International Journal of Computer Science and Engineering Communications.
Vol.3, Issue 3, 2015, Page.1062-1068, ISSN: 2347–8586

Robust Text Detection and Voice Conversion

J Diana[1], M Vanitha[2], S Pradeep Kumar[3], B Udaya[4], Saravanan.A[5]
Department of Computer Science and Engineering, Rajalakshmi Institute of Technology, India.
diana.j.2011.cse@ritchennai.edu.in[1], vanitha.m.2011.cse@ritchennai.edu.in[2],
pradeepkumar.s@ritchennai.edu.in[3], udaya.b@ritchennai.edu.in[4]

Abstract

Text detection in natural scene images is significant for many content based image analysis tasks. In this paper an accurate method is used for detecting texts in natural scene images. An effective pruning algorithm and the Tesseract algorithm is designed to extract Maximally Stable Extreme Regions (MSERs) as character candidates. Character candidates are grouped into text candidates by the single link clustering algorithm. Distance weights and clustering threshold are learned automatically by novel self training distance metric learning algorithm. The text candidates corresponding to non-text are identified by a character classifier. Non-text candidates are eliminated by a text classifier. In our system the documents will be scanned as images and after the scanning process, the data from the image is identified by Tesseract algorithm and the text is extracted automatically. The Text To Speech (TTS) Engine will convert the extracted text to voice. The translation will translate the text into user defined language. The text and the images are stored in a database. The stored images can be retrieved from the database for further use.

Keyword: *Pruning algorithm, MSER, Tesseract Algorithm, Leptonica, TTS Engine.*

1. Introduction

TEXT in images is very important for many content based image analysis. Text extraction is used in many mobile based image analysis. Extraction of this information has detection, localization, tracking, extraction, enhancement, and recognition of the text from a given image. Variations of text due to differences in size, style, orientation, and alignment, and complex background create a challenging in text extraction. Text within an image is of particular interest as it is useful for describing the contents of an image, easily extracted compared to other semantic contents and it enables applications such as keyword-based image search. Text detection refers to the determination of the presence of text in a given image. Text localization is the process of determining the location of text in the image and generating bounding boxes around the text. Text tracking reduce the processing time for text localization and to maintain the integrity of position across frames. The text detection stage is indispensable to the overall system for reducing the time complexity.

2. Literature Review

Connected component based method extracts the character candidates by connected component analysis [12] and Hybrid method [8]. Sliding window technique used region based method to search for images in text and to identify the text it used machine learning technique. Scale-space feature extractor is implemented to detect and tract text in digital video [6].The lexicon reduces the word recognizing error by reducing the dependency [9]. By using toggle mapping line segmentation and character and non-character classification SNOOPERTEXT are located. Novel coarse fine algorithm is used to locate text lines by using multiscale wavelet feature [4]. Fast and robust vanishing point detection is presented based on robust detection of text baselines [2]. For video text detection laplacian technique is used [10]. An adaptive binarization and extension algorithm is used to detect and read the text [3]. Region detector detects the text candidate and local binarization extracts the connected components into [5]. Connected component based method is used to extract a character candidate from images. Hybrid method is used. Conditional Random Fields model eliminates the non-character candidates [7]. Maximally Stable Extreme Regions (MSERs) based method categorized the connected components. MSER method have been used in many related works [8]. ICDAR 2011 method is widely used in many Robust reading competition database. The repeating components are eliminated by the MSER Pruning [9]. For text candidates rule based and clustering techniques is used but it takes more time consuming [1]. The number of character candidates are removed with more accuracy. Tesseract is an OCR engine to identify the text. It identifies the multilingual characters [10].

3. Existing System

In the Existing system they used MSER based text detection method [8]. The hierarchical structure of MSER is explored and the MSER pruning algorithm was designed for text detection. The character candidates are reduced by this method with moderate accuracy. Novel self- training distance metric learning algorithm was used. By this, distance and weights of the text are identified. Single link clustering algorithm is used to cluster the character candidates into text candidates. To identify the text candidates, text classifier is used. It eliminates the non-text candidates.

