Text Localisation for Roman Words from Shop Signage

Nurbaity Sabri[1], Noor HaziraYusof, Zaidah Ibrahim[2], Zolidah Kasiran[3], Nur Nabilah Abu Mangshor[4]

[1]Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA, Kampus Jasin, 77300 Melaka, Malaysia.
[2]Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia.
E-mail: [1]nurbaity_sabri@melaka.uitm.edu.my,[2]zaidah@tmsk.uitm.edu.my,
[3]zolidah@tmsk.uitm.edu.my, [4]nurnabilah@melaka.uitm.edu.my

Received: 7 February 2017
Accepted: 9 October 2017

ABSTRACT

Text localisation determines the location of the text in an image. This process is performed prior to text recognition. Localising text on shop signage is a challenging task since the images of the shop signage consist of complex background, and the text occurs in various font types, sizes, and colours. Two popular texture features that have been applied to localise text in scene images are a histogram of oriented gradient (HOG) and speeded up robust features (SURF). A comparative study is conducted in this paper to determine which is better with support vector machine (SVM) classifier. The performance of SVM is influenced by its kernel function and another comparative study is conducted to identify the best kernel function. The experiments have been conducted using primary data collected by the authors. Results indicate that HOG with quadratic kernel function localises text for shop signage better than SURF.

Keywords: *HOG, SURF and SVM*

INTRODUCTION

Research in text localisation for scene images is receiving a growing interest among researchers due to the vast advance of digital devices [1].Text data in images carry useful information where once it has been localised and identified, the text can be analysed, recognised and interpreted for applications such as license plate recognition as in [2-4], and road sign recognition as in [5]. This research focuses on the localisation of text on shop signage. However, text localisation in natural scene images is still a challenging task due to its complex background and variations of text font size, font type and colour.

Various techniques have been proposed for text localisation from scene images, and they can be categorised into two categories, namely connected component-based as in [6-8] and texture-based approach [9-10]. A connected component in an image is constructed based on intensity, edge or colour and is sensitive to noise as the complex background may consist of non-text components that have similar shapes as the character connected components. This leads to mis-localisation. Texture features are suitable for localising dense characters as those that are used for shop signage as applied in [11] since text usually has a different texture compared to the background. Two of the popular texture features utilised in text recognition are a histogram of oriented gradient (HOG) [9-10] and speeded up robust features (SURF) [12].Thus, a comparative study is conducted between these two texture features that act as input to support vector machine (SVM) classifier to investigate which is better. SVM is chosen as it is a popular classifier for text recognition [13-14]. The next section discusses the related work to text localisation followed by explanation about the research methodology, experimental results and conclusion.

RELATED WORK

The text consists of strings of characters, and a character has the following properties:

(i) It has a pair of boundaries or edges that are usually parallel to each other, and have similar length, width, directions and curvatures;
(ii) It has closed boundaries; and
(iii) It has denser pixels in the edges compared to the background.

Some researchers have used the above properties for text localisation. Based on the first properties of a character, Epshtein *et al.* [15] propose Stroke Width Transform (SWT). In this technique, after detecting the edges using Canny edge operator, SWT searches for pixels that belong to the same stroke by comparing the width of a pair of pixels. Then, connected component algorithm is being applied where the stroke width variance is computed. A connected component with a big stroke width variance is considered as non-text. However, this approach fails in text localisation for very strong highlights and very excessive blur images since there is not much uniformity of the colour distribution. These limitations have been improved by Zhao *et. al.* where they propose Stroke Unit Connection (SOIC) to detect Chinese characters. The strokes are being extracted in smaller segments where the image is being segmented into NxN patches. The stroke information is being fed into SVM which classifies whether it is a text or non-text patch.

Based on the third property of a character, six directions of Log-gabor filter is being applied to extract the edges [17]. For every direction filter, {0 o, 30 o, 60 o, 90 o, 120 o, 150 o}, the magnitude of the filter is extracted to produce an edge map. Then, all the edge maps are integrated to form stroke maps. Since text component usually has many strokes, the stroke map's coarseness is computed to detect the text lines. However, this technique is not suitable for images whose text has very low-contrast compared to the background. Canny edge detector is being used by Zhang and Kasturi [18] to detect the edges, and the edges that are too long or too short are removed since they are assumed do not belong to the text edges. Then, HOG is applied to capture the similarity of stroke edges. Four types

of edge points are defined according to the gradient direction and based on the second and third properties of a character; candidate character regions are formed. However, non-text objects that have similar properties as the text may be detected as text regions that leads to false positive detection. All the techniques mentioned above cannot handle very well for images with strong highlights, and yet these are common occurrences for shop signage images.

