ISSN:0975-9646

Microcontroller Based Mechanised Videographing of Text and Auto-Generation of Voice Text in Real Time

P. Nagabhushan , Vimuktha Evangeleen Jathanna

Department of Studies in Computer Science

University of Mysore, India

***Abstract*— The easy access to portable cameras as in mobile phones, has made it possible to video capture even text documents and carry out reading the text in leisure. This also motivates to provide readability to blind. With this backdrop in this paper we have devised a microcontroller based mechanised videographing of text. It is also proposed to voice text the videographed text in real time. Subsequently for the purpose of archival, video text is stored as a text file circumventing the large memory requirement because of text video.**

***Keywords*— Text reading system, video text, microcontroller, block scan.**

I. Introduction

Text extraction and its recognition has become extremely popular research in the vision community. Text is the most valuable and useful source of semantic information for content retrieval, indexing and content understanding. Also, due to the incorporation of low cost, small, handy and portable cameras in mobile phones, tablets, PDA's and other devices we can utilize these cameras for text acquisition from documents, natural scene and text superimposed on movie sub titles, sports scores or movie credits. These type of devices acts as an assistive text reading frame work for visually impaired. The text captured from these devices results in huge amount of text data for processing and demands an effective text reading system.

The text readers which are available in the market with mobile phones, PDA's and PDF readers are able to recognize and read text aloud. The mode of operation of these devices are: it works on static images i.e. text is extracted and recognized from a single source image captured by the camera and manual interaction for text localization in the image as in case of Android OCR. It is obvious that these type of devices used for text reading are relatively less used by visually impaired for reading text as it consumes lot of manual effort for capturing, localization and extraction of text. It will be convenient for the user if we consider developing a dedicated text reading system that operates with minimum manual interaction.

In this paper we have developed a microcontroller based text reading system with mechanized video capturing and auto generation of voice text for real time application. This device can independently read text aloud from documents, news papers and books which are captured as videos.

The remainder of this paper is structured as follows: In section II we have done a survey on the available literature and devices. In section III we present the motivation for developing the text reading system. Section IV presents the proposed design, development and working of the system using mechanical, electronic and programming modules. Section V describes the implementation of the system design and in section VI the results obtained are discussed.

II. Literature Survey

As stated in [1], there are several ongoing researches on camera based document analysis which include text detection, extraction, enhancement, recognition and its applications. Few text reading system models for assistive text reading for visually impaired is also presented in literature. A brief appraisal on these system and the text processing methods stated in the literatures is given below.

Rajkumar N et.al [2] proposed a camera-based assistive text reading framework to help blind persons to read text labels and product packaging from hand-held objects. The method determined the region of interest (ROI) in the video and extracts moving object region by a mixture-of-Gaussians-based background subtraction technique. In the extracted ROI, text localization and recognition are conducted to acquire text details. To automatically focus the text regions from the object ROI, text localization algorithm by learning gradient features of stroke orientations and distributions of edge pixels in an Adaboost model was employed.

Gade.Bhagya Lakshmi et. al [3] and R. Infant Abinaya et.al [4] both proposed camera based assistive text product label reading and image identification for hand-held objects using Raspberry Pi microcontroller board. Programming for text extraction and recognition was done by OpenCV on Linux platform

Nanayakkara, S et.al [5] developed a finger worn device containing a microcontroller and a button camera. This device assists the visually impaired by reading paper-printed text. The approach is novel since it utilized, local-sequential manner for scanning text which enables reading single lines, blocks of text or skimming the text for important sections while providing real-time auditory and tactile feedback.

Majid Mirmehdi et.al [6] developed a mobile head-mounted device for detecting and tracking text. The tracking system comprise of a remote controlled microcontroller combined with webcamera is encased in an ordinary flat-cap hat the entire sustem is connected as acquisition device to the laptop. A near to real-time text detection algorithm which uses Maximal Stable Extremal Regions (MSERs) for image segmentation is utilized for text detection and extraction

An automatic video text detection, localization and extraction approach in videos was proposed by Chengjun Zhu et.al [7]. In this approach Support Vector Machine was employed to partition video blocks into text and non text regions. Both artificial text and scene text were considered in this approach. Also due to the employment of correlation of videos using special symbol, detection of false regions were effectively reduced.

