International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

*Research Article*

# Localization and Recognition of Text with Perspective Distortion in Natural Scenes

**Annmaria Cherian†* and Sanju Sebastian†**

†Department of ECE, St. Joseph's College of Engineering and Technology, Palai, Kerala, India

Accepted 03 Nov 2015, Available online 10 Nov 2015, **Vol.5, No.6 (Dec 2015)**

***Abstract***

*Recognizing text in natural scene images refers to the problem of identifying words that present on it. Scene text recognition is very difficult due to some reasons such as, images contain very little amount of linguistic context, interpreting versions of letters and digits are required for scene text recognition and also scene text can appear in any orientation. Most of the existing works are focused on the recognition of texts which are frontal parallel to the image plane. We formulate a novel method which is used to recognize text in natural scene images which are perspectively distorted. Perspective distortion is avoided using Hough transform. Each character are recognized from cropped word image. Connected component analysis is used to detect the components that present on the cropped word image. Non text components are filtered using SVM classification. After that text components are recognized by Optical character recognition. We introduce a new dataset called Scene Text-Perspective, which contains scene images of the name boards placed in the road sides which are perspectively distorted. Experimental results on the proposed dataset shows that our method is simple and outperforms the existing methods.*

***Keywords:*** *Scene text recognition, connected component analysis, SVM.*

## 1. Introduction

Detection of texts and identifying the character in the scene images is a challenging visual recognition problem. If such texts can be recognized, they can be used for a wide range of applications like intelligent driving assistance, navigation aid for the visually-impaired and robots. Compare to the more structured recognition of text from machine-printed documents, known as optical character recognition (OCR), scene text recognition is more difficult for several reasons. First, viewing angle and lighting are typically not carefully controlled. Second, there is usually very little linguistic context are present in the images, other objects like buildings and vehicles can interfere the recognition process. Finally, many signs are meant to be eye-catching, and are designed with stylized or with unusual fonts. In other words, the scene text recognition task frequently requires interpreting versions of letters and digits that are significantly different than those seen in training.

Although there are existing works to recognize text in natural scene images but their scopes are limited to the texts which are frontal parallel to the image plane. However, in practice, scene texts can appear in any orientation, and with perspective distortion. The presence of perspective distortion, distract the human readers and makes image analysis operations such as scene text recognition, layout analysis and compression become slower or less reliable. In this project, we attempt to focus on the recognition of texts in natural scene images with perspective distortion. Figure 1 shows the natural scene image which are prospectively distorted.

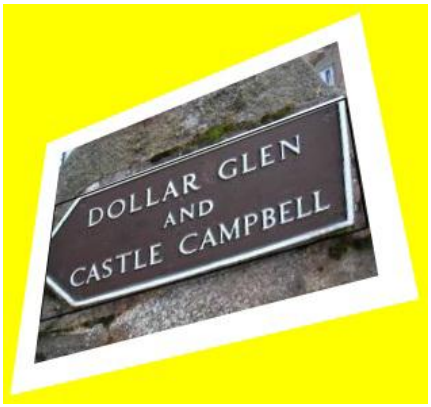


**Fig. 1** Natural scene images with perspective distortion

Weinman *et al*. and Smith *et al*. proposed the recognition method to take the label which were visually similar to the characters. However, these methods were tested only on simple sign images which

*Corresponding author: **Annmaria Cherian**

are appeared on plain backgrounds. Dance *et al*. and Myers *et al*. proposed the recognition method for perspective texts which requires high quality binarized character shapes. Although they work for texts on plain backgrounds, this method cannot handle texts with cluttered backgrounds (as in street images). Gandhi *et al*. rectified perspective texts in image sequences by utilizing the motion information. This method requires camera calibration, and does not work for still images.

In this paper a novel method is introduced to detect the text in natural scene images with perspective distortion. Perspective distortion can be avoided by finding the longest line in the image using hough transform. By using the longest line, angle of the image is rotated and distortion is avoided. Then cropped word image is found out using text localization. Different components present in the cropped word image are extracted using connected component analysis. Non text regions are filter out from the text region using SVM classification. HOG feature descriptors are used for the SVM classification. Features which are able to handle the different character poses are not required for the classifications, since perspective distortion is already avoided. The major advantage of this work is it does not require to collect enough training samples for a large number of character classes (62 classes for English characters and digits). Only frontal character samples are required for training. After that each text component is classified for recognition. Optical character recognition is used for character recognition.

