Automatic Text Recognition Using Optical Character Recognition Technique

P.Jayachandar M.E., Swetha S, Sneka J, Suryaprakash SM, Sowmiya R
Department of Electronics and Communication Engineering, Velalar College ofEngineering and Technology

ABSTRACT - Automated data extraction may significantly improve the accuracy and efficiency of all different business operations. Guide records access, which can be a time-consuming and error-prone process, is routinely required to feed business enterprise data structures with information from files. In light of this, automating information access can different tempos and times. Executing company organization procedures enables staff to concentrate on the more important and fundamental components of their daily sporting activity. Such files are routinely translated to text format after being digitally captured as images. The text popularity mission for identity papers is addressed in this study. we discovered a text recognition method that primarily relies on transfer learning and scene text popularity (STR) networks. Text is recognized even it is not present in training set images. Take less time even the image size is huge. Lagging while working with high definition images is ignored here.

I.INTRODUCTION

The technology known as optical character recognition (OCR) enables a system to recognize the scripts or alphabets embedded in the customers' vocal communication without the need for human participation. Optical person recognition has developed into one of themost successful uses of knowledge are in the areas of artificial intelligence and sample detection. In our survey, we looked into the various OCR methods. In this study, we examine and solve the fictitious and numerical methods of optical identification of a person. The recognition of patterns or alphabets is typically achieved using optical character recognition (OCR) and magnetic character recognition (MCR) techniques.The alphabets in the standard are expanded into pixel pix and can be written or stamped in any series, shape, or combination of these and other things. Instead, with OCR, the alphabets are stamped using magnetic ink, and the learning tool organises the alphabets according to the distinctive magnetic field each alphabet creates. Banking and specialised exchange appliances use both OCR and Optimized OCR. There are no obstacles to the script approach in handwritten text, according to earlier research on optical person detection or reputation. Due to the variety of human handwriting styles and the differences in calligraphy's perspective, size, and form, handwritten letters can be challenging to read and understand.This article discusses a variety of optical person identification techniques from conception to completion. One of the most captivating and traumatizing facets of pattern reputation, with a proliferation of practical applications, is optical man's or woman's reputation. The examples given above enable us to recognize that OCR knowledge has been developed by numerous researchers over a lengthy period of time, unconditionally comprising of an excellent global human research network.

Humans have made efforts to advance research through "antagonism and collaboration" in such an unnoticeable discourse. Worldwide symposiums and inductions are being scheduled in this way to encourage advancement in the field.For instance, the global conversation on article psychoanalysisandpopularity determination and the worldwide induction on Frontiers in Handwriting Detection both perform important roles in the intellectual and rely-of-fact sector.

II.LITERATURE SURVEY

FilippoAttivissimo, Given that this task still requires using and is frequently completed manually, wasting money and time, Documents Identity automatically analyse and confirm is a desirable invention for today's business industry. This follows the pattern of an individual automatic document analyser that distributes identification. The system is thought to be able to extract information on the identity of the most popular Italian document from photos of the appropriate quality, exactly like those that are frequently required of online subscribers to various services.

Jeonghun Baek1 Geewook Kim2∗Junyeop Lee1 Sungrae Park1, This essay offers three crucial contributions that address this issue. We start by examining the discrepancies in the datasets used for education and evaluation, as well as their effects on overall performance. The majority of the current STR styles fit into our unifying four-degree STR framework, which we introduce in the second step. This framework enables the creation of hitherto undiscovered

module combinations as well as the extensive review of previously proposed STR modules. Third, we look at how well-designed modules contribute to overall performance in terms of speed, accuracy, and memory demand, using a typical set of training and assessment datasets. Those evaluations eliminate the difficulty in identifying the overall performance advantage of the current modules in comparisons made today.

YoungminBaek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee∗Clova AI Research,we propose a novel scene textual content identification method to efficiently locate textual content by examining each character and their relationships to one another. Our suggested framework uses both the existing character-level annotations for synthetic images as well as the anticipated individual-degree floor truths for real images that were discovered through the discovered period inbetween version to overcome the lack of character individual degree annotations.

The community is familiar with the recently proposed graphic for affinity, which can be used to gauge character affinity. Comprehensive tests on six benchmarks, as well as the TotalText and CTW-1500 datasets that contain highly curved texts in botanical images, show that our individual-level text content detection works noticeably better than the leading detectors.

