TEXT LOCALIZATION IN NATURAL SCENE IMAGES USING SURF AND SVM CLASSIFIER FOR MULTILINGUAL SCENE TEXT

[1] Sankhya N. Nayak, [2] Dr. Nirmala C.R
[1]Asst.professor, [2]Professor & HoD
[1]CS & E,
[1]JNNCE, Shivamogga, India

Abstract: Text extraction from natural scene images is being carried out from many years but still flourishes due to the challenges it possesses. In this work we extract features from testing images using Speed Up Robust Features and train Support Vector machine. The testing image regions classified as text are further refined using mathematical morphology and connected component analysis to remove outliers.

IndexTerms - Speed Up Robust Features (SURF), connected components, Feature Descriptors, Support Vector Machine.

I. INTRODUCTION

The amount of information embedded in natural scene images is large and studies have shown immense interest in extracting data from images. Natural scene Images often contain different objects, colours, shapes and textures. Images give us the semantic information that can be used for content based image retrieval and will be also helpful in indexing and classification. The problem of identifying the region of interest, in our case, text, becomes a challenging task since the text could be embedded in an image in many sizes, orientations, colours, font styles and with a complex background.

Text however also have distinct characteristics in terms of frequency, orientation and spatial cohesion information. These characters can be used to identify the region of interest and further filtered by using a good classifier. Extraction of text has applications in the areas of analyzing of document, vehicle license plate detection, article analysis that has tables, maps, charts, diagrams, keyword based image search, part recognition in automation industry, content based retrieval, street signs, object recognition, text based video indexing and also helps in assistance of visually impaired people.

Text information extraction system is designed to extract text information from the images. Fig. 1 shows the architecture of text information extraction (TIE) system.

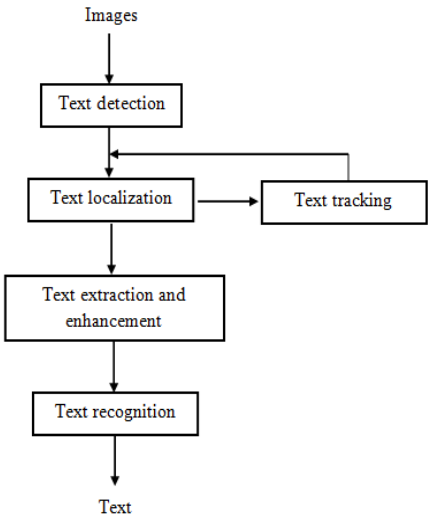

Fig. 1: Architecture of TIE system

The text detection is the identification of the presence of text in the image. Text localization refers to detection of areas where the text is actually present in the images. Bounding boxes are drawn around detected text areas. Text tracking is carried out to deal with the stability of text position over neighbouring frames and used in context of text detection in video. Text extraction phase refers to where texts in image are extracted from the background. Enhancement is necessary in order to segment text accurately for low resolution and noisy text image. OCR technology is used with extracted text image to transform them into plain text.

Scene Text are the one that can be noticed in images clicked in natural environments without concentrating on text exclusively. Natural scene image samples with text are depicted in Fig. 2.

Fig. 2: Samples of Multilingual Natural scene images

In this work, we convert RGB training images into binary images. Feature points are detected using Speed Up Robust Features (SURF) [15] methods. The strongest features are selected and their feature descriptors are extracted and stored as knowledge base. The testing images are binarized using adaptive binarization methods. Feature descriptors are extracted from the image. These features are used to train the SVM. Testing images also include complex background and hence the strong chosen descriptors may include outliers. After classifying, connected component analysis is used for outliers filtering.

II. LITERATURE SURVEY

Many algorithms have been proposed for localizing text data in an image. Each method gives robust results for specified set of images and concentrates on specific applications. Generic methods mostly handle English horizontal text in images.

Shivananda V. Seeri, J. D Pujari and P. S. Hiremath [1] [12] [13], proposed a new methodology to localize text and to remove non-text part from natural scene images that has complicated background. In this context, a new hybrid procedure is proposed that identifies multiple language texts. The promising text areas are be derived from the image depending on edge features in wavelet transformed image, which are tested to check whether it is a text data or a non-text data with the help of GLCM features and SVM classifier. To locate text in natural scene images, text localization algorithm has been designed.

