TEXT DETECTION AND RECOGNITION FROM IMAGES AS AN AID TO BLIND PERSONS ACCESSING UNFAMILIAR ENVIRONMENTS

Sherine Sebastian and Priya S.
Department of Computer Engineering, Model Engineering College, Thrikkakara, Ernakulam, Kerala, India
E-Mail: Sherinesebastian91@gmail.com

ABSTRACT

Independent travel is a well known challenge for blind or visually impaired persons. The text reading algorithm has proved to be robust in many kinds of real-world scenarios, including indoor and outdoor places with a wide variety of text appearance due to different writing styles, fonts, colors, sizes, textures and layouts, as well as the presence of geometrical distortions, partial occlusions, and different shooting angles that may cause deformed text. In this paper, we propose a method to detect panels and to recognize the information inside them. The proposal extracts local descriptors at some interest key points after applying color segmentation. Then, images are represented as a bag of visual words (BOVW) and classified using support vector machines. Finally, text detection and recognition method is applied on those images where a panel has been detected, in order to automatically read and save the information depicted in the panels. A language model partly based on a dynamic dictionary is also used.

Keywords: image processing, segmentation, bag of visual words, text detection, text recognition.

INTRODUCTION

Sign board detection and recognition using computer vision techniques has been an active area of research over the past decade. The text reading algorithm has proved to be robust in many kinds of real-world scenarios, including indoor and outdoor places with a wide variety of text appearance due to different writing styles, fonts, colors, sizes, textures and layouts, as well as the presence of geometrical distortions, partial occlusions, and different shooting angles that may cause deformed text. The proposal is to apply the text detection and recognition algorithm only on those images in which there is a sign board, in order to increase the efficiency of the system.

First, features at some key points are extracted in the train images and converted into feature descriptors, which are high dimensional vectors. The sampled features are clustered in order to quantize the space into a discrete number of visual words using fuzzy C-means clustering. The visual words are the cluster centers and can be considered as a representative of several similar local regions. The image can be represented by the histogram of the visual words, which counts how many times each of the visual words occurs in the image. The classes or categories of the input train images are learned by an SVM classifier.

Once the BOVW method finds that there is a sign board in an image, a text location and recognition method is applied on the image. The text location method is applied only on those areas of the image given by the color masks. The text detection is a texture based method. Then, character and word recognition is applied. A language model used to recognize single words, which is partly based on a fixed dictionary and partly based on a dynamic dictionary that depends on the province where the image was taken.

BACKGROUND

First Sign board detection and recognition using computer vision techniques has been an active area of research over the past decade. A good survey about the main vision-based proposals of the state of the art for intelligent driver assistance systems can be found in [1], where a discussion about the future perspectives of this research line is there included. Additionally, the work in [2] presents a recent contribution about an intelligent road sign inventory based on image recognition, which is related to the application proposed in this paper, but for traffic signs instead of traffic panels and using images taken from a vehicle instead of images served by Google Street View. Apart from a previous work of the authors in [3] where an automatic traffic signs and panels inspection system using active vision at night is presented, only two works have been developed in this matter.

The work proposed in [4] extracts candidates to be traffic panels, using a method that detects blue and white areas in the image using the hue and saturation components of the hue-saturation intensity space. Then, candidates are classified according to their shapes. This is done through a method that correlates the radial signature of their fast Fourier transform with a pattern corresponding to an ideal rectangular shape. Then, panel reorientation is carried out using a homography that aligns the four vertices of each blob. Segmentation of the foreground objects from the background of the panel is done by analyzing the chrominance and luminance histograms. Connected components labeling and position clustering is finally done for the arrangement of the different characters on the panels. This algorithm is invariant to translations, rotations, scaling, and projective distortion, but it is severely affected by changing lighting conditions. Recognition is applied at character level, but no language model is applied to correct misspelled words. There is not any information on where and how the images are extracted.