Christian Wolf et.al. in [8] proposed a method for content based image and video retrieval in multimedia documents. The algorithm for this approach used morphological post processing to detect the text and a measure of accumulated gradients by multiple frame integration for artificially embedded text in videos.

Many approaches for text extraction by mobile cameras from natural scene have been proposed. A focus mechanism using a hint on the location of target text was adapted by Egyul Kim et.al. [9]. This method adapts pixel sampling and a mean shift algorithm within the focused section and a distance measure is applied to compare target seed color.

Nirmala Shivananda and Nagabhushan [10] proposed a hybrid method for separating text from color document images which combines connected component analysis and an unsupervised thresholding for separation of text from the complex background. The proposed approach identifies the candidate text regions based on edge detection followed by a connected component analysis.

Different approaches for text processing in video frames with complex background was proposed by Fang Liu et.al.[11] and Shuicai Shi et. al.[12] in which density based histogram segmentation and block change rate approach for clustering candidate characters by integrating spatial connectivity and color feature of character pixels is employed.

A method for segmenting and recognizing text embedded in video and images was proposed in [13] by Jean-Marc and Datong Chen. This method used multiple segmentation of same text region producing multiple hypotheses based on Markov random field and eliminating background by connected component analysis and by enforcing a more stringent constraint called GCC by using ID-median operator.

An edge based text verification approach for content retrieval for video caption text was proposed by Jian Zhang in [14]. The verification was based on the investigation of the relation between the candidate blocks and the neighbor areas. The proposed method detects and removes background edges from both background and text blocks and also refines the text boundaries.

III. BACKGROUND

The main objective behind the design of the text reading system model is to imitate the reading pattern of human beings. Given a text document the normal human tendency is to read the text from left to right (few cases where it is dependent on the language of use). Once the end of the line is reached, automatically the eyes rolls back to the left first character of the second line and continue. The eyes adjust themselves to this pattern of reading in spite of images, tables or diagrams present in the document.

To design such a system, we shall assume that we have a document that contains multiple lines. In general, if we are capturing this document as a video using a hand held camera, we will start the scan from the left most corner of the first line i.e. the first character of the page and then slide till the end of the line and then by the reverse motion come to the start point of the second line. This is continued till the end of the last line of the page is reached. But this type of text capturing method with regard to visually challenged people poses certain intricacies in implementation and exhibits several artifacts like camera shake, illumination changes and blur. With this backdrop we have designed and developed a mechanized system that automatically videographs text as described above and generates voice text of the document. The detailed description of the design and its working is presented in the subsequent sections.

IV. PROPOSED SYSTEM

The design of the text reading system comprises of the mechanical, electronic and programming components. The advantage of this type of system is the stability of the camera without camera shakes, constant speed fluctuations and acquisition of text in video mode which helps to capture longer text lines. Figure 1 depicts the proposed model of the text reading system and figure 2 depicts the developed text reading system designed and developed for the purpose.

