Extraction and Identification of Manipuri and Mizo Texts from Scene and Document Images

Loitongbam Sanayai Meetei(✉), Thoudam Doren Singh, and Sivaji Bandyopadhyay

Department of Computer Science and Engineering, National Institute of Technology Silchar, Silchar, Assam, India
loisanayai@gmail.com, thoudam.doren@gmail.com, sivaji.cse.ju@gmail.com

Abstract. The content inside an image is exceptionally compelling. As such, text within an image can be of special interest and compared to other semantic contents, it tends to be effectively extracted. Text detection within an image is the task of detecting and localizing the portion of an image that contains the text information. Manipuri and Mizo are respectively the lingua francas of two neighboring northeastern states of Manipur and Mizoram in India. While Manipuri, is currently written using Meetei Mayek script and Bengali script, Mizo is written in Roman script with circumflex accent added to the vowels. In this work, we report the task of text detection in natural scene images and document images in Manipuri and Mizo. We made a comparative study between Maximally Stable Extremal Regions (MSER) coupled with Stroke Width Transform (SWT) and Efficient and Accurate Scene Text Detector (EAST) for the text detection. The detected text portion of both the languages is subjected to Optical Character Recognition (OCR) and a post OCR processing of spelling correction. In our experiment of the text detection, EAST outperformed the other method.

Keywords: Text detection · SWT · MSER · EAST · OCR · Manipuri · Mizo

1 Introduction

Manipuri and Mizo are the official languages of two neighboring northeastern states of India, namely Manipur and Mizoram. Manipuri is considered to be a Tibeto-Burman language though a clear cut boundary is not drawn and Mizo is a Kuki-Chin family language. Further, both Tibeto-Burman and Kuki-Chin languages are considered to be the sub-family of Sino-Tibetan languages. Manipuri is written using Meetei Mayek (also known as Meitei Mayek) along with Bengali script in the field of academics and online news websites. In the case of Mizo,

B. Deka et al. (Eds.): PReMI 2019, LNCS 11941, pp. 405–414, 2019.
https://doi.org/10.1007/978-3-030-34869-4_44

25 letters of the Roman script are used (excluding the letter 'q') along with circumflex accent added to the vowels. Also, a character Ṭ (minuscule: ṭ) is used in the orthography of Mizo which is pronounced as 'tr'. The usage of Manipuri and Mizo can be found in both the states and other neighboring northeastern states like Assam and Tripura. Their usage can also be found in the neighboring countries: Myanmar and Bangladesh. The number of speakers for Manipuri is around 3 million and that of Mizo is around 1 million.

The amount of natural language text available in the digital or electronic format is increasing day by day but mostly unstructured. While the data available in the textual format are somewhat consumable and ready for processing, the text in the image or video frames need further processing to be made ready for use. In whatever form the text appear in any multimedia format, text wants to be found or noticed. The text is displayed in a printed or any multimedia format for readability, meaning the contrast between the text and background is high and the strokes are regular enough that any normal person can detect them. The text can provide a great amount of useful information such as describing the theme of the image or other useful information e.g. name of a location, license plate number, etc. The extracted text information can be used for a variety of applications such as text-based image indexing, keyword-based image search, etc. The overarching goal is to convert the text data appearing in an image into high-quality information.

The type of dataset for performing text detection can be of a wide variety. Some of them are as follows: 1. Document images in the gray-scale format and multicolor format. This includes single or multiple column text from a book or news articles, textbook covers, etc. 2. Images with the caption, the text could be overlaid on the background or inside a frame for better contrast. Such images are mostly found in video frames, newspaper, etc. 3. Scene text, where the text appears on the image captured by an electronic device in an environment. Unlike the text in the document images, text in scene images tends to suffer from variations in skew, perspective, blur, illumination, alignment, etc.

