Application of Image Processing and Convolution Networks in Intelligent Character Recognition for Digitized Forms Processing

Shailendra Singh Kathait
Co-Founder & Head of Analytics
Valiance Solutions Pvt. Ltd.
Noida, Uttar Pradesh

Shubhrita Tiwari
Data Scientist
Valiance Solutions Pvt. Ltd.
Noida, Uttar Pradesh

ABSTRACT
Image processing is a rapidly evolving field with immense significance in science and engineering. One of the latest applications of Image processing is in Intelligent Character Recognition (ICR), that is the computer translation of handwritten text into machine-readable and machine-editable characters. ICR is an advanced version of Optical Character Recognition system that allows fonts and different styles of handwriting to be recognized during processing with high accuracy and speed. ICR, in combination with OCR and OMR (Optical Mark Recognition), is used in forms processing. Forms processing is a process by which one can capture information entered into different data fields filled in forms and convert it to an editable text. Forms processing systems can range from the processing of small application forms to large scale survey forms with multiple pages. The Recognition Engine, designed using Image Processing and Convolution Networks helps save time, labor and money in addition to the increase of accuracy.

Keywords
Image Processing, Intelligent Character Recognition, Optical Character Recognition, Optical Mark Recognition, Recognition Engine, Convolution Networks.

1. INTRODUCTION
Image processing is widely used nowadays to get insights from Image Data. Large abundance of Image Data present everywhere demands for analysis of this data. Using Image Processing techniques, different models can be developed to automate different processes.

One such application of Image Processing is in Intelligent Character Recognition which is automated extraction of data from handwritten forms in scanned jpg/png/tif format.

A Recognition Tool is developed that takes a scanned form as input, applies pre-processing techniques to extract handwritten characters and then uses Trained Convolutional Neural Network Models for recognition of these characters and then finally writes the extracted data to MYSQL database.

With ICR technology, the text is directly entered to database post classifying all the segments in whole document after doing proper character recognition through OCR. ICR operates by capturing handwritten text from image files and converting them into text searchable files thereby giving users the ability to search through the files with text strings and capture information from them by using the copy/paste function. Using ICR, there is no longer a need for multiple people to enter information to get the job done. This eliminates human errors.

This recognition engine improves the interaction between man and machine in many applications like mail sorting, cheques verification, office automation and a large variety of banking applications like accounts and credit card data, questionnaires, insurance claims etc. Form types, in use at most companies, include orders, applications, claims, change requests and survey forms received from customers as well as internally generated forms such as expense claims, request forms, time sheets and HR records. The data on these forms is often business critical so if it can be captured and transferred to internal systems quickly and accurately it is a huge benefit to the organization. At almost every firm, data is recorded in the forms filled by humans.

2. LITERATURE REVIEW
Different approaches have been used for forms processing and text recognition. The character recognition methods vary with the way the pixels of image are visualized and processed further. The different approaches can be statistical, semantic, neural network, pattern recognition etc.[1]

Different levels of accuracy are achieved by different softwares developed using different types of machine learning models.The tool developed gives an average accuracy of 90 percent.

A feedback loop is inserted into the tool that generates feedback of the characters recognized incorrectly which is corrected again by the user and then given to the model which gets retrained accordingly.

3. PROPOSED METHOD

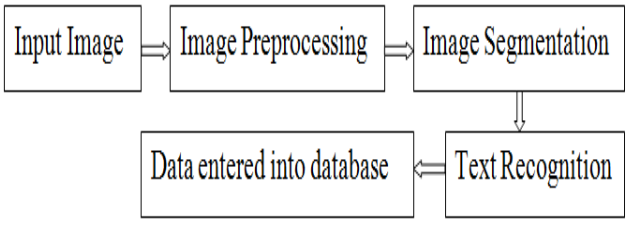


Fig 1: Proposed Methodology

The very first step in the entire recognition process is Image processing, that implements segmentation methods and extract the required fields from the input for recognition.