III. PROPOSED WORK

The proposed system effectively extracts the text from the image file using an optimal optical character recognition technique. This improved method improves text recognition accuracy over the previous one. To work with high resolution photos without lagging to produce the output, pixel-based comparison is employed. In order for the text prediction from image file to be more accurate than the prior system.Even with a large image size, it takes less time. A technique for optimised OCR is created.

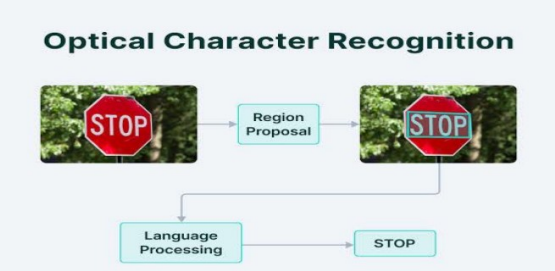


Figure 1 Image Acquisition

In image acquisition, the pattern pictures are accumulated, which might be required to educate the classifier set of rules and assemble the classifier model. Images have been taken at unique angles, showing the unique environmental and lights. The preferred JPG and PNG layout are used to shop those pics. In this study, pictures had been accumulated from farms in unique font types.

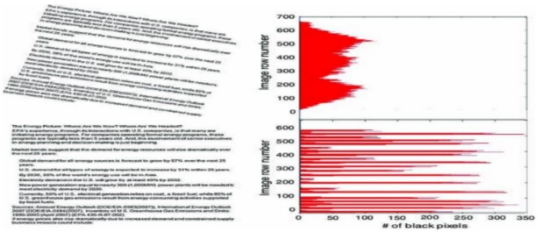


Figure 2 Image Preprocessing

The original photographs have all been kept in a single folder. The pictures were given names by us, and we are open to any number of prices. Just horizontal photographs were resized using 200x300 pixels and grew to be spherical using a method of 90 levels.

Vertical photos should be 200x300 pixels, and since the image graph's breadth and peak are equal, those images were downsized to 250x250 pixels. When the length of the picture graph is unquestionably excessive, processing tasks take longer. After that, one of the noise reduction techniques was applied to reduce noise in photographs and improve their clarity.

The preprocessed images were later placed in a folder.

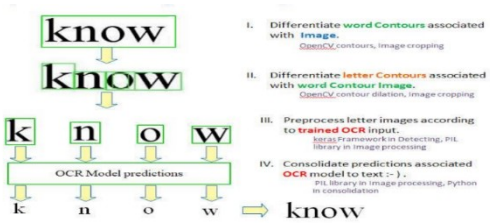


Figure 3 Image Segmentation

Photo segmentation makes about 1/3 of the method. All preprocessed pictures were first stored in their original format and converted to grayscale. So one of the outcomes of this research is the identification of the optimal shade version for preprocessing. The image is then converted to a binary layout. The OCR approach had been used to cluster these layout values. A picture segmentation and its associated set of rules have been carried out in this step.