On the other hand, Wu *et al.* [5] proposes a method to detect text on traffic panels from video. First, regions of the same color are extracted using a k-means algorithm, and traffic panel candidates are detected by searching for flat regions perpendicular to the camera axis. This method needs an accurate tracking method to detect corresponding points in successive frames. Furthermore, a multi-scale text detection algorithm is performed on each candidate traffic panel area. The text detection method integrates edge detection, adaptive searching, color analysis using Gaussian mixture models, and geometry alignment analysis. A minimum bounding rectangle is fitted to cover every detected text line. A feature-based tracking algorithm is then used to track all detected areas over the timeline as they are merged with other newly detected texts in the sequence. Finally, all detected text lines are extracted for recognition, but the authors do not comment how the recognition is carried out. In terms of text detection, this method provides good results under different lighting conditions, and it is not affected by rotations and projective distortions.

A focus of attention based system for text region localization has been proposed by Liuand Samarabandu in [6]. The intensity profiles and spatial variance is used to detect text regions in images. A Gaussian pyramid is created with the original image at different resolutions or scales. The text regions are detected in the highest resolution image and then in each successive lower resolution image in the pyramid. The approach used in [7, 8] utilizes a support vector machine (SVM) classifier to segment text from non-text in an image or video frame. Initially text is detected in multi scale images using edge based techniques, morphological operations and projection profiles of the image [8]. These detected text regions are then verified using wavelet features and SVM. The algorithm is robust with respect to variance in color and size of font as well as language. The existing techniques are categorized into the following:

- Edge based detection method: Edge based method focus on high contrast between the background and text and the edges of the text boundary are identified and merged. Later several heuristics are required to filter out the non - text regions.

- Connected-component based detection method: Connected component based methods use bottom up approach to group smaller components into larger components until all regions are identified in the image. A geometrical analysis is later needed to identify text components and group them to localize text regions.

- Texture based method: Texture based method is a feature based algorithm which involves the construction of gray-level co-occurrence matrix. This matrix is used to calculate the features like contrast, homogeneity, dissimilarity and which are the results for feature extraction in texture based method.

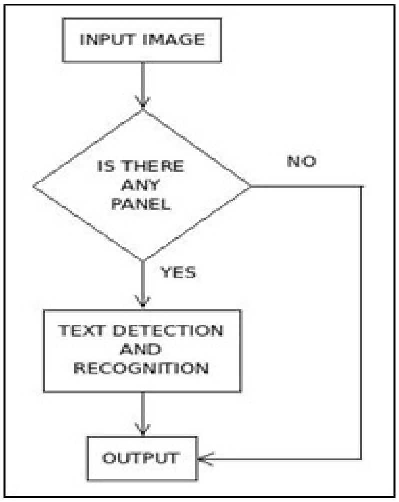

Figure-1. Flow chart of the proposed system.

In the paper, a novel approach to model sign boards using visual appearance, specifically a BOVW technique from local descriptors extracted at interest key points, unlike the typical methods in the state of the art that use other features such as edges or geometrical characteristics is used[9]. A previous color segmentation stage guides the key points searching in the image. A method based on color segmentation and the BOVW algorithm is applied on each frame to detect a sign board. In addition to the color segmentation proposed in [9], a method to find green mask is also proposed. In case a panel is detected, the text detection based on texture based method is proposed. Later, a text recognition algorithm is applied. Then a language model is used to correctly identify the words in the sign board.

METHODOLOGY

Sign board detection and recognition using computer vision techniques has been an active area of research over the past decade. Sign board detection and recognition has been out of the scope of researchers because, on the one hand, they are informative signs and then they have less priority than the regulatory or warning signs. On the other hand, there is a wide diversity of information contained in sign boards, which is difficult to analyze. The text detection and recognition algorithm is applied only on those images in which there are a sign board, in order to increase the efficiency of the system. For this purpose, a sign board detection method has been used, which is based on color segmentation, a BOVW approach and a texture based text detection method. The flow chart of the proposed system is as shown in the Figure-1.