3.1 Image Processing

The input given to the recognition engine is a scanned image in jpg/png/tif format, where scanning needs to be done at good resolution (300 dpi) for maximum accuracy. The root step is the pre-processing of this image.

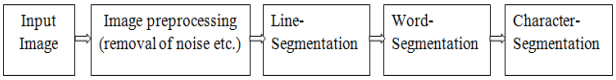


Fig 2: Image Processing Methodology

3.1.1 Image Preprocessing

Pre-processing of image involves removal of noise from image. It is a common name for operations with images at the lowest level of abstraction, both input and output of which are intensity images. The aim of pre-processing is an improvement of the image data that suppresses unwanted distortions or enhances some image features important for further processing. Different categories of image pre-processing methods exist according to the size of the pixel neighbourhood that is used for the calculation of a new pixel brightness: pixel brightness transformations, geometric transformations, pre-processing methods that use a local neighbourhood of the processed pixel, and image restoration that requires knowledge about the entire image.

The different pre-processing filters (median, symmetric etc.) can be used to improve the quality of the image and reduce distortion.After the pre-processing operations have been performed on the image, it is now ready for processing further for segmentation.

Forms are filled and mailed from all over. As a result they are received folded and are often dogeared and smudged. Moreover, use of stapling pins, paper clips etc. introduces a lot of noise in the form image. Due to this and given that a large number of forms have to be processed in a short time, the form registration algorithm needs to be robust to noise and highly efficient. The form registration also has to be very accurate since the accuracy of the field image extraction and hence field recognition depends on it.[2] When a document is scanned, some amount of skew inevitably occurs in the document. This poses a problem for the storage and analysis of these documents. To compensate, the skew is detected and corrected using skew-detection algorithm. One method of removal of noise is removal of border and extraction of a rectangular portion of the input image that contains relevant information to be extracted. After this required skew and shift correction is applied for obtaining image in a form that can be processed directly. Other noise removal methods include Erosion and Dilution of Image. For example, for removal of salt and pepper noise, median filter is used. Median filter replaces the value of a pixel by the median of gray levels in the neighborhood of that pixel. Depending upon the type of input image, different noise removal techniques are implemented. In order to remove all the noise, a rectangular portion of the input image is bordered and extracted for further processing.

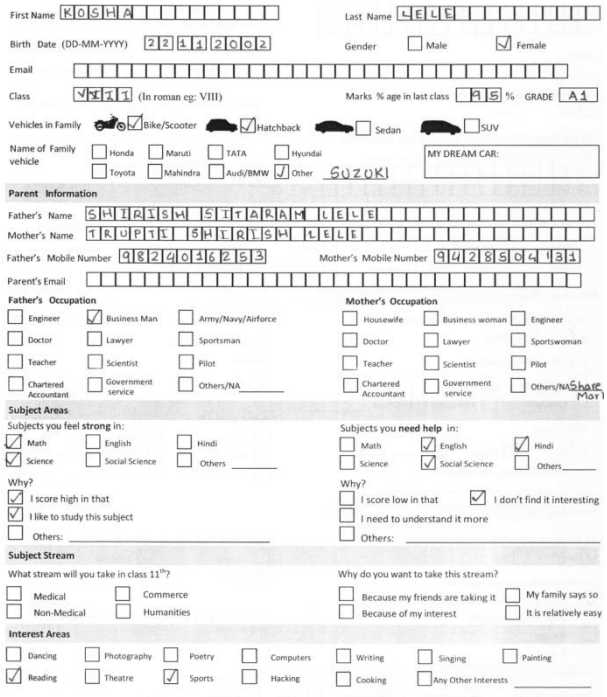


Fig 3: Input Image

3.1.2 Image Segmentation

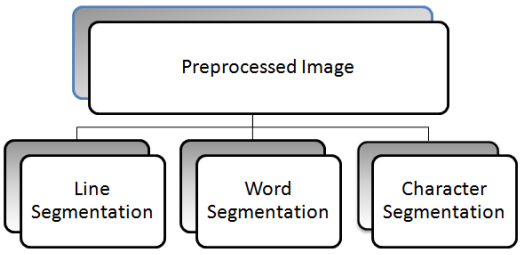


Fig 4: Image Segmentation

The division of an entire image into individual characters, image segmentation, is the most essential step in Recognition engine.