The organization of this paper as follows : Section 2 illustrate the system model for Character detection and recognition. Section 3 gives the idea about the dataset used in this work. The experimental results is discussed in section 4, and the conclusion is arrived at Section 5.

## 2. Character detection and recognition

Block diagram of the proposed methodology is shown in figure 2.

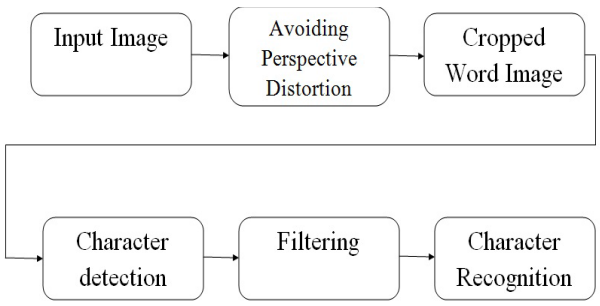


**Fig.2** Block Diagram

### *2.1 Hough Transform*

The Hough transform is a feature extraction technique used in computer vision, image analysis, image processing. This technique is used to isolate features of a particular shape within an image. The main advantage of the Hough transform technique is that it is tolerant of gaps in feature boundary descriptions and it is unaffected by image noise. The classical Hough transform was concerned with the identification of lines in the image, but later the Hough transform has been extended to identifying positions of arbitrary shapes, most commonly circles or ellipses. Detecting straight lines is the simplest case of Hough transform. The straight line y = mx+b can be represented as a point (b;m) in the parameter space. However, vertical lines have a problem. They would give rise to unbounded values of the slope parameter m. Thus, for computational reasons, Hesse normal form is used and the parametric representation of a line is:

$$r = x \cos\theta + y \sin\theta \quad (1)$$

where r is the distance from the origin to the closest point on the straight line, and θ is the angle between the x axis and the line connecting the origin with that closest point.

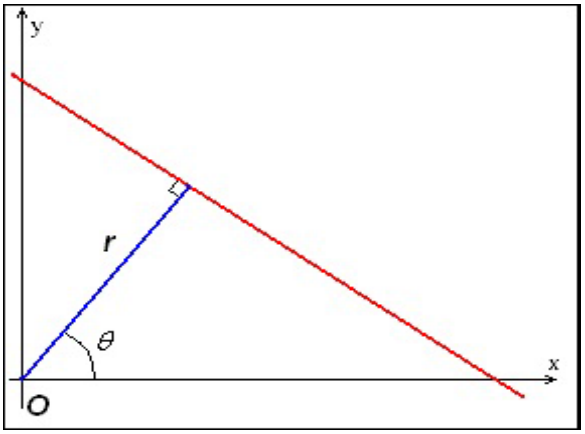


**Fig.3** Hough transform

It is therefore possible to associate with each line of the image a pair (r,θ). The (r,θ) plane is sometimes referred to as Hough space for the set of straight lines in two dimensions. After finding the hough transform, Peaks in the Hough transform is identified and the line segments are extracted. The longest line segment is found out and angle is calculated. The perspective distortion of the image is avoided by rotating the image based on the angle.

### *2.2 Text Localization*

Text localization is used to find the cropped word image. This work is limited to cropped word image. In order to obtain cropped word image we want to find the presence of text in the image. Canny edge detection is used for the text localization. The Canny edge detector is an edge detection operator that uses a multistage algorithm to detect a wide range of edges in images. The Canny algorithm is adaptable to various environments.

