I.J. Image, Graphics and Signal Processing, **2016, 1, 15-26**
Published Online January 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2016.01.02

Extraction of Scene Text Information from Video

Too Kipyego Boaz, Prabhakar C. J.
Department of Computer science, Kuvempu University, Shivammoga, Karnataka, India
E-mail: kipyego3@yahoo.com.au, psajjan@yahoo.com

Abstract—In this paper, we present an approach for scene text extraction from natural scene video frames. We assumed that the planar surface contains text information in the natural scene, based on this assumption, we detect planar surface within the disparity map obtained from a pair of video frames using stereo vision technique. It is followed by extraction of planar surface using Markov Random Field (MRF) with Graph cuts algorithm where planar surface is segmented from other regions. The text information is extracted from reduced reference i.e. extracted planar surface through filtering using Fourier-Laplacian algorithm. The experiments are carried out using our dataset and the experimental results indicate outstanding improvement in areas with complex background where conventional methods fail.

Index Terms—Natural Scene, Text Information Extraction, Stereo Frames.

I. Introduction

Text information extraction from scene images is an important topic in the field of computer vision. Text in scene images has been generally defined as text existing naturally in the image, where text is written on trucks, t-shirts, buildings, billboards, etc. Text information extraction from scene images have been a well-studied topic with many researchers such as [1], [2] achieving substantive and promising results, but this area still face several difficulties and challenges, thus a sustained research still upholds. Generally, scene text is often affected by camera parameters such as illumination, focus and motion, perspective distortion and image blurring. In addition to these, other challenging factors are arbitrary text layouts, multi-scripts, artistic fonts, colors, complex and variable background.

Scene text information extraction techniques can be classified into three categories such as Connected Component analysis (CC) method, edge-based method, and texture-based methods. The survey on the techniques lying within these categories is presented in [3]. The main difficulty in scene text extraction process is segmentation part and most of the earlier approaches had greater challenges and limitations in a complex scene. Therefore, discriminative approaches have been proposed which are capable to perform segmentation with considerable accuracy.

Shiva kumara et al. [5] proposed a two-step Fourier-Laplacian filtering technique. The authors employed a low-pass filter to smooth the noise, while a laplacian mask is used to detect text regions by generating Maximum Gradient Difference (MGD). These procedures are performed in frequency and spatial domains respectively. Their analysis of the MGD results revealed that text regions have larger values compared to non-text due to larger magnitudes of the positive and negative peaks. The k-means clustering method is used to cluster pixels belonging to text region against those belonging to non-text region.

Many researchers including [6[, [7], [8] have used stereo disparity which is estimated from stereo images in order to navigate the mobile robot based on detection of planar objects. The property that planar surfaces can be represented as linear functions in disparity space and thus have constant spatial gradient [7] provides a platform for the detection and extraction of planar surface based on the statistical features of the estimated disparity map. In order to generate a seed point, the authors, generated boundary pixels based on the approximated gradient magnitude. Jeffrey et al. [9] detected planar surfaces by performing Principal Component Analysis (PCA) on a local neighborhood to approximate local surface normal within the sampled points. Random Sample Consensus (RANSAC) is used to cluster these points into subsets to fit planar model.

Konolige et al. [8] integrated appearance and disparity information for object avoidance and used AdaBoost to learn color and geometry models for ideal routes of travel along the ground. Zhang S. et al. [10] address the problem of low efficiency and unsatisfactory matching of uniform texture regions in binocular stereo vision based on rapid window-based adaptive correspondence search algorithm using mean shift and disparity estimation. They combined color aggregation and local disparity estimation into matching cost aggregation, in order to reduce the color dynamic range of the original image and make complex pixel regions simple with uniform texture areas.

Outdoor images containing sign or advertisement boards, walls, sidewalks, roads, roofs and other objects like vehicles can appear planar when viewed from a distance. Normally, the text information is usually written on these planar surfaces in order to read and interpret information easily. This motivated us to propose a scene text extraction technique based on detection and extraction of planar surface, which is followed by extraction of scene text within the extracted planar surface. In this paper, we attempted to address the recent techniques developed to extract the scene text from the video sequence.

The remainder of the paper is structured as follows: An

Overview of related work is given in section II. Section III and IV describe our proposed approach, while section V presents experimental evaluation and conclusion is drawn in section VI.

II. Related Work

Text information extraction from scene images have received a great attention from researchers all over the the world when compared with video and images of document. Shahab et al., [4] proposed a technique to read scene text in ICDAR competition. The adjacent characters are grouped together so that candidate image patches are obtained and used to localize text regions. The authors extracted Haar features from both gradient and stroke orientation maps using block pattern method. The extracted features are used to train a classifier based on Adaboost model. The input of the classifier is then analyzed to determine text regions which are then merged into rectangular blocks.

Huang et al. [11] presented a Stroke Feature Transform (SFT) method based on the Stroke Width Transform (SWT). They employed this method in order to address text extraction problems based on shape and color, with the re-constraint of the relations among local edge points. Lu et al. [12] designed text-specific features based on contrast, shape structure and paired edges. Their approach detects candidate text boundaries, where, at least one character per boundary is extracted using a local threshold. When all candidate characters have been extracted, refining is done through support vector regression model trained using bags-of-words representation that removes false characters that do not belong to text.

The recent promising directions based on hybrid methods have become the focus of several recent works. They combine the advantages from a number of extraction algorithms. Yi et al. [13] presented a hybrid method to localize scene text by using regions as well as component information, where the neighbor component relationship, together with the unary component property, are used to construct a conditional random field model for connected components. The model parameters are optimized with Minimum Classification Error (MCE) learning and graph cuts inference algorithms.

Other methods such as Maximally Stable Extremal Regions (MSER) based approaches have recently become the focus. The integration of MSER approaches has demonstrated significant improvements in real-world applications. MSER algorithms have proved better in detecting candidate character features, irrespective of their quality in terms of noise levels and contrast.

Yin et al. [14] proposed a robust and accurate MSER based scene text detection method. They explored the hierarchical MSER structure in order to design a pruning algorithm based on simple generated features, where pruning significantly reduces the number of candidate characters to be processed. The authors employed a self-learning algorithm to learn distance weights based on distance metric. These learned parameters and the estimated posterior probabilities of text candidates are incorporated into a character classifier for clustering.

The above mentioned MSER-based approaches still have insufficient text regions with limited number of features of a candidate character region. In order to minimize such limitations, Iqbal et al. [15] proposed a method that uses Bayesian network score obtained through K2 algorithm, this is possible by establishing causal link on extracted regions pixel intensity. The low posterior probability candidate character regions are considered for grouping with the use of selection rules, perceptual grouping filters and repulsion scores from pair-wise filters.

III. Our Approach

As we mentioned earlier, it is assumed that text is contained in planar surface, this drives our interest to identify region within the video frames which represent planar surface on the 3D world. In the paper [16], we proposed an approach to extract planar surface from natural scene stereo frames for the purpose of extracting the saliency region, which helps to assess the quality of frames in reduced reference. The approach presented in the above paper comprises three major steps: a) estimation of disparity map using stereo frames, b) detection of candidate planar surface from the disparity space using gradient derivatives and c) segmentation of candidate text block by mapping connected component analysis of homograph image with detected candidate planar surface. The main drawback of this method is that segmentation based on the mapping of connected component to the disparity map produce low accurate results and contains some non-planar regions.

