Improving Text Translation from Images with MSER Algorithm

Sruthi. N[1] and Kamal Bijlani[2]
Amrita E-Learning Research Lab
Amrita Vishwa Vidyapeetham
Amritapuri, India
[1]sruthi.n0209@yahoo.in
[2]kamal1008@gmail.com

ABSTRACT

Translators play an important role in conveying ideas and thoughts from one language to another. A contextual based approach to translation of English words from images to the equivalent Hindi words is described in this paper. Text segmentation is an integral and critical process in text extraction which have the major control over the accuracy of any translator. The paper describes a study on text segmentation algorithm with Maximally Stable Extremal Regions (MSER) which is implemented as a preprocessing module of the translation system. Initially the system learns through user disambiguation of appropriate words in the sentence. However, once the system has learned enough contextual information, it performs the automatic translation. The experimental results of segmentation are compared and analyzed. Such a system has a number of uses, including e-learning of the English language as well as for Indian tourists from overseas countries. It can be applied for learning any courses. The system can also be extended to include other Indian languages.

KEYWORDS

Text Extraction, Image Processing, Language Translation, Optical Character Recognition, MSER, Text Segmentation.

1 INTRODUCTION

Images are a part of daily life. They are also extensively used in e-learning materials. Images also contain text which is important for learning the course. Usually e-learning materials may be available in different languages, but the text in the images is available in English only. If there is an automatic translation of these words to a language understandable by the user, then learning can be improved. Translating the text in images is also useful for tourists in foreign destinations. Most of the people in the rural areas of India know only Indian languages. Translating the text in images to Indian languages will help them read sign boards while travelling nationally as well as internationally. To facilitate this translation the text in the image should be extracted. Initial step before extraction being segmentation of text from the image.

Automatic segmentation is an important research area in image processing. There are several algorithms for automatic segmentation of images based on pixel intensity values, connected components, etc. Text segmentation is another important area in automatic segmentation. Segmenting the text area from other parts of the image helps in better recognition of the text. [13] states that there are two approaches for text segmentation: pixel-based and region-based segmentation. The algorithm [11] studied in this paper is a region-based segmentation. It uses maximally stable extremal regions (MSER) to find the text regions and edge detection to compensate for the effect of blurring of images on MSER. The algorithm is implemented as a preprocessing step in a system which can extract text from images and translate it to different Indian languages.

India is a country where each state has its own language, though Hindi is the national language and English is used in official proceedings. But the information available on the internet is usually in English. In order for the people who cannot understand English, it will be quite useful to provide automatic translation into their known language. Hence an automatic translator from English to

Indian languages is useful not only for learners but also for tourists.

Although a number of works and products have been produced for different pairs of natural language, we seek to develop fast translation of words in images using the context of the images. We also present a study of the text segmentation algorithm which is the preprocessing module for the translation system. The main purpose of the system is to help people, hence the option for customization to make the system more user friendly. The system aims to provide a fast translation customized according to the context of the image. The paper proceeds to give an overview of work done in the field of image segmentation and translation in section II. It is followed by a detailed description of the system in section III. Section IV gives the results obtained after implementing the system. The last section, section V gives a conclusion with future extension of the system.

2 RELATED WORKS

Segmentation is one of the most challenging problems in the field of image processing. There many algorithms in pixel-based and region-based segmentation. Thresholding techniques are one of the most commonly used techniques for text segmentation. But in case of blurry images, it is difficult to find a threshold which can be used to segment the text.

One of the latest algorithms developed by Xu-Cheng Yin et.al. in [12] segments text in multi-orientation. It is a vast improvement over segmenting text only in the horizontal direction. The method uses a new concept distance-metric learning with morphology-based clustering, orientation-based clustering and projection-based clustering one after the other to find multi-orientation text. One of their earlier papers [14] is a special case of this algorithm. But the algorithm needs initial training.

There are different methods of text segmentation enumerated in [10]. But the most important advantage of the algorithm using MSER by Huizhong Chen et.al. [11] is that it is simple and efficient. This method is also associated with thresholding.