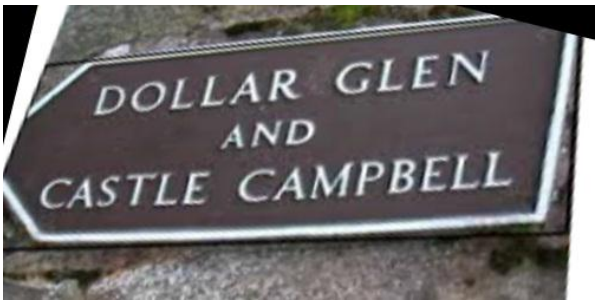


**Fig. 4** An example of cropped word image

Its parameters allow it to be tailored to recognition of edges of differing characteristics depending on the particular requirements of a given implementation. Canny edge detection is a four step process.
1) A Gaussian blur is applied to clear any speckles and free the image of noise.
2) A gradient operator is applied for obtaining the gradient's intensity and direction.
3) Non maximum suppression determines if the pixel is a better candidate for an edge than its neighbours.
4) Hysteresis thresholding finds where edges begin and end.
After obtaining the edge filtered image the holes are filled and find the region having maximum area. Finally locate the region having text.

### 2.3. Connected Component Analysis

Most of the scene text detection algorithms in the literature can be classified into region-based and connected component (CC)-based approaches. Region-based methods adopted a sliding window scheme, which is basically a bruteforce approach which requires a lot of local decisions. Therefore, the region-based methods have focused on an efficient binary classification Algorithm used for extracting components using connected components is given below.

1) An image A is convert it into binary image.
2) Define a structuring element B.
3) Initialize the Label matrix with zeros.
4) Find a non zero element position in the input matrix A.
5) Initialize a matrix X with zeros and place 1 in the non zero element position found in the previous step.
6) Perform dilation using the structuring element B on matrix X.
7) Perform intersection with the matrix A.
8) Check whether Y==X. If no, then X=Y and perform steps 6 and 7 again else stop the iteration.
9) Find the non zero elements position in the Y. In matrix Label place a number N in those positions. N is for labeling the connected components.
10) Similarly, place zero in those positions in the input matrix A.
11) Again find a non zero element position in the matrix A. If found, go to step 5 else stop the iteration.
12) Using the labels the connected components can be extracted.

Connected components containing text and non text components.

### *2.4. Filtering and Character recognition*

After obtaining each character, non text regions want to be filtered. Support Vector Machine (SVM) is used for filtering non text region from text region. In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non probabilistic binary linear classifier. An SVM model[8] is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. Given some training data D, a set of n points of the form,

$$D=\{(x_i,y_i)|\ x_i \varepsilon R^p\ ,\ y_i \varepsilon_{\{-1,1\}}\ i=1 \text{ to } n \qquad (2)$$

where the $y_i$ is either 1 or -1, indicating the class to which the point $x_i$ belongs. Find the maximum margin hyperplane that divides the points having $y_i$= 1 from those having $y_i$ =-1. Traditional feature like Histogram of Oriented Gradients (HOG) is used for SVM classification. The essential thought behind the histogram of oriented gradients descriptor is that local object appearance and shape within an image can be described by the distribution of intensity gradients or edge directions. The image is divided into small connected regions called cells, and for the pixels within each cell, a histogram of gradient directions is compiled. The descriptor is then the concatenation of these histograms. For improved accuracy, the local histograms can be contrast normalized by calculating a measure of the intensity across a larger region of the image, called a block, and then using this value to normalize all cells within the block. This normalization results in better invariance to changes in illumination and shadowing.

Character recognition is done using optical character recognition. Template matching is used in the optical character recognition. Template matching is a technique in digital image processing for finding small parts of an image which match a template image. It can be used in manufacturing as a part of quality control, a way to navigate a mobile robot, or as a way to detect edges in images. Its procedure is given below,

- Read the image and convert it into gray scale image.

- Convert the gray scale image into a binary image based on a threshold.
- Store the matrix word from image.
- Opens text.txt as file for write.
- Template is created by using the letters A.....Z,0,....9. Each template is a binary image of 42 x 24 pixels.
- Compute the number of letters in template file.
- The image is cropped to fit the text
- Next step is to separate each line.
- Count connected components and letters are extracted.
- Then each letter is normalized to a size of 42 x 24 pixels, which is the size of the template that will perform the correlation.
- Concatenate the letters.
- Open 'text.txt' file.