Sign Board Detection

It is based on color segmentation and a BOVW approach. We have chosen this technique since it has become one of the most popular in terms of classifying images. The BOVW method stems from text analysis, wherein a document is represented by word frequencies without regard to their order. These frequencies are then used to perform document classification. The BOVW approach to image representation follows the same idea. The visual equivalent of words is local image features. Therefore, the BOVW technique models an image as a sparse vector of occurrence counts of the vocabulary of local image features. In other words, it translates a very large set of high-dimensional local descriptors into a single sparse vector of fixed dimensionality (a histogram) across all images.

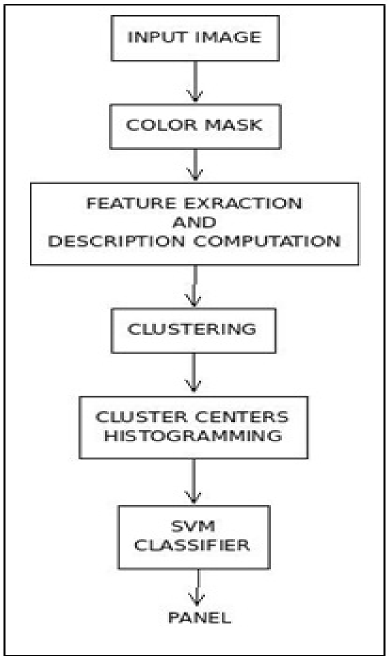
INPUT IMAGE
COLOR MASK
FEATURE EXRACTION AND DESCRIPTION COMPUTATION
CLUSTERING
CLUSTER CENTERS HISTOGRAMMING
SVM CLASSIFIER
PANEL

Figure-2. Flow chart of the proposed system.

As specified in the paper [9], a segmentation stage is done in order to guide the searching of key points over the potential areas to be panels in the image. In this way, we maximize the panels modeling again other areas of the image. The color masks are hypotheses to be confirmed through the extracted key points and bag of features approach. Then, using a prior color segmentation mask and a BOVW technique in cascade is an alternative approach for sign board detection without using edges or other geometrical features, as it has been done up to now in the literature. This technique can be generalized to detect any object characterized by a uniform color background in the image.

A slight modification is done in the method used in [9] to segment the green regions in the image as a combination of three independent methods using a logical AND operation as in

$$\text{Green Mask} = g_1(x, y) \text{ AND } g_2(x, y) \text{ AND } g_3(x, y) \quad (1)$$

$g_1(x, y)$ is computed using

$$g_1(x, y) = \begin{cases} 255 \text{ if } R(x, y) \le T_r \\ 0 \text{ otherwise} \end{cases} \quad (2)$$

as it is proposed in [10]. $R(x, y)$ is the red channel of the image, and T_r = 90 is the optimum value according to the source article. This method has been proved to be really useful to discard the green regions corresponding to the tree, while keeping the green regions corresponding to the panels. On the other hand, this method has the disadvantage that it is not able to reject dark regions in the image (black, gray, and dark colors). This is solved using the next two methods.

On the other hand, $g_2(x, y)$ is computed using

$$g_2(x, y) = \begin{cases} 255 \text{ if } H(x, y) \ge T_1 \text{ and } H(x, y) \le T_2 \\ 0 \text{ otherwise} \end{cases} \quad (3)$$

H(x, y) is the hue component of the image, and T_1 =80º and T_2 = 160º are the optimum values of the thresholds as we proposed. Unlike the previous method, this method is not able to distinguish between the green regions in the trees and the green regions in the panels, and it is not able to discard white regions in the image, but it is very useful to reject colors whose tonality is completely different to green, such as blue, red, or orange.

