A Survey Paper on Speech Captioning of Document for Visually Impaired People

Mr. Pritam S. Langde[1], Dr. Shrinivas A. Patil [2], Mr. Milind N. Potdar[3]

[1]Asst. Prof E&TC Engg. Department , ATS's SBGI Miraj, India

[2]HOD& Prof. E&TC Engg. Department, ATS's SBGI Miraj, India

[3] Asst. Prof E&TC Engg. Department, DKTE Society's Textile & Engineering Institute, Ichalkaranji.

Abstract

In today's world, information is the key in defining success and there are multiple ways like Books, Newspapers, magazines, etc. to get the information for the sighted people. But there is limited scope for virtually impaired people to grab the reading information without using the braille system. This paper presents the survey of an efficient model which are more helpful to virtually impaired people for reading the text present in the captured image document proposed by researchers. The survey elaborated on various performance evaluation strategies for reading the document which is combination of text and images. For capturing the results many researchers mainly uses high-quality web camera to capture images, Tesseract OCR (Optical Character Recognition) software to convert the image into text and TTS (Text to Speech) engine For the transformation of text into speech. also, they used some open-source software for processing of images.

The survey highlights use of Deep learning-based strategic understanding techniques which provide better solutions as well as the platform for further research in the same field.

Keywords: - *OCR (Optical Character Recognition), python, Text recognition,TTS*

I. Introduction

In our world, information is generally available in the form of books and documents. It is fully usable for the sighted people. From ancient time, information resembled in the aural format as no other representation of it is founded in printing format. When an era has come of printing it facilitates the sighted people partially to acquire knowledge. A major problem for a blind or visually impaired person (BVI) to interact with the world to share knowledge. For them, information has to be in a special tactile language or in voice format. The most difficult task for them is reading text from the books or documents. For the blind or visual impaired (BVI) person, it is a very difficult job to acquire information from the world. One feasible way in order to perform that job is that someone will help him to read aloud the context. Another way to get the information is by giving a feeling of the information. The latter technique is built through a representation of the information on a paper or a substantial surface so that a blind person can feel and make all the books or documents available for the blind people.

Automated caption generation of images can make the books library a more invitingplace for visually impaired surfers. Being able to automatically describe the content of an image using properly formed English sentences is a very challenging task. This task is significantly harder, for example, the well-studied image classification or object recognition tasks, which have been a main focus in the computer vision community. Indeed, a description must capture not only the objects contained in an image, but it also must express how these objects relate to each other as well as their attributes and the activities they are involved in. Also, book page contents are not fixed and other than text and image, tabulated contents may also appear.

II. Literature survey

In the past, a Visually impaired person was dependent on the braille scripting language for document reading. But definitely, there are more limitations while converting every document into a braille script language format. To solve these limitations various automated devices are designed. Several studies have been conducted on the concept of solving the problems of visually challenged people.

There are a variety of techniques developed to date for document reading task. The method involves text recognition. The literature is surveyed and few of them are addressed here by considering a specific application for document reader along with their advanced methods and success in results. The main focus of the survey is image description and speech synthesis along with text processing methods.

M. Perera, C. Farook, A. P. Madurapperuma, 2017 [1] have shown the method for action description using video data for visually impaired people. The method shown uses a feature set is extracted for each frame and is obtained from the projection histograms of the foreground mask. The number of moving pixels for each row and column of the frame is used to identify the instant position of a person. The person region segmentation is done to extract exact features of the body. Support Vector Machine (SVM) is used to classify extracted features of each frame. The final classification is given by analyzing frames in segments. The classified action is associated with the text which then is converted into speech. The experimentation shows a good levelof results which involve the processing of frames with a single person. There is no focus on data with multiple people in frames.