Canny edge detector and adaptive thresholding gradient has been applied by Ma *et al.* [19] to produce edge map for each sequence of images in the image pyramid. Then, text edges are being classified by SVM based on three features, that are, HOG, local binary pattern and statistical features that include mean, standard deviation, energy, entropy, inertia, local homogeneity and correlation. HOG is also utilised in [18-21] as it is less sensitive to image noise and can capture the characteristics of the text regions in the natural scene image. However, the text localisation result is affected if the text is on complex background. Neumann and Matas [14] applied Maximally Stable Extremal Regions (MSER) that detects character regions with SVM classifier. The connected component approach does not perform well for images with various illuminations. SURF has shown good performance in object detection [22-23], and this research tends to apply it for text localisation.

DATA SET AND METHODOLOGY

The primary data used in this study are images of shop signage that were captured by the authors using a smart phone. A total of 40 images for training and 40 images for testing have been acquired in this research. The captured images of the shop signage come in various font colour, size and type with different backgrounds. The text was in horizontal orientation. From personal observation, the region of interest is usually focused in the middle of the captured image. Thus, the text region is located in the centre of the whole captured image. Figure 1 illustrates some sample images of the captured shop signage.

Text localisation process usually involves two main steps that are, feature extraction and text classification. Feature extraction is a special form of dimensionality reduction in pattern recognition and image processing while text classification is the step of assigning predefined categories to free-text documents. Figure 2 illustrates the flow of a process for text localisation. The experiment was conducted by using MATLAB R2013a.

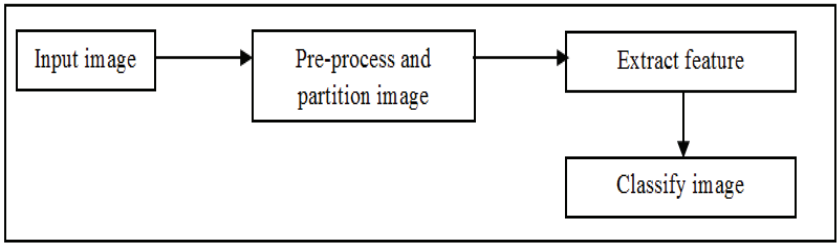

Figure 1: Flow of Process for Text Localisation

Pre-process and Partition Image

At the pre-processing stage, the images were converted to gray-scale images since HOG and SURF work with gray-scale images, and the image is resized to 256x256 due to memory constraints. Since the text of interest is in the middle of the captured image, the gray-scale images were automatically partitioned into three sub-blocks. These sub-blocks were then processed for feature extraction and classification. As a result, the location of the text is known. Figure 3a shows the original image, while Figures 3b until 3d illustrates the result of the partitioning of the image into three sub-blocks.

Feature Extraction

The general idea of HOG is that local object appearance and shape within an image can be represented by the distribution of intensity gradient as explained in [24]. The implementation of HOG starts by dividing the image into small regions, called cells, and HOG counts occurrences of edge orientations for each cells. The final HOG result is the combination of these histograms. The detail explanation of HOG is given below.

In this algorithm, the gradient values are calculated using a common method which is 1-D centre point discrete derivative mask. This method is implemented for both directions: - horizontal and vertical of an image. To implement this method, the image needs to be in grayscale and filtered with the kernels as in Equation 1.

$$Dx = [-1\ 0\ 1] \text{ and } Dy \quad \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \qquad (1)$$

Then, the next operation involves creating the cell histograms. Each pixel within the cell casts a weighted vote for an oriented-based histogram channel. This histogram is produced using the values obtained from the gradient operation. The histogram channel are distributed from 0 to 180 degrees or 0 to 360 degrees, depending on whether the gradient is 'unsigned' or 'signed' where each cells is in rectangular shape.

The gradient strength is locally normalised to cope with the differences between lighting and contrast of an image. These gradients are collection of cells into larger, spatially linked blocks. The shape descriptor (HOG) is the vector of the components of the normalised cell histograms from all of the block regions. These blocks typically overlap, meaning that each cell contributes more than once to the final descriptor.