The general steps for text detection from a multimedia document can be grouped into the following four stages: (1) Text detection, where the presence of text content is checked in the input data. If text content is detected the system proceeds to the next step of localization. (2) In text localization step, the area where the text appears in an image is localized. It can be carried out in either texture based or connected component based approach. (3) In the third step, after tracking the text area present in the image, the particular segment is extracted. The extracted portion is then enhanced for better visibility. (4) Finally, the segmented image portion of the text area from step 3 is processed for character recognition. The process is also known as optical character recognition (OCR) where the recognized characters are converted to machine-encoded text.

Detection of text from multimedia document images has a wide variety of applications in different domains. With smartphone devices becoming ubiquitous and playing a vital role in our daily routine, a consumer can use it to fetch the information about a product by sending the image captured from their smart-

phones to a remote server [16]. It can also be helpful in minimizing laborious work such as building book inventories [3]. The application of text detection in a scene images can provide aid to a visually impaired person in their daily commute [5,8]. Globally, the number of vehicles is increasing steeply. The detection of a vehicle license plate from a natural scene image [2,10,11] can be utilized in monitoring the traffic system, building an automated parking and other law enforcement activities. To the best of our knowledge, there is no report on text detection of Mizo from an image. While the research work of text detection in Manipuri is very recent activity. The main objective of our work is to detect the text written in Manipuri and Mizo in natural scene image and document image. And further, carry out the process of OCR on the segmented portion of the image containing text. The recognized text is again subjected to the post-OCR processing step of spelling correction. The work can be helpful in empowering various research area such as categorization of images, text mining, etc. The remaining of this paper is structured as follows. Section 2 describes the related works, we present our approach model in Sect. 3. The experimental results are discussed in Sect. 4, with Sect. 5 summarizing the conclusion and future work.

2 Related Works

A work on automatically detecting text in complex color images was reported by Zhong et al. [17]. The authors proposed two methods for detecting the text portion: (1) By segmenting the image into connected components with homogeneous color or gray value, and then applying several heuristics, the area which is likely to contain text are identified. This method cannot identify characters with blurred boundaries and touching characters such as cursive printed text. (2) The second method was devised for the horizontally aligned handwritten text by locating the area with high variance after the computation of the local spatial variation in the gray-scale image. A combination of the above two methods was reported to perform better on detecting text on a set of test image from a video frame, CD and book covers.

Epshtein et al. [7] proposed a method called Stroke Width Transform (SWT) for text detection in natural images. The unique feature of text i.e. uniformly thick strokes separates it from other elements in an image. However, the method does not require a letter to have a constant stroke width. After the implementation of Canny edge detection [1] on the input image, it computes the width of each stroke that forms an object in the image. With F-measure of 0.66, the method was reported to perform better than other previously proposed methods while experimenting on the database of ICDAR 2003 and 2005.

A text detection method on natural scene images using neural network was proposed by Zhou et al. [18]. The model consists of two stages, initially, a fully convolutional network is used to predict a word or text-line. Then, the predicted output is fed into a Non-Maximum Suppression to produce the final output. Although the model is reported to get an F-score of 0.78 on the ICDAR 2015 [9] dataset, it has some limitations on its capability to detect long text lines and vertically oriented texts.

A work on Meetei Mayek script text detection is reported by Devi et al. [6]. MSER (Maximally Stable Extremal Regions) [13] features were used to detect the text area appearing in an image, however, because of its poor performance in a blurred image, Edge Enhanced MSER [4] was applied to overcome the drawbacks of MSER. To exclude the non-text element in the output of Edge Enhanced MSER, geometric and SWT filtering was used. The system is reported to achieved F-measure of 0.69.

3 Architecture

We employ two models wherein the first model, we use MSER for feature extraction and used the extracted features for finding the text segment in the form of boxes using SWT. In the second model, EAST [18] is used for text detection. The results obtained from both the model is then fed into the OCR system for text recognition. Figure 1 illustrates the working of our models. The following section describes the approach of our models.