$g_3(x, y)$ is computed using

$$g_3(x, y) = \text{Otsu } (|G(x, y) - B(x, y)| \quad (4)$$

which applies Otsu's segmentation method on the image obtained by subtracting the blue color component B(x, y) from the green color component G(x, y). Otsu's method reduces the input image to a binary image, assuming that the input image contains two classes of pixels or a bimodal histogram. It computes the optimum threshold that separates both classes so that their intraclass variance is minimal. Unlike the first method, this method is not able to discard the blue regions that correspond to the sky, but it improves the performance of the first method by rejecting dark regions in the image and it improves the performance of the second method by rejecting white regions in the image. The input image is shown in the Figure-4 and the image after applying the color mask is shown in the Figure-5 as the white region.

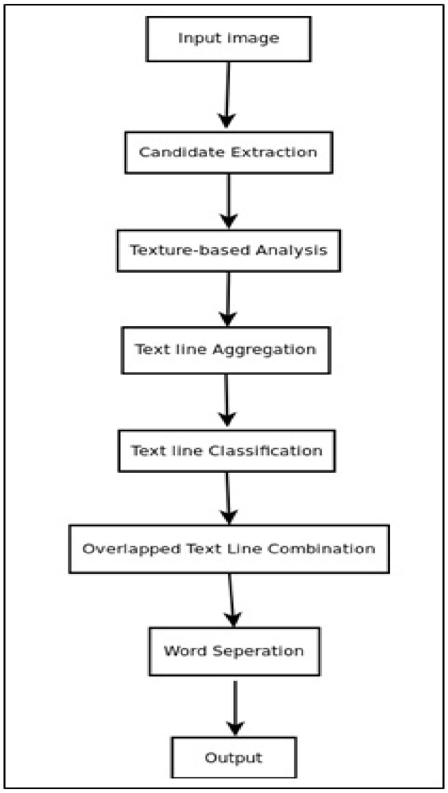

Figure-3. Flow chart of text location system.

Next, the features at some key points are extracted in the train images and converted into feature descriptors, which are high dimensional vectors. The features are extracted at some key points, which are obtained using the Harris-Laplace salient point detector. It uses a Harris corner detector and subsequently the Laplace operator for scale selection. Good descriptors should be able to handle intensity, rotation, scale, and affine transformations. SIFT descriptor is used as the feature descriptor, because of its highest value of the specificity. It means that the number of false positives for this descriptor is very low.

Figure-4. Input image.

Figure-5. Region corresponding to the sign board.

The sampled features are clustered in order to quantize the space into a discrete number of visual words using fuzzy C-means clustering. The visual words are the cluster centers and can be considered as a representative of several similar local regions. The image can be represented by the histogram of the visual words, which counts how many times each of the visual words occurs in the image. The classes or categories of the input train images are learned by an SVM classifier.

Given a test image, the nearest visual word is identified for each of its features. A BOVW histogram is computed to represent the whole image, and the classification decision is made by the classifier previously trained. Feature extraction, training, and testing are done separately on each region of interest to detect the sign board.

Text Detection and Recognition from the Sign Board

Once the previous method finds a sign board in the image, a texture based method is applied for text detection [11]. The texture based method is a feature based algorithm which involves the construction of gray-level co-occurrence matrix. This matrix is used to calculate the features like contrast, homogeneity, dissimilarity and which are the results for feature extraction in texture based method. The proposed method consists of five phases:(1) Background suppression in DCT domain, (2) Text feature extraction, (3) Texture classification, (4) Merging, (5) Refinement. It is explained in the algorithm Text Detection.

Algorithm 1: Text Detection

1. Divide the input image into 8x8 blocks and apply DCT for each block.
2. Suppress the background of image using high pass filter.
3. Perform inverse DCT on each block to obtain processed image.
4. Divide the processed image into 50X50 blocks.
5. Calculate the features homogeneity and contrast at 0^o, 45^o, 90^o, 135^o orientations for each block
6. Filter the non-text blocks using text features and discriminant functions.
7. Merge the obtained text blocks into text regions.
8. Refine the size of the detected text regions to cover the missed text present in undetected blocks and unprocessed regions.