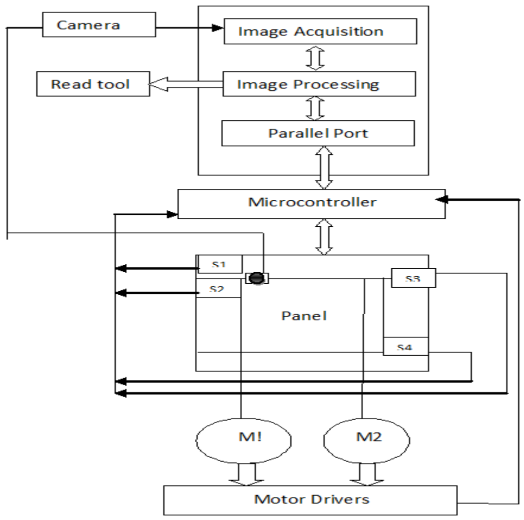

Fig. 1 Proposed design of the mechanised text reading system

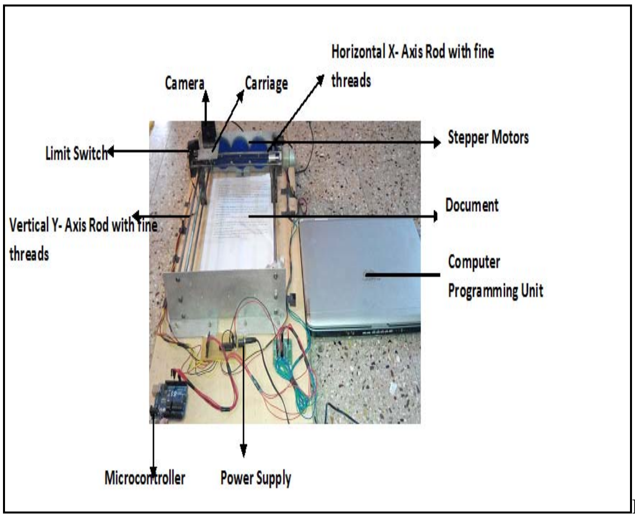

Fig. 2 Real time text reading system

A. Mechanical Unit Design

The mechanical unit of the text reading system comprises of a wooden panel, a carriage with a camera mounted on it, horizontal X-axis and vertical Y-axis rods with fine tuned threads on them. Two stepper motors M1 and M2 are used to move the camera along X axis and horizontal rod with camera along Y axis. Four limit switches are used as sensors for defining the starting and ending points of scan. The detailed description of each part and its design is presented below. Figure 3 depicts the mechanical unit.

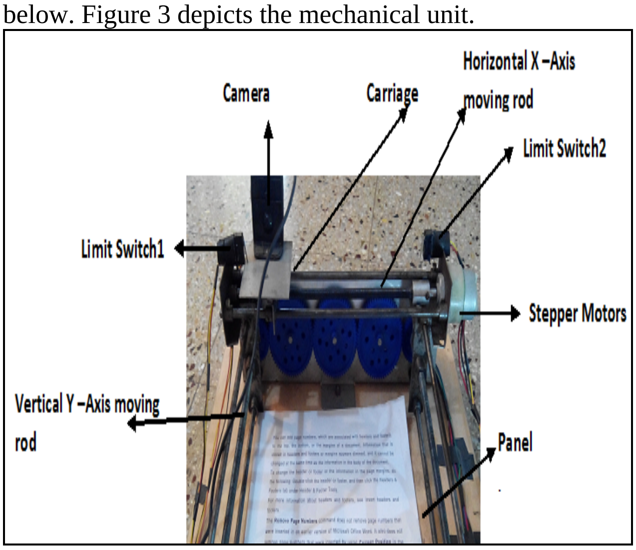

Fig. 3 Mechanical Unit of text reading system

1) Structure Design: A 30 cm x 35 cm wooden panel is used to develop the text reading system. A horizontal movable rod is fixed to the panel along X- axis. The two vertical ends of the panel is attached with two fixed rods V1 and V2 along Y –axis. The horizontal movable rod is fixed to a small movable metal plate on which a camera is mounted to capture the text. Hence the camera is placed parallel to the document. The height of the camera is adjusted depending on the quality of image captured. The rods contain the fine tuned threads on them. Each thread defines how much camera movement should occur. The equation of camera movement along X – axis and Y-axis is given below.

$$CM\_XAXIS = \Sigma\ (\text{Number of circular threads on X-axis rod}) \quad (1)$$

The horizontal rod movement with the camera fixed along y –axis i.e. vertical movement is given in equation 2.

$$HC\_YAXIS = \Sigma\ (\text{Number of circular threads on y-axis rod}) \quad (2)$$

2) Stepper Motors: In our system we have used two stepper motors M1 and M2 to drive the camera along X-axis i.e. left to right or right to left and the horizontal rod around the Y –axis i.e. top to bottom or bottom to top with the camera at the starting point. A stepper motor moves one step when the direction of current flow when the field coil changes, reversing the magnetic field of the stator poles. Hence this helps us to move the camera and the horizontal rod on the fine tuned threads. The motors drivers connected to the microcontroller are used to determine the direction of movement which is discussed in the later sections

3*) Limit switches*: Limit switches are used as sensors in our system to determine the start and end points of the camera movement. The limit switches are also called "micro switches". These switches give ON/OFF (or yes/no) binary outputs. The activation of the switch causes electrical contacts to either "break" ("normally closed" or NC switch) or "make" ("normally open" or NO switch) – or both NC and NO. We have used 4 limit switches S1, S2, S3, and S4. Switch S1 is used at the left most corner of horizontal rod along X axis. Switch S2 is used at the top most corner of vertical rod along Y axis. S1 and S2 define the starting point of the scan. Switch S3 is used at the top right most corner of horizontal rod along X axis which defines the end point of the line scan. Switch S4 is used at the bottom right most corner of vertical rod V2 which defines the end of the page. All these switches are connected to the microcontroller.