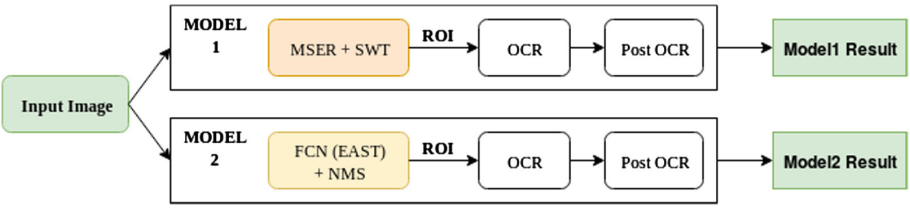


Fig. 1. Framework of our model.

3.1 MSER+ SWT

Maximally Stable Extremal Regions (MSER). One of the blob (a portion of an image with some similar properties) detection method, MSER, proposed by Matas et al. [13], is used as a feature detection technique in our experiment. The MSER extracts a set of co-variant extremal regions from an image by segmenting the image into a number of portions that are maximally stable. This feature tends to identify blobs that segregate themselves from their surroundings.

Stroke Width Transform (SWT). SWT, devised by Epshtein et al. [7], tries to detect strokes inside an image by computing per pixel the width of the most likely stroke containing the pixel. Stroke is defined as a contiguous element of an image with finite width bounded by two roughly parallel edges forming a constant width band. The transformation begins by detecting high contrast edges appearing inside the image. Traversing on each edge pixel in the orthogonal direction, the method attempt to detect a parallel edge that indicates a stroke. Each complete stroke is identified by the connecting adjacent cross-sections with a similar width.

Steps for First Model (Model1): The following steps are carried out in our first model:

1. The input image is first converted into a grayscale image (grayImg).
2. MSER is applied on the grayImg for feature extraction.
3. Apply Canny edge detection in the MSER regions and used it as input to SWT to get the final output: region of interest (ROI) in the form of a box.

3.2 EAST

EAST, an Efficient and Accuracy Scene Text, proposed by Zhou et al. [18] is a text detection technique build using a deep learning model.

Steps for Second Model (Model2): The following steps are carried out in our second model:

1. Resize the input image into a multiple of 32.
2. The resized image is then fed to the pre-trained model of Fully Convolutional Network (FCN) implementation of EAST.
3. Apply non-maxima suppression to suppress the weak and overlapping bounding boxes to get the final ROI.

The ROI obtained are the coordinates of bounding boxes in the form of a rectangle, detected as a region containing the text by the model.

The above two models are implemented using OpenCV (Open Source Computer Vision Library)[1], Python and its libraries. For the implementation of text detection, in Model1, the hyperparameters described in [15] for geometric and SWT elimination is used and for Model2, a pre-trained EAST model[2] is used.

3.3 OCR and Post-processing

OCR allows us to extract the recognized text from a region in the form of a machine-readable Unicode format. After getting the ROI from the above model, further steps of pre-processing are carried as follows:

1. Add padding on the surrounding of ROI by computing the dimensions of ROI.
2. Convert the region obtained from step 1 to grayscale format.
3. The grayscale image is then converted to a binary image format using Otsu's binarization method [14].
4. The output from step 3 is then resized by increasing its dimensions which is the final input to the OCR module.

[1] https://opencv.org/.
[2] https://github.com/argman/EAST.

The OCR is implemented by using an open-source tool: Tesseract v4's LSTM deep learning text recognition algorithm[3]. After getting the list of characters or word(s) from the OCR module, the text obtained is again subjected to a spelling correction process. The spelling correction is implemented using Symspell[4], a language-dependent spell correction module. The post OCR processing, spelling correction module, requires a dictionary of tokens along with its frequency generated from a corpus. After the collection of the corpus, special symbols are removed by tokenizing to build the word frequency dictionary module. For the word frequency dictionary, we have collected the dataset from news articles and a module to generate word frequency dictionary using Java is developed inhouse.

