ETHIOPIC AND LATIN MULTILINGUAL TEXT DETECTION FROM IMAGES USING HYBRID TECHNIQUES

Atirsaw Awoke[1] and Menore Tekeba[2]

[1]School of Electrical and Computer Engineering, Bahir Dar University, Bahir Dar, Ethiopia
[2]School of Electrical and Computer Engineering, Addis Ababa, Institute of Technology, Addis Ababa University, Addis Ababa, Ethiopia
Corresponding Author's Email: awokeeyrw@gmail.com

ABSTRACT

Caption and scene texts found in images contain valuable information. These texts can be used for many applications to answer questions like what, when, where, and by who to give context to the images. So, automatic text detection enhances the user's understanding of the media content. In Ethiopia, most street posts and promotional boards are written in multilingual characters such as Latin (English, Afaan Oromo etc.) and Ethiopic (Amharic, Tigrigna etc.). In this work, we have studied Ethiopic and Latin multilingual text detection from images for both caption and scene texts. After the images are pre-processed, maximally stable extremal region (MSER) algorithm, aspect ratio and stroke width transform (SWT) algorithm are used to extract text regions, respectively. Then texture features are computed using local binary patterns (LBP) from the extracted regions. Finally, the support vector machine (SVM) is used to classify text region vs non-text using the computed LBP features. We prepared a new multilingual Ethiopic and Latin script image dataset to evaluate our method.

Keywords*:* Images, LBP, SWT, MSER, SVM classifier

INTRODUCTION

The text embedded in an image provides clear and more obvious information about the content of specific media. It is necessary to detect and extract the text information from these images automatically for potential applications such as image retrieval and processing, in robotics, computer vision and intelligent transport systems. However, developing a robust system for extraction of texts from captured scenes is a great challenge due to several factors such as variations of style, colour, spacing, distribution, layout, light, background complexity, presence of multilingual scripts and fonts. In this modern era where we live today, multimedia technologies play an important role in transforming row data into digitally encoded information. Automatic detection of the region of a text area in images is an active research issue in the design of computer vision systems. Image text can broadly be classified into two categories: artificial or caption text and scene text.

Artificial text refers to those characters generated by graphic titling machines superimposed on video images, such as video captions, while scene text occurs naturally as a part of scene, such as text in images information boards/signs, nameplates, food containers, etc. The goal of this work is to develop a system which will efficiently detect

image regions that contain Ethiopic and Latin texts from images. The paper is organized as follows. The first part of the paper provides an overview of the general background and reviews of different works which helps to understand multilingual text detection techniques from images.

The next part basically included the design and development parts of the paper. Here the proposed model and algorithms used in the model are discussed. Then the results and the experiments conducted are discussed. Finally, the conclusion and the recommendation of a future work which are forwarded by the researchers are put together.

RELATED WORKS

Current text detection from images approaches can be broadly categorized into four groups: texture-based approach, region-based approach, hybrid approach of texture and region-based Methods and Morphological based text detection.

A. Texture Based Approach

These methods are based on the fact that texts in images have distinct textural properties which distinguish them from the background [1]. Wenge, et al, Proposed wavelet transform of an image is done to characterize the local energy variations (LEV) of pixels in the successive scale levels. In each scale level, the corresponding local energy variations are computed. The resulting binary map image in each scale level is subsequently analysed by connected component analysis (CCA) technique to label different objects and backgrounds. Finally, all text regions in the consecutive scale levels are fused into the original image and text regions are detected.

Kim Kwang suggested a technique that uses a combination of (Continuously Adaptive Mean Shift) CAMSHIFT and SVM for detection and extraction of text is proposed. They use a small window to scan the input image, classifies the pixel located at the centre of the window into text or non-text by analysing its textural properties using a SVM. Then the CAMSHIFT algorithm is used to verify text regions which are the result of the texture analysis. Ye, Qixiang, et al [4] applied coarse-to-fine detection framework by using different text properties in different detection stages. In the coarse detection, candidate text regions are firstly obtained using properties of dense intensity variety and contrast between text and its background. In the fine detection step, Texture property is used to discriminate text with other non-text patterns. Texture features such as wavelet moment features, wavelet histogram features, wavelet co-occurrence features and crossing count histogram features are used to identify text lines from the candidate ones. A forward search feature selection algorithm is used to find effective features and an SVM classifier is used to perform text/non-text classification tasks.