Block normalisation consists of different methods. Variable v represents the non-normalised vector consist of block of all histograms. The operation ||Vk|| where k-norm for k =1,2 and e is the minimun constant that will not influence the results. Then, the normalisation factor can be one of the Equations between 2 to 4.

$$\text{L2-norm: } f = \frac{v}{\sqrt{\|v\|_2^2 + e^2}} \qquad (2)$$

$$\text{L1-norm: } f = \frac{v}{\|v\|_1 + e} \qquad (3)$$

$$\text{L1-sqrt: } f = \sqrt{\frac{v}{\|v\|_1 + e}} \qquad (4)$$

SURF

The basic idea of the SURF algorithm is to find the interest point of an image and measure the distance between points in two separate images [22-23]. SURF used approximation of Gaussiansmoothing implemented in square shaped filters. The filtering processing will be faster using square shape defined in Equation 5.

$$S(x,y) = \sum_{i=0}^{x} \sum_{j=0}^{y} I(i,j) \qquad (5)$$

The image features produced by HOG descriptor are more robust and unique by describing the intensity distribution of the pixels within the neighbourhood of the point of interest. The dimensionality of the descriptor has a direct impact on both of its computational complexity and point-matching robustness and accuracy. A short descriptor may be more robust against appearance variations, but may not offer sufficient discrimination and thus give too many false positives.

First, the information produced from the surrounding region of the interest points are fixed. Then, the SURF descriptors are extracted from the constructed square region aligned with the selected orientation. In this step, the descriptor in one image is compared with the descriptor in several images. Irrelevant descriptors have to be searched. Finally, the distance between images interest points are calculated. By comparing the descriptors obtained from different images, matching pairs can be found as follows:

- Compare the length and width of each image;
- Compare the histograms of both images;
- Extract 'interest point' of each image;
- Match interest points between the two images; and
- Calculate the average distance between interest point.

Support Vector Machine (SVM)

Support vector machine (SVM) model is a representation of the examples as points in space, mapped so that the examples of the separate categories are split by a clear gap that is as broad as possible [25]. New cases are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. In SVM, the kernel function is used to map the training data into the kernel space. It also finds an optimal solution which maximises the distance between the hyper-plane and the difficult points close to decision boundary. In the final step of text localisation for each method is to feed the extracted features into SVM classifier. Since the performance of SVM depends on its kernel function, this paper conducts a comparative study of five kernel functions that are linear, Polynomial (POLY), quadratic, Gaussian Radial Basis Function (RBF) and Multilayer Perception (MLP).

The linear kernel is the simplest kernel function. It is given by the inner product <x, y> plus an optional constant c, illustrated in Equation 6.

$$K(x, y) = x^T y + c \qquad (6)$$

The Polynomial kernel is a non-stationary kernel. Polynomial kernels are well suited for aproblem where all the training data is normalised as in Equation 7. Adjustable parameters are the slope alpha, the constant term c and the polynomial degree d.

$$k(x, y) = (\alpha x^T y + c)^d \qquad (7)$$

The quadratic kernel is two-dimensional, quad-kernel for two-dimensional vectors as in Equation 8.

$$\vec{u} = (u_1 \;\; u_2) \;\; \vec{v} = (v_1 \;\; v_2) \qquad (8)$$

Radial Basis Kernel (RBF) is used in various kernelised learning algorithms in machine learning and commonly used in SVM classification. RBF is equivalent to mapping the data into an infinite dimensional Hilbert Space (a Hilbert space is a vector space closed under sot products), and so RBF cannot illustrate concretely.

A multiplayer perception (MLP) is a supervise machine learning where it provides a set of input data into a fitting output. The kernels proposed with some parameters associated with the use of the SVM algorithm that can impact the results.

RESULTS AND DISCUSSION

This research localises text from outdoor shop signage that consists of Roman words only. Images of the shop signage were captured by the authors using a smartphone. A total of 80 images for training and 40 images for testing have been used in this research. After the partitioning process, the total training images is 249 sub-blocks, and the total testing images are 120 sub-blocks. Table 1 shows the localisation rate (r) computed from each feature extracted from each sub-block and kernel function of SVM. The localisation rate (r) is computed by dividing the correctly classified sub-block with the total number of sub-blocks. By looking at Table 1, it can be seen that the combination of HOG and quadratic kernel function localises the text of the sub-block better than SURF and other kernel functions. Linear kernel function for HOG and Quadratic kernel function for SURF do not produce any results since they do not converge. This is due to the small number of training data. HOG produces better result compared to SURF because HOG is a gradient-based approach where it captures the information on the shape of the characters very well. On the other hand, SURF detects keypoints of the shape of the characters. Thus, it does not extract the complete information of the shape of the characters. Better text localisation rate can be achieved if more training data is provided for SVM.