4 Experimental Result and Discussion

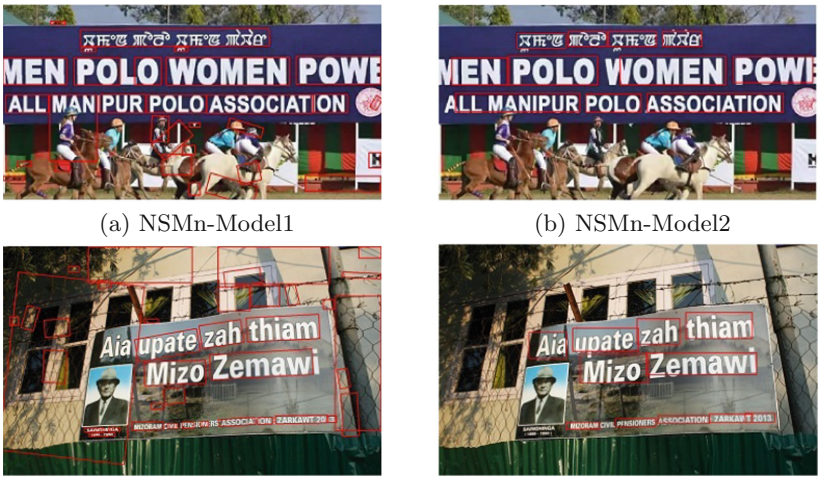
The proposed models are employed on a minimum number of input images: natural scene image[5] and images captured from the textbook, containing the text in Manipuri and Mizo separately. To evaluate our model, the ROI containing the Manipuri text and Mizo text on the input images are identified manually for our reference as ground truth. In Model1, we observed that certain non-text areas were also detected as the ROI. The Model2 is able to capture the ROI with better precision than the Model1. Sample outputs for each of the image containing text in Manipuri and Mizo are shown in Fig. 2, natural scene image in Manipuri (NSMn) and Mizo (NSMz), and Fig. 3, document image in Manipuri (DCMn) and Mizo (DCMz). Figures 2a and b shows the ROI detected by Model1 and Model2 in NSMn, while Fig. 2c and d shows the ROI detected by Model1 and Model2 in NSMz respectively. For the document image output, Fig. 3a and b shows the ROI detected by Model1 and Model2 in DCMn, while Fig. 3c and d shows the ROI detected by Model1 and Model2 in DCMz respectively.

From Figs. 2 and 3, it is observed that Model1 detects the image segment that does not contain text as the ROI. While the ROI detected by Model2 is the image segment where the text appears. To evaluate our model, the measurement system in [12] is used to compute the precision(P) and recall(R). F-score is calculated as the harmonic mean of P and R as follows:

$$F - score = \frac{2 \times P \times R}{P + R} \tag{1}$$

Table 1 shows the summarizes evaluation of our two models for text detection in the sample images. In Table 1, M1P, M1R, and M1F represent the precision, recall and F-score of Model1 respectively, and M2P, M2R, and M2F represent the precision, recall and F-score of Model2 respectively. The values in Table 1 are in terms of percentage. From Table 1, it can be observed that Model2 achieve better precision and F-score than the Model1.

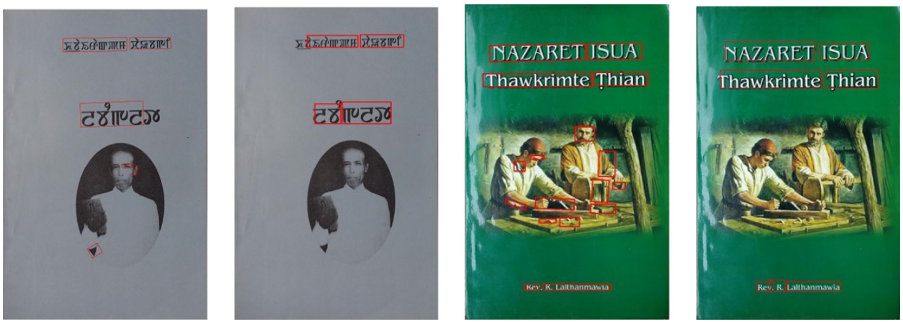
[3] https://github.com/tesseract-ocr/tesseract.
[4] https://github.com/wolfgarbe/SymSpell.
[5] Natural scene image source: http://e-pao.net; https://iias.asia/.