One system developed by Vinod, H. C., et al in [5] for detection of text makes use of Laplacian method based on wavelet and colour features. The Maximum Gradient Difference (MGD) map is obtained by moving the window over the image. Text regions typically have larger MGD values than non-text regions because they have many positive and negative peaks. Therefore, they convert the input frame into a binary frame and then Fuzzy C-means is applied to classify the feature into two clusters: background and text candidates.

B. Region Based Approach

Region-based approaches done by Zhang, Jing, and Rangachar Kasturi in [1] attempts to use similarity standard of text, such as colour, size, stroke width, edge and gradient information. These approaches usually have lower computation cost and the outputs can closely cover text regions. In paper [6] Liu, Xiaoqing, and Jagath Samarabandu proposed a method which has three stages: Candidate text region detection, text region localization and character extraction.

In its first stage, they used the magnitude of the second derivative of intensity as a measurement of edge strength, and this allowed a better detection of intensity peaks that normally characterizes text in the images. Edge detector is carried out by using a multiscale strategy in which the multiscale images are mainly produced by Gaussian pyramids after successively applying a low pass filter and down sample the original image reducing the image in both vertical and horizontal directions.

In the second stage, they assume texts are found in clusters arranged compactly. So, this characteristic of clustering is used to localize text regions. In the third stage, the existing OCR engines were used. For text detection mainly from video images, Anthimopoulos, Marios, et al [7] applied a two-stage methodology. In its first stage, the text lines are detected based on the canny edge map of the image. In the next stage, the result is refined using the sliding window and a SVM classifier. Feature vectors of the candidate regions are obtained usingLBP which describes the local edge distribution.

In research work [8], Chen, Huizhong, et al proposed Connected Component based (CC-based) text detection algorithm, which employs MSER as basic letter candidates. To improve the weakness of MSER they deploy the complimentary properties of canny edges and MSER is combined in edge-enhanced MSER. Further they propose to generate the SWT image of these regions using the distance transform to efficiently obtain more reliable results. Then they apply the geometric as well as stroke width information to perform filtering and pairing of CCs. Finally, letters are clustered into lines. In other work which uses a connected component-based approach for text extraction [9] is based on colour reduction technique. The Colour images are converted to gray scale images and edge image is generated using a contrast segmentation algorithm. Luminance value thresholding is applied to increase the contrast between the possibly interesting regions and the rest of the image. Then they applied the horizontal projection of the edge image in order to localize the possible text areas.

In the works of Zhang, Xin, et al [10], the Colour-Edge combined algorithm consisting of two stages is proposed for text detection and extraction. In the first stage, they detect the text area by applying the Colour-Edge combined algorithm.

In the second stage, the text background is removed and the left part is binarized for text location. The Transition Map method is used to filter the text from the background. The model makes use of the exponential changes of colour between the edge of text and background to detect the text area, and then the background is removed. To improve the efficiency of the method, canny edge detection and some morphology operation is performed

C. Hybrid Approach of Texture and Region Based Methods

Hybrid approaches [1] seek to introduce the textural property of text regions into region-based approach. In a research work done in [11] they try to combine texture and CC-based information to detect text. They propose first and second order statistical texture features to detect and localize the text in the image. CC extraction is used to segment candidate text components from the localized text region. Finally, morphological operations and heuristic filters are used to filter out non-text components. The other literature in this regard is the one mentioned in [12]. The authors created a text confidence map for a series of different scaling of gray scale images to represent the possibility of text on an area using the wald boost classifier trained by histogram of gradient (HOG) features. They adopted Niblack's algorithm to convert gray scale image into binary image, and uses the conditional random field to determine whether the candidate area contains text or not.

Finally, they applied a minimum spanning tree to connect the same line of text.