Kirti Bhure and J D Dhande [2] have proposed the object finding method to help visually impaired people. The SURF is used as it can extract distinctive features in an image to match different objects. The proposed recognition process begins by matching individual features of the user queried object to a database of features with different personal items which are saved in database. Their analysis indicates that SURF gives good results for the images without scale variations or rotations.

Reetika Verma and Rupinder Kaur [3] proposed solutions that focus on applying Neural Network Algorithm model for character recognition. The proposed SURF Feature and neural network has the capability of strong robustness performance and good distinction between feature points.

Sankhya Nayak and Nirmala C.R in [4], propose an adaptive method for detecting and extracting text from natural scene images. This method is robust against the shadows and uneven lighting conditions. The proposed method uses adaptive thresholding technique to binarize image and smoothen degradation factors. Canny edge detection is used to obtain edge image and Block operation of localization is used to remove non-text area from image. Connected component analysis is used for extracting text from image. The work is applied on images without and with shadows and uneven lighting conditions.

Rashmi V and Sankhya Nayak [5] in their work implemented a hybrid technique based on edge features in wavelet transformed image. Prewitt edge detector was used to identify strong, sharp edges in training phase. Haar wavelet transform is used in testing phase in combination with morphological operations, connected component analysis to identify text regions. SVM classifier is used to validate by grouping into text and non-text. Localized text was retained.

U. Elakkiya and M. Safa [6] have proposed a hybrid approach for finding and localizing text in natural scene images. Scene texts in noisy images can be detected using text area detector for the estimation of position of text and scale probabilities. This is carried out to separate components of candidate text with the help of local binarization algorithm. Conditional random fields (CRF) are used for conjoining unary and binary components. Next stage is word partition will be done by using word features like word count, centroid distances, bounding box distances, distances between words and ratio between centroid distances within separate words. Eventually, text that corresponds to sub-trees can be derived and the one with tiny components will be eliminated considering them as noises.

Poonam B Kadam, Latika R Desai [7] have proposed a framework to detect and recognize scheme efficient in urban scene text. Initially, preprocessing and segmentation is done. Next step is to extract feature set which then will be trained. Neural network will be used as a learning mechanism. Here, input character pattern will be matched with the stored character training set. With the help of labels given to the characters, network will itself learn many possible variation of a single pattern

and becomes adaptive in nature. This step is called character matching. Based on the calculated scores, system has been called efficient.

Kumuda and Basavaraj [8] have put forth a different approach for deriving text from image. This is done by joining texture and connected components. Identification of text is done with the help of first and second order statistical features. Then, connected components technique is utilized for isolating text from surroundings. Then heuristic filters are made used in order to remove non text components.

Pooja Singh [9] has developed a robust system in order to extract text from images. A detector has been designed to determine the presence of text in the image. A model has been developed that eradicates the non-text components which make use of unary and binary components relationships. Lastly, with the help of learning based method, text components will be grouped together.

Rashedul Islam, Md. Rafiqul Islam and Kamrul Hasan Talukder [11], have developed a novel and advanced approach of text derivation by conjoining few factors from edge-based and connected-components. System is tested for its efficiency and has displayed the median accuracy. The system is robust in extracting text from many different kinds of scene images.

III. SYSTEM DESCRIPTION

The proposed method localizes text from natural background. This method has two steps: creation of feature set knowledge base and testing phase. In first phase, Image is binarized using thresholding method. Strongest feature points are extracted and store into the knowledge base and used to train SVM. In testing phase, after extracting feature descriptors, SVM classification is used to select possible area that contains text. The outliers are removed with connected components analysis based on the geometrical properties. Then potential text regions are detected with the help of object matching. The architecture of the system is depicted in Fig. 3.

