Text Localization in Natural Images Through Effective Re-Identification of the MSER

Hanaa F Mahmood
Computer science Department
Loughborough University, UK
H.F.Mahmood@lboro.ac.uk

Baihua Li
Computer science Department
Loughborough University, UK
B.Li@lboro.ac.uk

Eran Edirisinghe
Computer science Department
Loughborough University, UK
E.A.Edirisinghe@lboro.ac.uk

ABSTRACT

Text detection and recognition from images have numerous applications for document analysis and information retrieval tasks. An accurate and robust method for detecting texts in natural scene images is proposed in this paper. Text-region candidates are detected using maximally stable extremal regions (MSER) and a machine learning based method is then applied to refine and validate the initial detection. The effectiveness of features based on aspect ratio, GLSM, LBP, HOG descriptors are investigated. Text-region classifiers of MLP, SVM and RF are trained using selections of these features and their combination. A publicly available multilingual dataset ICDAR 2003,2011 has been used to evaluate the method. The proposed method achieved excellent performance on both databases and the improvements are significant in terms of Precision, Recall, and F-measure. The results show that using a suitable feature combination and selection approach can can significantly increase the accuracy of the algorithms. Keywords—text detection; scene images; ICDAR; feature selection .

KEYWORDS

text detection; scene images; ICDAR; feature selection.

INTRODUCTION

Text detection and recognition from images could have numerous functional applications for document analysis such as assistance for visually impaired people, recognition of vehicle license plates, evaluation of articles comprising tables, street signs, maps, diagrams, keyword based image exploration, document retrieving, recognition of parts within industrial automation, content based extraction, object recognition, address block location as well as text based video indexing.(Ye, Q. & Doermann, D., 2015, Zhang, H. et al., 2013; Seeri, et al., 2015). Furthermore, Scene text detection techniques may be utilized in detecting text-based landmarks, vehicle license detection/identification, and object recognition as opposed

overall extraction and indexing. It is a challenge to detect, locate and retrieve scene text as there could be unlimited possible poses, sizes, colours and shapes, resolution, intricate backgrounds, uneven lighting, or blurring resulting from differing lighting, intricate movement and conversion, unfamiliar format, shadowing as well as differences in font size, style, alignment and direction. Further, texts may be in various scripts (Seeri, 2015).

Text detection distinguishes the text areas as extremal areas of an image and during the text recognition phase extracts the text information from such extremal areas. Text localization is refers to determining the text location in image and draw bounding boxes around the text. Although bounding boxes specify the accurate location of text in an image, segmented text from the background are still needed. This process includes transforming the image to a binary image and enhancing it before it is fed into an OCR engine. Text extraction is the stage of segmented text region from background. Usually the segmented area has different type of noise and low resolution. For this reason, it requires a number of enhancement operations. Furthermore, extracting the substance from images is considerably difficult, due to image quality and background noise. Thereafter, OCR is used to transform extracted text images to plain text (Wang, et al., 2015; Zhu 2015).

To tackle the problem of distinguishing text/non-text regions, researchers have used filters. There are a limited number of publications which deal with the classification algorithms of text/ non-text regions. They considered text/non-text classification problem as a texture classification problem. This approach led to using several texture descriptors and machine learning to discriminate between text and non-text regions. A variety of machine learning techniques have been used for text detection, including supervised and unsupervised feature learning, support vector machine (Kim, K.I., Jung, K. & Kim, J.H., (2003): Anthimopoulose et al, (2010) , multi-layer perceptron Chowdhury et al. (2012) Convolutional Neural Networks, deformable part based models, belief propagation , and Conditional Random Fields Pan, et al, (2011); Zhang et al. (2011). Kim, K.I., Jung, K. & Kim, J.H., (2003) identified text regions subsequently through the application of a continuously adaptive mean shift algorithm (CAMSHIFT) to outcomes of the texture analysis. Strong and efficient text detection is produced from the combination of SVMs and CAMSHIFT.