D. Morphological Based Approach

In work [13] by Thilagavathy, A., et al, the researchers go through the following steps to detect texts. They split the video into frames. Key frame is selected from the frames by edge comparison with the help of the Sobel edge detector.

Text candidate regions are generated again using Sobel edge detector. They assume texts in video frames found closer to each other. Thus, morphological dilation operation is performed to remove pixels that are far away from the candidate region.

PROPOSED METHODOLOGY

Among the many published region-based methods, we observe an increasing use of the MSER algorithm for character candidate's detection. Based on the review of previous methods, we identify the following text characteristics that are frequently used in text detection, localization, and extraction: contrast, colour, stroke density and aspect ratio. After pre-processing the target image, since text characters usually have consistent colour, we begin by finding regions of similar intensities in the image using the MSER region detector which is applied for character candidate extraction. Although the MSER algorithm picks out most of the text regions, it also detects many other stable regions in the image that do not contain text. For text detection either rule based or machine learning approach can be used.

We apply the combination of the rule-based methods and a machine learning approach to produce better results. False regions, which are detected as text regions by MSER, are minimized using the geometric property of texts which is the aspect ratio. Then SWT Algorithm is used to refine the false positives. Even after the SWT filtering is performed regions which are not a text are also detected as text region candidates. To eliminate those candidate regions which are not text, each candidate text regions are verified using a machine learning technique.

A. Pre-processing

Noise can seriously affect the quality of digital images. Different factors may be responsible for introduction of noise in the image. In this phase of the system, we apply three different filters used for smoothing, sharpening and denoising to pre-process images in order to obtain a picture with more stable text regions. Eliminating the noise without blurring the details too much and enhancing edges without amplifying noise is very difficult. So, when using more than one filter, special care should be taken in order to make sure their effect is important.

To get information which can describe images, we use these filters in the following order.

a) Bilateral filter: To smooth images.
b) High-pass filter: To sharpen images
c) Median filter: To filter out noise from images.