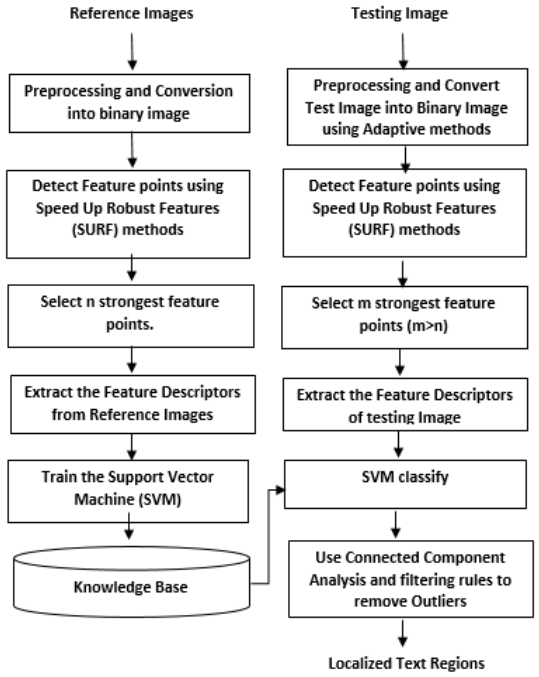

Fig. 3: System architecture

3.1 PRE-PROCESSING

In pre-processing stage, colour image is transformed into gray scale image. Average filter is applied for noise removal and for preserving sharp edges. Binarization is carried out using thresholding for reference images and adaptive binarization for testing images. The adaptive binarization handles even images with shadows, uneven lightening conditions. This is preferred over binarization with global thresholding for testing images as they contain complexities that are nor present in training images with reference to text and background.

3.2 DETECTION OF FEATURE POINTS

We have used SURF [15] to detect feature points. SURF is rotation-invariant detector and descriptor. It provides stable results that are independent of the angle of view. This method results performs better with respect to repeatability, distinctiveness, and robustness at lower computational cost.

Interest point detection is fulfilled by approximation of the Hessian matrix as represented in Eq.3. The concept of integral images in Eq.1 is utilized with this purpose: the entry of an integral image $I_\Sigma(x)$ at a location $x = (x, y)^T$ represents the sum of all pixels in the input image I within a rectangular region formed by the origin as in Eq.2.

$$I\Sigma = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos\frac{n\pi x}{L} + b_n \sin\frac{n\pi x}{L}\right) \quad \text{Eq. (1)}$$

$$(I\Sigma x) = \sum_{i=0}^{i\leq x} \sum_{j=0}^{j\leq y} I(i,j) \quad \text{Eq. (2)}$$

Given a point (x, y) in an image I, the Hessian matrix $H(x, \sigma)$ at position x and scale σ is defined as follows

$$\begin{bmatrix} Lxx(x,\sigma) & Lxy(x,\sigma) \\ Lxy(x,\sigma) & Lyy(x,\sigma) \end{bmatrix} \quad \text{Eq. (3)}$$

of the Gaussian second order derivative $\frac{\partial^2}{\partial x^2} g(\sigma)$ with the image I in point x, and similarly for $L_{xy}(x, \sigma)$ and $L_{yy}(x,\sigma)$. The approximated determinant of the Hessian in Eq. 4 represents the blob response in the image at the location x.

$$det(Happrox) = Dxx\ Dyy - (wDxy)^2 \quad \text{Eq. (4)}$$

where *Dxx*, *Dyy*, *Dxy* denote approximations of the second order Gaussian partial derivatives in x, y and xy-directions and w is a relative weight of the filter responses, that is used to balance the expression for the determinant of the Hessian. This is needed for the energy conservation between the Gaussian kernels and the approximated Gaussian kernels.

Interest points need to be found at different scales. The scale space is analyzed by up-scaling the filter size rather thus image size stays constant. The scale space is divided into octaves. An octave represents a series of filter response maps obtained by convolving the same input image with a filter of increasing size. Each octave is subdivided into a constant number of scale levels. In order to localize interest points in the image and over scales, a non-maximum suppression over a 3 × 3 × 3 neighborhood is applied. The maxima of the determinant of the Hessian matrix are then interpolated in scale and image space.