3.1 METHODOLOGY

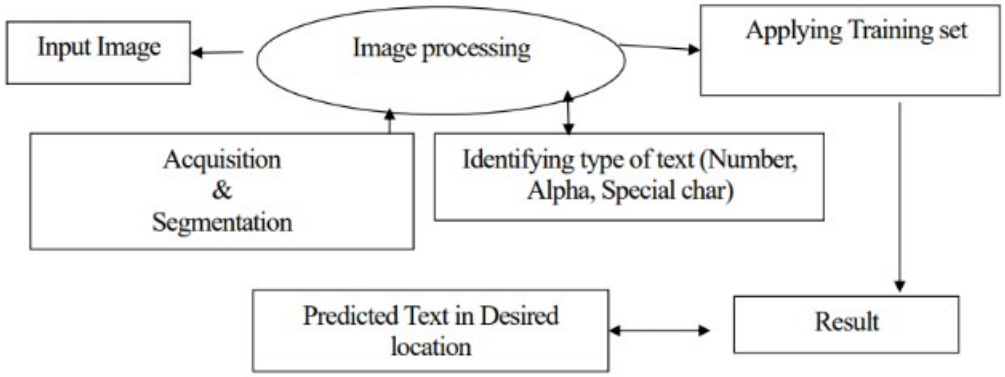


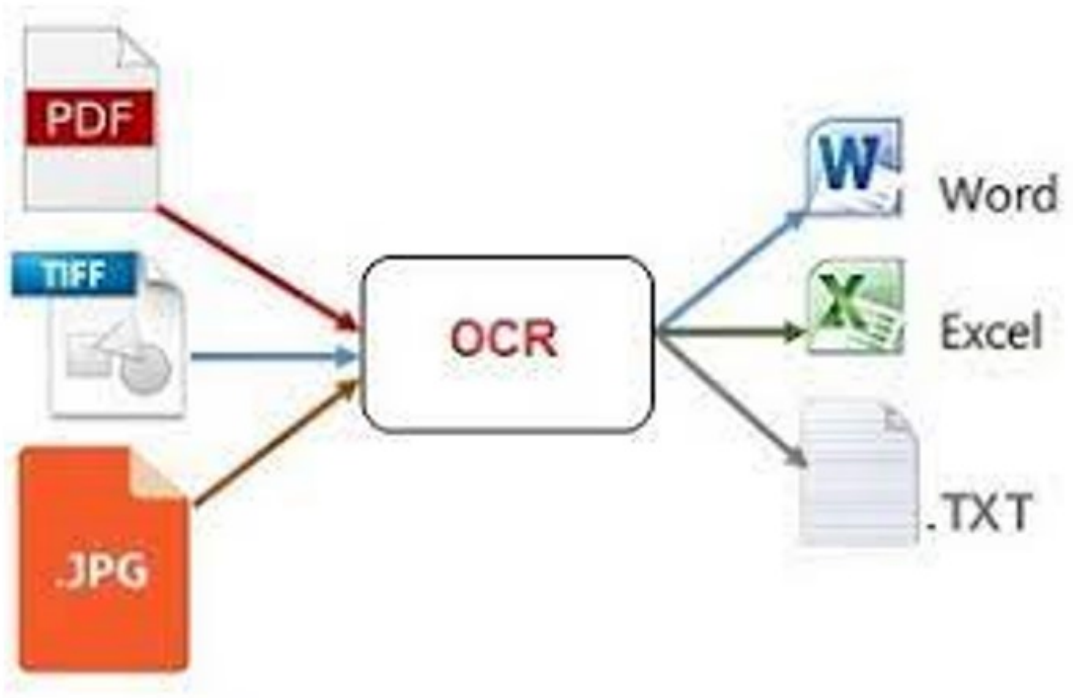
Figure 4 Steps Involved

Applying Training set pictures are used in this stage of the process. Execution of the segmented output, which was produced via characteristic extraction, has been completed. To conduct tests, photo sets have nevertheless been made. The education of those picture units is mentioned right here. For the purpose of categorizing the snapshots, a field information guide was used, and each photo was selected at random from the categorized units of a picture.

Optical character recognition (OCR) is a technique for identifying text in images, including images and files that have been subjected to analysis. Because taking pictures of text takes much less time than taking notes, we should have done so instead of taking notes or typing the text because we are too lazy to do either.

Figure 5 Diagram of the OCR

Fortunately, on today's smartphones, we can immediately observe OCR, allowing us to instantly replicate the textual content we took earlier without having to jot it down or retype it. In Python, we are able to try this too, simply through the usage of some strains of code. One of the pieces of OCR equipment that can be regularly exploited in Tesseract. Tesseract is an optical Character reputation mechanism for diverse running structure. It was initially advanced as proprietary software by Hewlett-Packard. Later, Google took over improvement.



In this way, global symposiums and inductions are being determined to stimulate the improvement in the domain. For example the global induction on Frontiers in Handwriting detection and the International discussion on article psychoanalysis and Recognition determination play an explanation task in the intellectual and matter-of-fact arena.

It is a common practice to digitize printed texts so they can be electronically edited, searched, stored more compactly, displayed online, and used in machine processes like cognitive computing, machine translation, (extracted) text-to-speech. This method of digitizing printed texts is used to enter data from printed paper data records such as passports, invoices, bank statements, computerized receipts, business cards, mail, printouts of static-data, or any suitable documentation. Pattern recognition, artificial intelligence, and computer vision are all areas of study in OCR.

Early versions needed to be trained with images of each character, and worked on one font at a time. Today, sophisticated systems with support for a range of digital image file formats inputs and a high degree of recognition accuracy for the majority of fonts are the norm. [2] Some systems are capable of reproducing formatted output that closely approximates the original page including images, columns, and other non-textual components.