Hanif S. M., Prevost, L., 2009 extracted Three types of features from text segment which are Mean Difference Feature (MDF), Standard Deviation (SD) and Histogram of oriented Gradient (HoG) to create big feature vector, AdaBoost algorithm was used to classify segments to text or non-text. Anthimopoulose et al, 2010 proposed a modification of Local Binary Pattern (LBP) called edge LBP. Their descriptor consists of 256 features

extracted from candidate text line by using a sliding window model and Support Vector Machines (SVM) to classify candidate areas.(Minelto et al 2011) extended the morphological operation (toggle mapping) to segment urban images. Shape descriptors. Fourier moment, pseudo Zernike moments and polar representation of candidate region used as descriptor and a hierarchical support vector machine as classifier. Gonzalez, et al 2012 filtered candidate text regions extracted by MSER by using a set of distinctive features then filtered regions were grouped into lines. Mean Difference Feature (MDF), Standard Deviation (SD) and HOG were used to train a SVM with linear kernel to classify lines into text or non_text. Zhang et al. 2011 used a mean-shift process to segment candidate text components and then build up a component adjacency graph. Integrating a first-order components term and a higher-order contextual term, a CRF (Conditional Random Fields) model was used to classify component as text or non-text. Trung et al 2012 proposed to use Gradient Vector Flow for the detection of candidate text regions. The detected regions were grouped into text lines by using sizes, positions and colors constraints. HOG and SVM were used to remove false positives using a learningbased approach.

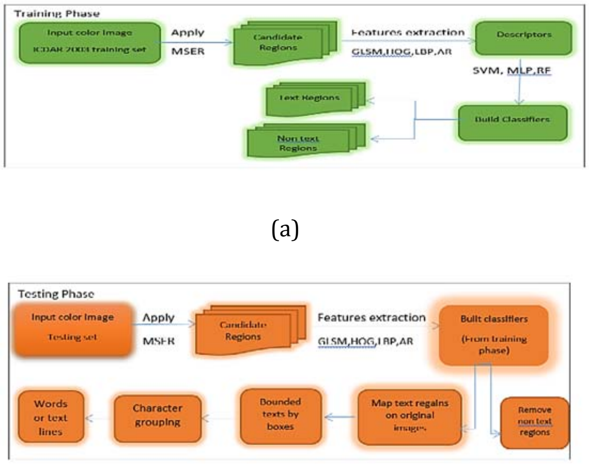
The advantage of MSER promoted researchers to use it for character candidate extraction that it can detect the majority of characters' regions regardless of their scale, noise, and to affine illumination variations. However, it detects non- text regions (Matas et al. 2004) (Neumann, L. & Matas, J., 2010). The method proposed (figure 1) here to overcome the problem of detecting non- text regions, deals with the classification algorithms of text/ non-text regions that are extracted by MSER from the grayscale to obtain text-region candidates. A feature descriptor is calculated using (GLCM, LBP, HOG) and Aspect ratio. The proposed scheme uses a small set of heterogeneous features which are spatially combined to build a large set of features. The selection and combination of features are the main contributions. Where all possible combinations between used features were tested to get the best detection accuracy. By using heterogeneous feature set, the combination of feature complexity in feature selection algorithm supports reducing the overall complexity of classifier. Furthermore, the computational load which is an important consideration in real time applications. Where the results show that using a suitable feature selection and combination approach can significantly increase the accuracy of the algorithms. Where the combination of HOG+lbp+GLSM+AR give the higher accuracy followed by the combination and selection of LBP+GLCM+AR, LBP+ GLCM+AR. The rest of the paper is organized as follows: Section 2 reviews the complete description of text detection and localization proposed method. Section 3 describes all details of the text region detection. Experimental result and evaluation are presented in Section 4, followed by Text localization in section 5 and the conclusions in Section 6. 2 EXPERIMENTAL AND COMPUTATIONAL DETAILS

2 Overview of the Proposed Method

MSER regions are used to predict text parts instead of having to create feature descriptors for every pixel which can be computationally expensive. MSER is used to obtain text-region candidates from the grayscale image. MSER detection delivers a list of possible text regions, and then a machine learning based classifier is employed to refine the detected regions. For each MSER region, image features are calculated using GLCM, LBP, HOG and aspect ratio descriptors. Figure 1 shows the flowchart of the algorithm. At the training phase, phase, a training set of text and non-text regions was collected from the ICDAR 2003 data set. and the resulting classification model is saved for the testing phase. At testing phase all MSER regions were extracted from each image in the testing set. Candidate regions were classified by using the classifiers built in the training phase based on the descriptors. All MSER regions that are reclassified as text will be mapped back onto the image. Pixels inside these boxes will be marked as text regions.

2.1 Candidate region detection using MSER

MSER is employed within computer vision as a technique of blob detection in images. Matas et al. (2002). It is adopted and has been widely used to solve text detection problems and won the first place in both ICDAR-2011and ICDAR-2013 competitions (Yin, X.X.-C et al , 2013),(Gomez, L. and Karatzas, D. 2013) (Chen, H. et al., 2011) . The MSER algorithm utilizes on the intensity information of the image. Since the text area in image tends to have connected equal intensity, that leads the output of this step to be candidate regions containing at least one symbol. The MSERs algorithm can detect the majority of characters, even in the presence of reduced quality (low contrast, strong noises, low resolution) although of their advanced performance a number of open challenges require addressing (Yin, X.X.-C et al , 2013).