Work in the field of translation has been going on in India for some time. Main institutions like IITs have been developing systems to translate English to Indian languages. Vishal Goyal and Gurpreet Singh Lehel of Punjabi University have developed a system to translate from Hindi to Punjabi [1].The system uses direct machine translation system. The system has additional modules for training the system, freeing the system from specific font dependency, text normalization, replacement of collocations and word by word translation.

Another English to Hindi translator is developed by Pushpak Bhattacharya [1] and team in IIT Bombay. It uses Universal Natural Language (UNL) as Interlingua. The system analyzes English sentences and converts it into UNL form which generates target Hindi sentences using a Hindi generator. Matra is a system developed by CDAC, Pune. All these systems use Transfer based machine translation. Matra system depends on heuristics to resolve ambiguities and requires human effort for analysis of the input. CDAC, Bangalore developed a system called Mantra which translates English to Hindi. It translates domain specific documents like gazette notifications using lexicalized tree adjoining grammar (LTAG).

In 2010, IIT Hyderabad developed English to Hindi translator by combining RBMT and phase-based SMT. Bengali-Assamese translator Vaasaanubaada uses EBMT technique. Translation of bilingual text occurs at the sentence level.

Google translate is a system which translates between some Indian languages and English. It is based on Statistical Machine Translation approach. Bing translator from Microsoft also uses Statistical Machine translation approach [2].

In the field of extraction much work has been done. Optical Character recognition (OCR) is one of the best techniques for text extraction from images. Using OCR text can be extracted from almost all types of images.Work done by Pawar Pooja, Rashmi Phalak, Waghmare Jayashri, and Shinde Yugandhara of K.K Wagh College of Engineering, Nashik [3] adds a preprocessing stage before giving the image to an OCR system for text extraction. They have done the work on English Comic Images. In the preprocessing stage, there is a block for grayscale conversion, smoothening the image and Connected Component Analysis (CCL).

There are a variety of ways for text extraction from images. Davod et.al [3] used Wavelet

Transform and Region of Interest (ROI) for image extraction. This technique is robust to noise. S. Audithan et.al [3] proposed the use of Haar discrete wavelet transform, Morphological dilation operators, a Canny edge detector for extracting text. The technique is independent of contrast. Zhan et.al [3] developed a technique which used Multiscale wavelet features, SVM classifier, Cubic interpolation, Gaussian filter, and K-means clustering algorithm. This technique is Robust to text color, font size, and languages.

There is an app which translates words from images in real-time to other languages. This app is called WordLens [9].WordLens was initially owned by Quest Visual and now acquired by Google. Its primary developer is Otavio Good. It translates from English to six other languages and vice-versa. It captures the images, recognizes the text and does the translation. An extension which can be added to the app is translating meaningful sentences and translating to Indian languages.

3 SOLUTION APPROACH

The segmentation algorithm is implemented as the preprocessing module of the system which extracts text from image. The proposed system is based on text extraction using OCR and the translation of that text to Indian languages as shown in Figure 1. First, input image passes through a preprocessing stage in which the text is segmented from the image. Then English text is identified from the image followed by extraction using Optical Character Recognition (OCR). OCR is a technique used to extract text from images and convert it to computer readable format. Tesseract is one of engines which use the OCR technique to extract text from the image. The text is extracted using Tesseract engine. In the last phase, the extracted text which is in English is converted to an Indian language using Google API and is displayed.

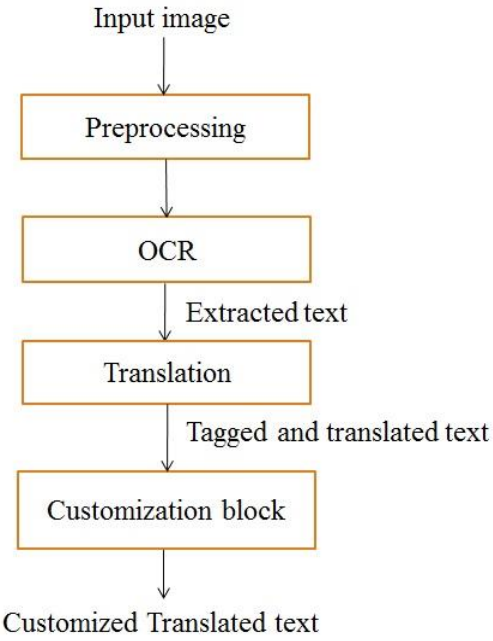

Figure 1. Block Diagram of Extraction and Translation system

3.1 Input Image

Input image can be any type. It can be JPEG, bmp, png, etc. But the text in the image should be printed text. Tesseract OCR can extract text from all the above image types.