In order to be invariant to image rotation, the Haar wavelet response is computed in *x* and *y* direction within a circular neighborhood of radius 6*s* around the interest point, where *s* is the scale at which the interest point is detected.

Once the wavelet responses have been computed and weighted with Gaussian centered in the feature point, the responses are represented as points in space and dominant orientation is estimated by computing the sum of all responses in horizontal and vertical directions. The two summed responses then yield a local orientation vector. For the extraction of the descriptor, the square region centered on the features point is constructed. The typical choice for the window size is 20s.

In our work, we have reduced the metric threshold for testing images which have complexities compared to reference images. We have set the value to 1000 for reference images as they contain only text information and blobs are clear and 800 to testing images. These will result in more matches than required and can be removed with filtering rules. Large octaves result in larger sized blobs and usually depends on the image size. We have resized the images to 512 X 512 for which the values between 1 and 4 are suitable. We have set the value to 4. Scale levels per octave are maintained at a default value of 3 for reference images and increased to 4 for testing images. These features are stored in objects. The most important factor in choosing this method is the speed of the detector in comparisons to other feature detectors like Scale invariant feature transform(SIFT).

3.3 FEATURE EXTRACTION

From the output of the SURF algorithm feature objects are extracted. These contain information about location of object, count, scale, orientation etc., The points with the strongest metric or uniformly distributed set of feature points are selected. The location of these feature points are taken as centroids of blobs. We extract these blobs and use them to extract the features contrast, local homogeneity and entropy, kurtosis, skewness from GLCM and additionally area and compactness of the blobs are used to generate feature vector to train the SVM.

3.4 SVM CLASSIFICATION

After extracting features from testing images, SVM classification is carried and possible candidate areas for text are retained. These images may still contain outliers. Outliers are the objects that have features similar to text.

3.5 CONNECTED COMPONENT ANALYSIS AND FILTERING

In this stage of testing, connected component analysis is carried out. To identify every object clearly, labelling is done. Then, region properties are used to characterize every labelled region in the label matrix. These connected components are grouped depending on the set conditions. The condition are, area of the components which are greater than or equal to 45, eccentricity more than 0.2 and extent greater than 0.01 are retained [4]. Text region that are extracted from connected component analysis method and filtering are retained as localized areas and bounding boxes are drawn around them.

IV. RESULTS AND ANALYSIS

The work is carried out using more than 600 images as reference images and 400 images as testing images. The reference images consists of images with Kannada, Telugu, Tamil and Malayalam text in addition to English. For English images we have used ICDAR datasets and for multilingual text, we have created our own database and images are also downloaded from Google pictures. For verifying our work, we have also used images referred by other authors. The testing images are multilingual and have additional challenges of text in different sizes, shapes and orientations with moderately complex backgrounds with shadows, uneven light illumination and pictures taken from different angles.

4.1 RESULTS FOR INPUT IMAGE WITH MULTILINGUAL TEXTS

Input to the system is RGB image. Here, image that consists of multilingual text are considered. Sample test images are shown in Fig. 4.

Fig. 4: Input RGB image

The binarized image with detected feature points are shown in Fig. 5. The feature points appear as the center of the blobs. It is observed that features points are found even for some background regions that match properties of text. The bars in the gate matches properties of English letters and the design in woman's clothing matches properties of Dravidian characters.

Fig. 5: Strong SURF feature points selection

The results after SVM classification are shown in Fig.6. It is observed that they contain outliers. However we can also observe that not all feature points selected in previous stage are used to extract features from blobs. Only strongest feature points are selected.

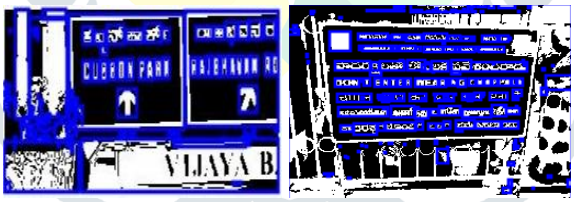

Fig. 6: Localized Text with Outliers

After applying filtering rules, the final text localized area is shown in red bounding boxes in Fig. 7. The filtering rules use geometrical properties such as area, aspect ratio and uniformity compared to neighbouring areas.