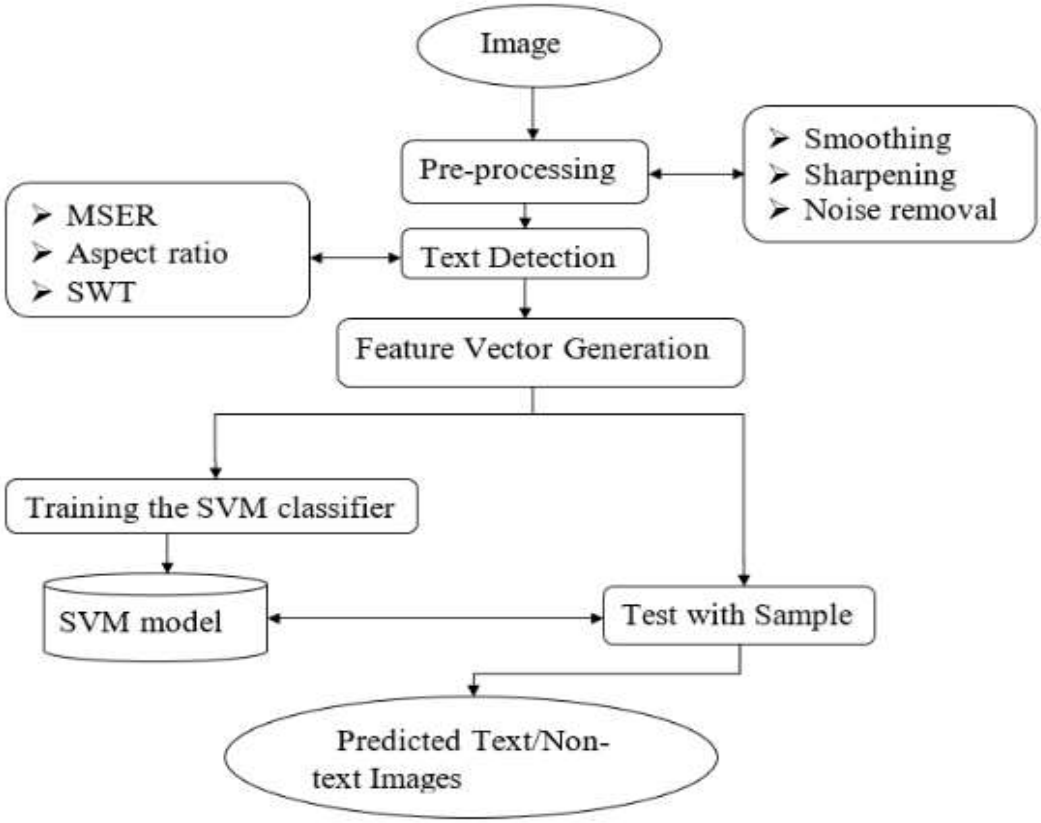

Fig. 1: System architecture for text detection

B. Extracting Character Candidates

In order to refine the character candidate information in images, we use MSER to retrieve the edges and local features of characters. The word extremal refers to the property that all pixels inside the MSER have either higher (bright extremal regions) or lower (dark extremal regions) intensity than all the pixels on its outer boundary. The MSER works fine for finding text regions. Since the consistent colour and high contrast of text regions with the background leads to stable intensity profiles.

C. Constructing Text Candidates

We choose the character candidates according to structural features called aspect ratio as given in equation 1. Aspect ratio: the shapes of characters are found to be roughly rectangular connected components. But in English" i"," I"," l"," j"," J" are comparatively thin texts, with arrangements similar to cable in an image. The below threshold is used to remove the background noises while it preserves the thin text at the same time.

$$Aspect\ Ratio = \frac{Max(Width,Height)}{Min(Width,Height)} \quad (1)$$

The biggest threshold is used to eliminate big candidate regions while the small threshold is used to discard non-text regions like wires and leaves.

D. Stroke Width Transform

Stroke width of a text remaining is almost the same as in a single character; however, there is significant change in stroke width in non-text regions as a result of their irregularity. The initial step of stroke width extraction is to get skeletons of the remaining MSERs. On every foreground pixel on the skeleton, distance transform is applied to compute the Euclidean distance from this pixel to the nearest boundary of the corresponding MSER.

Then we obtain a skeleton-distance map. If the standard deviation of the stroke width of the character candidates is large, it is less likely to be a true character. A value of 0.5 is chosen as a threshold for removal of non-text regions using equation 2.

$$Thresh. = \frac{Standard\ deviation_{stroke\ width}}{Mean_{Stroke\ width}} \quad (2)$$

Even after the stroke width filtering is performed, regions which are not text are also detected as text region candidates.

We first merge the individual characters into a single connected component using a small bounding box threshold which is 0.02. To eliminate those candidate regions which are not text, for each of the candidate text regions within the bounding box, their respective feature vector is computed with LBP which is the input for SVM. Finally depending on the feature vector, the SVM classifier classified each region as text and non-text. Those regions which are classified as text will be the output of the system.

E. Feature Extraction

LBP which is a feature extraction algorithm is chosen because one of the aspects of text is that it has a unique pattern for each character. These can be used for constructing a feature descriptor for a window. The LBP operator assigns a decimal number for each pixel of an image, called LBP codes. This pattern is used to encode the local structure around each pixel. Each pixel is compared with its eight neighbours in a 3 x 3 neighbourhood by subtracting the centre pixel value. The results which are less than the centre pixel value is encoded with 0 and others with 1. A binary number is obtained by concatenating all these binary codes in a clockwise direction starting from the top left one and its corresponding decimal value is used for labelling.

Table 1 LBP Computation

1	2	2		0	0	0	Binary:
9	5	6	-	1	1	1	00010011
5	3	1	>	1	0	0	Decimal:19

F. Classification

The preliminary aim of the classification phase of our system is to distinguish candidate text regions which come from as output from rule-based filtering as text or non-text. Finally, it is determined if the text candidates actually contain a text string using a SVM classifier.After text candidate combination; we use SVM to classify potential text regions into text and non-text using the extracted feature vectors. Those text regions which are classified as text will be the output of the system.