3.2 Preprocessing Stage

In the preprocessing stage, the image is segmented using MSER algorithm. This is done so that the image is much more enhanced for the OCR. The algorithm proposed in [15] uses MSER to detect text regions. Canny edge detection is used to remove the effect of blurring on the images. Edge filtered image is then subjected to filtering again using eccentricity, solidity and area threshold. Again the image is transformed using stroke width transformation. Those text candidate with wide variation in stroke width are eliminated. Finally the remaining text candidates are segmented. The output is a cropped version of the original image with only the text region.

Maximally Stable Extremal Region (MSER) is a concept introduced by J. Matas et. Al. in their paper [15]. They formally define it as shown in Figure 2 [15]. MSER has many desirable properties which make it ideal for text segmentation. They are:

- Invariance to affine transformations
- Stability

Image I is a mapping $I : \mathcal{D} \subset \mathbb{Z}^2 \rightarrow \mathcal{S}$. Extremal regions are well defined on images if:

1. $\mathcal{S}$ is totally ordered, i.e. reflexive, antisymmetric and transitive binary relation $\leq$ exists. In this paper only $S = \{0, 1, \ldots, 255\}$ is considered, but extremal regions can be defined on e.g. real-valued images ($\mathcal{S} = R$).
2. An adjacency (neighbourhood) relation $A \subset \mathcal{D} \times \mathcal{D}$ is defined. In this paper 4-neighbourhoods are used, i.e. $p, q \in \mathcal{D}$ are adjacent (pAq) iff $\sum_{i=1}^{d} |p_i - q_i| \leq 1$.

Region $\mathcal{Q}$ is a contiguous subset of $\mathcal{D}$, i.e. for each $p, q \in \mathcal{Q}$ there is a sequence $p, a_1, a_2, \ldots, a_n, q$ and $pAa_1, a_iAa_{i+1}, a_nAq$.

(Outer) Region Boundary $\partial\mathcal{Q} = \{q \in \mathcal{D} \setminus \mathcal{Q} : \exists p \in \mathcal{Q} : qAp\}$, i.e. the boundary $\partial\mathcal{Q}$ of $\mathcal{Q}$ is the set of pixels being adjacent to at least one pixel of $\mathcal{Q}$ but not belonging to $\mathcal{Q}$.

Extremal Region $\mathcal{Q} \subset D$ is a region such that for all $p \in \mathcal{Q}, q \in \partial\mathcal{Q} : I(p) > I(q)$ (maximum intensity region) or $I(p) < I(q)$ (minimum intensity region).

Maximally Stable Extremal Region (MSER). Let $\mathcal{Q}_1, \ldots, \mathcal{Q}_{i-1}, \mathcal{Q}_i, \ldots$ be a sequence of nested extremal regions, i.e. $\mathcal{Q}_i \subset \mathcal{Q}_{i+1}$. Extremal region $\mathcal{Q}_{i^*}$ is maximally stable iff $q(i) = |\mathcal{Q}_{i+\Delta} \setminus \mathcal{Q}_{i-\Delta}|/|\mathcal{Q}_i|$ has a local minimum at i^* ($|.|$ denotes cardinality). $\Delta \in \mathcal{S}$ is a parameter of the method.

Figure 2. Definition of MSER

- Multi-scale detection [15]

MSER regions in an image are essentially the connected components in an image. But a single threshold is not considered for finding these regions. A wide range of thresholds are tried and stable regions are found across these thresholds.

The MSER algorithm proposed in [15] is implemented excluding the stroke width transformation. In this algorithm, the image is initially converted to a grayscale image to find the MSER. It is followed by Canny edge detection on the original image. The mask formed from edge-detection is used to filter the MSER regions. Then it is again filtered using the parameters solidity, area threshold and eccentricity. This gives the candidate text regions. These regions are cropped off. This is shown in the block diagram in Figure 3.