B. Electronic Unit Design

The electronic unit design consists of an Arduino board with ATmega328 microcontroller, driver circuit for two stepper motors. A voltage regulator circuit is utilized to overcome power surges and regulate power supply. The circuit diagram of the text reading system with motor interface circuit and voltage regulator circuit is a given in figure 4.

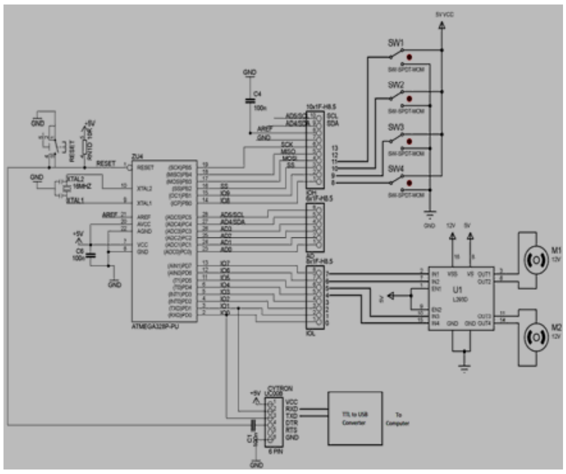

Fig.4 Circuit Diagram

1) Microcontroller: In our system we have used the Arduino microcontroller board with ATmega328 microcontroller. The microcontroller works with 5V power supply. It has 14 digital input/output pins (of which 6 can be used as PWM outputs), 6 analog inputs, few dedicated pins for power supply, ground, interrupts, LED and SPI communication, a 16 MHz ceramic resonator, a USB connection, a power jack, an ICSP header, and a reset button. It is connected to the computer with a USB cable and it appears on the operating system as a virtual com port without utilizing any external drivers. The Arduino can be programmed with the Arduino software and works with simple textual commands. Figure 5 depicts the Arduino board with the microcontroller".

Fig.5 Circuit Diagram

In our text reading system model we have used pins 8,9,10 and 11 for switches S1,S2, S3 and S4 respectively. We have used pins 4, 5, 6 and 7 for the connection from motor interface circuit. And GND pin is connected to the voltage regulator circuit.

2) Motor Interface Circuit: The motor interface circuit utilizes L293D IC's to drive the motors M1 and M2 along X and Y –axis. The motor interface IC works with 12 V power supply. The output pins 3 and 6 are used to drive motor M1 in clock wise and anticlockwise direction. For motor M2 out pin 11 and 14 are used drive motor in clock wise and anticlockwise direction. The input pins 2, 7, 10 and 15 are interfaced with the microcontroller pins 7,6,5,4 respectively. The block diagram of the motor interface circuit is given in figure 6.

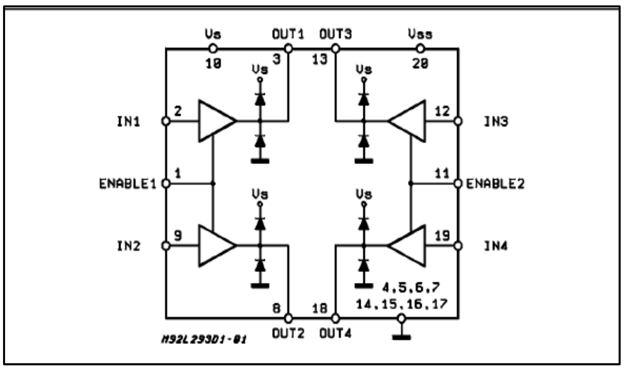

Fig 6 Motor interface circuit

3) Voltage Regulator Circuit: The LM7805CT is a three terminal positive voltage regulator which is available with fixed output voltages. These IC's are useful in a wide range of applications. The IC employs internal current limiting, thermal shut down and safe operating area protection, making it essentially indestructible. If adequate heat sinking is provided, the IC can deliver over 1A output current.