Hence, to solve the problems associated with the approach presented in the above paper, we adopted technique proposed by Jeffrey et al., [9]. This technique is used to detect and extract planar surface from natural scene video frames. After extraction of planar surface from complex background, further processing can be done by considering the extracted planar surface (called as text block) instead of whole image area. The extraction of planar surface from the complex background reduces the complexity involved while extracting the text from complex background by considering the whole image as processing area. Fig. 1 shows the flow diagram of the proposed approach.

To increase the segmentation accuracy, we introduce plane fitting technique by constructing planar model based on local surface normal computed through PCA and RANSAC. The image labeling is done by employing MRF with Graph cuts algorithm where planar surface is segmented from other regions based on the labels assigned to it. The process is further extended to extract scene text by filtering the extracted text block (planar surface) with Fourier-Laplacian algorithm to generate points which are classified using k-means as either belongs to text region or non-text region.

A. *Stereo Disparity*

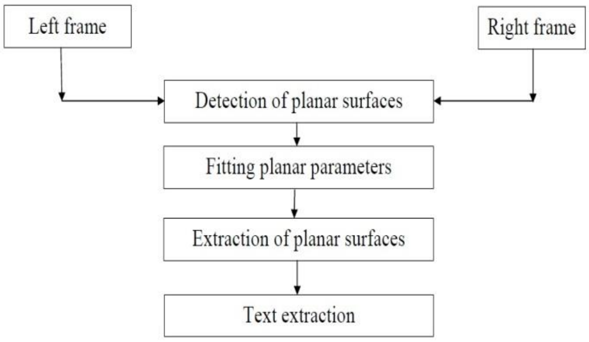

Left frame
Right frame
Detection of planar surfaces
Fitting planar parameters
Extraction of planar surfaces
Text extraction

Fig.1. Flow Diagram of Our Approach.

Given a pair of left and right video frames shown in the Fig. 2(a) and Fig. 2(b) respectively, which are rectified using un-calibrated rectification technique (fundamental matrix) from stereo vision toolbox. The rectification reduces the searching area from $2D$ to $1D$ along x-axis. We estimated the disparity map using a high speed stereo block matching technique based on Sum of Absolute Differences (SAD) Algorithm.

Generally, the disparity offset D_d of a pixel p_d in the disparity map is generated by computing the differences between pixel components in left reference frame p_l and a corresponding pixel components match of right frame p_r as shown in the following equation:

$$p_d(u,v) = p_l(u_l,v_l) - p_r(u_r,v_r) \Rightarrow D_d, \qquad (1)$$

where $\forall\ u_l = u_r$ d, $\forall\ v_l = v_r + i$ and i is disparity range.

It becomes computationally costly to match each and every pixel $p(u,v)$ in a reference frame as there is a requirement to search all pixels in the pre-defined parameter. To reduce these high computations, the ground control points are incorporated into the matching process to cut back the algorithmic complexity and sensitivity to occlusion-cost assigned to unmatched pixels. This manages the biasing process required to smoothen the solution and therefore, the task of choosing crucial previous potentialities describe the image formation [6]. This technique for estimating the disparity map effectively reduces the cost value and search time from maximum disparity time to relatively lower range where the end results of the optimal disparity map that can be achieved at point $p(u,v)$ is given below:

$$D_p = \arg\min_{d\in S_d} \{E(p(u,v), p_d(u,v))\}, \qquad (2)$$

where S_d is a possible range of disparity and E is cost. The S_d can be defined as:

$$S_d = \overline{D}_p \pm K, \qquad (3)$$

where D_p is the estimated disparity at point p and K is the disparity estimation threshold.

By a priori knowledge, it is well known that simple textured regions are in the same depth, therefore, this rule is employed to match the points between a pair of frames to reduce computational time that could have been spend computing the whole disparity map. The only requirement is to check if the point D_p is within the simple-textured region. A point of reference $p(u,v)$ is inspected to find out if it's two neighboring pixels $p(u-1,v)$ and $p(u,v-1)$ belong to the same texture region supported by their color intensity. It is when solely the corresponding pixels are within the same color with their corresponding disparity indices, $D(u-1,v)$ and $D(u,v-1)$ are equal, and then only we can assign:

$$D_p(u,v) = D_p(u-1,v). \qquad (4)$$

The computed disparity offsets from the pair of video frames, produces result that forms homogeneous regions clustered according to the number of pixel difference. The disparity regions are labeled according to the results of component offsets shown in Fig 2(c). The labeled disparity map is solely integer-valued with regions having no smooth transitions. The transitions are then smoothed using a Gaussian filter.

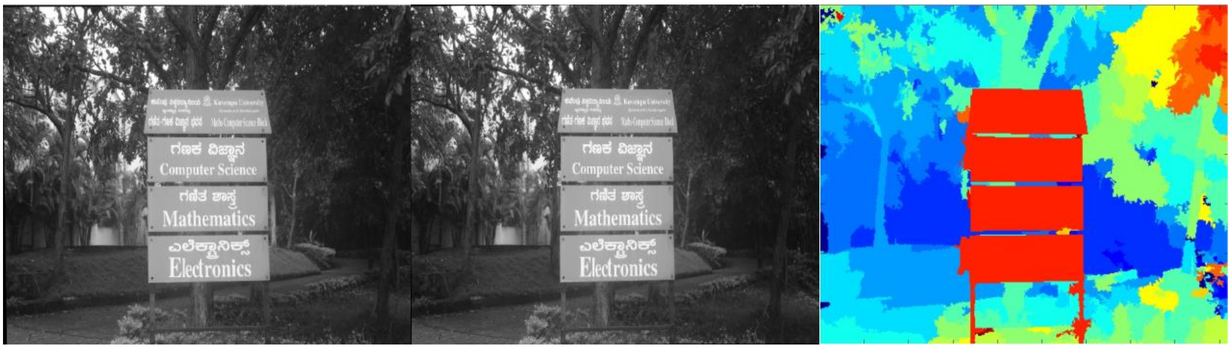
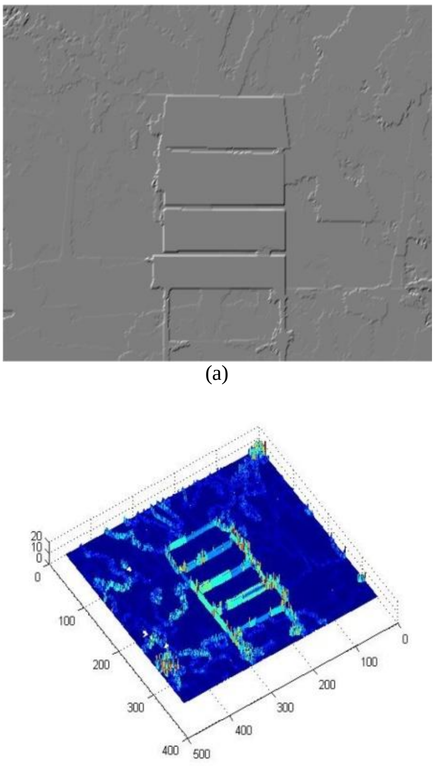

(a) (b) (c)

Fig.2. Results of Estimated Disparity Map: (a) Rectified Left Frame, (b) Rectified Right Frame and (c) Estimated Disparity Map.