T. Vania Tjahja, A. Satriyo Nugroho et.al, 2011 [2] conducted research to accommodate the needs of visually impaired people through an intelligent system, which reads textual information on papers and produces corresponding voice. Indonesian Automated Document Reader (I-ADR) is operated via a voice-based user interface to scan a document page. Textual information from the scanned page is then extracted using Optical Character Recognition (OCR) techniques. In this system, A user can then choose to have the system read the whole page, or they can opt to listen to a summary of the information on the page. SIDoBI (Sistem Ikhtisar Dokumen untuk Bahasa Indonesia) is integrated into the system to provide summarization feature. The result of either the whole-page reading or summarization is converted to speech through a text to- speech synthesizer. This whole system is developed under the Free Open Source Software policy and will be distributed openly to all users in need without any cost. Authors have focused on the text segmentation algorithm implemented in I-ADR to extract text from documents with complex layout. Author implemented IADR text segmentation module using Enhanced CRLA and propose an improved algorithm for text extraction. Evaluation of the proposed system with various page layouts showed promising results.

S. Musale and V. Ghiye, 2018 [3] have shown a smart reader, an effective system for visually impaired. They used OCR (Optical Character Recognition) functions of MATLAB for converting the image to text. Authors have given a method audio-tactile user interface that supports the user to read the information printed on the paper.

S. Das, Lalit Jain, Arup Das, 2018 [4] introduced the concept of deep learning for military image captioning. They developed the concept of "deep fusion" based on deep learning models adopted to process multipath modalities of big data. They demonstrated the deep fusion concept for image captioning using hybrid CNN/RNN deep network models. Also, they evaluated the results both subjectively and objectively with BLEU score.

A. Manikandan, Shouham Choudhury, Souptik Mujumdar, 2017 [5] "Text Reader for Visually Impaired People: ANY READER" They proposed a cost-effective solution to help the visually impaired people in reading out texts easily with full depth analysis of the textual content through an android application - ANY READER. That application notifies the user about the headings, subheadings, and paragraphs in the text through voice notifications. Also, that application is optimally designed to run from mid-range to high-end android phones which make it readily usable and accessible to various kinds of users irrespective of their possessed smartphones. During the implementation of whole process OCR (Optical Character Recognition), OpenCV Image processing Library and Android Studio are used.

P. Zientara, S. Advani, N. Shukla, et.al, 2017 [6] presented a multitask grocery assistance system for the visually impaired article. In this article, they developed interfaces, algorithms, and hardware platforms to assist the visually impaired with a focus on grocery shopping. For assistive technology, they focused on two main modes of providing feedback and guidance to the user. These modes are auditory. To provide this feedback to the users, they use the glove and the glasses. Here they introduced the latest architectures and emerging devices that are being explored to further improve the capabilities of such systems.

R. Dhar, S. Mukherjee, 2018 [7] presented paper Android-based Text Reader for Partial Vision Impairment people. Here the author designed a new Android-based Text Reader application/system that assists partially sighted and elderly people to understand the meaning of the object particularly text presented to them in any formats. Also it presents an application that assists people with partial vision to identify and recognize the surrounding text through a user friendly interface and helps them to develop a sense of awareness about the environment. System design includes Mobile Vision, Mobile Vision Text Detector and Text To Speech processing. While designing the system Google Mobile Vision Text API is used to recognize the text that segments the text into blocks, lines, and words correspondingly. Here also OCR detector processor was used to convert camera image into text.

S. Anzarus Sabab and Md. Hamjajul Ashmafee 2016 [8] explores an Intelligent Assistant for Blind person. As Braille is one of the methods which is used to read a book or document for blind persons so author focused on developing an algorithm for any document which has to be converted to braille format. They made a smart device with a multimodal system that can convert any document to the interpreted form to a blind. A blind can read document only by tapping words which are then audibly presented through a text to speech engine. This Blind Reader, an android application is a cost effective solution which converts any document (.docx, .pptx, .pdf) to accessible form to a BVI (blind or visually impaired). Which gives the BVI a total mental concept of reading context.

