CONTENTS

LIST OF FIGURES

Chapter 1

INTRODUCTION

The practice of writing is the oldest practice which is used as a means for connecting as well as broadcasting the information in such a way that it can be kept and retrieved. Some of the oldest practice included the use of palm leaves, clay plates, stones, papyrus and metal plates. This was used in the initial era. But in the year 105 A.D., paper documents started to be used for writing all types of documents and it became widely popular and easier to write on as compared to the other such methods. Along with the arrival of the computers and other such storage technologies, there is an increase in the request of the preservation of the precious documents. Such paper documents have been a widely in use and are termed as 'document images'. At present we use scanners, fax machines, smartphones as a means of getting the document images. For the purpose of moving forward in a paperless office, the documentation that is the digitization of the printed documents are carried out which are further archived as the document images in the internet or over the database of the image. Since the amount of documentation to be done is huge so it cannot be manually performed and thus we require the use of an automated process for the Document Images Analysis (DIA). This led to the automatic readings well as the documented image analysis by the OCR that is optical character recognition since the 1950's. Further it has led to the explosive growth in many areas such as form processing, invoice reading and bank processing tasks.

The main objective of the image documentation analysis is the identification of the graphics and the texts which are engraved in the images, and also to excerpt the envisioned information in a way that human could. Image documentation analysis can be defined in two classes. These are termed as the textual and the non-textual processing.

1. Textual processing of the documented images mainly considers the text apparatuses. Some of the tasks involved in the textual processing are skew

defining that is the tilt which appears when scanning of the documented image is done in the scanners, finding columns, paragraphs, text lines, and words and lastly recognizing the text by the use of the optical character recognition.

2. The Non-Textual is used as a component that frames the line diagrams, eliminates the straight lines between the text sections. The logos that are present in the images are handled by the graphics processing. The third important component of the documents are the images.

1.1 Requirement of document processing subareas

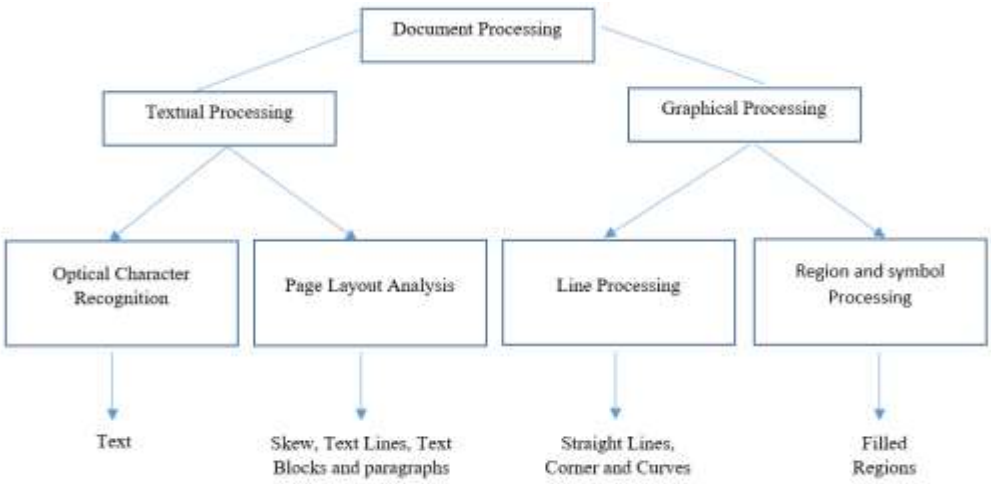

Figure 1.1: A Hierarchy of document processing subareas listing the types of document components dealt within each subarea

The three most explicit instance of the of the requirement of the document analysis are as follows:

1. Most of the compute generated documents are distinctive but then also it is unavoidably due to different types of computers and the software such that their electronic set-ups are inharmonious. Very few consists of the formatted texts and the tables and also the handwritten admissions. Different functions such as differentiation of the types of documents, allowing the functional parts abstraction as well as deciphering the formats which are produced

from one computer to another fashioned format.

2. For numerous years, the cataloguing and the acknowledgement of the address is performed by the robotic mail-sorting machines. But with the advancement there is an urgent necessity for the processing of the additional mail.
3. In the conventional library, there are multiple complications such as loss of the documents, portability issue, accessing from the remote areas, degradation of the materials and all such complications can be dealt by the documentation analysis techniques.

