Conversion of Image to Text Using Tesseract OCR Engine

1 L. Suriya Kala, 2 Dr. P. Thangaraj
1 Research Scholar, Mother Teresa Women’s University, Kodaikanal, Tamilnadu, India
2 Computer Science and Engineering, Bannari Amman Institute of Technology, Sathiyamangalam, Tamilnadu, India

Abstract - This proposal is to create a solid way to convert the un-modifiable document images into a more flexible text format here the system will accept the raw scanned image of a document and produce the desired text which can be used in any way. To do that we need a strong OCR engine which will do that for us now using the traditional way which is algorithms are time consuming and also need expert knowledge but in this system the whole process is aimed to benefit the novice personalities who are working in common institutions.

Keywords : OCR, EmguCV, tessdata

I. CONCEPT
To get the document which is a hard copy of the data that is collected over a period and then by using the system convert it into a soft copy of the documents not only they are converted into a more easy to use format they must also be manipulated, thus going one step by converting the scanned documents into a raw and editable text format which can be edited, copied or deleted

II. COMPONENTS
To do all of these that are said to be present in the system we need the following components

III. WORKING
How the system will work to produce the expected outcome is the major question here the front end is designed in a GUI based environment and the back end is also a stable one which can perform very well in given tasks, the thing is we have the scanned document, user friendly front and solid back end but how to connect all of this, here in this system the front end is embedded with the tessdata which is a package that comes with the Emgu CV. The tessdata contains the procedure and reference to the character of the language which is selected to be converted, when the system is launched the instance of the language is initiated in the tessdata.
Finally all that left is to feed the input the scanned document into the system; it is done by using the read file function in the visual studio. It provides a way to bring the image into the system to be analyzed and converted into text; whenever the image is given the OCR engine will analyze it and produce the identified characters in a text format.

But still there is one problem is not solved which is the quality of the input, what if the scanned image is not the correct size or not a good quality image. The size of the images must be ideal to produce a good result, to do that the image is preprocessed to fit the range of the ideal image. But as for the quality of the image is concerned it is all in the hands of the user, if you provide a moderate level of image the result will also be moderate.

IV. OUTPUT

The output is produced in the well know text format so it can be used in any way possible most of the times in every field there are enormous amount of data that are collected for a long period of time it is in a old note book format, But it can’t kept safe for a long time and moreover using it to find out some details is nearly impossible so to use those details in a effective way it has to be digitalized, here comes the hardest part the process of digitalizing the data is not that easy it will require an enormous amount of time and manpower even with all of that it is tedious work so the proposed system is the easiest way to do all of that.

V. CONCLUSION

The system is aimed to help those who need the assistance of technology in converting or digitalizing the documents but it is only a sample of it is uses, in a wide range it can also be used to ID the number plates or the lettering in the vehicles to solve some complicated cases in the crime field. If possible a translator can be embedded to convert the text into speech. Well the reach of this proposed system is in the view of the end users.

AUTHORS PROFILE

L. Suriya Kala had completed MCA., M.Phil and working as Assistant professor in Don Bosco College, Dharmapuri, Tamil Nadu India. She is a research Scholar in the specialization of Digital Image processing at Mother Teresa Women’s University Kodaikannal, Tamil Nadu, India.  
Dr P. Thangaraj Ph.D., Professor and Head ,Department of Computer Science and Engineering Banuari Amman Institute of Technology, Sathiyamangalam, Tamil Nadu, India.