C. Software Design

The main objective of the programming unit is to read the text present in the document. But it is obvious that the document cannot be read directly from a video. Hence an intermediate step is designed to segment, extract and recognize the text present in video. The proposed framework for frame acquisition for extraction and recognition of text from video is as given in the figure 7.

1) Video Acquisition: As discussed in the above section, with the aid of the mechanical unit and its interface with the electronic components we will capture the document as video using a web camera. The video acquisition method is depicted in figure 7.

2) Block Scan: The text reading system is supposed to capture the individual text lines and then concatenate them to form complete sentence. The camera used to capture the text due to its field of view, is able to capture multiple lines at a time. Hence a block here refers to multiple lines. We have set four lines per block. Instead of parsing individual lines the block scan helps to speed up the recognition process.

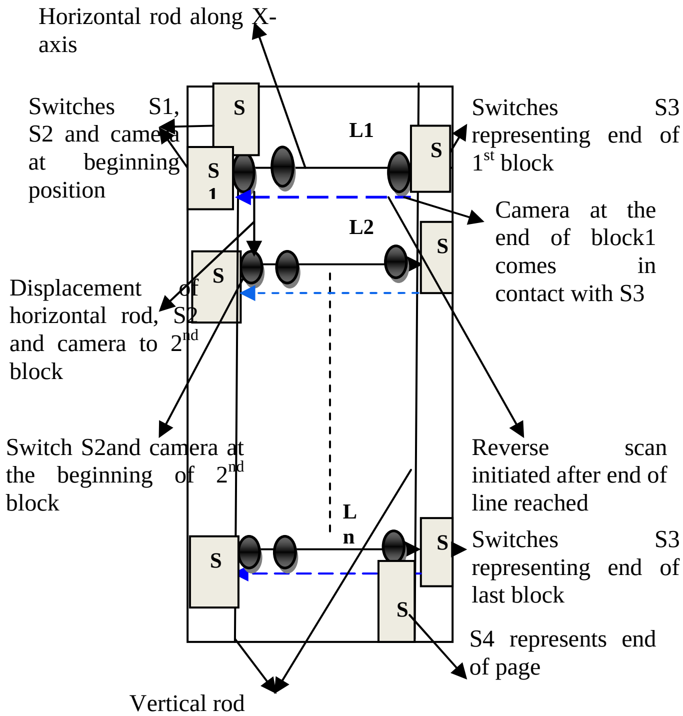

Fig 7 Video Acquisition

3) Acquisition Procedure: The text reading system consists of a horizontal rod to which a movable carriage is mounted. The camera is fixed to this carriage and is connected to the computer through the USB port. When the text reading system is switched on, the carriage is set to the starting position of the document i.e. left top most corner. The top and left corner of the carriage will come in contact with switch S1 and S2. This closes both the switches and initiates the scan. The motor M1 starts to move in clockwise direction allowing the carriage to move on the fine tuned threads of horizontal X-axis rod. The camera that is fixed on carriage captures the document as individual frames containing text. All these frames are stored in the temporary buffer for further processing. Once the carriage reaches the right most corner, switch S3 is closed stating that the end of block is reached. To capture the text between switch S1 and switch S3 we have set 30 frames per trigger using the selected camera source. The motor M1 starts rotating in the anticlockwise direction indicating the reverse scan i.e. from right to left and stops video acquisition. By now the frame registration, frame text extraction, recognition and text reading modules are executed

Once the camera comes back to the starting position closing the switch S2 the reading of the first block is completed and the buffer is emptied. The vertical Y-axis rod moves the horizontal rod together with the camera and the carriage to the next block and initiates the new scan. The frame acquisition, frame selection, frame registration, text processing and text reading modules are executed independently for each block of horizontal scan containing single / multiple lines of text till the end of page is reached.

4) Frame Selection: Each frame that is acquired by the web camera during horizontal block scan undergoes several artifacts like translational changes, linear blur, defocusing and noise due to camera movement. Hence before segmenting and extracting the text from the frame it is important to select the best frame depending on image quality metric. In this paper we have utilized no reference perceptual blur metric as described in [16]. The blur metric finds the sharpest intensity frame present among the frames present in the temporary buffer. The blur metric returns the value between 0 and 1. 0 refers to the sharpest image and 1 refers to the blur frame. A threshold t is set to a value that decides which blurry frame needs to be eliminated. Though some information is lost, it is seen experimentally that consecutive frame contains similar information hence remaining frames can be used for retrieving information. The algorithm for frame selection is described as given below.