Si-Woo Kim, Jae-Kyun Lee, Boo-Shik Ryu and Chae-Wook Lee 2008 [9] have given an embedded system for visually-impaired people. In this system, author present the analog-digital Code (AD) code and an effective embedded system which can transform text information into voice using the 2D AD code and Text To Speech (TTS). This voice information can also be transmitted to visually impaired people by capturing the AD code on paper or in books. The implementation of the embedded system is done using TI DSP chip (TMS320C6711) that can convert captured images by CMOS image sensor to voice. Here the AD code is clearly deciphered by real time within 1 second and can support error correction rates up to 50%.

P. Thakare, S. Kote, A. Pawale, A. Rajguru , O. S , 2018 [10] presented an Interactive reader and Recogniser system for visually impaired people. It presents three modules viz., Text to Speech, Face Recognition & Object Recognition. The implementation is done on the android device. In face recognition, the major part is to detect the face at first using cascaded classifier and then compare the detected face with the data present in the database using Euclidian distance. In object recognition, Limited set of objects which will be stored in the internal storage will be compared to the real time images using connected component feature vectors. Text to Speech extraction is performed using the predefined Google API with optical character recognition. The system consists of a wearable glass embedded with wireless cameras which are being connected to the android phone. As per authors experimentation, this system gives accuracy up to 95%. The prototype consists of normal glasses embedded with a wireless camera and an android phone with basic configurations.

E. Veera Raghavendra, et al. ,2010 [11] have shown a method of a screen reader to help visually impaired people to use or access the computer and the Internet. The development is shown for Indian languages and this development is limited by availability of Text-to-Speech (TTS) system in Indian languages, support for reading glyph-based font encoded text, Text Normalization for converting nonstandard words into standard words, supporting multiple languages. Google TTS is used for language identification and speech conversion.

Dr. I S Akila et al , 2018 [12] have given prototype development method of text reader using a raspberry pi. This is done using Raspberry Pi 3 model B and a camera module with the concepts of Tesseract OCR [Optical Character Recognition] engine and Google Speech API [Application Program Interface] which is the textual input to speech engine. The model is programmed using Python language. It is portable and easy to use thus providing a better reading experience to the visually challenged people.

T. Kornsingha et al, 2011 [13], have given a method for developing a voice system, reading medicament label, for visually impaired people. Understanding the texts or messages that appear on the label by using the sense of hearing a voice emitted from the label reader. Labeling pharmaceutical products provide medicinal information, type of medicament, description of treatment and period of use. RFID (Radio Frequency Identification) technology and microcontroller are used. RFID tag based system information is pre-stored in the database which then is extracted when the tag is scanned and then converted into speech.

S. Sulaiman et al, 2016 [14], have addressed shortcoming by investigating the difficulties faced by visually impaired Internet users to understand spoken texts when using screen readers with a non- BM language narrator or speaker. This study aims to reduce the time taken for the visually impaired to understand information written in the BM language and narrated by a screen reader. This paper highlights the processes involved in developing a prototype screen reader, which will read in the BM language and in a Malaysian accent. To achieve all the objectives, preliminary interviews and testing sessions were conducted to collect data to test the hypotheses made. The findings were then used as the main source of data to develop a prototype screen reader. From the built prototype, user-test was conducted with a sample group consisting of the visually impaired to test the functionalities and evaluate the effectiveness of the software. Results and recommendations are shared at the end of the paper as milestones for future enhancements

A. Singh, et.al, 2018 [15], have given the device developed for reading the text for visually impaired people. The device works offline using a single push button which initializes the four serially integrated operations namely scanning, image pre-processing, text extraction from the image and its narration through the earphones attached. High-quality images in the focus window of 25-45 cm could be captured even in dark/low-light conditions wherein a series of image processing algorithms are then executed to de-skew, Hough line transform and binarize theimage removing unwanted noise to fetch the best output to be fed to the OCR. The text format output is then processed through TTS for narration. The key features of the device include its lightweight design (180 grams), portable nature, non-tedious in use, near real-time operation, wide scanning field, provision of the rechargeable battery with high run time, low cost of development and no requirement of external aid in its operation. Further, the device can also be operated while commuting.