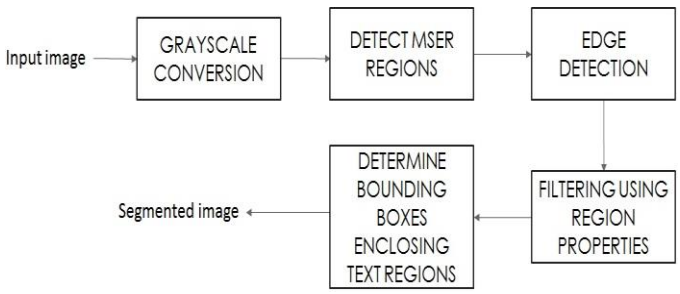

Figure 3. Block diagram of MSER algorithm.

3.3 Extraction Stage

This is the stage in which the text is extracted from the image. Tesseract OCR is used here. It is an open source OCR engine available for Windows, Mac and Linux. It only extracts simple one column text. There are several steps involved in text extraction. Initially Adaptive Thresholding is performed on the image. In this step, the image is converted to binary image by dynamically calculating the threshold for the conversion. Connected Component Analysis is done on the binary image to get the outline of the components. The end result of Connected Component Analysis is groups of words separated by spaces and fuzzy spaces. Then it goes through two passes in which Tesseract recognizes the words. In the first pass recognized words are given to Adaptive Classifier as training data. Words not recognized in the first pass are recognized in the second pass [8].

3.4 Translation Stage

Google Translate API is used to translate the extracted text to Hindi. Google API supports 64 different languages among which Bengali, Gujarati, Hindi, Malayalam and Tamil are the Indian languages supported. It can translate English to Hindi preserving the meaning of the original English sentence. Unlike the Matra system which

uses the transfer based approach, Google Translate uses statistical machine translation. It finds patterns among millions of documents to give the best translation. Google translate also provides translation of an entire web page. In this system Google API is used to translate English into 10 Indian languages- Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, Telgu and Urdu. In the initial phase only translation to Hindi is implemented.

3.5 Customization

The text translated from English to Hindi is then customized according to the context. Tagging is a part of the customization module. The text is tagged before translation. Only those nouns which have a probability above a certain threshold are tagged. The probability of a word to be a noun is found out using OpenNLP tool. The threshold is dynamically determined by the system for each image. As a part of the customization module, a dictionary is maintained in which different words and their synonyms are stored as categories. The categories are formal, informal, movies, weather and sports. The tagged words are compared with words in the dictionary and their synonyms are found. All these words are displayed as suggestions. An option is also given for the users to choose anyone of the above mentioned categories. The users can choose the suitable word and category and the words are replaced with users' choice. User's word choice and category choice are noted in the dictionary and used to train the system.

The initial phase consisted of implementing the extraction, translation and customization modules. The system successfully extracts simple text from grayscale images without preserving its format. In the initial phase only one translation, English to Hindi, is implemented. Image segmentation was implemented in the second phase. Fig. 6 shows the user interface for the system. The user has an option to choose a category from the given list of categories. Categories include formal, informal, movies, sports, education and food. These are the most common categories we encounter in the field of education and tourism. More categories can be added if required. User can input an image file and get the translation along with the suggestions (synonyms) for nouns in the sentence. Common image formats like JPEG, png, tiff, and bmp can be given as input. In the user interface in Fig. 6, the translated text is shown in the space provided below the title 'Translated text'. The suggestions can be found below 'Alternatives'. In 'Alternatives', the system shows the word for which the suggestions are given, the suggestion with the highest user count and all alternatives of the word.

A test input to the system without preprocessing module for the initial phase is shown in Fig. 4. It is a grayscale image.

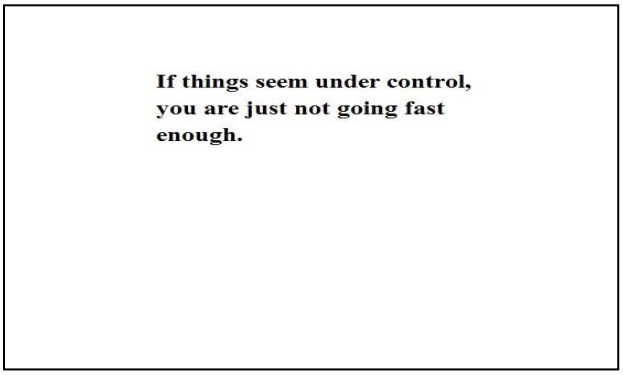

Figure 4. Input image without going through preprocessing

The text from this image was extracted, but the format and style of the text were not preserved. It was then translated to Hindi by the system.