// Algorithm Frame selection

Step 1: Start
Step2 : for I = 1 < number of frames
Step 3: IM = read(frames(I))
Step4: blur = blur_metric(IM)
Step 5 if blur == T
Step 6 Store IM in new_frames
Step 7 end if

End for

5) Frame Registration: The frame registration involves the process of finding the similarity between the consecutive video frames and eliminating redundant frames. As we know when we capture a video it contains lot of content wise similarity. When the camera is moved along X-axis i.e. from left to right the camera is able to capture 60 frames per second. But we have set frame trigger to 60 frames for a minute thereby reducing lot redundancy. Also for the remaining frames we have used the frame differencing algorithm to see the shift in the contents of the frame. It is observed that, there exists shift in the text contents at some time instance 't' due to camera movement.

If we visualize the frames in a horizontal scan it appears as if the text content present at the left hand side of the i-1 frame is moved out in i frame and some text from i+1 frame is moved in at the right hand side. Hence we can first find the displacement of text between the frames in time instance t using frame differencing which also depicts the motion. We calculate frame differencing as follows:

Frame _difference = abs(average(color of frame_i) -- average(color of frame_i-1)) (3)

where, Frame_difference is a matrix that quantify the amount of displacement of each pixel with respect to motion. The average of the pixels in the frame difference matrix is calculated.

Initially we set the first frame acquired as the reference frame and average of the frame difference is calculated with the next frames. A threshold T is set and compared. If the comparison score is greater than the threshold, it implies lot of shift has occurred in the current frame from that of the reference frame. Hence change the reference frame to the current frame and start calculation of frame differencing. All the reference frames are stored in a directory for text segmentation. The algorithm for the proposed module is as given below.

// Frame Registration for block scan
Step 1: Start
Step 2: Previous =1, curr=2
Step3 : for I = prev < new_frames
Step 4: for j = curr < new_frames
Step5: Frame_difference = avg(abs(average(color of frame_i) - average(color of frame_j)))
Step 6: If(Frame_difference > Th)
Step 7: Pre = j;
Step 8: Curr = j + 1
Step 9: Store frames(j)
Step 10: break; got to Step 3
End if, End for, End for

6) Text Segmentation: Each frame in the directory is subjected to segmentation. The segmentation is based on morphological segmentation. The documents we have considered contain uniform white background with black text with single or multiple lines. As the block scan we have considered is strictly horizontal and is free of skew we have considered horizontal projection profile for line segmentation. We then classify the text regions based on the connected component in each line. A character cannot be too small in the video. Hence we consider that if the height and width of a connected component are lesser than that of the smallest character in the video, then the region is considered as noise.

Assume Hc and Wc as the height and width of a connected component and Areac be the area of pixels in connected component. Then we consider the connected component as noise if

Height of connected component < threshold height (4)

Width of connected component < threshold width (5)

Area of connected component < threshold area (6)

Also, we consider the connected component as a text if the height to width ratio Rw/h lies within the range.

Rmin < Rw/h < Rmax (7)

Once the text is segmented the text contents are extracted using the bounding box around the connected components. The result of segmentation is given in figure 9