The following categories of segmentation can be used:

1. **Threshold based segmentation:** Histogram thresholding and slicing techniques are used to segment the image. They may be applied directly to an image, but can also be combined with pre- and post-processing techniques.
2. **Edge based segmentation:** With this technique, detected edges in an image are assumed to represent object boundaries, and used to identify these objects
3. **Region based segmentation:** Where an edge based technique may attempt to find the object boundaries and then locate the object itself by filling them in, a region based technique takes the opposite approach, by starting in the middle of an object and then

"growing" outward until it meets the object boundaries.[3]

The images that are taken as input, consist of 2 types of fields:

One that is filled with alphabets and numbers in square boxes and the other one that is filled with right and wrong marks in check boxes.

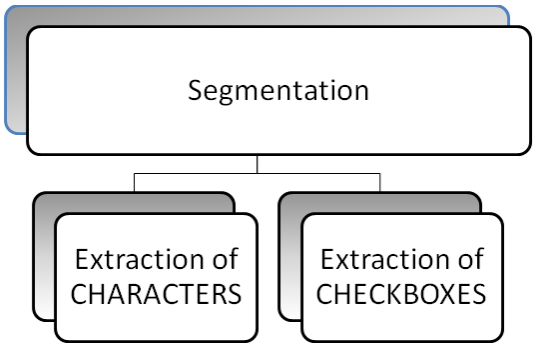


Fig 5: Extraction Process

Therefore their segmentation is done in 2 stages:

- **Extraction of Characters(alphabets and numbers):** Segmentation involves extraction of characters from image and obtaining their matrix structure. Extraction of words is achieved through Image Segmentation. Segmentation subdivides an image into its constituent regions or objects. It includes Line-segmentation (extraction of lines from image) followed by Word-segmentation (extraction of words from lines) followed by Character-segmentation (extraction of characters from words). The concept of different segmentation techniques like Marr-Hildreth Edge Detector, Canny edge detector, Edge linking, Boundary Detection etc. are combined together to get the maximum efficient technique. The approach followed for extraction of characters is based on Horizontal and Vertical Projection.

  Horizontal Projection: Matrix containing number of black pixels per row of image.

  Vertical Projection: Matrix containing number of black pixels per column of image. The lines in the input image is extracted using the concept of Horizontal Projection. The blank fields between the images have zero or very less number of black pixels so they can be separated out on this basis. Once the lines are extracted, these lines are separated into characters on the basis of Vertical Projection The columns that include spacing between the different characters in a particular line, has zero or very less number of black pixels so they can be extracted on the basis of this concept.The individual characters are now obtained in the form of images, that are to be passed to the next step for further recognition and to be written into database.

- **Extraction of Check-boxes:** Same concept of Horizontal and Vertical Projection, as used for characters, is used for extraction of check-boxes. The extra step included after extraction is as follows: In order to determine whether there is a right or wrong mark inside the box, the total number of black pixels of the extracted check-box is calculated. For empty check-box, the number of black pixels is less than that for the one which is filled with some mark. Few samples of the empty and filled-in check-boxes are taken and the total numbers of black pixels in them are calculated by creating a function and calling it as many times as required. Using the obtained values, an optimum threshold value is selected for the number of black pixels in an empty check-box and any value greater than this threshold value will imply that the check-box is filled with a mark. Now only those check-boxes that are filled with marks, obtained above, are separated and passed on to the trained models for further recognition.