B. Estimation of Gradient Map

The planar boundaries can be detected as discontinuities on the disparity map along x and y -axes [7]. Based on the property that disparity map of a planar surface exhibits a continuing spatial gradient, we estimate and detect candidate planar surface that represents text block. We detect the planar surface through the identification of characteristic significant region within the disparity map by computing two separate $1D$ gradient derivatives along the horizontal and the vertical directions. The absolute values from these two gradient maps are added together to approximate gradient magnitude G^t (Fig. 3(a)).

In order to identify the boundary for detected planar surface, the contour map's magnitude (Fig. 3(b)) is calculated by segmenting the approximated gradient map by a threshold value T with all values above it, is considered to be on a boundary and are set to $n_{(e)}$ (a significant negative value) proportional to the area under a convolution kernel σ and again, all the invalid or unsatisfactory values obtained within the gradient image reconstruction are set to the same value $n_{(e)}$. These boundary discontinuities comprise of positively valued pixels separating a wide homogeneous regions.



(b)

Fig.3. Intermediate Results of Our Approach: (a) Detected Planar Surface within the Gradient Map and (b) Its Corresponding Contour Map.

The whole contour map is convolved with a signum kernel, counting all the valid gradient values. During the computation of Σ^t , the point (u_Σ, v_Σ) depicting the maximum value is estimated to define a planar surface's seed point. The area $\sigma_x \times \sigma_y$ covered by kernel σ is the acceptable minimum size of a planar surface in the frame.

$$\Sigma^t = \sum_{y=v-\frac{\sigma_y}{2}}^{v+\frac{\sigma_y}{2}} \sum_{x=u-\frac{\sigma_x}{2}}^{u+\frac{\sigma_x}{2}} \mathrm{sgn}(G^t(x,y)+1). \qquad (5)$$

C. Fitting planar Model

The contour map shows the boundary of planar surface, and it helps in the extraction of planar surface. In order to extract the planar surface, first, we need to fit the planar model using RANSAC. We estimated the planar equation of the detected planar surface using surface normal parameters of the seed point (u_Σ, v_Σ) using PCA. A planar surface becomes a linear function mapping pixels from one image to its corresponding matching image. Let (x, y, z) be a point in world coordinate and (u,v) be a point in pixel coordinates then, a point w with a depth d in the $3D$ space from the viewing position can be represented by the plane equation as:

$$ax_w + by_w + cz_w = d. \qquad (6)$$

And for a non-zero depth it can be rewritten as:

$$a\frac{x_w}{z_w} + b\frac{y_w}{z_w} + c = \frac{d}{z_w}. \qquad (7)$$

The disparity of a planar surface can be estimated by mapping the above expression into image coordinate as:

$$au + bv + c = D(u,v), \qquad (8)$$

where $u = \frac{x_w}{z_w}$ and $v = \frac{y_w}{z_w}$, when the camera focal length $f = 1$.

Here, by using $2D$ video frames captured through the video cameras, it is attainable to estimate the distance of objects based on component offsets, where near objects portray higher component offset value, whereas farther objects have lower component offsets. By analyzing these disparity differences and grouping neighbor pixels with equal number of offsets, we can get pixels that belong to the same group, except region points that cannot be viewed by both cameras. These sets of pixels which cannot be captured by both cameras are called occluded regions and can be estimated partially. The computed sub-pixels produce extremely reliable and most correct values which can be exploited for further processing.

The planar surface that have large homogeneous regions based on constant gradient derivative values have local surface normal estimated throughout the region.

This property is used to estimate a local surface normal through PCA on the seed neighborhood with a valid disparity, and thereafter, RANSAC is used to fit a model from an array of $3D$ points based on the plane parameters. The main aim is to find large homogeneous regions in a disparity map, where we can fit a planar model and is done by generating a number of gradient regions from the disparity map. The regions are labeled as per the gradient magnitude values with plane support points extracted. The plane support points define local regions in the image that support planes of various orientations. We seek to maximally cluster similar labeled support points, which will classify and define the largest spatial region that may correspond to a single plane.

A local surface normal is computed for every homogeneous region. Every maximum valued point (from equation 5) within each labeled regions corresponding to the $p(u,v,D(u,v))$ is selected in each iteration to be a seed point of that region. The 3×3 neighborhood kernel centered on the seed point $p(u,v)$ is extracted as it gives a better overview of the region. Using the extracted kernel values, the PCA components' row-wise variance are computed from their corresponding means $\bar{p}_i$ and the row-wise means are calculated as:

$$\bar{p}_i = \frac{1}{N}\sum_{j=1}^{N} p(u+j-1,v), \tag{9}$$

where $i = 1,2,3,\ldots,9$.

The covariance matrix of the point is computed through the variance results using the equation shown below:

$$COV = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{N}(p_{i,j} - \bar{p}_i)^2. \tag{10}$$

A diagonal matrix of generalized eigen values are obtained from the above computed covariance matrix COV . The vector containing the eigen values are further processed to generate eigenvectors. The eigenvectors corresponding to the two biggest eigen values define the normal vector of the local neighborhood. To compute the normal vector $\hat{n}$ of the neighborhood, we calculate the cross product of the two eigenvectors' value.

$$\hat{n} = (\hat{\alpha},\hat{\beta},\hat{\varepsilon}). \tag{11}$$

We use RANSAC algorithm to robustly fit a plane to a set of 3D world data points generated from the disparity map. Random sample set $S = 2000$ points are taken from the disparity map with the in-lier distance threshold $t = 0.05$, in each iteration, a subset of points marked to support a local surface normal parameter $\hat{n} = (\hat{\alpha},\hat{\beta},\hat{\varepsilon})$ to fit a planar model using the following form of plane equation:

$$\hat{\alpha}u + \hat{\beta}v + \hat{\varepsilon}D(u,v) + \phi = 0. \tag{12}$$

The sample points S from the left reference video frame is used to find out if those points belongs to the most supported planar model and if true it is marked as an in-lier point (i), where $\forall i \in S$. When all the sampled points have been tested and all supporting points marked as in-liers, they are removed from the sampled points and an array of points are generated from their corresponding values to fit the planar model. The process is then repeated for the next planar model with again calculating the local surface normal of the neighborhood of the next largely supported planar model, the process iterates until 80% of sampled points have been fitted or when the RANSAC fails to find a consensus of the sampled points. A well supported plane is extracted since RANSAC finds a largest consensus [17], which can now be further processed by labeling all other pixels support these planar models through MRF technique discussed in the next section. Fig. 4 shows the fitted planar model.

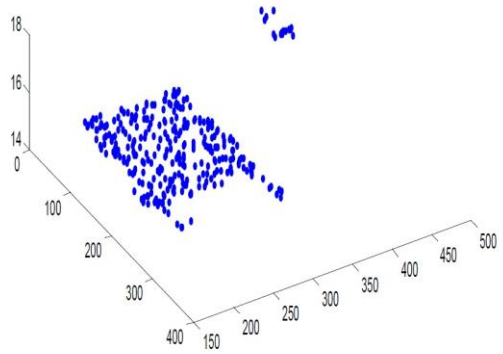


Fig.4. The Fitted Planar Model on the Disparity Image Shown in the Fig. 2(c).

