ISSN : 2231-0711

L.Suriya Kala et al | International Journal of Computer Science Engineering and Technology(IJCSET) | Dec 2014 | Vol 4, Issue 12,347-348

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

Fig. 1. The basic concept of the proposed system

To do all of this we must have a stable method which will assure a good result, but by adding all these to the agenda now the final product will become somewhat difficult to create and more complex to be used by a novice person. A product is successful when the expected outcome is produced and more importantly the customer is satisfied. And this product will do the both by producing a simple and user friend front end and a good and solid Back-End which will meet all the needs from the customer.

II. COMPONENTS

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

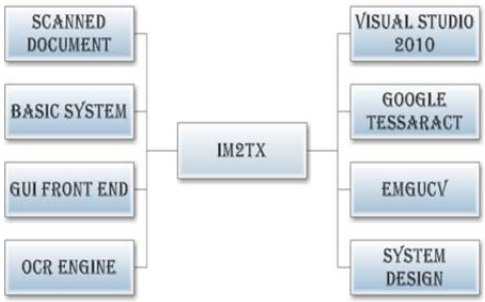

Fig. 2. The components need and used in this system

The basic scanned document

- A well power OCR Engine
- Good and user friendly front end
- A solid connection between front end and back end
- Easy to use functionalities

Now all of this can be achieved by the only one way which is the windows based IDE, even though the there are numerous amount of applications that are available in worldwide we need the one which can support the OCR engine, because front end designing is only to those who don't have the extreme knowledge in the OCR field but the back end is as complex as any other methodologies that are used to do the job.

Taking all of this into concern we have decided to use the *visual basic studio* which provides us with the windows based front end designing and also supports the OCR engine the newer version are not stable and compatible with the OCR engine the stable version *visual studio ultimate 2010* is selected as the IDE but is means that only a 25% of the job is done the major 75% is still waiting.

There are several ways to do the OCR function by using some well known products such as MATLAB, but the only problem is that they are too costly and very hard to use, for someone who doesn't have the mathematical knowledge up to the grade and they also need long period of time to be created by using algorithms and Fourier series etc.

So another option is to use the open source Google project tesseract which is available for *n* number of languages, now the new problem arises, how to use this tesseract in our system. That is where the *EmguCV* (*Open Source*) comes in. It provide a way to embed the tesseract into the system and it is also compatible with the *visual studio ultimate 2010*, so the problems are solved now all that is left is to design a system which can do all of these jobs.

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 *EmguCV* . 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.

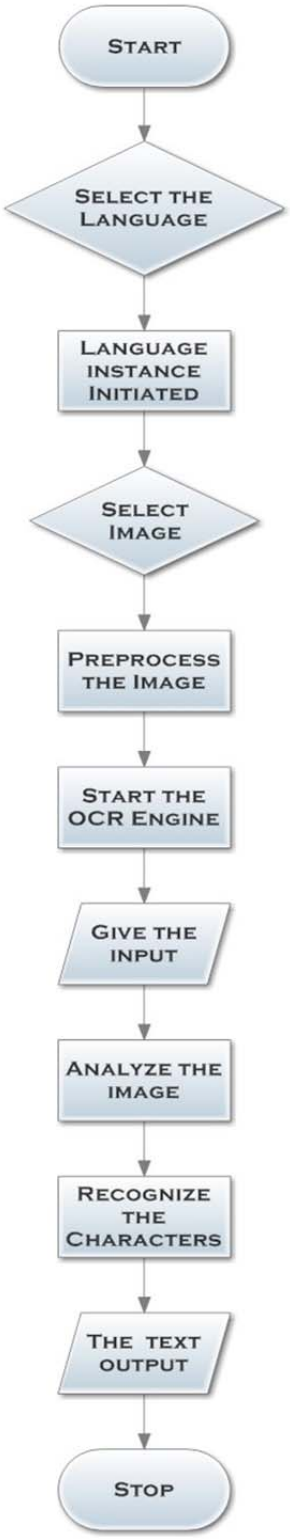

Fig 3.The flow of data and work in the proposed system

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.

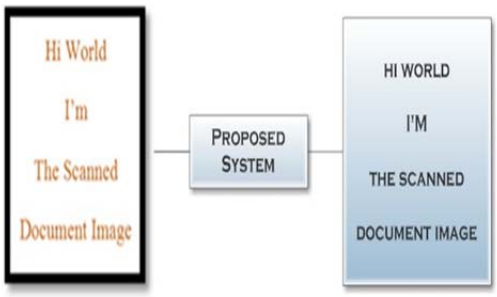

Fig. 4. The basic input and output

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 Bannari Amman Institute of Technology, Sathiyamangalam, TamilNadu, India.