (as explained in the next section) using the gap-filling process and use this as a reference to further eliminate non-text regions from the output.

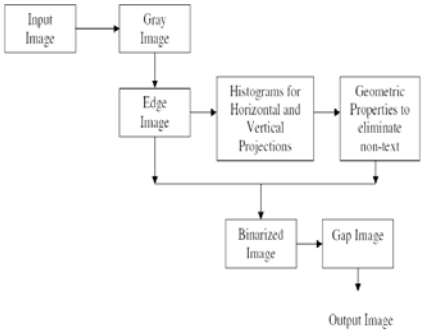


Figure 3: Basic Block diagram for Connected Component based text Extraction**.**

### *A. Preprocessing*

The input image is pre-processed to facilitate easier detection to brightness or intensity of the image where as, the U and the V channels refer to the actual color information. Since text present in an image has more contrast with its background, by using only the Y channel, the image can be converted to a grayscale image with only the brightness / contrast information present.

Content-based image retrieval calculates visual similarities between a query image and images in a database. Accordingly, the retrieval result is not a single image but a list of images ranked by their similarities. The result is not a single image, but a list of images that have been developed for image retrieval based on empirical estimates of the distribution of features in recent years. Different *similarity/distance measures* will affect retrieval performances of an image retrieval system significantly.

### *A. Detection of Edges*

In this process, the connected-component based approach is used to make possible text regions stand out as compared to non-text regions. Every pixel in the edge image is assigned a weight with respect to its neighbors in each direction. As depicted in Figure, this weight value is the maximum value

between the pixel and its neighbors in the left (L), upper (U) and upper-right (UR) directions. The algorithm proposed uses these three neighbor values to detect edges in horizontal, vertical and diagonal directions. The resultant edge image obtained is sharpened in order to increase contrast between the detected edges and its background, making it easier to extract text regions. Figure below shows the sharpened edge image for the Y Channel gray image G from above Figure, obtained by the algorithm proposed.

The algorithm for computing the edge image E, as proposed is as follows:

1. Assign left, upper, upper Right to 0.

2. For all the pixels in the gray image G(x, y) do
   A. left = (G(x, y) – G(x-1, y))
   B. upper = (G(x, y) – G(x, y-1))
   C. upper Right = (G(x, y)-G(x+1,y-1))
   D. E(x, y) = max (left, upper, upper Right)

3. Sharpen the image E by convolving it with a sharpening filter.

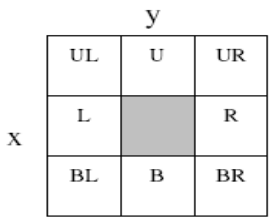


Figure 4: Weight for pixel (x,y)

### *B. Localization*

In this step, the horizontal and vertical projection profiles for the candidate text regions are analyzed. The sharpened edge image is considered as the input intensity image for computing the projection profiles, with white candidate text regions against a black background. The vertical projection profile shows the sum of pixels present in each column of the intensity or the sharpened image. Similarly, the horizontal projection profile shows the sum of pixels present in each row of the intensity image. These projection profiles are essentially histograms where each bin is a count of the total number of pixels present in each row or column

Candidate text regions are segmented based on adaptive threshold values, $T_y$ and $T_x$, calculated for the vertical and horizontal projections respectively. Only regions that fall with in the threshold limits are considered as candidates for text. The value of threshold $T_y$ is selected to eliminate possible non text regions such as doors, window edges etc. that have a strong vertical orientation. Similarly, the value of threshold $T_x$ is selected to eliminate regions which might be non text or long edges in the horizontal orientation.