D. Markov random Field Labeling

The problem that arises here is how to assign each pixel with a label that represents some local quantity, such as disparity, for the estimation of MRF which gives a framework to add the prior to the observation. This pixel labeling is naturally represented in terms of energy minimization where energy function has two terms: one term to penalize a solution inconsistent with observed data, while the other term to enforce some kind of spatial coherence [18].

We sought to infer the correct configuration of labels based on the mid-level and high-level representations by leveraging the work done by Jeffrey et al. [9] where the mid-level representations tries to infer the correct configuration of assigning a label either as a planar or a non-planar surface, while the high-level representations seeks to infer the labeled planar pixels into their corresponding candidate planar models based on their orientations through the computation of local surface normal. The initial labeling field is usually not optimal, and it can be possibly improved with the spatial-temporal constraints. This is performed through the energy minimization method based on MRF with Graph cuts technique as optimization algorithm for pixel labeling.

Optimization in MRF problem is finding the maximum of the joint probability over the graph, usually with some of the variables given by some observed data, or alternatively, it can be done by minimizing the total energy, which in turn requires the simultaneous minimization of all the clique potentials. In addition with the set of prior and likelihood information from the MRF labeling we estimated a maximum a posteriori by incorporating both of this information into a Bayesian inference.

$$pb(f/r) = pb(r/f)pb(f)/pb(r), \tag{13}$$

where $pb(f)$ of configuration f and the likelihood densities $pb(r/f)$ of the observation r.

Our representation consists of two coupled MRFs to label pixels belonging to the planar surface based on the confidence allocated to each pixel. The pixels that are contained in the image must be labeled according to their configuration. Each pixel must have a neighborhood that will determine the pixel labeling, by the prior identification of planar pixels which satisfy the planar equation. All other pixel neighbors are tested to see if it satisfies any of the plane equations and if true, a label corresponding to that plane candidate is assigned to it.

In the first MRF, the labeling of pixels takes $\{0,1\}$. All the pixels which have been positively identified belong to a plane are labeled 1 and pixels which do not belong to any plane are labeled with a zero (0). While in the second MRF the number of labels depends on the identified plane candidates thus $l = \{0,1,2,...,m\}$, candidate planes $c = \{1,2,...,m\}$ each shall be labeled. The zero is a label assigned to pixels which fail to belong to any candidate plane and forms the non-planar regions within the image lattice. We perform foreground (planar) and background (non-planar) separation using energy minimization through α-expansion of the Graph cuts [19] as shown in Fig. 5.

Fig.5. Labeling of Planar Surface from the Image Shown in Fig. 2(c) using MRF Image Segmentation.

IV. Text Information Extraction

The text information extraction process of our approach consists of a number of steps, which are employed to segment text regions from the background. The result of previous step consists of two types of regions which have been positively identified as planar region and non-planar regions. We mapped this result on the original left reference frame to extract all the features, which only falls under the planar region and discard the non-planar region by assigning a zero value to it.

The image map shown in Fig. 6(a) contains a number of regions which comprise of text candidate region (white portion) and two types of background: first, the part we shall name it global background (dark portion) which previously was part of non-planar region, and the second, named local background (orange portion) which is the planar surface portion that encloses the probable candidate text region, but its pixels are not part of candidate text.

We adapted the method proposed by Shivakumara et al. [5] to highlight and differentiate between text and non-text regions. Because of resultant image map (Fig. 6(a)) having low-contrast, we convert to gray-scale, then, transform it from spatial domain to frequency domain. An ideal low-pass filter is applied to smooth the noise which forms part of high-frequency components. A second order derivative of Laplacian operator is applied to produce a stronger response with fine details. It highlights the difference between text and non-text regions as its row-wise profile result reveals that text regions have higher number of positive and negative peaks when compared to non-text regions. The frequent zero crossings arising from positive and negative peaks correspond to transitions between text and background. Ideally, there should be the same number of text-to-background and background-to-text transitions.

The resultant filtered map is then transformed back to spatial domain for further processing to extract the maximum gradient difference (MGD) as shown in Fig. 6(b), which is defined as the difference between maximum and minimum values contained in rows within a window. The MGD is obtained by moving a local $1 \times N$ window over the filtered image and computing the differences between the maximum Mx_{val} and minimum Mn_{val} values as shown below:

$$Mx_{val} = \max_{\forall t \in [-N/2, N/2]} g(x, y-t), \tag{14}$$

$$Mn_{val} = \min_{\forall t \in [-N/2, N/2]} g(x, y-t), \tag{15}$$

$$MGD = Mx_{val} - Mn_{val}. \tag{16}$$

Text regions have larger MGD value as compared to non-text regions due to the larger magnitude of the positive and negative peaks. The k-means clustering is applied to classify all pixels into two clusters, candidate text C_1 with a cluster mean M_1, and non-text C_2 with the cluster mean M_2 based on the Euclidean distance of MGD values as shown below:

$$TextCL = \begin{Bmatrix} C_1 & if \;\; M_1 > M_2 \\ C_2 & Otherwise \end{Bmatrix}. \tag{17}$$

The pixel points from MGD results belonging to text regions are used in the color clustering part to identify the text cluster. We label all pixels within the region of Fig. 6(e) based on color clustering using k-means classification where they are clustered into two classes: 1) background (non-text) and 2) foreground (text). The two types of background we named local and global are then normalized to give a single background overview, then a binarization technique is applied to get a binarized image and a geometric analysis based on horizontal-vertical aspect ratio is further applied to the binarized image to filter out non-text components which are clustered as text but does not belong to candidate text. Fig. 6(f) shows the accurate text extraction result using our approach.

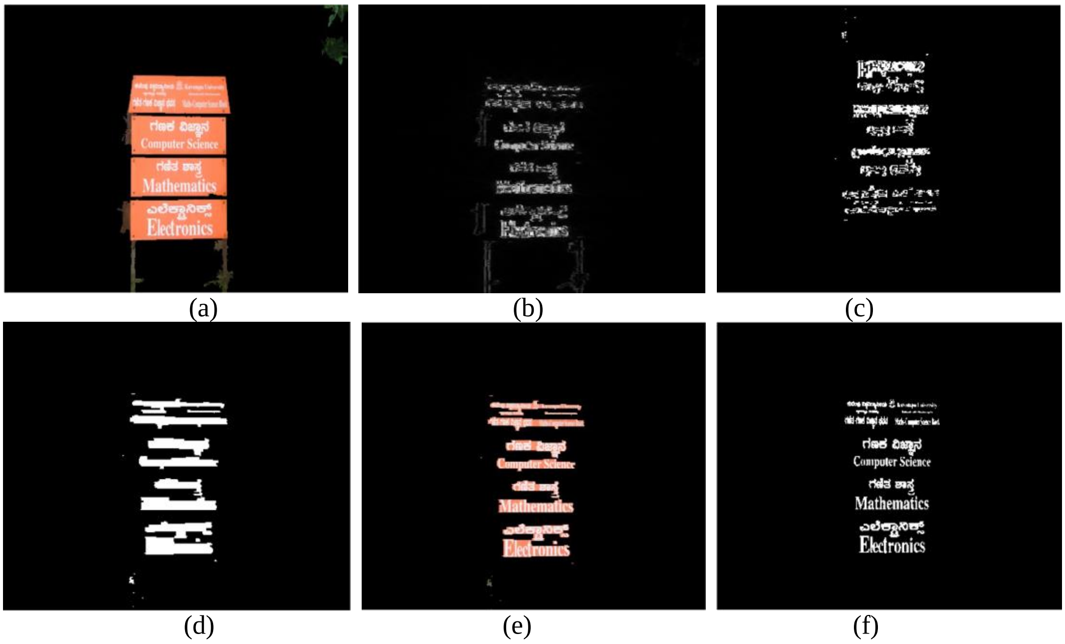


Fig.6. Intermediate Results of Proposed Method: (a) Mapping of Original Left Frame With Planar Surface Extracted from the Map Shown in Fig. 5, (b) MGD Results, (c) Candidate Text Cluster Result, (d) Candidate Text Cluster Result after Morphological Operation, (e) Text Localization and (f) Extracted Text.