Table 1: Results of Text Localisation

Feature	SVM Kernel Function	r%
HOG	Linear	0
	Polynomial	70.40
	Quadratic	72.00
	RBF	49.60
	MLP	52.00
SURF	Linear	43.20
	Polynomial	40.80
	Quadratic	0
	RBF	51.20
	MLP	63.20

CONCLUSION

Text localisation from shop signage is important prior to text recognition and even translation. It is a process that identifies the location of the text. It can be applied in various applications such as indexing, mapping and navigational purposes. However, it is a challenging task since shop signage consists of text with various font type, size and colour, and the background is complex with various figures. This research conducts a comparative study between two popular texture features, namely HOG and SURF that act as input to SVM classifier. The localisation rate is computed and from the results obtained, HOG with Quadratic Kernel function for SVM achieves higher localisation rate compared to other kernel functions with SURF. This shows that gradient-based approach is better than key point-based approach. For future work, localising multi-lingual text from shop signage will be examined with a connected component approach for text recognition.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the help of the Ministry of Education (MOE) and Universiti Teknologi MARA (UiTM) for sponsoring this research under the National Grant No 600-RMI/FRGS 5/3 (165/2013).

REFERENCES

[1] X. Liu, and D. Doermann, 2008. A Camera Phone Based Currency Reader for the Visually Impaired. In *Assets 2008 Proceeding ACM SIGACCESS Conferences Computers and Accessibility*, pp. 305-306. DOI: 10.1145/1414471.1414551.

[2] A. E. Rashid, 2013. A Fast Algorithm for License Plate Detection. In *2013 International Conference on Signal Processing, Image Processing & Pattern Recognition* (ICSIPR), Coimbatore, India, 7-8 Feb 2013. DOI: 10.1109/ICSIPR.2013.6497956.

[3] S. J. Yang, J. B. Jiang, M. K. Wu and C. C. Ho, 2013. Real-Time License Plate Detection System with 2-level 2D Haar Wavelet Transform and Wiener- Deconvolution Vertical Edge Enhancement. In *2013 9th International Conferences on Information, Communications & Signal Processing* (ICICS), Tainan, Taiwan, 10-13 Dec 2013. DOI: 10.1109/ICICS.2013.6782805.

[4] G. S. Hsu, S. D. Zeng, C. W. Chiu and S. L. Chung, 2015. A Comparison Study on Motorcycle License Plate Detection. In *2015 IEEE International Conferences on Multimedia & Expo Workshops* (ICMEW), Turin, Italy, 3 July 2015. DOI: 10.1109/ICMEW.2015.7169772.

[5] T. Zhang, J. Lv and J. Yang, 2013. Road Sign Detection Based on Visual Saliency and Shape Analysis. In 2*013 IEEE International Conferences on Image Processing*, Melbourne, Australia, 15-18 Sept 2013. DOI: 0.1109/ICIP.2013.6738756.

[6] H. I. Koo and D. H. Kim, 2013. Scene Text Detection via Connected Component Clustering and Non-Text Filtering. *IEEE Trans. on Image Processing, Vol. 22*(6), pp 2296-2305. DOI: 10.1109/TIP.2013.2249082.

[7] L. Neumann and J. Matas, 2012. Real-time Scene Text Localisation and Recognition, Proc. In *2012 IEEE International Conferences Computer Vision Pattern Recognition*, pp 3538-3545. DOI: 10.1109/CVPR.2012.6248097.

[8] D. Ding, J. Yoon and C. Lee, 2012. Traffic Sign Detection and Identification Using SURF. In *International Soc. Design Conference* (ISOCC), Korea, 4-7 June 2012. DOI: 10.1109/ISOCC.2012.6406907.

[9] S. Tian, S. Lu, B. Su & C. L. Tan, 2013. Scene Text Recognition Using Co-Occurrence of Histogram of Oriented Gradients. In *2013 12th International Conference on Document Analysis and Recognition*, pp. 912-916. DOI: 10.1109/ICDAR.2013.186.