(a)

(b)

Figure (1): Block Diagram of the Text Localization Module of the Proposed Method. (a) Training Phase, (b) Testing Phase

One such challenge is the detection of many false positives (non- text regions) that do not contain characters. Therefore, it is important to apply additional checks to eliminate non-text regions. Grayscale image is the input to the MSER algorithm and a sequence of images (It)255 are the output of the algorithm where t=0. The output is generated by successively binarized input image with t starting from 0 to 255. I0 which is the first image in the series is completely black then white regions appear and grow in the next images in the series. The latest image, I255, is completely white. The white regions in the series are called extremal regions, which calculate by how many successive images in the series this area stays the same. While Maximally Stable Extremal Regions are, the regions can be selected by choosing a threshold value R, which are the regions completely the same in at least R successive images of the series. Figure 4 shows the result of detected MSER regions. It shows clearly that MSER algorithm detects a large number of false positives – nontext regions

2.2 Candidate region re-identification using learnt descriptors

The following presents the theoretical and conceptual background of methods that have been used to extract features from candidate regions and build descriptors, which are Gray level *co-occurrence matrix (GLCM), Local Binary Patten (LBP) , Histogram of Gradient (HOG) and Aspect Ratio (AR)*

2.2.1 Gray level co-occurrence matrix (GLCM)

Gray level co-occurrence matrix (GLCM) based features have been widely used in image analysis. Given an image, the GLCM computes how often different combinations of gray levels cooccur in the image or a section of the image. The texture information is captured by computing four traditional features from the GLCM: energy(ASM), contrast(CON), correlation (COR) and homogeneity(HOM) (Haralick et al. 1973)(Soh and Tsatsoulis 1999). These features have been employed to detect the text regions and eliminate the text-like false positives. Equation (1) shows the probability measure GLCM described by (Clausi 2001) where the grey level quantised number is G, and given a certain orientation (θ), inter-pixel distance (δ) and (δ , θ) pair

$$Pr(x) = \{C_{ij}|(\delta,\theta)\}; C_{ij} = \frac{P_{ij}}{\sum_{i,j=1}^{G} P_{ij}} \qquad (1)$$

2.2.2 Local Binary Pattern (LBP)

Local Binary Pattern (LBP) has proven to be highly discriminative for texture segmentation and its advantages on invariance to monotonic gray level changes and computational efficiency make it suitable for image analysis tasks (Huang et al. 2011). Zhang et al. (Zhang 2006) have shown that LBP features can not only capture texture characteristics, but also localize structure characteristics, which is suitable for text detection. The idea of using the LBP features are that texts be composed of strokes which are very similar to the patterns produced by an LBP operator. Therefore, the LBP operator is very effective to be represented by using LBP generated patterns. This fact motivated the use of LBP for text detection and adjust it to the specific problem. LBP is a simple descriptor that generates a binary pattern code by comparing the gray level of a pixel and its local neighborhood. "Then it creates a histogram using the binary pattern codes, The bin in the histogram corresponds to a unique binary code (Kwak, J. T. et al 2015). LBP enable the use of different distance between the center pixel and its local neighborhood although the standard LBP use 8 pixels in a 3 × 3 pixel block. But this basic formula helps to. find best performance by applying different cell size.

2.2.3 Histogram of Gradient (HOG)

The working philosophy behind HOG is local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge directions (Dala et all 2005).

The image is divided into small regions called cells then HOG features were extracted from this cell. For each cell, a histogram of gradient directions or edge orientations for the pixels within that cell is calculated. The descriptor represents the combination of these histograms and produces different feature sets of different length.

The image convolves into two directions horizontally and vertically with 1D [−1, 0, 1] mask to detect image edge after applying gamma normalisation and colour normalisation. process and the switching fields of the distinct dots, longitudinal minor loops were also measured.

$$Dx=[-1\ 0\ 1] \qquad Dy=\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \tag{2}$$