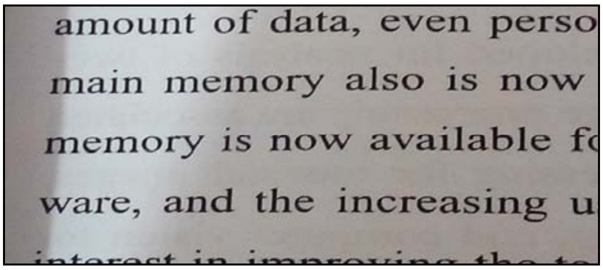

Fig 8 Text frame captured by acquisition device

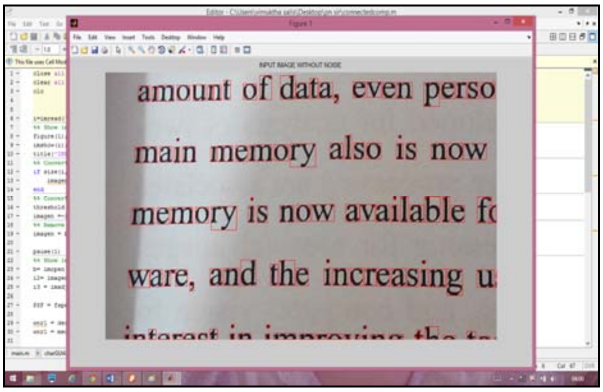

Fig 9 Text segmented and extracted in the frame using bounding box

7) Text Recognition: The next step of text extraction is recognizing the extracted text from the frame. We have exploited the Tesseract OCR engine to recognize the text. The recognized text is stored in the notepad file. The individual frame doesn't contain the complete line of text. As text comparison is easier, effective and efficient we developed an algorithm that compares text present in the two notepad files and merges the two files by appending the extra contents.. The algorithm for concatenation is as described below.

// Text recognition for a block
Step 1: Start
Step2: IM = read(frames(1))
Step 3: ocrtext = OCR(IM)
Step4: Write ocrtext into notepad.txt
Step 5 for I = 2 < number of frames
Step 6 IM = read(frames(I))
Step 7: ocrtext = OCR(IM)
Step 8; Write ocrtext into notepad1.txt
Step 9: Compare notepad.txt with notepad1.txt
Step 10: Merge the extra content to notepad.txt
Step 11: empty notepad1.ttxt
End for, End

The notepad at the end of one block scan will contain the OCR recognised text of the full block. Once the merging of all the frames is completed the text reading software is invoked.

8) Text Reading: The main goal of developing the text reading system is to read the contents of the document aloud. We have utilized the espeak software. The read tool module is executed at the end of writing each block to the notepad file. Once the reading of one block is completed, the program sends a text command the microcontroller to move the carriage and the camera along the Y-axis to initiate the next block scan. Image acquisition module, frame registration, selection, text segmentation, extraction and recognition modules are again invoked. This process is repeated till the end of the page is reached. The notepad at the end of last block scan in the page contains the full page contents

V. IMPLEMENTATION

The software module for the text reading system is implemented by integrating the Aurduino, Java Netbeans IDE with image processing and video processing toolbox of Matlab 2014a. Once the device is started the arduino initiates the camera and its movement. The GUI of the Netbeans IDE starts capturing and storing the frames in the temporary buffer. All the text processing functions are in on the frames are implemented using Matlab 2014a. Finally the espeak software is used to read the text present in the notepad file

VI. RESULTS

We have tested this system for reading different documents. The documents contained single and multiple lines of text with different font styles, size and line spacing. The system was able to read around 98% of text as we have utilized the Tessaract OCR engine. The efficiency of system is mainly dependent on how fast the camera is moved to capture the text. The delay in efficiency of text re efficiency of text reading system is due to the movement of camera motors. If we consider using motors with higher torque we can overcome the efficiency issues. The graph for time taken to read number of line is plotted below. The X axis in the graph refers to the number of lines read and Y axis refers to time taken

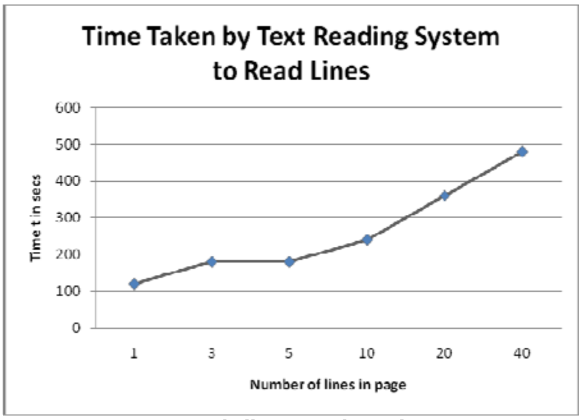

Fig 10 Graph illustrating the reading time

VII. CONCLUSIONS

In this paper we have designed and developed a microcontroller device that is mechanised for videographing of text and auto-generates voice text in real time. This text read system utilizes mechanical, electronic and programming components. The text segmentation and recognition is performed on live videos. The novelty of this system is that the document is captured as video in a sequential manner and text is extracted and recognized from the video. The entire content in the video is presented in the form of notepad file. The system is also capable of reading text lines and the end of block scan. This system can be utilized for visually impaired for reading text books, notes and other documents.