Extracted text: If things seem under control, you are just not going fast enough.

Translated text: बातें को नियंत्रण में लग रहे हैं, तो आप सिर्फ काफी तेजी से नहीं जा रहे हैं।

Figure 6 shows the final output of the system for the test input of Figure 4. In the above example, alternatives for the word "बातें" are also suggested. If the given translation is not appropriate, the user can choose from the suggested alternatives, there by the system will learn the context. If a user chooses an alternative, the system records it for the given context. The next time when the same word comes for the same context, the system shows the 'Most Used' word in place of the original translated word.

The input to the image segmentation module is shown in Figure 5. The segmentation algorithm takes the image in Figure 5 as input and segments the text from the image. The output of the segmentation algorithm is shown in Figure 7.

Figure 5. Test input to the preprocessing module

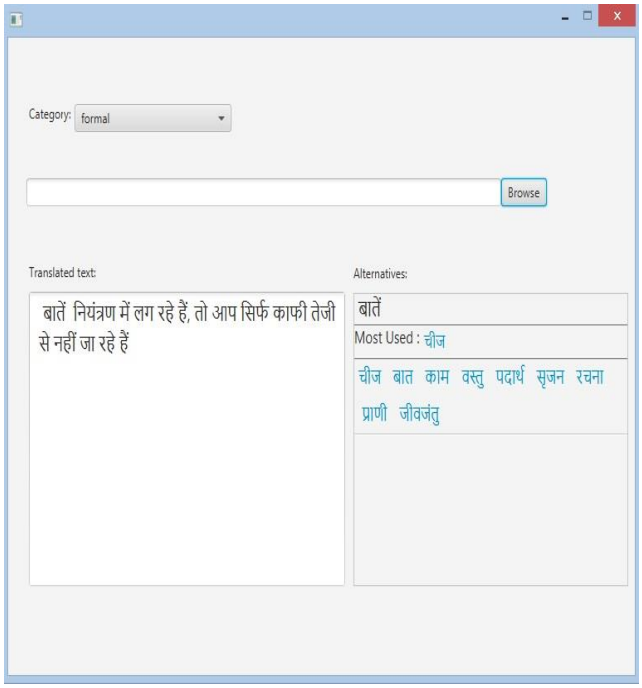

Figure 6. Final output of the system

Figure 7. Segmented image

4 EXPERIMENTS

The dataset used for the experiment is the ICDAR 2003 dataset. 40 images were chosen for the initial experiment. The images were of signs we commonly find in streets like sign boards of shop, street signs, etc. The segmentation algorithm has different parameters like eccentricity, solidity, and area threshold. These different parameters help in identifying the text region and avoid background regions. The experiment is conducted by varying eccentricity from 0.95 to 1 and the result is recorded. Solidity and area threshold are maintained 0.5 and 5000 respectively.

Another experiment conducted studies the effect of segmentation on extraction of text from the images. The same dataset is used for the experiment. This experiment tests whether including text segmentation as preprocessing stage increases accuracy of text extraction.

5 RESULTS

The study reveals that the value of the parameters set initially gives good segmentation results. The result of the experiment is shown in Table 1. The algorithm gives the best precision and recall rate when eccentricity is 0.995, solidity is 0.5 and area threshold is 5000. The values of eccentricity, solidity and area threshold changes with types of images. Here the images used are shop signs, street signs and other commonly seen signs in the street. The values for different types of images can be found by extensive experimentation using different type of images. The system can extract text from images with text font size 16, bold and above accurately. Below 16, accuracy decreases. The system gives suggestion for those nouns which have a probability above the threshold. A comparative study was conducted between our system and the app WordLens. Figure 8 shows the result of the study.