By integrating these methods they build a robust scene text detection method. They detect 76% of text using this method. MSER method still have some difficulties to detect the text in repeating components. The MSER detects the text even in low quality. Connected Component based method and hybrid method is absent in effective text candidate construction algorithm.

3.1 Existing architectural model

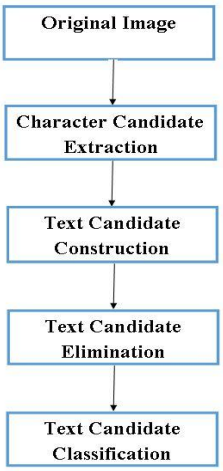

Figure 1. Flowchart of Existing System

Figure 1 represents the existing text detection methods [8]. Text detection method includes the following stages.

1. Character candidate's extraction: Using MSER algorithm character candidates are eliminated and the repeating components are removed by the proposed MSER pruning algorithm by minimizing regularized variation.
2. Text candidate's construction: Novel Self Training Distance Metric Learning algorithm is used to identify the distance and weight of the text. By using proposed metric learning algorithm Clustering threshold are learned simultaneously. By the single-link clustering algorithm character candidates are clustered into text candidates.
3. Text candidate's elimination: Text candidates corresponding to non-texts are eliminated by using the character classifier and non-text probabilities are removed with text candidates.
4. Text candidate's classification: Text classifier identifies the text candidates corresponding to true texts. To test whether the text candidate corresponding to true text or not by using Adaboost classifier.

3.2 Pruning algorithm overview

The disadvantage of MSER algorithm is it does not removes the repeating component. By using the MSER pruning algorithm the MSER tree is pruned by parent children elimination. Let *a* be the aspect ratio and v be the variation of a MSER, the aspect ratios of characters are expected to fall in [*Amin, Amax*], the regularized variation is defined as;

$$V = \begin{cases} V + \theta 1\ (a - amax) \text{ if } a > amax \\ V + \theta 2\ (amin - a) \text{ if } a < amin \\ \text{V otherwise} \end{cases}$$

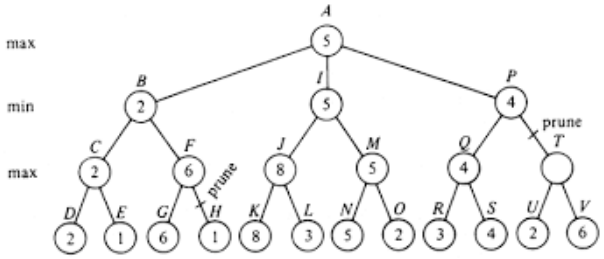

Figure 2 The process of MSER pruning.

Figure 2 shows the MSER pruning tree according to variation. If any variations increases, it prunes the tree by using linear reduction. By calculating the variation it reduces the tree. It identifies the repeating component by constructing a tree [8].

3.3 Linear reduction

The linear reduction algorithm is used where MSERs has only one child. The algorithm chooses the minimum variation and discards the other from parent and child [8]. This process is done recursively to the whole tree. The MSERs tree, returns the root of the processed tree whose linear segments are reduced.

3.4 Text Candidate's Construction

3.4.1 Text Candidate Construction algorithm

By using single-link clustering text candidates are constructed by clustering character candidates. It is suitable for text candidate's construction task. Single link Clustering algorithm belong to hierarchical clustering. Each data is treated as a singleton cluster and the clusters are merged until it reaches the single remaining cluster. The two closest member have the smallest distance are merged in each step in Single link clustering algorithm [8].

3.4.2 Distance Metric Learning Algorithm

Many clustering algorithms depends on a good distance metric. The main task is to learn a distance metric that satisfies the labels and constrains in the supervised data by the given the clustering algorithm [14]. By minimizing distance between points the strategy of metric learning is to learn the distance function pairs in c while maximizing distance between point pairs in *M*, where *C* specifies pairs of points in different clusters and pairs of points in the same cluster is specified by m. clusters are formed by merging smaller clusters, and the final cluster will forms the binary cluster tree.