3.1.3 Character Recognition

The output of the aforementioned step is the individual extracted characters (alphabets, digits, right and wrong marks), in the form of jpeg images that need to be recognized and written into database.

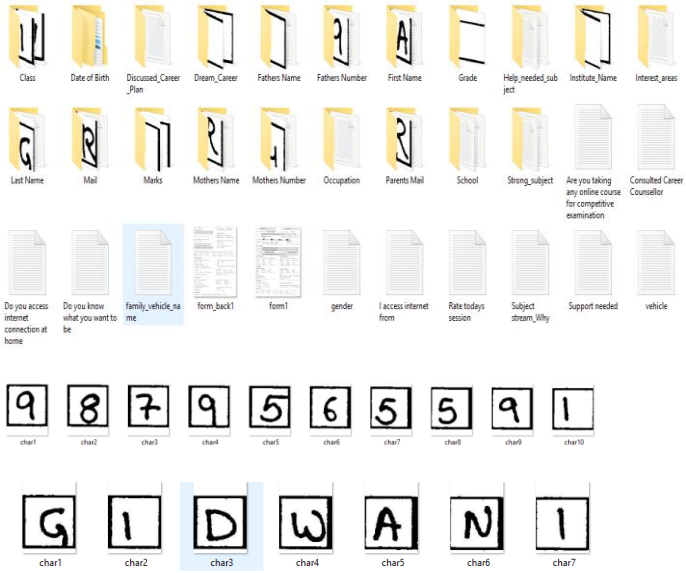


Fig 6: Segmentation Results

3.1.3.1 Text Matching

The simplest method to recognize the characters is using the concept of correlation.

In this a library is prepared containing the samples of all types of different characters that need to be recognized.

After the library is created, the extracted images are matched with all the images in the library and the one with which it is having the maximum similarity that is highest value of correlation, is recognized as the correct match and written to database. This method is efficient when there is predefined knowledge of the type of characters that needs to be recognized, that is the characters to be recognized will be of a standard font and format. Therefore it cannot be used for different handwritten characters and tick marks.

Therefore neural network approach is used for recognition.

3.1.3.2 Artificial Neural Networks

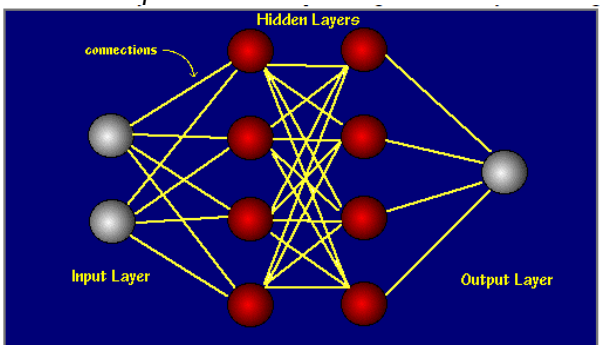


Fig 7: Artificial Neural Networks

The idea of ANNs is based on the working of human brain. The human brain is composed of 100 billion nerve cells called neurons. They are connected to other thousand cells by axons. Stimuli from external environment or inputs from sensory organs are accepted by dendrites. These inputs create electric impulses, which quickly travel through the neural network. A neuron can then send the message to other neuron to handle the issue or does not send it forward.ANNs are composed of multiple nodes, which imitate biological neurons of human brain. The neurons are connected by links and they interact with each other. The nodes can take input data and perform simple operations on the data. The result of these operations is passed to other neurons. The output at each node is called its activation or node value.Each link is associated with weight. ANNs are capable of learning, which takes place by altering weight values.So neural network can be defined as:

A neural network is a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs.[5]

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. This is true of ANNs as well. {Features of Neural Network}}

1. **Adaptive Learning:**
An ability to learn how to do tasks based on the data given for training or initial experience.