Table 1. Precision and Recall Rate for Different Values of Eccentricity

Eccentricity	Precision	Recall
0.96	0.725	0.814
0.97	0.7344	0.8176
0.98	0.7197	0.8136
0.99	0.705	0.8375
0.995	0.868	0.752
1	0.667	0.8645

Factors	WordLens	Proposed System
Translation to Indian Languages	✗	✓
Sentence Translation	✗	✓
Customization	✗	✓
Web Dependency	Independent	Dependent
System Format	Mobile app	Desktop
Input	Inplace Image	Image file

Figure 8. Comparison of WordLens and Proposed System

The table in Figure 8 shows that our system has three major advantages. Our system is very useful in the Indian context since it can translate English to Hindi. Another major advantage is that our system can translate the text preserving its meaning. WordLens gives accurate translation of each word, but cannot preserve the meaning of the sentence. Using our system users can customize their experience. WordLens is a mobile app while our system is a desktop app. WordLens can do the text translation in the image itself without taking its digital image while our system needs an image file as input.

But WordLens cannot take image file as input. WordLens mainly focus on travelers while our system can be used in both e-learning as well by travelers when it is developed into a mobile app. Overall our system is better for the Indian context.

Figure 9 shows the graph obtained by calculating the accuracy of the system with different number of words. The graph clearly reveals that the accuracy of the system increases with segmentation. The graph also shows a dip in accuracy when number of words increases. The reason is found to be the poor quality of the image.