By the definition of single-link clustering, we must have

$$D(u, v; w) > \_ \text{ for all } (u, v) \in C,$$
$$D(u, v; w) \leq \_ \text{ for all } (u, v) \in M.$$

3.5 Text Candidate's Elimination

ICDAR 2011 training database shows that only 9% of the text candidates correspond to true texts by using the text candidate's construction algorithm. Most of the non-text candidates need to be removed before training. It is hard to train an effective text classifier using an unbalanced database.

4. Proposed Design

In proposed system, an android application is created to extract the text from the scanned image and provide the text to speech conversion for the extracted word. The proposed Tesseract algorithm and leptonica for text extraction. Tesseract is an optical character recognition engine to identify the text. Leptonica performs the following operation Binaries, pixel wise mapping the image, rotating the image, scaling the image and enhancing the image. After extracting the text from the image we perform two operations text to speech conversion and translation. Text to speech conversion will convert the extracted text into voice by using TTS Engine (Text To Speech Engine). Translation will convert the text into user defined language. The text and the image will stored in the database. We can retrieve the text from the database. The blurred text with low resolution can also be processed through this method. No online facility is needed to scan the image.

4.1 Tesseract algorithm

Tesseract is an open-source OCR engine that was developed by HP. Figure 4.1.1 shows the Data Flow Diagram of Tesseract algorithm [11]. Tesseract assumes the input is a binary image with optional polygonal text regions defined. The outlines of the components are stored by connected component analysis. The number of child and grandchild outlines is simple to detect and recognize it as black-on-white text. Outlines are gathered together and nesting into blobs. Blobs are organized into text lines. The lines and regions are analyzed for fixed pitch or proportional text. Fixed pitch text is chopped immediately into character cells. Proportional text is broken into words using definite spaces and fuzzy space. Recognition is done by two-pass process. In the first pass, an attempt is made to recognize each word in turn. Each word is send to an adaptive classifier as trained data. In the second pass the words which are not recognized well are identified again. Tesseract chops the words into characters using the pitch and disables the chopper for word recognition [11].

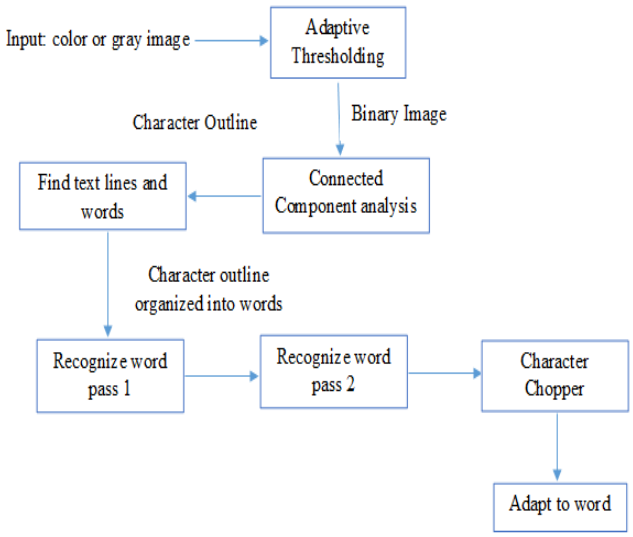

Figure 3. Data Flow Diagram of Tesseract algorithm.

4.2 Proposed Architecture

The object with text is captured by using camera and then the text is extracted from the image by the following phases.

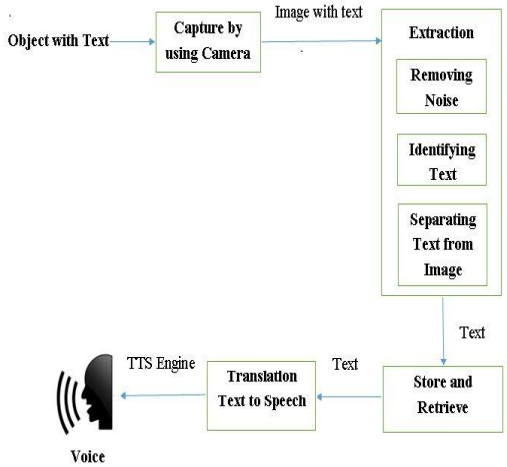

Figure 4. Architecture of proposed system

1. Removing Noise: To remove the noise from the image the first approach is to smoothen the image. It is done by replacing each pixel by the average of the neighboring pixel value. Sharp boundaries that make up the letters have been smeared due to averaging.
2. Identify the text: Rotate the image about 90 degree and perform the bitmapping between the original image and the rotated image. If any differences found then extract the text. This bitmapping operation will be repeated until it identifies the text by rotating the image about 180 degree and 270 degree and so on.