In the second step the image patch is subdivided into cells or rectangular regions, and computations are made within each cell for the gradient for each pixel. For each channel there are separate computations for the gradient in colour images, and the gradient chosen for a pixel is the largest gradient. Therefore, a convolution operation is used to obtain x and y derivatives within an image I:

$$I_x = I \times D_x \text{ and } I_y = I \times D_y \tag{3}$$

The gradient magnitude is:

$$|G| = \sqrt{I_x^2 + I_y^2} \tag{4}$$

The gradient orientation is shown by:

$$\theta = \arctan \frac{I_y}{I_x} \tag{5}$$

The cell orientation is determined by a weighted vote computed by each pixel within the cell for the following step, and the L2 norm or gradient magnitude provides a weighting for the vote. Orientation bins accumulate these votes, so dependent on whether a vote is a signed gradient or an unsigned gradient, votes with a 0 to 360 degree range or 0 to 180 degree range are cast into the closest bin. Therefore, the 0 to 180 degree range shows that the gradient is unsigned in this algorithm, and a histogram stores these gradients. Dalal and Triggs suggest that this algorithm enables better performance due to the unsigned gradients in this histogram based on the use of a conjunction with nine channels. Dalal, N., and Triggs, B. (2005).

2.2.4 Aspect Ratio (AR)

Aspect ratio is simply a ratio to describe the proportional relationship between the width and height of any image. (The ratio of width to height equation), (Yin, X.X.-C., Huang, K. & Hao, H.-W., 2013)((Yao, et al., 2013 a) Most letters in English have aspect ratio being close to 1, so this feature can be useful to filter out false character candidates. Therefore, several heuristics are used to filter out non-text components based on aspect ratio. (Yao, et al., 2013 a)(Yao, C. et al., 2012.) (Yin, X.X.-C., Huang, K. & Hao, H.-W., 2013).

Aspect ratio = max (width, height)/ min(width, height) (6)

2.3 Training phase (Initial detection of text region candidates (Result from training set)

The ICDAR 2003 dataset (Lucas, S.M. et al 2003)(Lucas, S. M. , et al 2005) has been widely used as a benchmark for researchers in the field of text detection. There are 509 completely annotated text images included in this dataset. 251 of these images are employed in testing and 258 are for training. The texts in this database vary greatly in fonts, sizes, styles and appearance. The dataset provides targets with the images, which are the ground truth locations for text that are available in the images. The target used to calculate a precise evaluation of the results of text detection techniques where the text detection methods provide estimates, which is a form of a rectangle that bound a text area in the image (Lucas, S.M. et al 2003; Lucas, S.M. et al 2005; Mosleh, A., et al 2012). In the training phase, 7423 regions have been extracted from ICDAR 2003 dataset to train and test different descriptors. 6353 positive patches and 1070 negative patches randomly sampled from training set of ICDAR 2003 dataset. Figures 2,3 show examples of text non-text regions respectively.

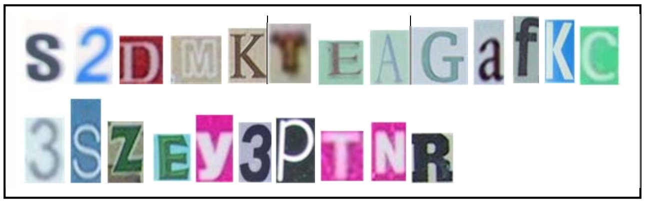


Figure 2: Examples of text samples in the ICDAR dataset

Figure 3: Examples of non-text regions

For all descriptors, the features were set as the positive samples represented by 1 in the classifier. Features were set as the negative samples, represented by 0 in the classifier. For this purpose, we use a classifier based on SVM with linear kernel, Multilayer Perceptron MLP, Random Forest RF.

2.3.1 GLCM descriptor

To classify text and non-text regions GLCM have been calculated in four orientations (0º,45º,90º,135º). This is because GLCM is not direction invariant, texts are locating in different directions in the images. To solve this invariant problem four main directions were defined to detect text. Three texture features for GLCM detection were selected. Then the mean and variance of correlation, entropy and homogeneity were also calculated for the text and non_text regions in dataset. The result show that multilayer perceptron and Random forest give almost the same result table 1.

Table 1: The results of detection using the GLCM feature and SVM, MLP, RF

Classifier	TP	FP	Precision	Recall	F Measure
SVM	0.956	0.055	0.956	0.956	0.956
MLP	0.971	0.029	0.971	0.971	0.971
RF	0.978	0.017	0.979	0.978	0.979

2.3.2 LBP descriptor

Three different cell sizes of 32, 16, 8 have been used to find the best for text classification. Experiments found that the best cell size for classification is 8X8 . The results of using 8x8 cell LBP feature can be found in table 2.