RESULTS AND DISCUSSIONS

As far as we know, there is no work done in detection of Ethiopic text from images. Due to this there is no available Ethiopic text image database used for conducting an experiment. So, we prepare a multilingual scene and caption text dataset. The dataset contains 400 images, which contains 50 caption images and 350 scene images.

The second dataset used is the International Conference on Document Analysis and Recognition (ICDAR) 2003 dataset which evolves from a series of robust reading competitions held by ICDAR. The dataset consists of 462 images including 229 for training and 233 for testing. For each word within each image, the ground truth is fully annotated. We set two types of evaluation units in the dataset; objective and subjective evaluations.

Objective evaluation, the word level as shown in Figure 2.1(a), is used in the ICDAR dataset. However sometimes, it is hard to partition text regions within an image into individual words based on their spacing; it is almost impossible and non-trivial to perform word partition. Therefore, we consider subjective evaluation, the evaluation metric to use regions which contain the text rather than a word as shown in Figure 2.1(b).

(a) Sample ground truth for objective evaluation

(b) Sample dataset for subjective evaluation

Fig 2: A multilingual scene text example.

The performance of our approach was evaluated by measuring its detection rate, based on precision, recall, and F-Measure. The three evaluation values were computed by using equations 3-5, respectively. Precision is defined as the number of true positives (TP) over the number of true positives plus the number of false positives (FP).

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

Recall is defined as the number of true positives (TP) over the number of true positives plus the number of false negatives (FN).

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

These quantities are also related to the (F-Measure) score, which is defined as the harmonic mean of precision and recall.

$$F - Measure = 2 * \left(\frac{(Precision*Recall)}{(Precision+Recall)}\right) \quad (5)$$

1) ***Objective Evaluation***: This evaluation unit is based on the ICDAR word level evaluation mechanism. DetEval software is used for evaluation. The first experiment is conducted on our prepared dataset. For both the training and testing 400-word blocks are extracted using our system. All the dataset is fully annotated. Finally, DetEval software is used for objective evaluation and the following results in Table II are obtained.

Table 2: Text detection result

Precision	Recall	F-Measure
81%	74%	77.5%

As far as we know there is no work to compare our method with it by conducting an experiment with the same dataset. So, to compare with other methods we conduct an experiment on the publicly accessible ICDAR 2003 dataset. We compare our approach with existing methods which conduct experiments on ICDAR 2003 dataset. The comparison is made with the state-of-the-art method developed by Boris Epshtein.

Table 3: Text detection result in ICDAR dataset

	Criteria for Assessment			
Method		Precision	Recall	F-Measure
	Ours	78%	70%	74%
	Method[14s]	73%	60%	66%

The experimental results in Table 3 demonstrate the effectiveness of the proposed method. Our method performs better than the results obtained by the state-of-the-art method given in [14]. We believe the improvement in the performance comes from the usage of the pre-processing stage, MSER and classification stage of our system.

2) Subjective Evaluation: For most images as shown in Figure 2.1(b), it is hard to partition and prepare a ground truth for a word text in images.

As a result, even if the detector successfully identified the text in the image, the match scores between a bounding box for an entire block of text and bounding box for a single word tended to be very low. In this evaluation mechanism there are no ground truths consisting of bounding boxes for individual words, while our detection system outputs bounding boxes for the entire group of text regions.

We consider the evaluation metric to use regions which contain the text rather than a word level objective evaluation mechanism. The performance measure in the subjective evaluation is its detection rate, defined as the ratio between the number of detected text images and all the given images containing text. The experiment is conducted using our prepared dataset.

Table 4: Text detection result

Precision	Recall	F-Measure
97.28%	84.12%	90.7%

(a) MSER detection result

(b) After removing non-text regions based on aspect ratio

(c) After removing non-text regions based on stroke width transform

(d) Candidate text regions

(e) Final detection result

Fig. 3: Sample step by step text detection.