V. Experimental Results

We believe that, this is the first work done directed to building a TIE technique based on stereo frames. Since, there is no standard dataset, which includes stereo frames of natural scene with text information. Our stereo camera setup consists of two similarly calibrated video cameras positioned horizontally 3-4 inches apart with a frame rate of 29.99 frames per second. The video sequences were captured only focusing on outdoor environment containing planar objects with text information. The video cameras are positioned just orthogonal to the objects of interest. The captured pair of video sequences of same scene has a length of between 5 to 10 seconds.

We created our own dataset from a pair of video sequences, which consists of high quality stereo frames. The stereo frames selected from video sequence must be of the same scene and contain text information. Our dataset contains one pair or more than one pair of frames per scene object and evaluation of our algorithm was carried out by giving input as one pair of frame per execution. The stereo frames are between 320×270 and 450×350 resolutions.

In our experiments, we extracted two types of planar orientations namely: 1) large regions that have a zero constant gradient value both vertically and horizontally in disparity space, and 2) regions that have zero constant gradient value vertically and a varied gradient value horizontally. These regions are assumed to be text-rich areas as they portray uprightness property which characterizes text planar surfaces.

Measuring the performance of text extraction is extremely difficult and until now there has been no comparison of the different extraction methods [3]. The performance of text extraction can be inferred from the text recognition results. Since, our work is focusing on text extraction rather than text recognition. We evaluated the performance of text localization process, which is one of the important steps in text extraction. Two metrics were adapted for the evaluation, including: 1) precision-the fraction of detections which are positives, 2) recall-the fraction of positives which are detected rather than missed.

A. *Perfomance Evaluation*

We have carried out experiments using dataset which contains stereo frames with horizontal text and experimental results were compared with other popular text extraction methods. The dataset contains 179 pairs of video frames, which were captured in outdoor scene with horizontal English and regional Kannada script. In the pair of video frames, only left frame were used for employing other popular methods such as edge-based method [20], CCA method [21], and gradient-based [5].

Ground truth was marked by hand on frames of dataset. Given the marked ground truth and detected result by the algorithm, we can automatically calculate the recall and precision. The precision and recall rates have been computed based on the area ratio r (in the equation below) of the bounding box between ground truth and result of our algorithm as shown in Fig. 7.

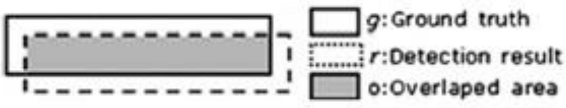


Fig.7. Illustration of the Overlap of A Ground Truth Box and Detected Bounding Box (Courtesy of Ye et al., 2005)

$$ratio(r) = \frac{Area\,(DetectedBox \cap GroundTruthBox)}{Area\,(DetectedBox \cup GroundTruthBox)}. \tag{18}$$

The following definitions were used for results evaluation process:

Truly Detected Box (TDB): A detected box truly detects the texts if the area ratio r, defined is at least 50%.

False Detected Box (FDB): The false detected box detects the texts if the area ratio r, defined is less than 50%.

Ground Truth Box (GTB): manually marked the text box by hand on test samples.

$$recall = \frac{(\#TDB)}{(\#GTB)}, \tag{19}$$

$$precision = \frac{(\#TDB)}{(\#TDB + \#FTB)}, \tag{20}$$

$$f - measure = \frac{(2 \times recall \times precision)}{recall \times precision}. \tag{21}$$

The Fig. 8 shows the result of our approach for extraction of planar surface from video frames of natural scene captured in complex background. The experimental result shows that the proposed method detects and extracts the planar surface accurately. The Table 1 shows the precision, recall and f-measure obtained for stereo frames of our dataset using existing methods and proposed method. The results shown in the table are the average value of precision, recall and f-measure computed for all the stereo frames of dataset.

Table 1. Comparison of Text Localization Results using Recall, Precision and F-Measure

method	precision	recall	f- measure
Edge-based method [20]	0.64	0.67	0.65
CCA method [21]	0.73	0.69	0.71
Gradient-based [5]	0.59	0.82	0.69
Proposed method	0.93	0.96	0.94

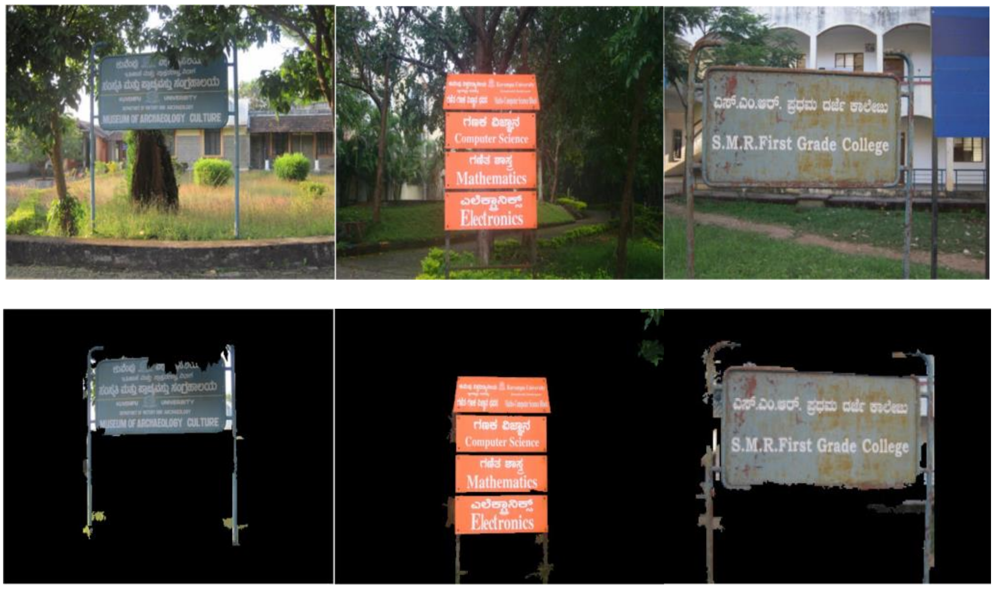


Fig.8. The Results of Planar Surface Extraction Obtained by Our Method on Video Frames of Our Dataset: First Row-Original Left Frames and Second Row-Extracted Planar Surface.

The proposed method has the highest recall, precision and f-measure compared to other existing methods. The experimental results show that the proposed method achieves high accuracy for scene text extraction for the video frames with complex background includes trees, building and other objects.