### 3. Scene Text-Perspective (ST-Perspective)

Most of the standard datasets for scene text recognition, are limited to frontal texts. For example, the annotators of the Street View Text (SVT) dataset were instructed to minimize skew when choosing the angles of texts. Recently, there are more challenging datasets:NEOCR and MSRATD500, which include texts of arbitrary orientations and perspective texts. However, these two datasets are still not ideal for evaluating perspective text recognition because they are not specifically designed for perspective texts, many of the words in these datasets are still frontal. And also we are selecting the images which helpful for finding the way of visually impaired persons and also for driving assistance. However, this information has been discarded in these datasets. We introduce a new dataset called Scene text Perspective (ST-Perspective), which is specifically designed for evaluating perspective text recognition. It also preserves the name boards present in the road sides which helps to find the way or navigation aid for visually impaired persons and also it helpful for driving assistance.

### 4. Experimental Results

In this section, we present some experimental results to detect the text present in the natural scene images with perspective distortion. We performed experiments on our own dataset. The first step in the project is to avoid the perspective distortion. This is done by Hough transform. Edge detection is the preprocessing step of Hough transform. Perspective distortion is avoided by finding the longest line segment and find the angle using the two endpoints of longest line. Figure 6 shows the image in which red line shows the longest line present in the image. This image is rotated based on the angle. The image in which perspective distortion is avoided as show in figure 7.


**Fig. 5** Input image which is perspectival distorted


**Fig.6** Red line shows the longest line


**Fig.7**Perspective distortion is avoided

After avoiding the perspective distortion, cropped word image is obtained using text localization. In text localization canny edge detection is used. The scope of this project is limited to cropped word recognition. Inorder to obtain the cropped word image, we want to locate the text in the image. Input image is a frontal image. This image is converted into gray scale image and canny edge detection is applied to this image and the edge detected image is shown in figure 8. The Canny edge detector is an edge detection operator that

uses a multistage algorithm to detect a wide range of edges in images. The Canny algorithm is adaptable to various environments. Its parameters allow it to be tailored to recognition of edges of differing characteristics depending on the particular requirements of a given implementation. After canny edge detection holes are filled and the results is shown in figure 9.

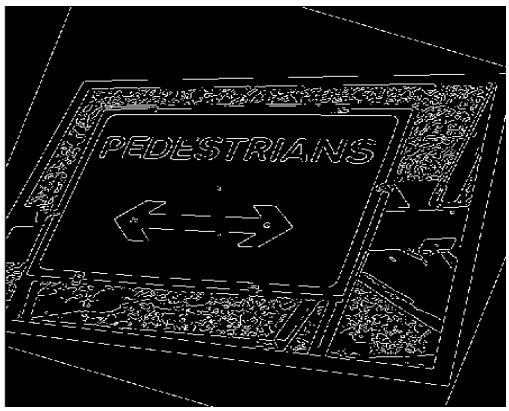


**Fig. 8** Images after canny edge detection

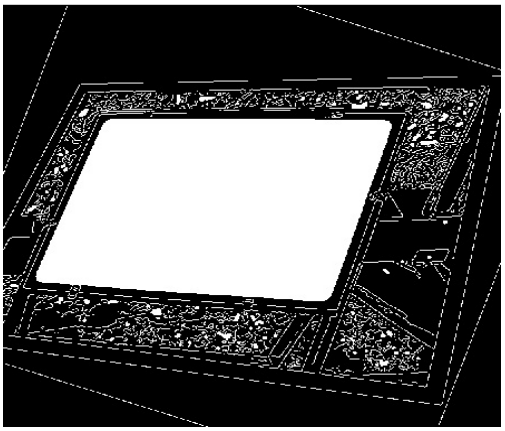

**Fig. 9** Images after holes filled


**Fig. 10** Image after finding the region with maximum area

Then find the region with maximum area and multiply this image with input image and the results is shown in figure 10. An edge filtering method is applied to this image and image with text is cropped from this image as shown in figure 11. Each character is detected using connected component analysis. Each components is labelled. Each text and nontext characters extracted are shown in figure 12. Support Vector Machine classification is used for filter the nontext region from the text region. HOG features are used for training the SVM classification. After filter the non text region each text character are recognized using optical character recognition. Correlation between the test image and trained images are calculated and its mean value is taken. The text file obtained after recognizing the text is shown in figure 13.