These all instances give a hint on how important the document image analysis is in the present era. It will develop more in the everyday document structures. An example can be taken of the optical character recognition that is there will be an increase in the OCR use as for searching, hoarding and excerpting the documents based in a certain paper. The analysis of the page-layout will help in the differentiation of the format of page or a particular form and will also help in its replication. The diagrams will be edited and passed in from pictures or by hand. Handwritten admissions will be deciphered into electronic pamphlets by the use of pen-constructed computers. Paper based pamphlets that are present in the companies and the libraries will be electrically transformed so that it can be stored and can be delivered to the office or home address immediately. But this technique of residing the pamphlets data electronically in the computers will be slow as there are numerous protocols and systems and also that the paper is an easy medium for us. So, the paper documents is going to be present with us for same gradation of our life and ultimately with the progress in technologies this documentation will take place in the future.

OCR that is Optical Character Recognition is being applied to the entire spectrum of industries and thus has led to the revolutionization of the document management process. Deep learning has persuaded in every component of computer vision and so is true for document images, character recognition and handwriting recognition. By the use of deep learning we have been able to obtain phenomenal text recognition accuracy as compared to the feature extraction and machine learning techniques. By the use of OCR not only have we been able too convert the

documents to image file but also fully searchable documents such that the text can be recognized by the computers. By the presence of OCR, there is no need for the manual typing of the important documents for saving it into the electronic databases. OCR has the property that it extracts the relevant information and passes it automatically. It produces accurate results and completes it an optimal timing duration. Different areas which uses the OCR techniques are including education, finance, and government agencies. Due to OCR many texts are available online which has led to the money as well as time saving for the students. Invoice imaging applications is used for keeping the financial records of the businesses and for preventing the backlog of the payments. In the government agencies This OCR technique simplifies the data assemblage among the processes. With the advancement of the technology there is an increase in the use of the OCR technology, which includes the use of the handwriting recognition and text recognition in document images. For the purpose of transcribing the archives of the handwritten documents Deep Learning is used for the automatic handwriting recognition (HWR). OCR related techniques such as barcode recognition is used over a daily basis in retail and other industries. Accuracy can be checked by the two criteria that is:

1. On character level
2. On word level

If the human eye can see the image clearly then the accuracy of the OCR will be good but if the original source image is not visible clearly by the eye then OCR is liable to produce some errors. Thus, if we say that if the characters in image are easily visible then the OCR can give much better accuracy. Training of the OCR is done via the images that are of very high resolution thus their performance deteriorates when the noisy low-resolution images are provided as an input to it. For the low-resolution images super resolution techniques can be applied. But the drawback is the intensification of the noise content in the low-resolution images. However, there are various pre-processing and post-processing techniques available. The most basic pre-processing technique is the de-noising of low-resolution image. But it has its own disadvantages it leads to the loss of the texture details and the high frequency information. It causes inaccuracies in the document binarization, image segmentation and text recognition. By using the semantic and

syntactic rules or by using a lexicon for correcting the spelling of the words, this post processing techniques can be used. The output given by the OCR consists of candidate character and also the distance between candidate characters thus it plays an important role in the OCR post processing. At present the distance of the candidate is traditionally transformed to reliability of the corresponding candidate character to be used. Speaking traditionally if the reliability of the candidate is big then small is the candidates corresponding distance. This led to the proposal of the statistical approach which is used for the calculation of the reliability with respect to the candidate's characters distribution and correct characters with different candidate distances. Statistical characteristics is reflected by it and its complexity is low thus leading to the good results for some applications. But still the candidate's distance is limited in the optical character recognition. Two divisions of the handwritten text recognition are online and offline recognition. And offline handwritten text recognition can also be a part of text detection in document images. Time series of coordinates are present in online handwriting recognition, representing the progress of the pen-tip, is captured and only an image of the text is available in the case of offline handwriting recognition. Online recognition yields much better result as it easily extracts the relevant atures [4]. Recognition of the isolated characters or words is one of the extensive divergences. This latter is much harder thus the excellent characters has been obtained for the recognition of digits and characters [5]. In the documented images the text in it can be broken down into the case where the style of writing is constrained that is only hand written characters are to be allowed and further the challenging case is where it is unconstrained.