2. **Self-Organisation**:
An ANN can create its own organisation or representation of the information it receives during learning time.

3. **Real Time Operation**:
ANN computations may be carried out in parallel, and special hardware devices are being designed and manufactured which take advantage of this capability.

4. **Fault Tolerance via Redundant Information Coding**:
Partial destruction of a network leads to the corresponding degradation of performance. However, some network capabilities may be retained even with major network damage.

Neural networks process information in a similar way the human brain does. The network is composed of a large number of highly interconnected processing elements (neurons) working in parallel to solve a specific problem. Neural networks learn by example. They cannot be programmed to perform a specific task. An artificial neuron is a device with many inputs and one output. The neuron has two modes of operation; the training mode and the using mode. In the training mode, the neuron can be trained to fire (or not), for particular input patterns. In the using mode, when a taught input pattern is detected at the input, its associated output becomes the current output. If the input pattern does not belong in the taught list of input patterns, the firing rule is used to determine whether to fire or not.

Feed-forward networks:
The signals to travel one way only, from input to output. There is no feedback (loops) i.e. the output of any layer does not affect that same layer. Feed-forward ANNs tend to be straight forward networks that associate inputs with outputs. They are extensively used in pattern recognition.

Feedback networks: Feedback networks can have signals travelling in both directions by introducing loops in the network. Feedback networks are very powerful and can get extremely complicated. Feedback networks are dynamic; their state is changing continuously until they reach an equilibrium point. They remain at the equilibrium point until the input changes and a new equilibrium needs to be found. Every neural network posseses knowledge which is contained in the values of the connections weights. Modifying the knowledge stored in the network as a function of experience implies a learning rule for changing the values of the weights.[6]

Deep Learning Algorithms:
Deep learning is a branch of machine learning based on a set of algorithms that attempt to model high level abstractions in data by using a deep graph with multiple processing layers, composed of multiple linear and non-linear transformations.Theoretical results suggest that in order to learn the kind of complicated functions that can represent high-level abstractions (e.g. in vision, language, and other AI-level tasks), one may need deep architectures. Deep architectures are composed of multiple levels of non-linear operations, such as in neural nets with many hidden layers or in complicated propositional formulae re-using many sub-formulae. Searching the parameter space of deep architectures is a difficult task, but learning algorithms such as those for Deep Belief Networks have recently been proposed to tackle this problem with notable success, beating the state-of-the-art in certain areas.``Deep Learning" algorithms can automatically learn feature representations (often from unlabeled data) thus avoiding a lot of time-consuming engineering. These algorithms are based on building massive artificial neural networks that were loosely inspired by cortical (brain) computations.Below image shows comparison of deep learning feature discovery process among other algorithms.Different deep learning libraries like Theano, h2o, mxnet etc. were used for recognition of alphabets and digits. The results obtained were analyzed and compared.

The deep neural nets have a number of hidden layers between the input and output layer. The different neural network algorithms were used for character recognition and accuracy was compared. The highest accuracy was obtained in case of Convolution Networks (``Tensorflow").