In this experiment, we used our ST-Perspective dataset for evaluation. In particular, words with less than 3 characters or containing non-alphanumeric characters were ignored.


**Fig. 11** Image of text region

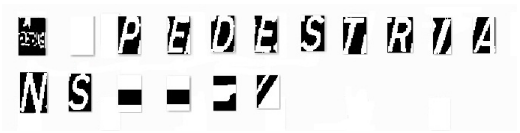

**Fig. 12** Characters obtained after connected component analysis


**Fig. 13** Image of .text file

The second column of Table 1 shows that our method significantly outperformed the other methods.

**Table 1** Recognition accuracy on perspective words (in percentage).

| Method | ST-Perspective word |
|---|---|
| K. Wang *et al.* | 44.5 |
| Mishra *et al.* | 27.8 |
| T. Wang *et al.* | 20.8 |
| Trung *et al.* | 62.3 |
| Our method | 72 |

The increase in accuracy from 62.3 percentage to 72 percentage (of our method) represents a relative improvement of 10 percentage. Here recognition of character is done on cropped word image. It increase the efficiency. Trung *et al.*[9] is also used cropped word image and which is obtained manually. But here cropped word image is obtained using text localization method. perspective distortion present in the image is

avoided in the first step itself so only frontal character samples are used for training, and thus does not require collecting samples of perspective characters.

This drastically reduces the cost of data collection. From the results it is clear that our method outperforms the other methods.

## Conclusions

We have described a method for recognizing perspective scene texts of arbitrary orientations. Our work serves as a step towards practical applications (of scene text extraction) two aspects. First, most existing works make the simplistic assumption that text is horizontal and frontal parallel to the image plane. However, in many real-world scenarios, this assumption does not hold. Thus, by handling perspective texts, this work has attempted to address an important research gap. Second, perspective distortion present in the image is avoided in the first step itself so only frontal character samples are used for training, and thus does not require collecting samples of perspective characters. This drastically reduces the cost of data collection.

Another contribution is the ST-Perspective dataset, which we propose to evaluate perspective text recognition. On this dataset, our method compares favourably to the state of-the-art. Therefore, our method serves an improved method for scene text recognition which are perspectively distorted.

## References

J. J. Weinman, E. Learned-Miller, and A. R. Hanson (2009), Scene Text Recognition Using Similarity and a Lexicon with Sparse Belief Propagation, *IEEE TPAMI*.

D. L. Smith, J. Field, and E. Learned-Miller. (2011) Enforcing Similarity Constraints with Integer Programming for Better Scene Text Recognition., *In CVPR*

G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray.,(2004) Visual Categorization with Bags of keypoints *In ECCV*

R. Nagy, A. Dicker, and K. Meyer-Wegener. (2011) NEOCR: A Configurable Dataset for Natural Image Text Recognition *In CBDAR*

T. Gandhi, R. Kasturi, and S. Antani.,(2000), Application of Planar Motion Segmentation for Scene Text Extraction. *In ICPR* .

Duda, R. O. and P. E. Hart, (January, 1972). Use of the Hough Transformation to Detect Lines and Curves in Pictures,Comm. *ACM, Vol. 15, pp. 11.*

Cristianini, Nello and Shawe Taylor, John(2000). An Introduction to Support Vector Machines and other kernel based learning methods, *Cambridge University Press.*

Huang, TeMing Kecman, Vojislav and Kopriva, Ivica. Kernel Based Algorithms for Mining Huge Data Sets, in Supervised, Semisupervised, and Unsupervised Learning, *SpringerVerlag, Berlin, Heidelberg, 260 pp. 96 illus., Hardcover, ISBN 3540316817.*

Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian and Chew Lim Tan, (2013)Recognizing Text with Perspective Distortion in Natural Scenes,*ICCV.*

K. Wang, B. Babenko, and S. Belongie. (2011)End-to-End Scene Text Recognition, *ICCV.*

A. Mishra, K. Alahari, and C. V. Jawahar(2012). Top-Down and Bottom-up Cues for Scene Text Recognition,*CVPR*.

T. Wang, D. J. Wu, A. Coates, and A. Y. Ng. (2012)End-to-End Text Recognition with Convolutional Neural Networks, *ICPR*.