Table 2 : The results of text detection using LBP feature with 8x8 cellsize

Classifier	TP	FP	Precision	Recall	F Measure
SVM	0.915	0.088	0.918	0.915	0.915
MLP	0.912	0.139	0.912	0.912	0.911
RF	0.901	0.167	0.912	0.901	0.897

It can be concluded that a small cell size gives more LBP feature information, which can achieve greater classification model accuracy. The three classifiers (SVM, MLP, RF) which were used in experiments give almost the same result when the cell size is smaller.

2.3.3 HOG descriptor

Different cell sizes 50 × 50, 32 × 32, 25 × 25 have been used to find the best for text detection for each cell size, block size set to 2×2 , Block overlap to 1 × 1 ,the number of orientation histogram bins was set to nine, which provided a reasonably low dimensional feature vector that delivered good descriptive power and resulted in better classification accuracy. Because the HoG features are texture-based, the lengths of the HoG descriptors vary depending on the cell size and block size. The cell size 25x25 achieved overall best performance. Table 3 shows the classification accuracy of the three classifiers at the cell size 25 × 25.

Table 3: Text detection accuracy using the HOG with cell size 25 × 25

Classifier	TP	FP	Precision	Recall	fMeasure
SVM	0.856	0.159	0.860	0.856	0.857
MLP	0.889	0.120	0.893	0.889	0.890
RF	0.898	0.149	0.897	0.898	0.897

2.3.4 Aspect Ratio descriptor

The performance on text region re-identification from MSER candidates using Aspect Ratio features are compared with SVM, MLP and RF as shown in Table4. Random Forest gives the best accuracy.

Table 4: text detection accuracy using the Aspect Ratio

Classifier	TP	FP	Precision	Recall
SVM	0.671	0.671	0.450	0.671
MLP	0.698	0.587	0.705	0.705
RF	0.759	0.278	0.767	0.767

2.3.5 Combination of Multiple types of features

The GLGM, Aspect Ratio, LBP Histogram and HOG features were extracted from the MSER regions to form a combined feature vector for reclassification of text candidates. The following combinations of feature sets were studied to determine the possible best feature set : 1) AR and GLCM(AGLCM); 2) AR and LBP (ALBP); 3) AR and HoG (AHoG); 4) GLCM and LBP (GLBP); 5) GLCM and HOG(GHOG); 6) HOG and LBP (HLPB).

The results from table 5 show that the combination of all features give the best accuracy and SVM is the best classifier. Where the combination of LBP+GSLM+AR and the combination of LBP+GSLM give the second-best result.

the possible best feature set : 1) AR and GLCM(AGLCM); 2) AR and LBP (ALBP); 3) AR and HoG (AHoG); 4) GLCM and LBP (GLBP); 5) GLCM and HOG(GHOG); 6) HOG and LBP (HLPB).

Table 5: The accuracy of using combination of features

Combination of features	TP	FP	P	R	F	Best classifier
GLCM + AR	0.983	0.015	0.983	0.983	0.983	SVM
LBP+GLCM	0.987	0.023	0.987	0.987	0.987	SVM
LBP + AR	0.905	0.105	0.907	0.905	0.906	SVM
LBP+GLCM +AR	0.987	0.023	0.987	0.987	0.987	SVM
HOG+AR	0.912	0.124	0.914	0.912	0.911	RF
HOG+LBP	0.937	0.069	0.937	0.937	0.937	SVM
HOG+GLCM	0.983	0.021	0.983	0.983	0.983	SVM
HOG+GLCM +AR	0.983	0.021	0.983	0.983	0.983	SVM
HOG+lbp+GLCM +AR	0.993	0.012	0.993	0.993	0.993	SVM, RF

2.3.6 Combination of selected from multiple types of features

To reduce the feature space and speed up the processing cycle, we used the Correlation-based Feature Selection CFS approach as a feature selector. CFS algorithm helps to rank feature subsets according to the correlation based on the heuristic "merit" as reported by Boukharouba et al (Boukharouba 2014). This reduced the original feature attributes obtained from the descriptors of text candidate regions to the minimal.

CFS is a filtering algorithm that evaluates subsets of features based on the predicting power of the individual features of a class label. CFS is defined by Boukharouba et al (Boukharouba 2014) as:

$$Merit_{S_k} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}. \qquad (7)$$

Here, Sk is the number of features selected in the current subset, $\overline{r_{cf}}$ is the average value of all feature-classification correlations, and rff is the average value of all feature-feature correlations. It begins with an empty set of features adds one feature at a time that holds best discriminative value.

Table 6 shows the classification accuracy of the above 6 feature combinations give good results but the combinations of HOG , GLCM and AR produce the best accuracy with SVM as classifier.