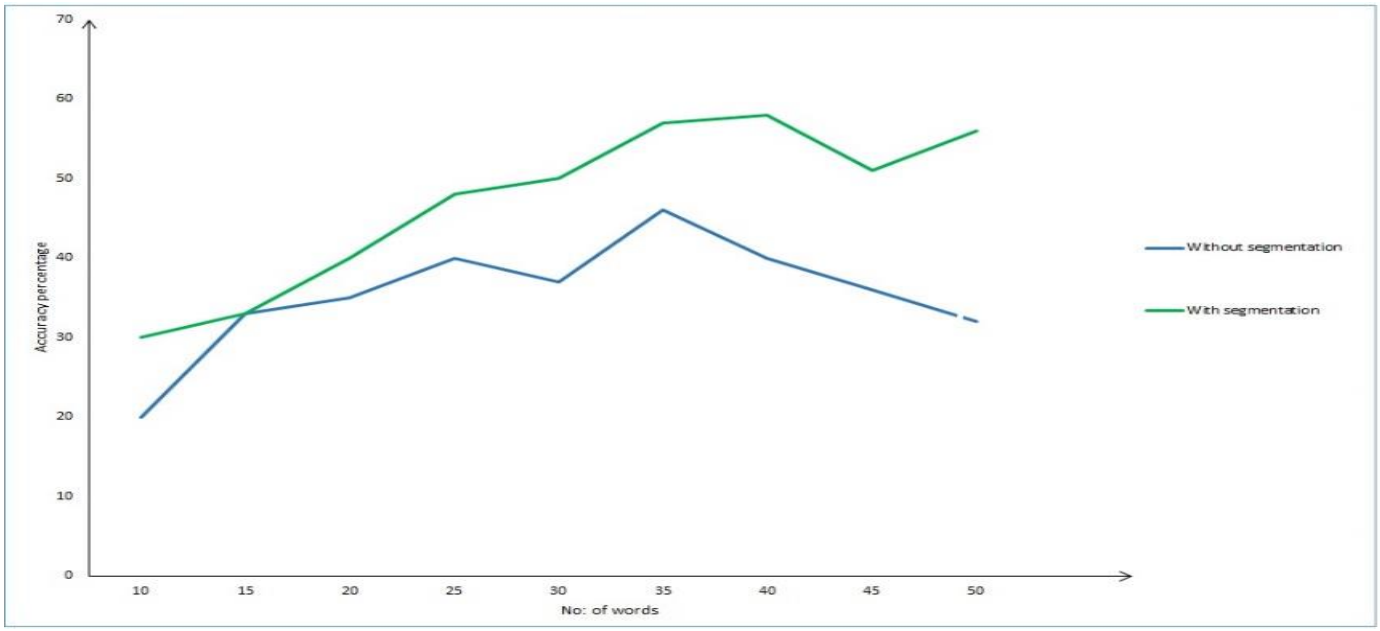

Figure 9. Graph of accuracy of the system with and without segmentation

6 CONCLUSION

The proposed system describes the extraction of English text from images and translates it to Hindi. The system gives an appropriate enhanced translation of text from an image while incorporating text segmentation algorithm. Using MSER segmentation algorithm the system gives a reasonable translation while preserving the meaning of the text. The accuracy of the system to extract words increases with segmentation algorithm included in the preprocessing module. Customization provides a way for people to understand the text in the image even though they may not understand the initial translated text. It can be further extended by including color images also. Future works include evaluation of the system. The Tesseract engine does not recognize handwritten text. It recognizes only printed text. So the system can be extended to recognize handwritten text also. A mobile application can be developed which will be useful for travelers.

ACKNOWLEDGEMENT

We thank Amrita E-Learning Research Lab and Amrita Vishwa Vidyapeetham University for providing the facilities and support necessary for this project.

REFERENCES

[1] Latha R. Nair, and David Peter S., "Machine Translations for Indian Systems," International Journal of Computer Applications, vol. 39– no.1, February 2012.

[2] Annena George, "English to Malayalam Statistical Machine Translation System," International Journal of Engineering Research and Technology (IJERT), vol. 2, no. 7, July 2013.

[3] C.P. Sumathi, T. Santhanam, and G.Gayathri Devi, "A Survey On Various Approaches Of Text Extraction In Images," International Journal of Computer Science & Engineering Survey (IJCSES),vol.3, no.4, pp. 27-42, August 2012.

[4] Pawar Pooja**,** Rashmi Phalak, Waghmare Jayashri, and Shinde Yugandhara, "Text Extraction From English Comic Images Using Connected Component Algorithm," in Proceedings of 4th IRF International Conference, pp. 166-169, 16th March, 2014.

[5] M. ˇCadík, "Perceptual Evaluation of Color-to-Grayscale Image Conversions," Proceedings of Pacific Graphics Computer Graphics Forum, pp. 1745-1754, 2008.

[6] G V Garje, and G K Kharate, "Survey Of Machine Translation Systems InIndia," International Journal on Natural Language Computing (IJNLC),vol. 2, no.4, pp. 47-67, October, 2013.

[7] Ray Smith, "An Overview of the Tesseract OCR Engine," Ninth International Conference onDocument Analysis and Recognition, 2007(ICDAR 2007), vol. 2, pp. 629 - 633, 23-26 September, 2007.

[8] Ravina Mithe, Supriya Indalkar, and Nilam Divekar, "Optical Character Recognition," International Journal of Recent Technology and Engineering (IJRTE), vol. 2, pp. 72-75, March, 2013.

[9] Jordan Golson, "Google picks up incredible visual translation app Word Lens and makes it free," http://www.techrepublic.com/article/google-picks-up-incredible-visual-translation-app-word-lens-and-makes-it-free, May 20,2014.

[10] Qixiang Ye, David Doermann, "Text Detection and Recognition in Imagery: A Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PP, November 3, 2014.

[11] Huizhong Chen, Sam S. Tsai, Georg Schroth, David M. Chen, Radek Grzeszczuk and Bernd Girod, "Robust Text Detection in Natural Images with Edge-Enhanced Maximally Stable Extremal Regions," Image Processing (ICIP), 2011 18th IEEE International Conference, 2011.

[12] Xu-Cheng Yin, Wei-Yi Pei, Jun Zhang Hong-Wei Hao, "Multi-Orientation Scene Text Detection withAdaptive Clustering,"IEEE Transactions On Pattern Analysis And Machine Intelligence, vol. PP, January 1, 2015.

[13] Asha D, Dr.Shivaprakash Koliwad, Jayanth.J, "A Comparative Study of Segmentation in Mixed-Mode Images," International Journal of Computer Applications (0975 – 8887), vol.31, October 2011.

[14] Xu-Cheng Yin, Xuwang Yin, Kaizhu Huang, Hong-Wei Hao, "Robust Text Detection in Natural Scene Images," IEEE Transactions on Pattern Analysis And Machine Intelligence, vol. 36, May, 2014.

[15] J. Matas, O. Chum1, M.Urban, T. Pajdla, "RobustWide Baseline Stereo from Maximally Stable Extremal Regions," 13th British Machine Vision Conference, September 2002.