(a) NSMn-Model1 (b) NSMn-Model2

(c) NSMz-Model1 (d) NSMz-Model2

Fig. 2. ROI detected by Model1 and Model2 in the sample natural scene images



(a) DCMn-Model1 (b) DCMn-Model2 (c) DCMz-Model1 (d) DCMz-Model2

Fig. 3. ROI detected by Model1 and Model2 in the sample document images

Table 1. Evaluation of Model1 and Model2 on the sample input images

Image Ids	M1P	M1R	M1F	M2P	M2R	M2F
NSMn	38.71	100	**55.8**	100	91.67	**95.6**
NSMz	36.8	100	**53.8**	100	78.5	**87.9**
DCMn	60	100	**75**	75	100	**85.7**
DCMz	35	85	**49**	85	85	**85**

After the completion of the text detection task, the ROI obtained is subjected to the OCR module. To get a better result in OCR, we added padding on the bounding box of ROI, especially on the height of the ROI detected. We performed the OCR for the Meetei Mayek script of Manipuri (NSMn and DCMn) and the Roman scripts of Mizo (NSMz and DCMz). The OCR result using the ROI detected from Model1 contains a lot of noisy character(s) as compared to the one from Model2 because of the detection of the non-text image segment as the final ROI. After the application of OCR, a post OCR processing of spelling correction is carried out. Spelling correction module requires a word frequency dictionary for each of the languages. For building the word frequency dictionary of Manipuri, a dataset of 93000 words with 18500 unique word forms is collected from a local newspaper[6] and for the Mizo, we collected the dataset from the Mizoram local newspaper[7] from April 2013 to February 2019. The Mizo text corpus consists of 21 million words with 191736 unique word forms. As the text in images is less likely to contain any special characters, any such characters present in the OCR result is removed before applying the spelling correction. The OCR result of the ROI detected by the Model2 in Figs. 2 and 3 and the post OCR processing of spelling correction is shown in Table 2. The ground truth of the above sample is given below:

Table 2. OCR and spelling correction result of sample images.

Image Ids	OCR result	Spelling correction result
NSMn	[illegible]	[illegible]
NSMz	;Âm upate zah thiam Mizo. Zemawli MIZORAM PENSIONERS ASSOCIATION : ZARKAWT — T2013	*em* upate zah thiam Mizo *Zemawi* MIZORAM PENSIONERS ASSOCIATION ZARKAWT *2013*
DCMn	[illegible]	[illegible]
DCMz	NAZARET JSUA Thawkrimte Thian Rev. /. R. Lalthanmawia	NAZARET *ISUA* Thawkrimte Thian Rev R Lalthanmawia

1. NSMn: [illegible]
2. NSMz: Aia upate zah thiam Mizo Zemawi
   MIZORAM CIVIL PENSIONERS ASSOCIATION ZARKAWT 2013
3. DCMn: [illegible]
4. DCMz: NAZARET ISUA Thawkrimte Ṭhian Rev R Lalthanmawia

The final results obtained in our experiment are close to the ground truth except for the case of document image with text in Meetei Mayek (DCMn).

[6] http://hueiyenlanpao.com/.
[7] https://www.vanglaini.org/.