S. Mishra, S. M Banubakode et.al, 2014 [16] have addressed touch screen devices available in the market with various accessible techniques like using a screen reader, haptics and different input mechanisms. But they did not give any remarkable results in convenience handling and text entry speed. In this paper, the authors analyze different interaction techniques and their impact on text entry speed. They also proposed a new user interface for touch screen device to minimize accessibility barriers for visually impaired users.

Xiaoqiang , Binqiang Wang et. al, 2018 [17] presented Exploring Models and Data for Remote Sensing Image Caption Generation. In this paper they investigate to describe the remote sensing images with accurate and flexible sentences. Firstly, some annotated instructions are presented to better describe the remote sensing images considering the special characteristics of remote sensing images. And Secondly, in order to exhaustively exploit the contents of remote sensing images, a large-scale aerial image data set is constructed for remote sensing image caption The experimental results from this paper shows that the image caption methods for natural image can be transferred to remote sensing image to obtain only acceptable descriptions.

A. Karpathy and Li Fei-Fei 2015,[18] explores Deep Visual-Semantic Alignments for Generating Image Descriptions. Thay made a model which generates natural language descriptions of images and their regions. That model is based on a novel combination of Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks (RNN) over sentences, and a structured objective that aligns the two modalities through a multimodal embedding. Also, conducted large-scale analysis of RNN language model on the Visual Genome dataset of 4.1 million captions and highlight the differences between image and region-level caption statistics. Experimental results of the model shows that a model provides results superior to all previous work, controlling for the strength of the underlying convolutional network features. Also, a model described a Multimodal Recurrent Neural Network architecture which generates descriptions of visual data.

Kun Fu et. al 2018[19] presented Image-Text Surgery: Efficient Concept Learning in Image Captioning by

Generating Pseudopairs. a novel method for Image-Text Surgery has been proposed in this paper to synthesize pseudo image-sentence pairs. The Pseudopairs were generated under the guidance of a knowledge base, with syntax from a seed data set and visual information from an existing large-scale image baseVia pseudo data, the captioning model learns novel concepts without any corresponding human-labeled pairs. A model evaluate on a subset of the MSCOCO data set. The experimental results demonstrate that the model provides significant performance improvements over state-of-the-art methods in terms of F1 score and sentence quality.

J. Dong, Xirong Li, and Cees G. M. Snoek, 2018 [20] conducted research on Predicting Visual Features from Text Image and Video Caption Retrieval. Research shows the viability of resolving image and video caption retrieval in a visual feature space exclusively. It contributes Word2VisualVec, which is capable of transforming a natural laguage sentence to a meaningful visual feature representation. For state-of-the-art results, author suggest Word2VisualVec with multi-scale sentence vectorization, predicting the ResNet feature when adequate training data is available or the GoogLeNet-shuffle feature when training data is in short supply.

III. Conclusion

This paper provides a survey for strategic understanding of document reading systems for blind peoples . The most methods used for reading the text content are OCR based and some methods like image captioning are also used to read the other contents in the documents like images, tables etc. using deep learning approaches.

More work was carried out in the area of preparing devices like walking sticks, goggles, raspberry pi-based reader, RFID based labelling devices, etc. In existing systems Complex hardware system was used which is not suitable in every situation. Also, it is very difficult for a blind person to find the fault in the device if the device became not working properly. The paper may remain helpful for further research in the field of document reading for blind peoples.

References

[1] M. Perera, C. Farook and A. P. Madurapperuma, "Automatic video descriptor for human action recognition," 2017 National Information Technology Conference (NITC), Colombo, 2017, pp. 61-67.