3.2 INSTALLATION

Tesseract is currently functional on the macOS, Windows, and Linux platforms. Tesseract supports more than 100 languages and (UTF-8) Unicode. In this, we are capable to begin with the Tesseract OCR setup method and check \sthe removal of text from pics. First step is to position it inside the Tesseract. We must place the Tesseract library on our system before we can use it. If you are using Ubuntu, you can sincerely use apt-get. I was given to install Tesseract OCR: python –m pip install tesseract-OCR (or) pip install tesseract-OCR. For macOS users can use Homebrew to install Tesseract. Brew install tesseract.

3.3 IMPLEMENTATION

After the installation is complete, let's go back in time by using Python and Tesseract.

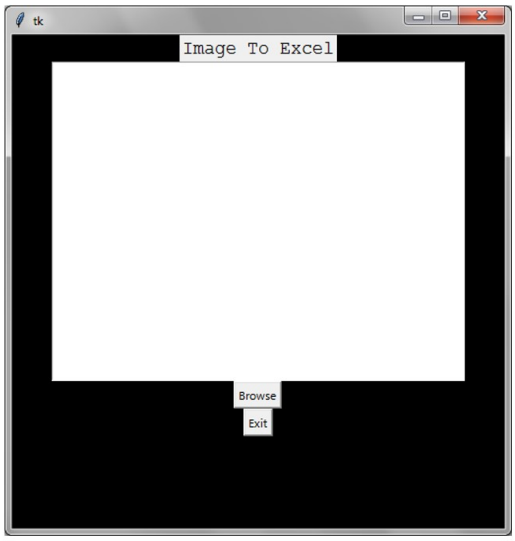
- We import the dependencies first.
- Import Tesseract and Image from PIL
- Import NumPy as NP

3.4 SAMPLE CODE

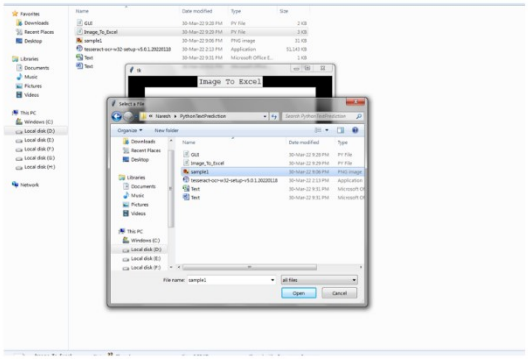
```
Import pytesseract
Import glob
Import os
Import cv2

path_="D:/Naresh/Image_To_String/2.png"

Source=path_
Img = cv2.imread (source)
cv2.imshow ("Input image", img)
file_path = ""
NP_list = []
predicted_NP = []

Forfile_path in glob. Glob (file path, recursive = True):

NP_file = file_path.split ("/") [-1]
Number chars, _= os.path.splitext (NP_file)

NP_list.append (number chars)

    '''
    Reading each characteres in image file using openCV-*
    '''
NP_img = cv2.imread (file path)

    '''
    We will then pass each character file
    To the Tesseract OCR engine utilizing the Python library
    Wrapper for it. We get back predicted char for
    String. We append the predicted char in a
    List and compare it with the characters shape
    '''
predicted_res = pytesseract.image_to_string (NP_img, Lang ='eng',
Config ='--oem 3 --psm 6 -c tessedit_char_whitelist = ABCDEFGHIJKLMNOPQRSTUVWXYZ0123456789')

filter_predicted_res = "".join (predicted_res.split ()).replace (":", "").replace ("-", "")
predicted_NP.append (filter_predicted_res)


defestimate_predicted_accuracy (ori_list, pre_list):

forori_chars, pre_chars in zip (ori_list, pre_list):
Acc = "0 %"
number_matches = 0
        If ori_chars == pre_chars:
Acc = "100 %"
        Else:
            If len (ori_chars) == len (pre_chars):
For o, p in zip (ori_chars, pre_chars):
                    If o == p:
number_matches += 1
```

```
Acc = str (round ((number matches / len(ori_chars)), 2) * 100)
Acc += "%"
forori_chars, pre_chars in zip (ori_list, pre_list):
Acc = "0 %"
number_matches = 0
        If ori_chars == pre_chars:
Acc = "100 %"
        Else:
            if len(ori_chars) == len(pre_chars):
For o, p in zip (ori_chars, pre_chars):
                    If o == p:
number_matches += 1
Acc = str (round ((number matches / len(ori_chars)), 2) * 100)  acc += "%" pytesseract.pytesseract.tesseract_cmd =
r'
C:\Program Files\Tesseract-OCR\tesseract'
    s=pytesseract.image_to_string (path_)
    Print
f=open ("d:/Naresh/Image_To_String/Text.doc","w")
Write
Print ("\n\these contents are saved to Text.doc in cuurent working directory...!")
Close ()
estimate_predicted_accuracy (NP_list, predicted_NP)
```