5 Conclusion and Future Work

In our work, we have created a prototype model for detecting and identifying the text in Manipuri and Mizo. We experimented with two types of models on different images (natural scene and document). The second model, EAST detect the text with better precision in all the cases. The techniques used in both the languages show that the performance of these systems doesn't depend on the language family but it's largely on the script used to represent them. The grapheme complexity of Meetei Mayek font makes it harder to achieve a quality result. The detection of the text for both languages are close in terms of F-score. However, the OCR made a difference in the overall performance. The spelling correction module of both languages is found to be effective in our experiment. In the future, work on enhancing the OCR module for better result and handling of mixed language texts of different scripts can be carried out. Also, increasing the corpus size of both languages will enhance the performance of the spelling correction module.

Acknowledgments. This work is supported by Scheme for Promotion of Academic and Research Collaboration (SPARC) Project Code: P995 of No: SPARC/2018-2019/119/SL(IN) under MHRD, Govt of India.

References

1. Canny, J.: A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. PAMI **8**(6), 679–698 (1986)
2. Cano, J., Pérez-Cortés, J.-C.: Vehicle license plate segmentation in natural images. In: Perales, F.J., Campilho, A.J.C., de la Blanca, N.P., Sanfeliu, A. (eds.) IbPRIA 2003. LNCS, vol. 2652, pp. 142–149. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-44871-6_17
3. Chen, D.M., Tsai, S.S., Girod, B., Hsu, C.H., Kim, K.H., Singh, J.P.: Building book inventories using smartphones. In: Proceedings of the 18th ACM international conference on Multimedia, pp. 651–654. ACM (2010)
4. Chen, H., Tsai, S.S., Schroth, G., Chen, D.M., Grzeszczuk, R., Girod, B.: Robust text detection in natural images with edge-enhanced maximally stable extremal regions. In: 2011 18th IEEE International Conference on Image Processing, pp. 2609–2612. IEEE (2011)
5. Chen, X., Yuille, A.L.: Detecting and reading text in natural scenes. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004, CVPR 2004, vol. 2, pp. II–II. IEEE (2004)
6. Devi, C.N., Devi, H.M., Das, D.: Text detection from natural scene images for manipuri meetei mayek script. In: 2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS), pp. 248–251. IEEE (2015)
7. Epshtein, B., Ofek, E., Wexler, Y.: Detecting text in natural scenes with stroke width transform. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2963–2970. IEEE (2010)

8. Ezaki, N., Bulacu, M., Schomaker, L.: Text detection from natural scene images: towards a system for visually impaired persons. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004, ICPR 2004, vol. 2, pp. 683–686. IEEE (2004)
9. Karatzas, D., et al.: ICDAR 2015 competition on robust reading. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 1156–1160. IEEE (2015)
10. Kim, K.I., Jung, K., Kim, J.H.: Color Texture-based object detection: an application to license plate localization. In: Lee, S.-W., Verri, A. (eds.) SVM 2002. LNCS, vol. 2388, pp. 293–309. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45665-1_23
11. Kim, S.K., Kim, D.W., Kim, H.J.: A recognition of vehicle license plate using a genetic algorithm based segmentation. In: Proceedings of 3rd IEEE International Conference on Image Processing, vol. 2, pp. 661–664. IEEE (1996)
12. Lucas, S.M., Panaretos, A., Sosa, L., Tang, A., Wong, S., Young, R.: ICDAR 2003 robust reading competitions. In: Seventh International Conference on Document Analysis and Recognition, 2003, Proceedings, pp. 682–687. Citeseer (2003)
13. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions. Image Vis. Comput. **22**(10), 761–767 (2004)
14. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
15. Özgen, A.C., Fasounaki, M., Ekenel, H.K.: Text detection in natural and computer-generated images. In: 2018 26th Signal Processing and Communications Applications Conference (SIU), pp. 1–4. IEEE (2018)
16. Tsai, S.S., et al.: Mobile product recognition. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 1587–1590. ACM (2010)
17. Zhong, Y., Karu, K., Jain, A.K.: Locating text in complex color images. Pattern Recogn. **28**(10), 1523–1535 (1995)
18. Zhou, X., et al.: East: an efficient and accurate scene text detector. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5551–5560 (2017)