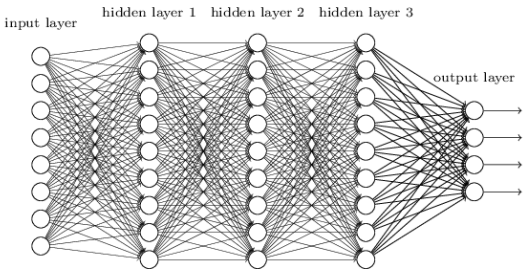


Fig 8: Deep Neural Networks

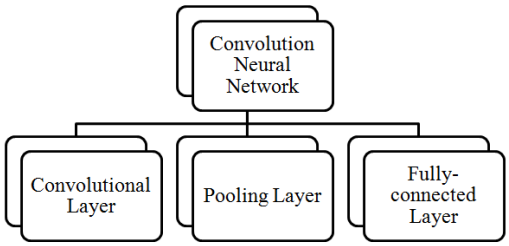


Fig 9: Convolution Neural Networks

Convolutional Neural Networks:
Convolutional Neural Networks (ConvNet) are very similar to ordinary Neural Networks, they are also made up of neurons that have weights and biases that can be trained. Each neuron receives some inputs, performs a dot product and optionally follows it with a non-linearity. The whole network expresses a single differentiable score function: from the raw image pixels on one end to class scores at the other. And have a loss function (e.g. SVM/Softmax) on the last (fully-connected) layer. ConvNet architectures make the explicit assumption that the inputs are images, which allows us to encode certain properties into the architecture. These then make the forward function more efficient to implement and vastly reduce the amount of parameters in the network. Neural Networks receive an input (a single vector), and transform it through a series of hidden layers. Each hidden layer is made up of a set of neurons, where each neuron is fully connected to all neurons in the previous layer, and where neurons in a single layer function completely independently and do not share any connections. The last fully-connected layer is called the "output layer" and in classification settings it represents the class scores. Convolutional Neural Networks take advantage of the fact that the input consists of images and they constrain the architecture in a more sensible way. In particular, unlike a regular Neural Network, the layers of a ConvNet have neurons arranged in 3 dimensions: width, height, depth. CNNs are analogous to traditional ANNs in that they are comprised of neurons that self-optimise through learning. Each neuron receive an input and perform operations (such as a scalar product followed by a non-linear function) - the basis of countless ANNs. The only difference between CNNs and ANNs is that CNNs are primarily used in pattern recognition within images. This allows us to encode image-specific features into the architecture, making the network more efficient in recognition. Convolution Networks are more efficient for images since they imply multiple layers inside the network and generate auto-features from the input vectors, by implementing pooling and patching operation.This increases the accuracy of the recognition. A CNN typically consist of 3 types of layers:

1. **Convolutional Layer:**
The convolutional layer is the core building block of a CNN. The layer's parameters consist of a set of filters (or kernels), that have the ability to learn, which have a small receptive field, but extend through the full depth of the input volume. During the forward pass, each filter is convolved across the width and height of the input volume, computing the dot product between the entries of the filter and the input and producing a 2-dimensional activation map of that filter. As a result, the network learns filters that activate when it detects some specific type of feature at some spatial position in the input.

2. **Pooling Layer:**
This layer performs pooling operation, that is a form of non-linear down-sampling. The max pooling operation is implemented that partitions the input image into a set of non-overlapping rectangles and, for each such sub-region, outputs the maximum. The function of the pooling layer is to progressively reduce the spatial size of the representation to reduce the amount of parameters and computation in the network, and hence controls overfitting. It is common to periodically insert a pooling layer in-between successive layers in a CNN architecture. The pooling layer operates independently on every depth slice of the input and resizes it spatially. The most common form, that is used is a pooling layer with filters of size 2x2 applied with a stride of 2 downsamples at every depth slice in the input by 2 along both width and height, discarding 75 percent of the activations.

3. **Fully Connected Layer:**
Finally, after several convolutional and max pooling layers, the high-level reasoning is done through fully connected layers. Neurons in a fully connected layer have full connections to all activations in the previous layer, as seen in regular Neural Networks. Their activations can hence be computed with a matrix multiplication followed by a bias offset.[7]