Fig. 9 shows the visual comparison of text localization results of our method with other popular existing methods for the stereo frames of our dataset. The connected component method locates the text area partially. It failed to locate the whole area of text. This method locates the exact localization of the text area for the first frame but includes background partially. The gradient based method locates the text area. It also locates the part of background and produces many false positives.

Edge based method failed to precisely locate text area, because scene images are rich in text like features such as tree-leaves which produces many edges and are classified as text. The second row of Fig. 9(c), which is an edge based result, is but just visually same as the original image, due to the overlapping of the bounding boxes which were used to enclose text area as shown also in Fig. 10(d).

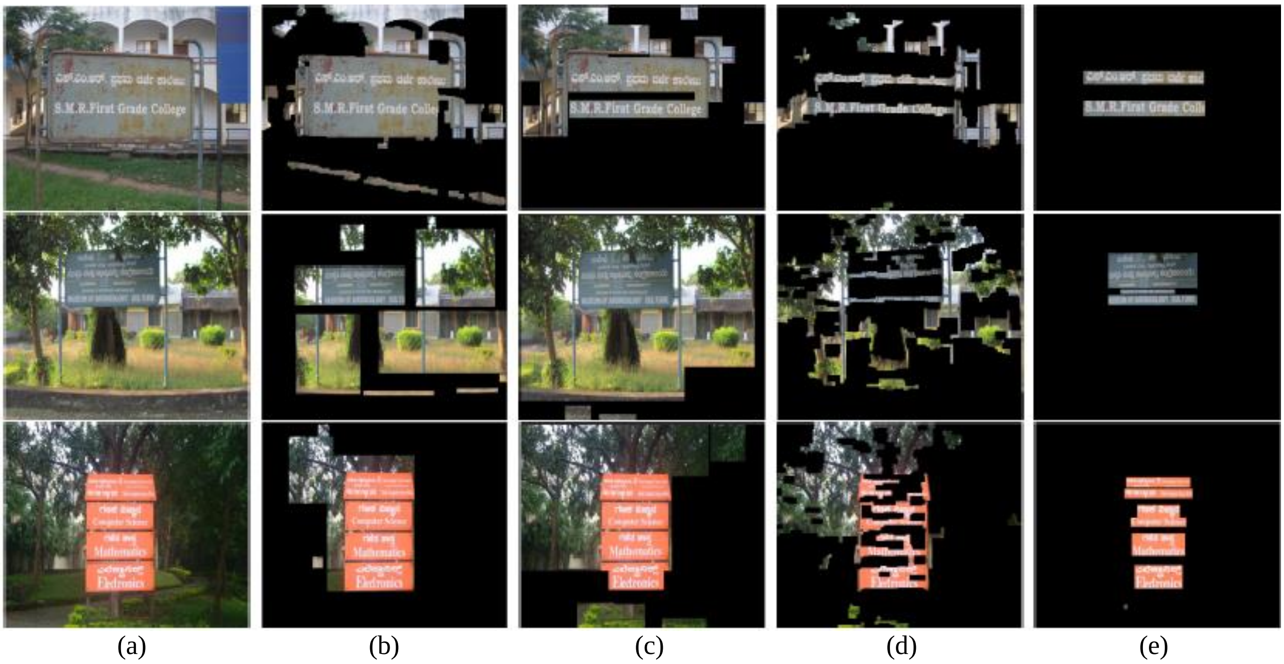


Fig.9. The Comparison of Text Localization Result on Our Dataset: Columns (a) Original Left Frames, (b) Results of Connected Component Method, (c) Results of Edge-Based Method, (d) Results of Gradient-Based Method And (e) Results of Proposed Method.

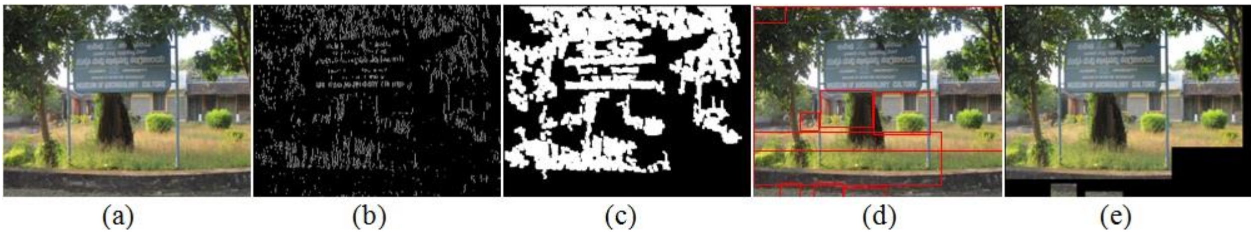


Fig.10. The Text Localization Results from Edge Based Method: (a) Original Left Frame, (b) Vertical Edge Results, (c) Morphological Operation Results, (d) Text Area Localization Based on Bounding Box And (e) Text Localization.

The proposed method first extracts the planar surface of Fig. 8, as the first part of text localization and thereafter, proceeds to process these regions by employing Fourier-Laplacian technique to accurately and precisely locate the text area. The Fig. 9(e) shows the text localization result of our approach for video frames of natural scene, which were captured in complex background. The proposed method locates all of the text area correctly for third frame, but it shows one false positive and two true negatives for the first and second frame respectively.

The true negative for the second frame is due to occlusion of tree-leaves. Our approach successfully locates the text area of the third frame, even though text is slightly oriented. The main limitation of the proposed method lies on its localization of the supporting poles of the sign board. This is due to the fact that, while extracting the planar surface, it considers the supporting poles as part of planar surface.

B. Disusssion

The literature survey reveals that majority of the images or video frames used in the experiments of TIE technique are having text information, which spanned the image from border to border. This means that the images or video frames are captured very close to the camera. Because of this property of the images or video frames, the edge-based or CC-based techniques perform well for this form of images. But this is not true in case of our dataset, because the edge-based or CC-based techniques show poor performance for our dataset. Our dataset consists of frames, which were captured far away from camera but at a considerable distance and text is not spanned from border to border.

Further, another complexity found in our dataset is complex background which includes trees, pathways and buildings. The experimental results were evaluated using recall, precision and f-measure.

Our approach achieves good performance compared to existing methods for our dataset. Based on the experimental results, we can conclude that our algorithm has excellent accuracy for localization evaluated through recall, precision, and f-measure.

Fig. 11 displays sample results of text extraction by our method for slightly oriented text information contained in video frames. The text extraction results for slightly oriented text demonstrates that our approach can be employed for both horizontal and slightly oriented text. Fig. 12 illustrates some examples of the localized text lines. It can be seen from the results that, most of the text is well detected and localized despite variations in character font, size, color, texture, script and orientation. We were able to get very good results for the video frames with complex background consisting mainly trees and some other non-planar objects, which were easily labeled as non-text planar properties. Fig. 13 illustrates some failure examples. In the first frame, the text is not localized accurately, the reason that, it contains multi-colored text. During color clustering it considers and clusters the text with high contrast color and text having low contrast color is treated as background at binarization stage. The false alarm in second frame is due to low contrast and text does not appear properly, because the frame was captured too far away from camera.

The proposed method was implemented using MATLAB, though the whole process of building TIE vision system based on outdoor scene image dataset took much time to complete on a PC with a 2.93 GHz core 2 duo processor and 2 GB memory.

