|  |  |
| --- | --- |
| Internship Project Title | **DINESH KUMAR.R** |
| Project Title | **Automate extraction of handwritten text from an image** |
| Name of the Company | **TCS iON** |
| Name of the Industry Mentor | **Debashis Roy** |
| Name of the Institute | **SRM INSTITUTE OF SCIENCE OF TECHNOLOGY** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Start Date | End Date | Total Effort (hrs.) | Project Environment | Tools used |
| 02/07/2020 | 12/08/2020 | 125 | google colab | * google drive * iam online dataset  pytesseract code from [Alexander](https://launchpad.net/~alex-p) [Pozdnyakov](https://launchpad.net/~alex-p) |
| Project Synopsis:  Optical Character Recognition (OCR) Systems aim to recognize text and bring it to editable form from the given document image, where the input text can be in machine printed, hand written or hand printed form. Many recognition systems have been developed for languages based on various scripts and digits all over the world, taking input in either of the online and offline modes, with considerable efficiencies. These systems have proved to be highly applicable in the fields of Banking, Education, IT systems and Postal Sector for digitization of processes and automated information retrieval. | | | | |
| Solution Approach:  the solution approach for this project is done with the   * Inverting images. * Rescaling. * Binarisation. * Noise removal. * Rotation/Deskewing. * Edge removal.   All these operations can be accomplished using OpenCV and/or numpy operations in python.  So all in all this story be the set of problems and the possible solutions related to OCR and HTR; implementing it is something you will have to look out for. I have implemented the whole of it; | | | | |
| Assumptions:   * The dataset provided contains the same language words as the input text’s language * The dataset i have used is an benchmark (well recognised) one. * The noise reduction is done the dataset which means the junk images. | | | | |
| Project Diagrams: | | | | |
| Algorithms:  1) ALGORITHM FOR ANN(ARTIFICIAL NEURAL NETWORK) BASED OCR:   * 1. Start with the input of a document. * 2. Formation of the network. * 3. Initialize the weights. * 4. Load the trainer set files into the machine. * 5. Start analysis of image * 6. Detect the characters   If(symbol available)  Goto step 5  Else  Read desired output text  If (character available)  Read the output text  Else Goto step 7   * 7. Calculate the network output * 8. Compute the errors and update the weights If(errors <= threshold)   End of process  Else if(Iteration= maximum epochs)  Goto step7  Else  End of process  2) DATA FLOW FOR ANN BASED ON OCR:   * Step 1: Input the document into Optical Character Recognizer. In this step input the hard document to convert into the soft. * Step 2: Scan the image to convert. * Step 3: Train the network for the appropriate working. In this we train our network using the concept of Artificial Neural Network using backpropagation algorithm of training. * Step 4: Load the image into the network. In this we load the scanned image into the tool which we developed to process it. * Step 5: Identify the lines present in the image. * Step 6: Detection of characters presented in the image. In this we detect the characters present in the document and match it with our database i.e. Unicode for which we trained our Artificial Neural Network. * Step 7: Map the character image into pixel matrix. If the match is found we map those characters into pixel matrix to show their existence. * Step 8: Get the output document which is in soft and editable format. | | | | |
| Outcome:  The outcome the project is :   * To extract the text from the given input image file . * To display the extracted text from that image with accuracy of 0.7-0.8 . | | | | |
| Exceptions considered:   * The accuracy considered is around 0.75 - 0.80 so more than that is model cant predict * The language considered for the text extraction is english other than that we cant use any other language. * The image should be captured properly without any disturbances. | | | | |
| Enhancement Scope:  The application of extraction of handwritten text from an image algorithm is extensive. Now-  a-days recent advancement in technologies has pushed the limits further for man to get rid of older  equipment which posed inconvenience in using. In our case that equipment is a keyboard. There are  many situations when using a keyboard is cumbersome like, | | | | |
| Link to Code and executable file:  <https://colab.research.google.com/drive/1KkndnD2DrTb7KBD1SCe3yAdepfNoM-wR?usp=sharing> | | | | |