Table 6: Classification accuracy results with selected features

Selected features	TP	FP	P	R	F	Bst classifier	No of selected features
HOG+LBP	0.925	0.113	0.925	0.925	0.924	RF	65+114
LBP+AR	0.923	0.120	0.926	0.923	0.921	RF	114+1
HOG+LBP+AR+GLCM	0.949	0.083	0.952	0.949	0.949	RF	65+114 +1+4
LBP+ GLCM	0.943	0.092	0.945	0.943	0.942	RF	114+4
LBP+ GLCM+AR	0.984	0.027	0.985	0.984	0.984	SVM	114+4+1
HOG+ GLCM+AR	0.956	0.051	0.956	0.956	0.956	RF	65+4+1

2.4 Candidate region re-identification (Test and Evaluate Phase)

The goal of the proposed method is to detect as many text components as possible and it is difficult or impossible to recover the missed characters in the following steps. This lead to set the threshold of MSERs to its lowest value 1 which made it possible to capture most challenging cases. For each image in testing phase the first step of the text detection module is to extract Maximally Stable Extremal Regions (MSER) in order to obtain text-region candidates. The MSER algorithm depends only on the intensity of the image. Since the text in the image tends to have equal intensity, that means the output of this step are candidate regions containing at least one symbol. Figure (4) shows the result of detecting MSER regions. It shows clearly that the MSER algorithm detects many false positives – regions (non-text). The classification model has been used to classify candidate regions to text and non text regains. Where the best descriptors and best classifier which are determined in training phase have been used (the combination of all features (HOG+lbp+GLSM+AR with SVM). All regions that are classified as non-text regions by the MSER are removed from the scene image. MSER regions that are classified as text are then mapped onto the image and bounded by boxes. The pixels inside these bounding boxes are marked as text pixels, as shown in Figure (5).

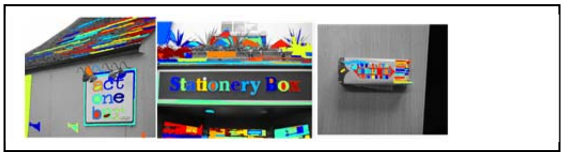


Figure 4 : The result of using MSER detector

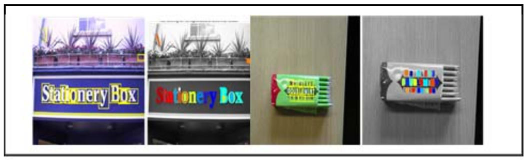


Figure 5: Non-text regions are removed using SVM classifier learnt from the combination of features, b: candidate characters bounded by boxes

2.5 Text localization

The character grouping module joins the detected characters into text regions – which may be words or text lines. This step is important to enable recognition of the actual words in an image, providing more meaningful information than just individual characters. It is also an essential step when using Optical character recognition (OCR) to recognize the words in largely connected text regions. One approach for merging individual text regions into words or text lines is to first find neighboring text regions and then form a bounding box around these regions. To find neighboring regions, the bounding boxes computed earlier, which are the result of the MSER regions classification to text and non_text, are expanded. This makes the bounding boxes of neighboring text regions overlap such that text regions that are part of the same word or text line form a chain of overlapping bounding boxes Figure (6).

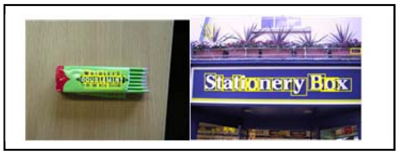


Figure 6: Results of expanded bounding boxes of texts

The overlapping bounding boxes can then be merged together to form a single bounding box around individual words or text lines. To do this, the overlapping ratios between adjacent bounding boxes were calculated. Therefore, non-zero overlapping ratios would indicate possible neighboring characters in words or at different text lines figure 7.

2.6 Performance Evaluation

To evaluate the performance of the proposed method, Precision p and recall r have been calculated. precision was defined as the ratio of correct estimated area to the whole detected region. A method has low precision if the number of text bounding rectangles is too large. Recall was defined as the ratio of the correct estimated area to the area of the ground truth regions. Where area of a region refers to the number of pixels inside it. Methods obtaining low precision mean that the methods over-estimated, while low recall means the methods under-estimated. Hence, the best match m(r;R) for a rectangle r in a set of rectangles R is defined as

$$m(r;R) = max\{m_p(r,r')|r' \in R\} \qquad (8)$$

$$Precision = \frac{\sum_{r_e \in E} m(r_e.T)}{E}, \qquad (9)$$

$$Recoll = \frac{\sum_{r_t \in E} m(r_t.T)}{T} \qquad (10)$$

Where E are a set of the estimated boxes and T are the sets of target (ground truth) boxes, respectively. These two measures are combined into a single quality measure f with a weight.