After finding the letter candidates by DCT and inverse DCT methods as described in the algorithm, filtering is done. An important cue for text is that it appears in a linear form. Text on a line is expected to have similarities, including similar stroke width, letter width, height and spaces between the letters and words. We consider each pair of letter candidates for the possibility of belonging to the same text line. Two letter candidates should have similar stroke width.

The distance between letters must not exceed three times the width of the wider one. Additionally, average colors of candidates for pairing are compared, as letters in the same word are typically expected to be written in the same color. At the next step of the algorithm, the candidate pairs determined above are clustered together into chains. Initially, each chain consists of a single pair of letter candidates. Two chains can be merged together if they share one end and have similar direction. The process ends when no chains can be merged. Each produced chain of sufficient length is considered to be a text line. Finally, text lines are broken into separate words, using a heuristic that computes a histogram of horizontal distances between consecutive letters and estimates the distance threshold that separates intra-word letter distances from inter-word letter distances.

Some modifications and new functionalities have been proposed in order to increase the efficiency and reduce the number of false positives. Instead of applying the text location method in the whole image, it is done only on those areas of the image given by the blue, green and white color masks. Once the text is detected the text recognition method proposed in [9] is applied. The character recognizer described in [12] was developed to recognize letters from “A” to “Z” and from “a” to “z” and digits from “0” to “9”. The character recognizer may fail when sign boards are far, as text is small and difficult to segment it and recognize. However, it is not necessary to recognize all the characters perfectly. They are just estimation, because a word recognizer is applied later. The word recognizer is based on a unigram probabilistic language model that constrains the output of the character recognizer to a set of meaningful words weighted to their prior probabilities. The flow chart of the text location method is shown in Figure-3. The output of the proposed system is as shown in Figure-7.

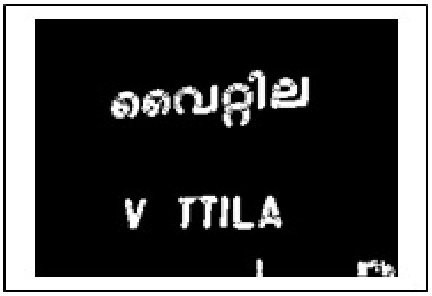

Figure-6. Detected text from the sign board.

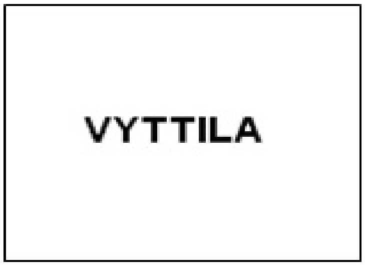

Figure-7. Output image.

CONCLUSIONS

A novel approach to model sign board using visual appearance, specifically a BOVW technique from local descriptors extracted at interest key-points, unlike the typical methods in the state of the art that use other features such as edges or geometrical characteristics is proposed. The sign board is modeled using a BOVW technique from local descriptors extracted at interest key points which is not an easy task due to the immense variability of the information included in sign board. An efficient segmentation method based on color masks has been used to guide the key points searching in the image. Texture based method, a feature based algorithm which involves the construction of gray-level co-occurrence matrix has been proposed for the text detection. Text recognition system which includes a symbol recognizer for sign board, a method to reduce the size of the dictionary to a limited geographical area is also used in the project.

ACKNOWLEDGEMENTS

The authors would like to thank de Campos, T.E., Babu, B.R., Varma M, for the char74K dataset, which have been used only for research purposes.

REFERENCES

[1] Mogelmose, M. Trivedi and T. Moeslund. 2012. “Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey ,” IEEE Trans. Intell. Transp. Syst., Vol. 13, No. 4, pp. 14841497.

[2] Wu, X. Chen and J. Yang. 2005. “Detection of text on road signsfrom video”, IEEE Trans. Intell. Transp. Syst., Vol. 6, No. 4, pp.378-390.