CONCLUSIONS

Text detection in images is a difficult problem. The difficulty is due to wide dissimilarity in fonts, sizes colour and textures of text regions embedded in scenes. The location of the text in the image also follows a random behaviour presence of repeating patterns and complex backgrounds in unconstrained images increase the difficulty for detecting text from images. Encoding all these variabilities in a rule-based approach is extremely challenging. So, in this work, we combine rule-based approach with machine learning methods to generate a model to discriminate between text and non-text regions from images. In order to minimize the effect of noise in images, pre-processing techniques have been applied in input images.

Then since text characters usually have consistent colour, we begin by finding regions of similar intensities in the image using the MSER region detector. Although the MSER algorithm picks out most of the text, it also detects many other stable regions in the image that are not text. Geometric properties of a text character which is aspect ratio followed by SWT are used to filter out non-text regions. After the geometric filtering and SWT is performed, regions which are not a text may be detected as text region candidates.

To eliminate these candidate regions which are not text, the respective feature vector of each candidate text region is computed by LBP. The feature vector is then used as input for the SVM classifier which decides whether the candidate regions are text or not.

FUTURE WORKS

1. Script identification can be done since the output of the system which is the localized text regions may contain multiple lines of text with different scripts which can be applied as input for script identifier.
2. Multilingual OCR can be developed using this work as a pre-processing step.

REFERENCES

[1] Zhang, Jing, and Rangachar Kasturi.” Extraction of text objects in video documents: Recent progress.” Document Analysis Systems, 2008. DAS’08. The Eighth IAPR International Workshop on. IEEE, 2008.

[2] Mao, Wenge, et al.” Hybrid Chinese/English text detection in images and video frames.” Pattern Recognition, 2002. Proceedings. 16th International Conference on. Vol. 3. IEEE, 2002

[3] Kim, Kwang In, Keechul Jung, and Jin Hyung Kim. ” Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm.” IEEE Transactions on Pattern Analysis and Machine Intelligence25.12 (2003): 1631-1639

[4] Ye, Qixiang, et al.” Fast and robust text detection in images and video frames.” Image and Vision Computing 23.6 (2005): 565-576.

[5] Vinod, H. C., S. K. Niranjan, and G. L. Anoop. ” Detection, Extraction and Segmentation of Video Text in Complex Background.” International Journal on

Advanced Computer Theory and Engineering 5 (2013): 117-123.

[6] Liu, Xiaoqing, and Jagath Samarabandu.” Multiscale edge-based text extraction from complex images.” Multimedia and Expo, 2006 IEEE International Conference on. IEEE, 2006.

[7] Anthimopoulos, Marios, Basilis Gatos, and IoannisPratikakis.” A two-stage scheme for text detection in video images.” Image and Vision Computing 28.9 (2010): 1413-1426.

[8] Chen, Huizhong, et al.” Robust text detection in natural images with edge-enhanced maximally stable extremal regions.” Image Processing (ICIP), 2011 18th IEEE International Conference on. IEEE, 2011.

[9] Gllavata, Julinda, Ralph Ewerth, and Bernd Freisleben.” A robust algorithm for text detection in images.” Image and Signal Processing and Analysis, 2003. ISPA 2003. Proceedings of the 3rd International Symposium on. Vol. 2. IEEE, 2003

[10] Zhang, Xin, Fuchun Sun, and Lei Gu.” A combined algorithm for video text extraction.” Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on. Vol. 5. IEEE, 2010.

[11] Kumuda, T., and L. Basavaraj.” Hybrid Approach to Extract Text in Natural Scene Images.” International Journal of Computer Applications 142.10 (2016).

[12] Pan, Yi-Feng, Xinwen Hou, and Cheng-Lin Liu.” A hybrid approach to detect and localize texts in natural scene images.” IEEE Transactions on Image Processing 20.3 (2011): 800-813.

[13] Thilagavathy, A., et al.” Tamil Text detection in videos.”International Journal of Engineering and Innovative Technology (IJEIT)Volume 3, Issue 9, March 2014

[14] Epshtein, Boris, EyalOfek, and Yonatan Wexler.” Detecting text in natural scenes with stroke width transform.” Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010.