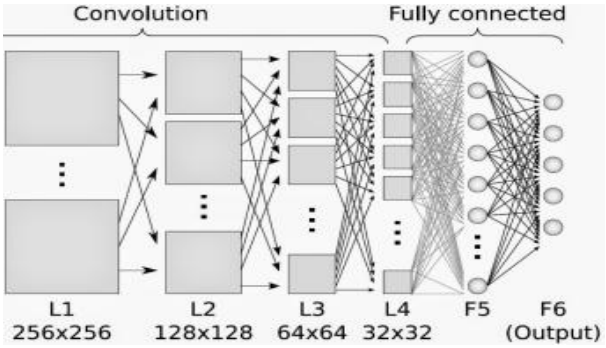


Fig 10: Convolution Neural Network Layers

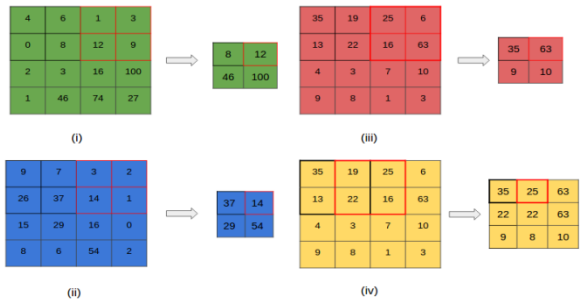


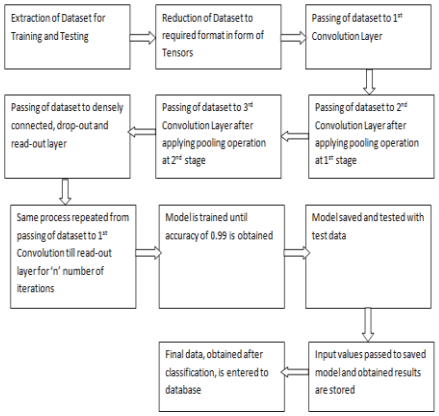
Fig 11: Pooling Layer

The problem statement in this case is a type of classification problem, where the input is to be classified into different classes on the basis of previously trained data-set.

Tensorflow:

TensorFlow is an interface for expressing machine learning algorithms, and an implementation for executing such algorithms. A computation expressed using TensorFlow can be executed with little or no change on a wide variety of heterogeneous systems, ranging from mobile devices such as phones and tablets up to large-scale distributed systems of hundreds of machines and thousands of computational devices such as GPU cards. [8]

TensorFlow is an open source software library for numerical computation using data flow graphs. Nodes in the graph represent mathematical operations, while the graph edges represent the multidimensional data arrays (tensors) communicated between them. The flexible architecture allows to deploy computation to one or more CPUs or GPUs in a desktop, server, or mobile device with a single API. The model obtained is tested on different set of data. The trained model identifies the characters with an accuracy of 90 percent. After the recognition of characters, they are written into text files and dumped into database.



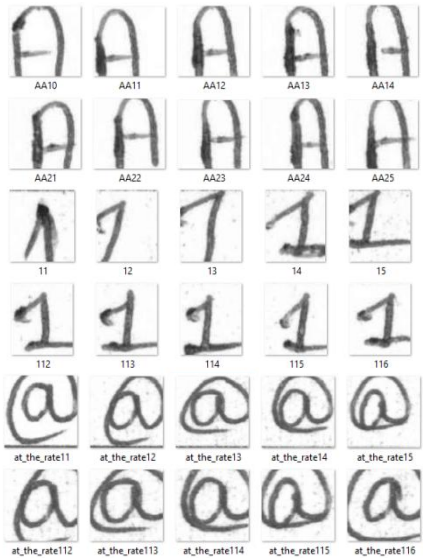


Fig 12: Tensorflow Algorithm with samples used

Tensorflow Key points:

1. The standard image size used during the entire processing is 28X28 that is 28X28 equivalent to 784 different features are generated for each image.
2. The number of iterations are accordingly increased in order to increase the accuracy.
3. The pixel values of images vary from 0 to 255.
4. The target labels, that are used for classification, are converted to one-hot vectors and then given to the network.
5. The entire training data-set is divided into training data (80 percent) and test data (20 percent).
6. The training data consist of handwritten samples of Alphabets, Digits, Special characters that can occur in e-mails like @,- etc., Right and Wrong marks
7. Different models for all the above training data-set are trained and saved which is loaded later for recognition.
8. The trained models perform classification operation and perform recognition.
9. Softmax Regression: If there are n number of classes for classification, the probability that the given input is of a particular class is suppose 80 percent and the rest 20 percent is divided between the probability of it being of different classes.Softmax regression is a natural, simple model. If the probabilities are to be assigned to an object being one of several different things, softmax is the thing to do, because softmax gives us a list of values between 0 and 1 that add up to 1.The final step in the training of model is the softmax layer. A softmax regression has two steps: first the evidence of input is added up being in certain classes, and then that evidence is converted into probabilities.To tally up the evidence that a given image is in a particular class, weighted sum of the pixel intensities is calculated. The weight is negative if that pixel having a high intensity is evidence against the image being in that class, and positive if it is evidence in favor.
10. The weights and biases are adjusted according to the data-set and as per the accuracy obtained.
11. The data is served in batches for better processing and high efficiency.
12. 2X2 patches are used for pooling operation in pooling layer in convolutional network that gives high accuracy.

13. The model is saved after the desired accuracy is achieved and then tested on test-data.
14. This saved model is further loaded and then the input data is passed for recognition.
15. Tensorflow also provides dynamic modeling, in which the state of the model is saved as checkpoints and then it can be again trained with new data-set from the saved checkpoint. On addition of more amount of training data, the accuracy can be further increased.

3.2 Writing into Database

After the recognition of characters, they are directly entered into created tables in database (MY-SQL used in this engine) and the entire process of Recognition engine is completed.

4. CONCLUSION

The described approach works well for one category of handwritten forms.

Accuracy of 90 percent can be increased further by increasing the amount of training data-set given to the model for training, that can further increase the accuracy of recognition of the characters, thereby increasing the overall accuracy of the recognition engine.

The recognition tool built, can be made to be used for any type of handwritten/printed form of any format, and the entered data can be written into database in much less time as compared to that done manually. So as to avoid Customization for each type of form, the printed characters can be extracted from Forms itself so as to improve the efficiency of the Recognition Tool.

5. REFERENCES

[1] Pasikanti Susruthi Divya Sruthi, S. Durga Devi,``Grid Infrastructure Based Intelligent Character Recognition: A Novel Algorithm for Extraction of Handwritten and Typewritten Characters Using Neural Networks",Department of IT-SNIST-Yamnampet Hyderabad AP India.

[2] Avinash Kumar Yadav, Krushna Rajbinde,Nalin Bhat,Vajid Khan ``Recognition, Formatting In Image Files By Using Image Processing",Moze College of Engineering, Pune.

[3] Dipti Deodhare, NNR Ranga Suri, R. Amit,``Preprocessing and Image Enhancement Algorithms for a Form-based Intelligent Character Recognition System", University of Southern California Bangalore India.

[4] Digital Image Processing, 2nd Edition, by Rafael C. Gonzalez and Richard E. Woods

[5] "Neural Network Primer: Part I" by Maureen Caudill, AI Expert, Feb. 1989

[6] Priyanka Wankar, Sonali. B. Maind, ``Research Paper on Basic of Artificial Neural Network", Sawangi (M) Wardha India.

[7] Alex Krizhevsky, Geoffrey E. Hinton, Ilya Sutskever, ``ImageNet Classification with Deep Convolutional Neural Networks", University of Toronto.

[8] Keiron O'Shea1and, Ryan Nash ``An Introduction to Convolutional Neural Networks", Aberystwyth University Ceredigion UK

[9] Ashish Agarwal, Eugene Brevdo, Mart´ın Abadi, Paul Barham, ``TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems" , Google Research.

[10] http://www.slideshare.net/ASHI14march/image-pre-processing

[11] TensorflowWebpage: https://www.tensorflow.org/tutorial/

[12] http://www.nowpublishers.com/article/Details/MAL-006

[13] http://www.datafinity.co.uk/forms_processing.html