Among our experiments, the worst experimental result was due to dominance of certain colors as black and white that made it difficult to generate pixel offsets as all the color channels had the same values and cannot give a reliable significance. The weakly textured planar surface also makes it hard for the algorithm to match points. The object targeted must be considerably not too close or too far from the camera in order to produce desired results.

VI. Conclusion

We presented a TIE technique for extraction of scene text from stereo frames acquired in the natural scene. We achieved a high accuracy rate both in planar surface extraction and text area localization. We performed planar surface extraction using MRF with Graph cuts technique by labeling regions based on estimated planar parameters. The text block (planar surface) is further processed to extract text by applying Fourier-Laplacian to generate text edges and classifying them using k-means clustering. The low false alarms rate ensure that our method extract more accurate text information in order to recognize the text properly. The fact that, our dataset is captured far away from the camera and our approach does not require any training of samples, therefore, our approach can be utilized for TIE applications without any further effort.

One of the important limitations to be noted in our approach lies on the computation time, as it becomes extremely expensive to incorporate all the algorithms described in this proposed approach. However, our approach is able to extract the text properly in complex background and is equivalent to human vision system, which can be incorporated to build a robust vision system application for mobile robot.

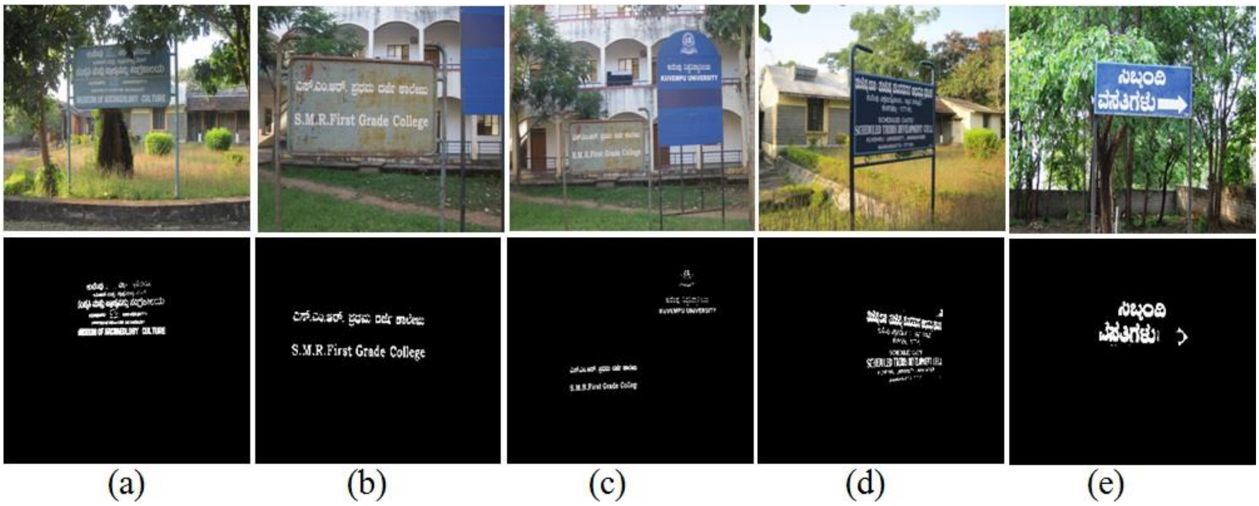


Fig.11. Sample Results of Text Extraction for Slightly Oriented Text by Our Approach: First Row-Original Left Frames and Second Row-Corresponding Extracted Text Information.

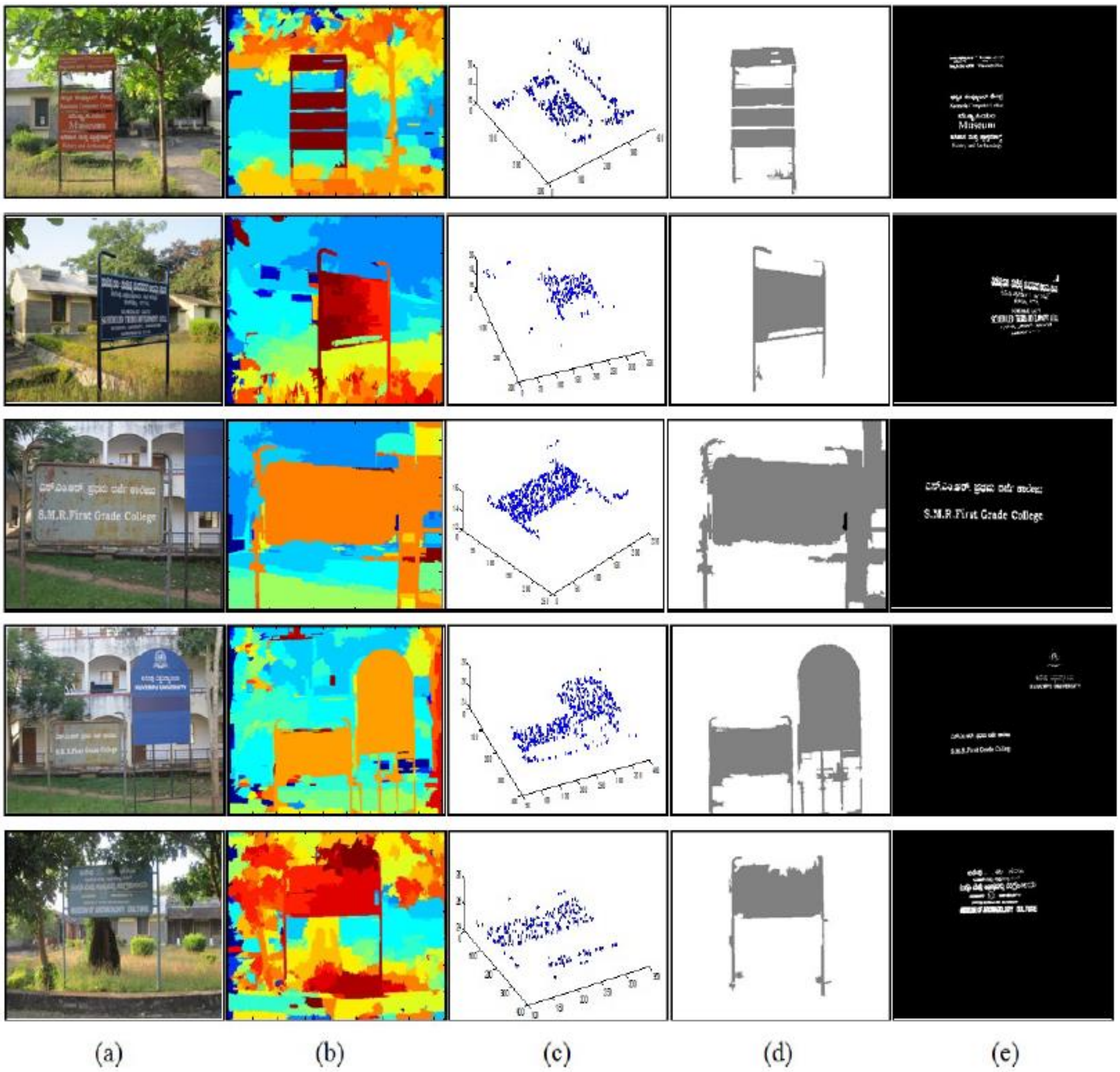


Fig.12. Examples of Text Extraction Results of Our Approach: Columns (a) Original Frames, (b) Disparity Maps, (c) Fitted Planar Models, (d) MRF Segmentation and (e) Text Extraction Results.