Fig. 7: Final Localized Text in Images

4.2 RESULT ANALYSIS

The results were analyzed after testing more than 400 images. It was observed that the system worked well with Dravidian and English languages and often even with Devanagari text. Efficiency of the proposed system was determined by calculating the Precision rate, Recall rate and F-score. The images were grouped into two categories. Category 1 refers to images where adaptive binarization was used before extracting SURF features and Category 2 refers to cases where grayscale images were used to extract features. The work was tested for images with shadows, uneven lightening conditions, text in different scales and multilingual text.

The metric used are:

True positives (TP): correctly detected text area.
False positives (FP): are detected as test regions but are not. Usually this happens when images contain objects in the backgrounds that resemble and possess properties of characters such as small window bars may look like character 'T' or semicircular or curved objects may possess properties that match Dravidian characters and coil like structures match Malayalam and Tamil characters. Though most of these are filtered out in connected component analysis and filtering some of them remain if they are close in vicinity of text.
False Negatives (TN): are actually text but are not a part of localized text. This usually is the case when characters are minuscule in nature or we loss them partially during binarization, especially when they have shadow or reflection on them. This also happens if the text is partially obscured by other objects.
Recall Rate (RR): is also referred to as the true positive rate or sensitivity-measures the proportion of actual positives that are correctly identified. It is calculated as

$$RR = TP / (TP+FN)$$

Precision Rate (PR): also referred to as positive predictive value. It is calculated using

$$PR = TP / (TP+FP)$$

We have used f-score to analyze the work.

F-score= 2*(Precision Recall)/Precision Recall.

The summary is given in Table. 1.

TABLE. 1: QUANTITATIVE ANALYSIS

	Gray Image for extracting SURF			Binarized Image for extracting SURF		
Images with	Precision Rate	Recall Rate	f-score	Precision Rate	Recall Rate	f-score
English Text	96.9072	96.9072	96.9072	97.9381	96.9387	97.4359
Dravidian Text	94.8979	97.8947	96.3730	96.9072	96.9072	96.9072
Multilingual Text	97.9166	95.9183	96.9072	96.9387	97.9381	97.4359

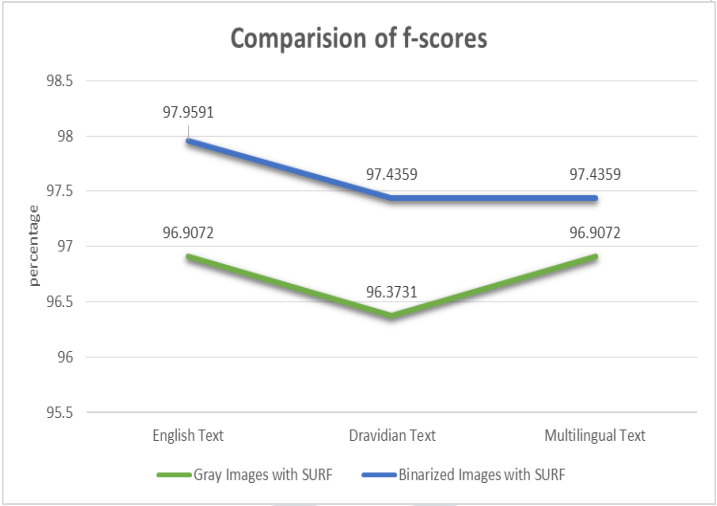

Fig. 8: Graph for comparison of f-scores

The comparison of f-scores for different categories of images using gray scale and binarized forms is shown in graph in Fig. 8. It is observed that the f-score of images for English images was approximately 97% whereas for Dravidian and multilingual images it averaged around 96.6 % for binarized images which is a slight improvement over gray scale images where they were around 96% for all the cases. The overall f-score when gray scale image is used to extract feature set is 96.7 % where as it improvised when binary image was used to 97.6%.