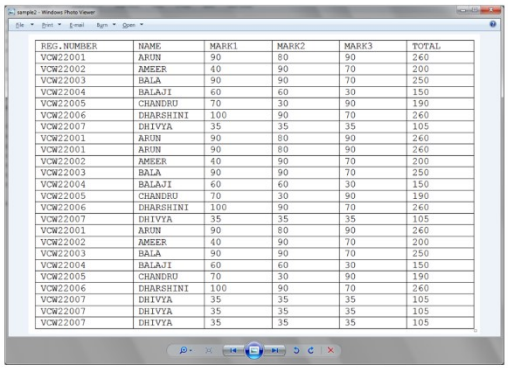
3.5 SAMPLE SCREENS



Browse To Choose the Image File

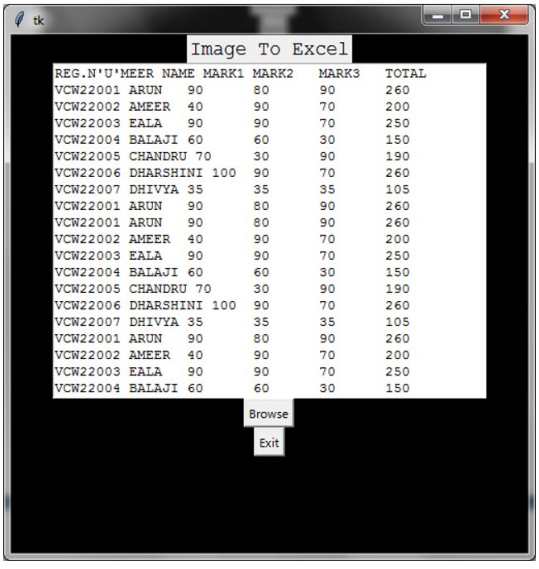


Input Image File



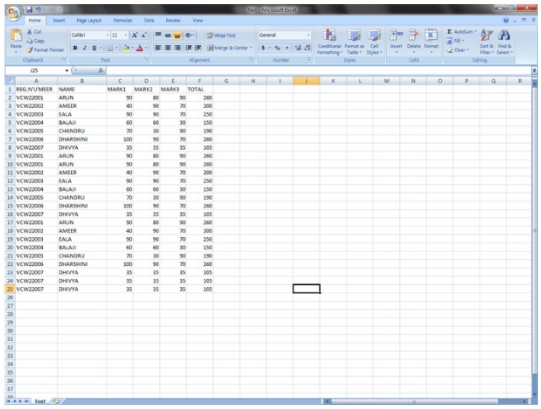
REG.NUMBER	NAME	MARK1	MARK2	MARK3	TOTAL
VCW22001	ARUN	90	80	90	260
VCW22002	AMEER	40	90	70	200
VCW22003	BALA	90	90	70	250
VCW22004	BALAJI	60	60	30	150
VCW22005	CHANDRU	70	30	90	190
VCW22006	DHARSHINI	100	90	70	260
VCW22007	DHIVYA	35	35	35	105
VCW22001	ARUN	90	80	90	260
VCW22001	ARUN	90	80	90	260
VCW22002	AMEER	40	90	70	200
VCW22003	BALA	90	90	70	250
VCW22004	BALAJI	60	60	30	150
VCW22005	CHANDRU	70	30	90	190
VCW22006	DHARSHINI	100	90	70	260
VCW22007	DHIVYA	35	35	35	105
VCW22001	ARUN	90	80	90	260
VCW22002	AMEER	40	90	70	200
VCW22003	BALA	90	90	70	250
VCW22004	BALAJI	60	60	30	150
VCW22005	CHANDRU	70	30	90	190
VCW22006	DHARSHINI	100	90	70	260
VCW22007	DHIVYA	35	35	35	105
VCW22007	DHIVYA	35	35	35	105
VCW22007	DHIVYA	35	35	35	105

Predicted Text from Image File



3.6 EXPERIMENTAL RESULTS

After using the education set images, one word record has been used for storing the extracted contents. These files are referred to as "word record format". This technique is referred to as the content material extraction technique. Rows of education files had been randomly shuffled at every and tryout time to enhance the model's accuracy. Each schooling file changed into installed and tested in 5 instances, and accuracy changed into taken. The common of those accuracies modified due to the precision of each model. All kinds of content material had been found.