[3] V. Reina, R. J. L. Sastre, S. L. Arroyo and P. G. Jimnez. 2006. “Adaptive traffic road sign panels text extraction ,” In Proc. 5th WSEAS ISPRA.

[4] Z. Hu. 2013.“Intelligent road sign inventory (IRSI) with image recognitionand attribute computation from video log, “ Comput.- Aided Civil Infrastruct. Eng., Vol. 28, No. 2, pp. 130-145.

[5] Yau-Chat Tsoi and Michael S. Brown. 2004. “Geometric and ShadingCorrection for Images of Printed Materials A Unified Approach Using Boundary”, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

[6] Qixiang Ye, Wen Gao, Weiqiang Wang and Wei Zeng. 2003. “A Robust Text Detection Algorithm in Images and Video Frames”, IEEE.

[7] de Campos T.E., Babu B.R., Varma M.. 2009. “Character recognition in natural images,” VISAPP, 05-08.

[8] Xiang Zhu, Scott Cohen, Stephen Schiller and Peyman Milanfar. 2013. "Estimating Spatially Varying Defocus Blur From A Single Image", IEEE Transactions on Image Processing, Vol. 22, No.12, pp. no. 4879-4891.

[9] Alvaro Gonzlez, Luis M. Bergasa and J. Javier Yebes. 2014. 'TextDetection and Recognition on Traffic Panels From Street-Level Imagery Using Visual Appearance," IEEE Transactions on Intelligence Transportation Systems, February. Gonzlez.

[10] N. Kulkarni. 2012. "Color thresholding method for image segmentation of natural images," Int. J. Image, Graph. Signal Process., Vol. 4, No. 1, pp. 28–34.

[11] M. Garrido, D. Llorca, M. Gaviln, J. Fernandez, P. Alcantarilla, I. Parra, F. Herranz, L. M. Bergasa, M. Sotelo and P. Revenga. 2011. "Use of the Hough Transformation To Detect Lines and Curves in Pictures, Automatic traffic signs and panels inspection system using computer vision," IEEE Trans. Intell. Transp. Syst., Vol. 12, No. 2, pp. 485-499.

[12] Gonzlez and L. M. Bergasa. 2013. "A text reading algorithm for natural images," Image Vis. Comput., Vol. 31, No. 3, pp.255-274.

[13] Datong chen, JuergenLuettin and Kim Shearer. 200. "A Survey of TextDetection and Recognition in Images and Videos", IDIAP – RR ,00-38, August.

[14] Arpit Jain, XujunPeng, XiaodanZhuang, Pradeep Natarajan and Huaigu Cao. 2014. "Text Detection And Recognition In NaturalScenes And Consumer Videos", IEEE International Conferenceon Acoustic, Speech and Signal Processing (ICASSP), pp.1245-1249.

[15] Xiaoqing Liu and JagathSamarabandu. 2005. "A Simple and FastText Localization Algorithm for Indoor Mobile Robot Navigation", Proceedings of SPIE-IS and T Electronic Imaging, SPIEVol. 5672.

[16] Qixiang Ye, Qingming Huang, Wen Gao and Debin Zhao. 2005. "Fast and Robust text detection in images and video frames", Image and Vision Computing 23.

[17] M. Swamy Das, B. Hima Bindhu and A. Govardhan. 2012. "Evaluation of Text Detection and Localization Methods in Natural Images," International Journal of Emerging Technology and Advance Engineering ISSN 2250-2459, Vol. 2, No. 6.

[18] FanmanMeng, Hongliang Li, Guanghui Liu, and King NgiNgan, "From Logo to Object Segmentation", IEEE TramsactionsOn Multimedia, Vol. 15, No. 8.

[19] Zhou Dengwen. 2010. "An Edge-Directed Bicubic Interpolation Algorithm",3rd International Congress on Image and Signal Processing (CISP2010).

[20] Adrian Ulges, Christoph H. Lampert and Thomas M. Breuel. 2006. "Document Image Dewarping using Robust Estimation of Curled Text Lines", IEEE.