[10] R. Minettoa, N. Thomeb, M. Cordb, N. J. Leitec and J. Stolfic, 2013. T-HOG: An Effective Gradient-Based Descriptor for Single Line Text Regions, *Pattern Recognition, Vol. 46*(3), pp 1078-1090. DOI: https://doi.org/10.1016/j.patcog.2012.10.009.

[11] Q. Ye and D. Doermann, 2015. Text Detection and Recognition in Imagery: A Survey, *IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 37*(7). DOI: 10.1109/TPAMI.2014.2366765.

[12] S. Ahmed, M. Liwicki and A. Dengel, 2012. Extraction of Text Touching Graphics using Surf. In *2012 10th IAPR International Workshop on Document Analysis Systems*, Gold Coast, Australia, 27-29 March 2012. DOI: 10.1109/DAS.2012.39.

[13] C. Jung, Q. Liu & J. Kim, 2008. A New Approach for Text Segmentation using a Stroke Filter. *Signal Processing, 88*(7), 1907-1916. DOI: https://doi.org/10.1016/j.sigpro.2008.02.002.

[14] L. Neumann and J. Matas, 2011. Text Localisation in Real-world Images using Efficiently Pruned Exhaustive Search. In *2011 International Conferences on Document Analysis and Recognition*, Beijing, China, 18-21 September 2011. DOI: 10.1109/ICDAR.2011.144.

[15] B. Epshtein, E. Ofek and Y. Wexler, 2010. Detecting Text in Natural Scenes with Stroke Width Transform. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR), San Francisco, USA, 13-18 June 2010. DOI: 10.1109/CVPR.2010.5540041.

[16] Y. Zhao, T. Lu and W. Liao, 2011. A Robust Color-Independent Text Detection Method from Complex Videos. In *2011 International Conferences on Document Analysis and Recognition* (DAS), Beijing, China, 18-21 September 2011. DOI: 10.1109/ICDAR.2011.83.

[17] X. Huang and H. Ma, 2010. Automatic Detection and Localisation of Natural Scene Text in Video. In *2010 20th International Conference on Pattern Recognition* (ICPR 2010), Istanbul, Turkey, 23-26 August 2010. DOI: 10.1109/ICPR.2010.786.

[18] J. Zhang and R. Kasturi, 2010. Text Detection Using Edge Gradient and Graph Spectrum. In *2010 20th International Conference on Pattern Recognition* (ICPR 2010), Istanbul, Turkey, 23-26 August 2010. DOI: 10.1109/ICPR.2010.968.

[19] L. Ma, C. Wang and B. Xiao, 2010. Text Detection in Natural Images Based on Multi-Scale Edge Detection and Classification. In *2010 3rd International Conferences on Image and Signal Processing* (CISP), Yantai, China, 16-18 October 2010. DOI: 10.1109/CISP.2010.5648158.

[20] Y. F. Pan, X. Hou and C. L. Liu, 2011. A Hybrid Approach to detect and Localise Texts in Natural Scene Images. *IEEE Transactions on Image Processing, 20*(3). DOI: 10.1109/TIP.2010.2070803.

[21] S. Muhammad Hanif and L. Prevost, 2009. Text Detection and Localisation in Complex Scene Images using Constrained AdaBoost Algorithm. In *10th International Conference on Document Analysis and Recognition* (ICDAR), Barcelona, Spain, 26-29 July 2009. DOI: 10.1109/ICDAR.2009.172.

[22] L. Jianguo and Z. Yimin, 2013. Learning SURF Cascade for Fast and Accurate Object Detection. In *2013 IEEE International Conference on Computer Vision and Pattern Recognition*, Portland, USA, 23-28 June 2013. DOI: 10.1109/CVPR.2013.445.

[23] H. Jun-Wei, C. Li-Chih and C. Duan-Yu, 2014. Symmetrical SURF and Its Applications to Vehicle Detection and Vehicle Make and Model Recognition. *IEEE Transaction on Intelligent Transportation Systems, Vol. 15*(1). DOI: 10.1109/TITS.2013.2294646.

[24] N. Dalal, and B. Triggs, 2005. Histograms of Oriented Gradients for Human Detection, In *IEEE Computer Society Conferences Computer Vision Pattern Recognition*, 2005, San Diego, USA, 20-25 June 2005. DOI: 10.1109/CVPR.2005.177.

[25] C. Cortes and V. Vapnik, 1995. Support-Vector Networks, *Machine Learning*, *20*, pp. 273-297, 1995. DOI: 10.1023/A:1022627411411.