factor α set to 0.5, which represents the relative weight between the two metrics (Lucas, S.M. et al 2003; Lucas, S.M. et al 2005)

$$f = \frac{1}{\frac{\propto}{percision}+\frac{1-\propto}{Recall}} \qquad (11)$$

The proposed method has been evaluated on several public test datasets and compared against several state-of-the-art text detectors described in the literature. Specifically, we compared it with the contestants of the ICDAR Challenge (Lucas S.M 2005), and also with the detectors of (Epshtein et al. 2010), (Chen ,H. et al. 2011)., (Pan et al. 2011) , (Neumann et al. 2012), (Yi,C. and Y.Tian 2012), and (Yao, C. et al., 2012). Tables 7 and 8 show the results obtained for each dataset.

Table 7: Text detection scores of proposed method and other detectors on the ICDAR 2003 dataset(%)

Algorithms	Precision	Recall	F Measure
Proposed Method	0.85	0.79	0.81
Ye Q. &Doermann, D (Ye, Q. &Doermann, D., 2013)	0.892	0.623	0.733
Yin et al (Yin et al 2013)	0.86	0.68	0.76
Neumann and Matas (Neumann, L. Matas, J. 2013)	0.85	0.68	0.75
Shi et al. (Shi et al. 2013)	0.83	0.63	0.72

Table 8: Text detection scores of proposed method and other detectors on the ICDAR 2011 dataset (%)

Algorithms	Precision	Recall	F Measure
Proposed Method	0.85	0.83	0.83
Yi,C. and Y.Tian (Yi,C. and Y.Tian 2012)	0.73	0.67	0.70
Pan et al. (Pan et al. 2011)	0.67	0.70	0.69
Yao, C. et al. (Yao, C. et al., 2012)	0.69	0.66	0.67
Chen ,H. et al. (Chen ,H. et al. 2011)	0.73	0.60	0.66
Epshtein et al. (Epshtein et al. 2010)	0.73	0.60	0.66

The training data on the ICDAR 2011 datasets were not applied for testing in experiments. Tables 7 and 8 show that the proposed method accomplished excellent performance on both datasets and the improvements are significant in terms of Precision, Recall, and F measure. The increased in terms of the proposed method is mainly due to the combinations of different types of features (HOG+ LBP +GLSM+AR).

3. Conclusion

In this paper a text detection and localization method is presented. The proposed method improved text detection using MSER through a re-identification step using classification models learnt from GLCM, LBP, HoG, Aspect Ratio and combinations of these features. The re-identification performances of SVM, MLP and RF classifiers are compared with regard to accuracy. A combination of HOG+lbp+GLSM+AR gives the best accuracy followed by the combination of LBP+GSLM on tested data set.

The ICDAR2003,2011 dataset has been used as a benchmark in our experiments. After text pixel regions are confirmed, character grouping based on overlapping ratio of bounding boxes is employed to join pixel regions into word regions or text lines, enabling fast text recognition when using off-the-shelf OCR. As our future work, we aim to improve the feature selection method using deep learning, thus to find more discriminative features and achieve better robustness

Acknowledgment

This work was implemented during Miss Hanaa Mahmood was a PhD student at Loughborough University, UK. The authors thank the Iraqi (MOHESR) for the scholarship funding.

REFERENCES

[1] Chen, H. et al., 2011. Robust text detection in natural images with edge-enhanced maximally stable extremal regions. Proceedings - International Conference on Image Processing, ICIP, pp.2609–2612.

[2] Chowdhury, A., Bhattacharya, U. & Parui, S.K., 2012. Scene text detection using sparse stroke information and MLP. 21st International Conference on Pattern Recognition (ICPR 2012), (Icpr), pp.294–297.

[3] Clausi ,D. A.,(2002). An analysis of co-occurrence texture statistics as a function of grey level quantization. Canadian Journal of remote sensing, 28(1):,pp45-62.

[4] Dalal, N., and Triggs, B.,. (2005) Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, PP: 886-893. IEEE.

[5] Epshtein, B., Ofek, E. & Wexler, Y., 2010. Detecting Text in Natural Scenes with stroke width transform. IEEE Conf. Comput. Vis. Pattern Recognit, (d), pp.2963–2970

[6] Gomez, L. , Karatzas, D., 2013 "Multi-script text extraction from natural scenes," in ICDAR,

[7] Gonzalez, Alvaro, et al. (2012) "Text location in complex images." Pattern Recognition (ICPR), 21st International Conference on. IEEE, 2012.