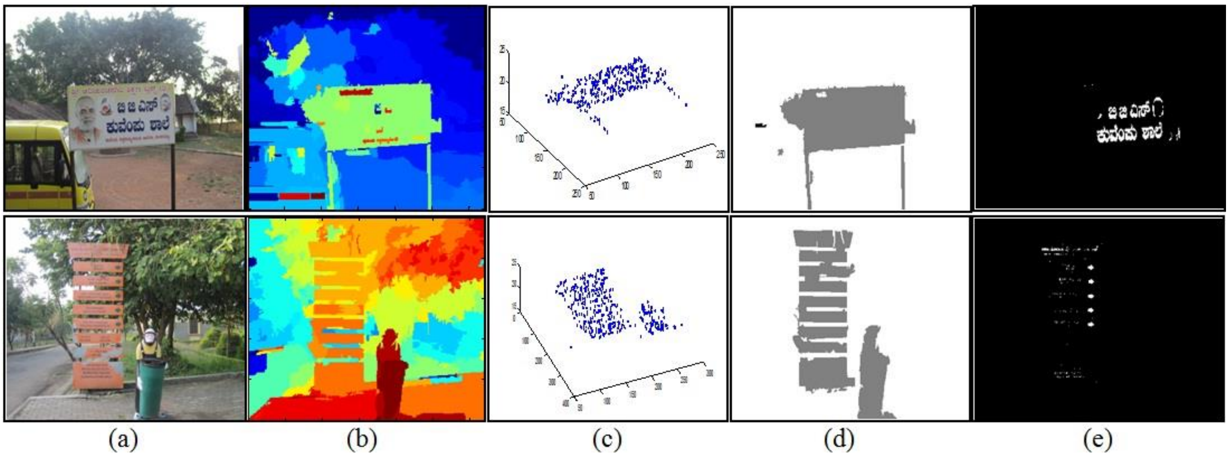


Fig.13. Frames with Failures and False Alarms of Our Approach: Columns (a) Original Images, (b) Disparity Map, (c) Fitted Planar Models, (d) MRF Segmentation and (e) Text Extraction Results.

REFERENCES

[1] Smith, Michael A., and Takeo Kanade, "Video skimming for quick browsing based on audio and image characterization*", School of Computer Science, Carnegie Mellon University*, 1995.

[2] Epshtein Boris, Eyal Ofek and Yonatan Wexler. "Detecting text in natural scenes with stroke width transform." *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010.

[3] Jung Keechul, Kwang In Kim, and Anil K Jain, "Text information extraction in images and video: a survey." *Pattern recognition* 37.5 (2004): 977-997

[4] Shahab A., Shafait F. and Dengel, A., "ICDAR 2011 robust reading competition challenge 2: Reading text in scene images*", International Conference on Document Analysis and Recognition (ICDAR), IEEE,* pages 1491-1496, 2011.

[5] Shivakumara Palaiahnakote, Trung Quy Phan and Chew Lim Tan, "A laplacian approach to multi-oriented text detection in video," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 33.2 (2011): 412-419.

[6] Bobick Aaron F. and Stephen S. Intille, "Large occlusion stereo." *International Journal of Computer Vision* 33.3 (1999): 181-200.

[7] Corso Jason, Darius Burschka and Gregory Hager. "Direct plane tracking in stereo images for mobile navigation." *Robotics and Automation, 2003. Proceedings. ICRA'03. IEEE International Conference on*. Vol. 1. IEEE, 2003.

[8] Konolige K., Agrawal, M., Bolles, R. C., Cowan, C., Fischle, M., & Gerkey, B. (2008, January). "Outdoor mapping and navigation using stereo vision", In *Experimental Robotics* (pp. 179-190). Springer Berlin Heidelberg.

[9] Jeffrey A. Delmerico,, Jason J. Corso, and Philip David. "Boosting with stereo features for building facade detection on mobile platforms." *Image Processing Workshop (WNYIPW), 2010 Western New York*. IEEE, 2010.

[10] Zhang, Shujun, Jianbo Zhang, and Yun Liu. "A Window-Based Adaptive Correspondence Search Algorithm Using Mean Shift and Disparity Estimation." *Virtual Reality and Visualization (ICVRV), 2011 International Conference on*. IEEE, 2011.

[11] Huang W., Lin Z., Jianchao Y., and Wang, J., "Text localization in natural images using stroke feature transform and text covariance descriptors," *IEEE International Conference on Computer Vision (ICCV),* pages 1241-1248, 2013.

[12] Lu S., Chen, T., Shangxuan, T., Joo-Hwee L., and Chew-Lim T., "Scene text extraction based on edges and support vector regression," *International Journal on Document Analysis and Recognition (IJDAR),* pages 1-11, 2015.

[13] Yi F. Pan, X. Hou and C.L. Liu, "Text localization in natural scene images based on conditional random field", *In ICDAR 2009, IEEE Computer Society, pages 6–10,* 2009.

[14] Yin X., Xuwang Y., Huang K. and Hong-Wei H. "Robust text detection in natural scene images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, issue 5, pages 970-983, 2014.

[15] Iqbal K., Xu-Cheng Y., Hong-Wei H., Sohail A. and Hazrat, A. "Bayesian net-work scores based text localization in scene images", *International Joint Conference on Neural Networks (IJCNN),* pages 2218-2225, 2014.

[16] Boaz T.K. and Prabhakar C.J., "Quality Assessment of Stereo Images using Reduced Reference Based on Saliency Region," *International conference on contemporary computing and informatics, IEEE,* pages 503-508, 2014.

[17] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[18] Szeliski, Richard, et al. "A comparative study of energy minimization methods for Markov random fields." *Computer Vision–ECCV 2006*. Springer Berlin Heidelberg, 2006. 16-29.

[19] Boykov, Yuri, Olga Veksler, and Ramin Zabih. "Fast approximate energy minimization via graph cuts." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 23.11 (2001): 1222-1239.

[20] Liu X., and Samarabandu J., "An edge-based text region extraction algorithm for indoor mobile robot navigation," *IEEE International Conference on Mechatronics and Automation*, Vol. 2, pages 701-706, 2005.

[21] Zhong Yu, Kalle Karu and Anil K. Jain. "Locating text in complex color images." *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*. Vol. 1. IEEE, 1995.

Authors' Profiles

Prabhakar C.J. received his Ph.D. degree in Computer Science and Technology from Gulbarga University, Gulbarga, Karnataka, India, in 2009. He is currently working as Assistant Professor in the department of Computer Science and M.C.A, Kuvempu University, Karnataka, India. His research interests are pattern recognition, computer vision, machine learning and video processing.

Too K. Boaz is a research scholar in Department of Computer Science and M.C.A, Kuvempu University, Karnataka, India. He received his Msc. Degree in Computer science from Bharathiyar University, Coimbatore, Tamil Nadu in 2011. Currently he is pursuing his Ph.D. in Computer Science in the area of Image Processing Applications. This paper is a part of his research work.