### *C. Enhancement and Gap Filling*

The geometric ratio between the width and the height of the text characters is considered to eliminate possible non-text regions. This ratio value will be defined after experimenting on different kinds of images to get an average value. In this project, regions with minor to major axis ratio less than 10 are considered as candidate text regions for further processing. Next a gap image will be created which will be used as a reference to refine the localization of the detected text regions. If a pixel in the binary edge image created is surrounded by black (background) pixels in the vertical, horizontal and diagonal directions, this pixel is also substituted with the background value. This process is known as gap filling

## IV. RESULTS

The experimentation of the proposed algorithm was carried out on a data set consisting of different images such as indoor, outdoor, posters etc. These test images vary with respect to scale, lighting and orientation of text in the image. The complete list of test images is shown in Table. The significance of testing the algorithms on variations of scale, lighting and orientation is to determine the robustness of each technique with respect to variance in these conditions, and also to determine where each technique is successful and where it fails. The performance of each technique has been evaluated based on its precision and recall rates obtained.

$$\text{Precision Rate} = \frac{\text{Correctly detected words}}{\text{Correctly detected words} + \text{False positives}} \times 100\% \quad (1)$$

$$\text{Recall Rate} = \frac{\text{Correctly detected words}}{\text{Correctly detected words} + \text{False Negatives}} \times 100\% \quad (2)$$

Precision rate takes into consideration the false positives, which are the non-text regions in the image and have been detected by the algorithm as text regions. Recall rate takes in to consideration the false negatives, which are text words in the image, and have not been detected by the algorithm. Thus, precision and recall rates are useful as measures to determine the accuracy of each algorithm in locating correct text regions and eliminating non-text regions.

### *A. Scale Variance*

The test images are varied with respect to the distance from the camera. This test is to evaluate the robustness of each algorithm with respect to size of text in an image. A total of 3 scale levels have been considered for each image type.

Scale variance test is to determine the robustness of each algorithm to detect text regions for changes in scale or distance from the camera. The precision and recall rates

obtained by both algorithms have been calculated for each of the three scaled test images as shown below.

Figure 5 Row 1 images, Row 2 results of EB,Row 3 results of CC

Table 1: Results from Edge Based Algorithm

| Image Type | Image distance From camera ( meters) | Precision Rate (%) | Recall Rate (%) |
|---|---|---|---|
| Indoor | 1 | 73.91 | 94.4 |
| | 1.5 | 63.63 | 93.33 |
| | 2 | 58.82 | 66.66 |
| Outdoor | 1 | 68.75 | 78.57 |
| | 1.5 | 68.42 | 92.85 |
| | 2 | 60.86 | 92.87 |

Table 2: Results from Connected Component based Algorithm

| Image Type | Image distance From camera (meters) | Precision Rate (%) | Recall Rate (%) |
|---|---|---|---|
| Indoor | 1 | 61.53 | 50.00 |
| | 1.5 | 57.14 | 47.05 |
| | 2 | 58.33 | 38.88 |
| Outdoor | 1 | 28.57 | 40.00 |
| | 1.5 | 44.00 | 84.60 |
| | 2 | 52.00 | 92.80 |

Tables 1 and 2 above show the results obtained by each algorithm for two different image types, varied with respect to distance from the camera, (1) being the closest and (3) being the farthest from the camera. In case of indoor images, the average precision rate obtained by the connected component based algorithm (59%) is lesser than that obtained by the edge based algorithm (65.45). Also, the recall rates obtained by the connected component algorithm (45.3%) are lesser than those obtained by the edge based algorithm (84.7%). In case of outdoor images also, the average precision (41.52%) and recall rates (72.46%) obtained by the connected component based algorithm are lesser than those obtained by the edge based algorithm (66%, 88.09%).

B. Lighting Variance

The lighting variance test is to determine the robustness or invariance of each algorithm to changes in lighting conditions. The precision and recall rates obtained from each algorithm have been shown below for three indoor and three outdoor Images

Figure 6 Row 1 : images, Row 2: results of EB, Row 3: results of CC

Table 3: Results from Edge based Algorithm

| Image Type | Image distance From camera ( meters) | Precision Rate (%) | Recall Rate (%) |
|---|---|---|---|
| Outdoor | Day light | 37.50 | 100 |
| | Evening light | 20.00 | 80 |
| | Night light | 60.00 | 100 |

Table 4: Results from Connected Component based Algorithm

| Image Type | Image distance From camera ( meters) | Precision Rate (%) | Recall Rate (%) |
|---|---|---|---|
| Outdoor | Day light | 33.33 | 83.33 |
| | Evening light | 20.00 | 66.66 |
| | Night light | 66.66 | 33.33 |

The results obtained show in the case of outdoor images, the average precision (39.19%) and recall (93.33%) rates obtained from the edge based algorithm are higher than those obtained by the connected component algorithm (39.9%, 61.1%).Thus the edge based algorithm is more robust in outdoor lighting conditions.