3. Separating the text from image: The text is separated from the image by using text classifier. The text classifier will eliminate the non-text candidate from the text candidate. By giving the training data, the text is compared with the trained data. If any match found it will identify as text and separate the text from image. After extracting the text from the image the text will be stored and retrieved from the database. The TTS Engine will convert the text to speech. As a result we will get the voice as output.

4.3 TTS Engine

Text to speech synthesis is converting the text to the synthetic speech it is close to real speech as possible according to the pronunciation norms of special language. Such systems are called text to speech (TTS) systems. Figure 4.3.1 shows the Data Flow Diagram of the TTS Engine [12]. First the text Analysis part will analyse the input text and organize into manageable list of words. It consists of numbers, abbreviations, acronyms and idiomatic and transforms them into full text. Second, the text normalization will transform the text to the pronunciation form. The four main phases of text normalization are number converter, abbreviation converter, acronym converter, word segmentation.

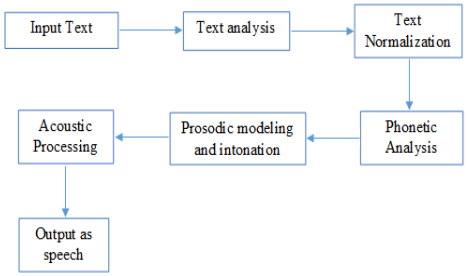

Figure 5. Data Flow Diagram of TTS Engine

Third, the Phonetic Analysis will converts the orthographical symbols into phonological ones using a phonetic alphabet. It is known as “grapheme-to-phoneme” conversion. Phone means sound it defines the sound wave. Fourth, the prosodic modeling describes the speaker’s emotion. Intonation is simply a variation of speech while speaking [12]. In Acoustic processing the speech will be spoken according to the voice characteristics of a person. There are three type of Acoustic synthesis available (i).Concatenate Synthesis (ii).Formant Synthesis (iii).Articulatory Synthesis. The concatenation of prerecorded human voice is called concatenate synthesis, in this process a database is needed having all the prerecorded words. Formant-synthesized speech is constantly intelligible. It does not have any database of speech samples. Articulatory synthesis techniques for synthesizing speech based on models of the human vocal tract are to be developed. Finally, we get the speech as output.

5. Results and Discussion

For implementation, Eclipse Android Developer Tool is used. The coding is implemented by Java language. SQLite Database is used to store the data. For mobile specification 5 megapixel clarity camera is used.

5.1 Registration Module

Figure 5.1 Registration Module

Fig 5.1 represents the Registration Module. For registration, the user has to provide their information by giving User name, email address, password and confirm password.

5.2 Login Module

Figure 5.2 Login Module

Fig 5.2 represents the login module. The user has to login to this application by giving username and password.

5.3 Capturing Module

Fig 5.3 Menu Module

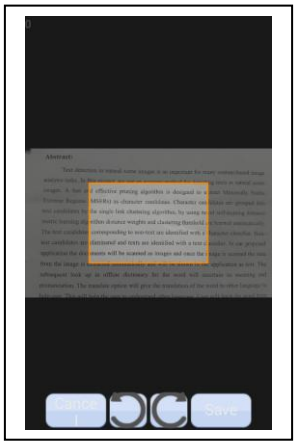

Figure 5.3.1 Capturing Module

Figure 5.3 represents the Menu Module. After login to this application the main menu will display. The user has to capture the image by using camera. Figure 5.3.1 represents the capturing module. After capturing the image the rectangular tool box is created to select the area to be extracted.