[2] Teresa Vania Tjahja et al., "Recursive text segmentation for Indonesian Automated Document Reader for people with visual impairment," Proceedings of the 2011 International Conference on Electrical Engineering and Informatics, Bandung, 2011, pp. 1-6

[3] S. Musale and V. Ghiye, "Smart reader for visually impaired," 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, 2018, pp. 339-342

[4] S. Das, L. Jain, A. Das, " Deep learning for military image captioning", International conference on information fusion (FUSION), Cambridge UK, ISIF 2018

[5] V. M. Manikandan, S. Choudhury and S. Majumder, "Text reader for visually impaired people: Any reader," 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, pp. 2389-2393, 2017.

[6] S. Advani et al., "A Multitask Grocery Assist System for the Visually Impaired: Smart glasses, gloves,

and shopping carts provide auditory and tactile feedback," in IEEE Consumer Electronics Magazine, vol. 6, no. 1, pp. 73-81, Jan. 2017.

[7] R. Dhar and S. Mukherjee, "Android-based Text Reader for Partial Vision Impairment," 2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), Gorakhpur, pp. 1-5, 2018.

[8] S. A. Sabab and M. H. Ashmafee, "Blind Reader: An intelligent assistant for the blind," 2016 19th International Conference on Computer and Information Technology (ICCIT), Dhaka, pp. 229-234, 2016.

[9] S. Kim, J. Lee, B. Ryu, and C. Lee, "Implementation of the Embedded System for Visually-Impaired People," 4th IEEE International Symposium on Electronic Design, Test and Applications (delta 2008), Hong Kong, pp. 466-469, 2008.

[10] P. U. Thakare, K. Shubham, P. Ankit, R. Ajinkya and S. Om, "Interactive Reader and Recogniser System," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, pp. 427-431 , 2018.

[11] E. V. Raghavendra and K. Prahallad, "A multilingual screen reader in Indian languages,"2010 National Conference On Communications (NCC), Chennai, pp. 1-5 , 2010.

[12] S. Akila, B. Akshaya, S. Deepthi and P. Sivadharshini, "A Text Reader for the Visually Impaired using Raspberry Pi," 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), Erode, pp. 778-782, 2018.

[13] T. Kornsingha and P. Punyathep, "A voice system, reading medicament label for visually impaired people," RFID SysTech 2011 7th European Workshop on Smart Objects: Systems, Technologies and Applications, Dresden, Germany, pp. 1-6 , 2011.

[14] N. F. B. M. Noh, S. Sulaiman and A. B. M. Noor, "Accessibility matters: The need of Bahasa Melayu (BM) screen reader for the visually impaired internet users," 2016 4th International Conference on User Science and Engineering (i-USEr), Melaka, pp. 11-16, 2016.

[15] Singh, S. Dcvnani, V. Kushwaha, S. Mishra, A. Gupta, and K. K. S. Pandian, "An Efficient Auxiliary Reading Device for Visually Impaired," 2018 International Conference on Smart City and Emerging Technology (ICSCET), Mumbai, pp. 1-5, 2018.

[16] S. Misra, S. M. Banubakode and C. A. Dhawale, "Novel user interface for text entry on touch screen mobile device for visually impaired users," 2014 Global Summit on Computer & Information Technology (GSCIT), Sousse, pp. 1-5, 2014.

[17] X. Lu, B. Wang, X. Zheng and X. Li, "Exploring Models and Data for Remote Sensing Image Caption Generation," in IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 4, pp. 2183-2195, April 2018

[18] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 664-676, 1 April 2017

[19] K. Fu, J. Li, J. Jin and C. Zhang, "Image-Text Surgery: Efficient Concept Learning in Image Captioning

by Generating Pseudopairs," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 12, pp. 5910-5921, Dec. 2018

[20] J. Dong, X. Li and C. G. M. Snoek, "Predicting Visual Features From Text for Image and Video Caption Retrieval," in IEEE Transactions on Multimedia, vol. 20, no. 12, pp. 3377- 3388,Dec.2018