IV.CONCLUSION

We give an appraisal of the presentationof the optical letter set test. In this review, weexamined the hypothetical and numericalforms of the most extreme hard difficulty inside the extent of optical letter setcharacter, which is changed through scale, decipher, and pivot in optical letter set recognition. The potential arrangements of the OCR methods are additionally examined. The accuracy also, personality aren't sufficient for pragmatic organization. It can likewise furthermore require a decent estimated improvement.

REFERENCES

[1] Ayush Wattal, AshuthoshOjha, Manoj Kumar "Impediment Recognition Belt for Outwardly Hindered Utilizing Raspberry Pi and Ultrasonic Sensors" Division of Data Innovation JSSATE, Noida, India. Public Meeting on Item Plan (NCPD 2016), July 2016.

[2] Amy Nordrum, Title: Pothole discovery for blind, IEEE, 30 May, 2016.

[3] G. Balakrishnan, G. Sainarayanan, R. Nagarajan what's more, S. Yaacob, Wearable Continuous Sound system Vision for the Outwardly Impaired, Engineering Letters, vol.14, no. 2, 2007.

[4] DuraisamySathya and Pugalendhi Ganesh Kumar,'Gotten Distant Wellbeing Checking Framework, IET Medicalservices Innovation Letters', vol. 4, issue. 6, pp. 228-232, 2017.

[5] Sudeep Gupta, IlikaSharma,AishwaryaTiwariandwhat's more, GauravChitranshi "High level Aide Stick forthe Outwardly Hindered Individuals" in Proc. FirstGlobal Meeting on FutureFiguring Innovations, India, 4-5 September2015.

a. Dodds, D. Clark-Carter, and C. Howarth, Thesonic PathFinder: an assessment, Diary of VisualDebilitation and Visual impairment, vol. 78, no. 5, pp. 206-207, 1984.

[6] Gbenga, D. E., Shani, A. I., and Adekunle, A. L. (2017). Shrewd strolling stick for outwardly disabledindividuals utilizing ultrasonic sensors and Arduino.Worldwide diary of designing andinnovation, 9(5), 3435-3447.

[7] Olanrewaju, R. F., Radzi, M. L. A. M., and Recovery, M. (2017, November).iWalk: Astutestrollingstick for outwardly debilitated subjects. In 2017 IEEEfourth Worldwide Meeting on SavvyInstrumentation, Estimation andApplication (ICSIMA) (pp. 1-4). IEEE.

[8] Ikbal, M. A., Rahman, F., and Kabir, M. H. (2018, September).Microcontroller based shrewd strolling stick for outwardly debilitated individuals.

[9] C.Nagarajan and M.Madheswaran - 'Experimental verification and stability state space analysis of CLL-T Series Parallel Resonant Converter' - *Journal of ELECTRICAL ENGINEERING,* Vol.63 (6), pp.365-372, Dec.2012.

[10] C.Nagarajan and M.Madheswaran - 'Performance Analysis of LCL-T Resonant Converter with Fuzzy/PID Using State Space Analysis'- *Springer, Electrical Engineering*, Vol.93 (3), pp.167-178, September 2011.

[11] C.Nagarajan and M.Madheswaran - 'Stability Analysis of Series Parallel Resonant Converter with Fuzzy Logic Controller Using State Space Techniques'- *Taylor & Francis, Electric Power Components and Systems*, Vol.39 (8), pp.780-793, May 2011.

[12] Nagarajan and M.Madheswaran - 'Experimental Study and steady state stability analysis of CLL-T Series Parallel Resonant Converter with Fuzzy controller using State Space Analysis'- *Iranian Journal of Electrical & Electronic Engineering*, Vol.8 (3), pp.259-267, September 2012. G.Neelakrishnan, K.Anandhakumar, A.Prathap, S.Prakash "Performance Estimation of cascaded h-bridge MLI for HEV using SVPWM" Suraj Punj Journal for Multidisciplinary Research, 2021, Volume 11, Issue 4, pp:750-756

[13] G.Neelakrishnan, S.N.Pruthika, P.T.Shalini, S.Soniya, "Perfromance Investigation of T-Source Inverter fed with Solar Cell" Suraj Punj Journal for Multidisciplinary Research, 2021, Volume 11, Issue 4, pp:744-749