V. CONCLUSION

Text localization from a natural scene image with complicated background is a challenging and significant problem. Here in this work, a hybrid technique is implemented for text localization that involve deriving potential text areas from an image depending on SURF features. We would like to highlight that at the time of paper submission we did not come across any paper that used SURF features in text localization or extraction. For testing purpose, the potential text regions obtained after SVM classification are further refined using connected components analysis and are filtered based on the set criteria to obtain text region. For text contents, bounding boxes are drawn around it yielding us text localization of the given image.

VI. ACKNOWLEDGEMENT

Authors would like to thank the principal and staff of JNNCE and BIET for their support in this work.

VII. REFERENCES

1. Shivananda Seeri J.D. Pujari, P.S. Hiremath, "Multilingual Text Detection In Natural Scene Images Using Wavelet Based Edge Features and SVM Classifier", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 5, Issue 11, Pg. 81-89, 2015.
2. Kirti Bhure, J. D Dhande, "Image Processing using SIFT", International Journal of Advance Research, Ideas and Innovations in Technology", Vol 5, Issue 2 ,Pg. 980-983,2017.
3. Reetika Verma ,Rupinder Kaur, "An Efficient Technique for CHARACTER RECOGNITION Using Neural Network & Surf Feature Extraction", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol 5, Issue 2 ,Pg. 1995-1997,2014.
4. Sankhya Nayak, Nirmala C.R, "Text Extraction From Natural Scene Images Using Adaptive Methods", International Journal of Pure and Applied Mathematics, Vol 120, Issue. 6, Pg. 11669-11682, 2018.
5. Rashmi V and Sankhya Nayak, "A Hybrid Approach To Localize Text in Natural Scene Images", International Journal of Engineering Applied Sciences and Technology, Vol. 3, Issue 1, Pg. 53-60, 2018.
6. Elakkiya U., Safa M., "Text Detection in Natural Scene Images", International Journal of Latest Trends in Engineering and Technology", Vol. 6, Issue, Pg. 48-55, 2015.
7. Poonam Kadam B., Lathika R. Desai, "A Hybrid Approach to Detect and Recognize Texts in Images", IOSR Journal of Engineering, Vol. 4, Issue 7, Pg. 13-19, 2014.
8. Kumuda T, Basavaraj L., "Hybrid Approach to Extract Text in Natural Scene Images", International Journal of Computer Applications, Vol. 142, Pg. 18-22, 2016.
9. Pooja Singh," Scene Text Detection using the Regression Tree Technique", International Research Journal of Engineering and Technology,Vol. 3, Issue 3, Pg. 1678-1682, 2016.
10. Shraddha Naik, Sankhya N. Nayak, "Text detection and Character Extraction in Natural Scene Images", International Journal of Emerging Technology and Advanced Engineering, Vol 5, Issue 2, 2015.
11. Rashedul Islam, Md. Rafiqul Islam and Kamrul Hasan Talukder, "An Approach to Extract Text Regions from Scene Image", International Conference on Computing, Analytics and Security Trends, college of engineering, Pune, Dec 2016.
12. Shivananda V Seeri, J.D. Pujari, P.S. Hiremath, "Text Localization and Character Extraction in Natural Scene Images using Contourlet Transform And SVM Classifier", I.J.Image, Graphics And Signal Processing, Vol 5, Pg. 36-42,2016.
13. Shivananda V Seeri, J.D. Pujari, P.S. Hiremath, "Multilingual Text Localization in Natural Scene Images using Wavelet Based Edge Features and Fuzzy Classification", International Journal of Emerging Trends and Technology in Computer Science, Vol. 4, Issue 1,Pg. 210-218, 2015.
14. Zhenyu Zhao, Cong fang, Zhouchen Lin, "A Robust Hybrid Method for Text Detection in Natural Scenes by Learning Based Partial Differential Equations", Neurocomputing, Vol 168, Pg. 23–34, Elsevier,2015.
15. Herbert Bay, Andreas Ess,Tinne Tuytelaars, Gool, "Speeded-Up Robust Features (SURF)", Computer Vision and Image Understanding, Volume 110, Issue 3, Pg. 346-359, Elsevier,2008.