5.4 Extracting Module

Figure 5.4 Preview of Extraction

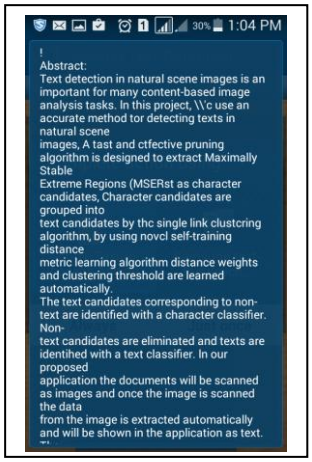

Figure 5.4.1 Extracting Module

Fig 5.4 represents the preview page for extraction. The captured image is displayed in the main screen to further process. Figure 5.4.1 represents the extracting module. The text is extracted by using Pruning and Tesseract algorithm. The extracted text can be editable.

5.5 Text To Speech Conversion Module

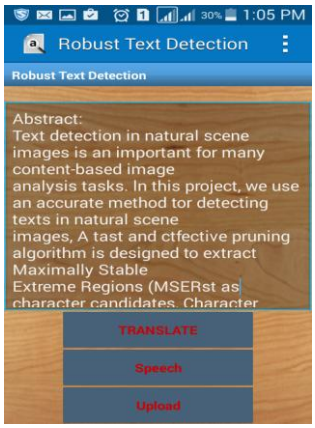

Figure 5.5 Text To Speech Module

Figure 5.5 represents the Text To Speech Conversion Module. The extracted text is translated to user defined language and it gives output as voice.

5. Conclusion

In this paper Tesseract Algorithm is used to extract the text. We proposed a Text To Speech conversion to help the visionless people. This conversion is done by using TTS Engine. By integrating the Maximally Stable Extreme Region (MSER) Pruning algorithm with Tesseract algorithm the best result has been achieved. The work has achieved 90 percentage for the text extraction. Several limitations for further research are empirically analyzed. The special symbols like bullets are not extracted properly. Detecting a multilingual text simultaneously is difficult.

References

[1] A hybrid approach to detect and localize texts in natural scene images in IEEE Trans. Image Process., vol. 20, no. 3, pp. 800–813, Mar. 2011.
[2] Ad boost for text detection in natural scene in *Proc. ICDAR*, Beijing, China, 2011, pp. 429–434.
[3] An Overview of the Tesseract OCR Engine Ray Smith Google Inc. theraysmith@gmail.com
[4] Automatic caption localization in compressed video. IEEE Trans. Pattern Anal. Mach. Intel, vol. 22, no. 4, pp. 385–392, Apr. 2000.982 IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 5, may 2014.
[5] Conditional random fields: Probabilistic models for segmenting and labeling sequence data in Proc. Int. Conf. Mach. Learn., San Francisco, CA, USA, 2001, pp. 282–289.
[6] Detecting and reading text in natural scenes in Proc. IEEE Conf. CVPR, vol. 2. Washington, DC, USA, 2004, pp. 366–373.
[7] Localizing text in scene images by boundary clustering, stroke segmentation, and string fragment classification in IEEE Trans. Image Process., vol. 21, no. 9, pp. 4256–4268, Sept. 2012.
[8] Robust Text Detection in Natural Scene Images
Xu-Cheng Yin, Member, IEEE, Xuwang Yin, Kaizhu Huang, and Hong-Wei Hao.
[9] Scene text recognition using similarity and a lexicon with sparse belief propagation. IEEE Trans. Pattern Anal. Mach. Intel., vol. 31, no. 10, pp. 1733–1746, Oct. 2009.
[10] Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm in IEEE Trans. Pattern Anal. Mach. Intel., vol. 25, no. 12, pp. 1631–1639, Dec. 2003.
[11] Text string detection from natural scenes by structure-based partition and grouping in IEEE Trans. Image Process., vol. 20, no. 9, pp. 2594–2605, Sept. 2011.
[12] The Main Principles of Text-to-Speech Synthesis System Ü.R. Aida–Zade, C. Ardil and A.M. Sharifova