[8] Hanif S. M., Prevost, L. , (2009) , Text Detection and Localization in Complex Scene Images using Constrained AdaBoost Algorithm , 10th International Conference on Document Analysis and Recognition.

[9] Haralick, R. M., Shanmuga, K., & Dinstein, H. (1973). Textural features for image classification. IEEE Transactions on Systems Man and Cybernetics, 3(6),pp:610-621.

[10] Kim, K.I., Jung, K. & Kim, J.H., 2003. Texture-Based Approach for Text Detection in Images Using Support Vector Machines and Continuously Adaptive Mean Shift Algorithm æ. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 25(12), pp.1631–1639.

[11] Kwak, J. T., Xu, S., & Wood, B. J. (2015). Efficient Data Mining for Local Binary Pattern in Texture Image Analysis. Expert Systems with Applications, 42(9), 4529–4539. http://doi.org/10.1016/j.eswa.2015.01.055.

[12] Lucas, S. M. , Panaretos, A. , Sosa, L. , Tang, A. , Wong, S., Young, R., Ashida, K., Nagai, H., Okamoto, M., Yamamoto, H. Miyao, H., J. Zhu, W. Ou, C. Wolf, Jolion, Todoran, J.-M., L. ,Worring, M., and Lin. X. ,(2005), ICDAR 2003 robust reading competitions: entries, results, and future directions. IJDAR, 7(2-3),PP:105–122.

[13] Lucas, S.M. Panaretos, A . Sosa, L. Tang, A. Wong, S. Young, R., 2003. ICDAR 2003 Robust Reading Competitions. In INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION (ICDAR).

[14] Matas, J. Chum, M. Urban, and T. Pajdla , 2002. Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. Proc. of British Machine Vision Conference, pp.384–393.

[15] Minetto, R. , Thome, N., Cord, M., Stolfi, J. Precioso F., Guyomard, J., and Leite, N. J.,(2011), "Text detection and recognition in urban scenes," in IEEE ICCV Workshops, 2011, pp: 227–234.

[16] Neumann, L. & Matas, J., 2010. a Method for Text Localization and Recognition in Real - World Images. In 10th Asian conference on computer Vision, ACCV. pp. 770–783.

[17] Neumann, L. Matas, J. 2013, On combining multiple segmentations in scene text recognition, in: Proceedings of the ICDAR, pp. 523–527.

[18] Neumann, L. Matas, J., 2012, Real-time scene text localization and recognition, in: Proceedings of the CVPR, pp. 3538–3545.

[19] Pan, Y., Hou, X. & Liu, C., 2011. A Hybrid Approach to Detect and Localize Texts in. IEEE TRANSACTIONS ON IMAGE PROCESSING, 20(3), pp.800–813.

[20] Seeri, S. V, Pujari, J.D. & Hiremath, P.S., 2015. Multilingual Text Localization in Natural Scene Images using Wavelet based Edge Features and Fuzzy Classification. , 4(1), pp.210–218.

[21] Serra, J., Simon (Ed.) J.C. , (1989), Toggle Mappings: From Pixels to Features, Elsevier ,pp: 61-72

[22] Shi, C. Chunheng, W. Baihua, X. Yang, Z Song, G., 2013. Scene text detection using graph model built upon maximally stable extremal regions. Pattern Recognition Letters, 34(2), pp.107–116.

[23] Soh L. K. and satsoulis, C. T, (1999), Texture analysis of sar sea ice imagery using gray level co-occurrence matrices. Geoscience and Remote Sensing, IEEE Transactions on, 37(2),PP:780-795.

[24] Wang, X. Song, Y.,Zhang, Y., Xin, J, 2015. Natural scene text detection with multi-layer segmentation and higher order conditional random field based analysis. Pattern Recognition Letters, 60-61, pp.41–47.

[25] Wang, X. Song, Y.,Zhang, Y., Xin, J, 2015. Natural scene text detection with multi-layer segmentation and higher order conditional random field based analysis. Pattern Recognition Letters, 60-61, pp.41–47.

[26] Yao, C. et al., 2012. Detecting Texts of Arbitrary Orientations in

[27] Ye, Q. & Doermann, D., 2015. Text Detection and Recognition in Images and Video : a Survey. IEEE transactions on pattern analysis and machine intelligence, 37(7), pp.1480–1500

[28] Yi, C. , Tian, Y., (2012), Localizing text in scene images by boundary clustering, stroke segmentation, and string fragment