ACKNOWLEDGMENT

We thank Mr. Vishwanath Thammanna and Mr.Vijay Thammanna of Glovish Technologies, Bangalore for fabricating the text reading system.

REFERENCES

[1] K. Jung, K.I. Kim, A.K. Jain, "Text information extraction in images and video: a survey", Pattern Recognition, 37 977-997S, 2004.

[2] Rajkumar N, Anand M.G, Barathiraja N, "Portable Camera-Based Product Label Reading For Blind People", International Journal of Engineering Trends and Technology (IJETT) – Volume 10 ,Issue No : 1 - Apr 2014.

[3] Gade.Bhagya Lakshmi , Mr.T. Sreenivasu," Portable Camera-Based Assistive System for Product Label Reading with voice output for Blind Persons", ISSN No: 2348-4845 International Journal & Magazine of Engineering, Technology, Management and Research, Volume No: 1, Issue No: 11 (November), 2014.

[4] R. Infant Abinaya, E. Esakkiammal, Mrs.Pushpalatha, "Compact Camera Based Assistive Text Product Label Reading and Image Identification for Hand-Held Objects for Visually Challenged People", International Journal of Computer Science and Information Technology Research, Volume. 3, Issue 1, pp: (87-92), Month: January - March 2015

[5] Nanayakkara, S., Shilkrot, R., Yeo, K. P., and Maes, P." EyeRing: a finger-worn input device for seamless interactions with our surroundings". In Augmented Human ,2013.

[6] Carlos Merino-Gracia,, Karel Lenc,Majid Mirmehdi, " A Head-Mounted Device for Recognizing Text in Natural Scenes "Lecture Notes in Computer Science Volume 7139, pp 29-41,2012.

[7] Chengjun Zhu, yuanxin Quyang, Lei Gao, Zhenyong Chen, Zhang Xiong."An Automatic Video Text Detection, Localization and Extraction Approach". School of computer Science and Technology, Beihang University, SITIS - Pages 166-175-2006.

[8] Christin Wolf, Jean-Michel Jolion, and Francoise Chassaing."Text localization, enhancement and binarization in multimedia documents". In Int. Conf. on Pattern recognition, pages 1037–1040, Quebec City, Canada., August 2002.

[9] Egyul Kim, SeongHun Lee, JinHyung Kim. "Scene Text Extraction using Focus of Mobile Camera". In 10^{th} IEEE Int. Conf. on Document Analysis and Recognition, pages 166–170 -2009.

[10] Nirmala Shivananda and P. Nagabhushan, "Separation Foreground Text from Complex Background in Color Images",IEEE Transactions Processing vol.10, 306-0

[11] Fang Liu, Xiang Peng, Tianjiang Wang, Songfeng Lu. A Density Based Approach for Text Extraction in Images[1]. IEEE Transaction, 978-1-4244

[12] Shuicai Shi,Tao Cheng, Shibin Xiao,Xuequiang LV. Text Processing in Video Frames with Complex Background. International Forum on Information Technology and Applications, IEEE pages 450-455,2009.

[13] Datong Chen and Jean-Marc Odobez. "Robust Video Text Segmentaion and Recognition with Multiple Hypotheses", IEEE, Proc of International Conference on Image Processing, vol [2], Pages -433-435,2002.

[14] Jing Zhang, Dimitry Goldgof, R.Kasturi. "A New Edge Based Text Verification Approach for Video", Department of Computer Science and Engineering, University of South Florida.

[15] Datong Chen and Jean-Marc Odobez. "Sequential Monte Carlo Video Text Segmentation", IEEE, Proc of International Conference on Image Processing, vol [1] ,2003

[16] R. Ferzli and L. J. Karam, "A No-Reference Objective Image Sharpness Metric Based on Just-Noticeable Blur and Probability Summation", ICIP'07, pp.445-448, 2007

[17] H. R. Sheikh, A. C. Bovik, L. Cormack. "No-reference quality assessment using natural scene statistics: JPEG2000", IEEE Trans. on Image Processing, 14(11),pp.1918–1927,2005.