C. Orientation/Rotation Variance

This test is to evaluate the robustness of each algorithm for orientation or rotation variance of text in images. Three indoor and three outdoor images have been considered with 0º, 45º and 135 º rotation angles. The precision and recall rates obtained from each algorithm have been shown below. Tables 5 and 6 above shows the results obtained from each algorithm when tested for orientation or rotation invariance. In case of indoor images, the average precision rate obtained by the connected component algorithm (77%) is slightly higher than the average precision rate obtained by the edge based algorithm (68%). Also, the average recall rate obtained by the edge based algorithm (98%) is higher than that obtained by the connected component based algorithm (91%).

Figure7 Row 1: images, Row 2 : results of EB, Row 3 results of CC

Table 4: Results from Edge based Algorithm

| Image | Rotation (deg) | Precision Rate(%) | Recall Rate(%) |
|---|---|---|---|
| Indoor | 0 | 69 | 95 |
| Indoor | 45 | 65 | 100 |
| Indoor | 135 | 75 | 100 |

Table5: Results from CC based algorithm

| Image | Rotation (deg) | Precision Rate(%) | Recall Rate(%) |
|---|---|---|---|
| Indoor | 0 | 70 | 83 |
| Indoor | 45 | 82 | 95 |
| Indoor | 135 | 79 | 95 |

## V. CONCLUSION

The results obtained by each algorithm on a varied set of images were compared with respect to precision and recall rates. In terms of scale variance, the connected component algorithm is less robust as compared to the edge based algorithm for text region extraction. In terms of lighting variance also, the connected component based algorithm is less robust than the edge based algorithm. In terms of rotation or orientation variance, the precision rate obtained by the connected component based algorithm is lower than the edge based, and the recall rate obtained by the edge based is lower than the connected component based The average precision rates obtained by each algorithm for the remaining test images are similar. Thus, the results from the experiments indicate that in most of the cases, the connected component based algorithm is less robust and invariant to scale, lighting and orientation as compared to the edge based algorithm for text region extraction

## 7. REFRENCES:


1. Dr.Fuhuri Long,Dr.Hongjiang and Prof.David Daga Xiaoqing Liu and Jagath Samarabandu, An Edge-based text region extraction algorithm for Indoor mobile robot navigation, Proceedings of the IEEE, July 2005.
2. Xiaoqing Liu and Jagath Samarabandu, Multiscale edge-based Text extraction fromComplex images, IEEE, 2006.
3. Julinda Gllavata, Ralph Ewerth and Bernd Freisleben, A Robust algorithm for Text detection in images, Proceedings of the 3rd international symposium on Image and Signal Processing and Analysis, 2003.
4. Keechul Jung, Kwang In Kim and Anil K. Jain, Text information extraction in images and video: a survey, The journal of the Pattern Recognition society, 2004.
5. Kongqiao Wang and Jari A. Kangas, Character location in scene images from digital camera, The journal of the Pattern Recognition society, March 2003.
6. K.C. Kim, H.R. Byun, Y.J. Song, Y.W. Choi, S.Y. Chi, K.K. Kim and Y.K Chung, Scene Text Extraction in Natural Scene Images using Hierarchical Feature Combining and verification, Proceedings of the 17th International Conference on Pattern Recognition (ICPR '04), IEEE.
7. Victor Wu, Raghavan Manmatha, and Edward M. Riseman, TextFinder: An Automatic System to Detect and Recognize Text in Images, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, No. 11, November 1999..
8. Qixiang Ye, Qingming Huang, Wen Gao and Debin Zhao, Fast and Robust text detection in images and video frames, Image and Vision Computing 23, 2005.
9. Rainer Lienhart and Axel Wernicke, Localizing and Segmenting Text in Images and Videos, IEEE Transactions on Circuits and Systems for Video Technology, Vol.12, No.4, April 2002.
10. Qixiang Ye, Wen Gao, Weiqiang Wang and Wei Zeng, A Robust Text Detection Algorithm in Images and Video Frames, IEEE, 2003.