For the text recognition in case of the isolated text in document images is the UNIPEN database, CEDAR, NIST, and CENPARMI data sets . The recognition of the characters is one of a bedrock task but as compared to it pattern recognition is one of the most gruelling task. As pattern recognition has its way of the interaction with the humans, thus being the potent field of research. Being more accurate character recognition is the task of investigating and spotting characters from the input characters and then converting into the ASCII or other such machine languages. Handwriting recognition system is defined as a technique by which an automatic data processing system will recognize the characters and other such symbols which is written by hand in the natural handwriting. Classification of the

handwriting recognition is done into two categories that is offline handwriting recognition and on-line handwriting recognition. If the PC can understand the scanned handwriting then it is referred to as offline handwriting recognition. In case of the online recognition, the recognition proceeds when we write via touch pad stylus pen. Classifiers perspective has two main categories for character recognition systems which are segmentation free (global) and segmentation based (analytic). The segmentation free is also termed as the holistic approach for acknowledging the characters without the segmentation into characters. As the name suggests in case of the segmentation-based approach segmentation of each and every word takes place which may be either uniform or non-uniform.

1.2 METHODOLOGIES

For the complete text detection and recognition systems there are two methodologies which are used and are termed as stepwise and integrated.

1.2.1 Stepwise methodologies

Different modules for the detection and recognition process are used in case of the stepwise methodologies. For the purpose of detection and recognition it makes use of a feed-forward pipeline to detect, segment and recognize text from document images. There are four primary steps involved in case of the stepwise methodologies which are as follows:

1. Localization.
2. Verification.
3. Segmentation.
4. Recognition.

The patches in the image are classified by the localization process and then the clustering into the candidate for the text regions is performed. Further these are classified into text or non-text regions which is done at the time of substantiation. The basic assumption taken in the substantiation is that most of the text parts can be viewed as a type of uniform pattern. For retaining the accurate outlines of image

blocks during the recognition step, the separation of the characters are done in the segmentation step.

Image blocks are converted into the characters in the recognition process. In some of the process further supplementary steps can be added or the validation and/or segmentation step can be omitted for achieving the text enhancement or super-resolution and rectification. Convolutional neural networks are used for the purpose of detection, tracking, segmentation, recognition, and correction in the stepwise methodology on raw pixel values. Then the components that are detected of the resident maximal responses are clustered as text.

For the determination of the start and the end frame of localized text a tracking process is unified. For the accurate character recognition based on the CNN segmentation step created on the shortest path is used. For removing the recognition obscurities and segmentation faults, language model is used.

1.2.2 Integrated Methodologies

In the case of integrated methodologies, it tries to classify or recognize the words where the both detection and recognition procedures consists of the common information and also use the cooperative regularisation or optimization approaches. There are other stepwise methods that use the feedback procedure for decreasing the falsely detected characters while some integrated methods make use of the pre-processing steps for segmenting out the regions of interest. One of the main differences lies in the fact that the integrated methodology makes use of the recognition as a main focusing point.

Considering the integrated methodology, recognition response or the character classification are considered as the primary cues and is also shared with the detection and recognition modules. Problems related to the complex multiclass occurs when the character classification responses are used as the primary feature. It requires the discrimination of characters from the background as well as that from each other. Solution not only requires the dynamic classification of the characters or the recognition models but it also requires suitable integration strategies, which can include holistic matching, joint optimization or decision delays.

Another approach is the **word spotting approach** which uses the nearest neighbour classifier and the histogram of oriented gradient (HOG) features. In the case of word spotting particular words are made to match with the given lexicon in the image patches by the word or the character models. Multi-scale sliding window classification is done for obtaining the character responses and further the character localization is done by the non-maximum suppression. The inputs considered for the optimum configuration of a particular word are the scores and the location of characters. It is done by employing the pictorial model from a small lexicon. For the training of the character models one of the process is the combination of the unsupervised feature learning along with multi-layer CNN. This is used in text recognition and detection procedures. For localizing the candidates text lines, sliding window character classification based on CNN is used. After this by the use of the beam search algorithm integration of the character responses with character spacings takes place.

Decision delay approach works by keeping the track of the multiple segmentation of all the characters. It continues to keep the track until it reaches the last stage so that each character's context can be attained. Segmentation of the characters are attained by the extremal regions. Based on the segmentation that is done, with the help of the character intervals, character classification scores and language priors a directed graph is created. For selecting the path on the graph through the highest score the algorithm used is the dynamic programming. The output consists of the sequence of sections and the labels attached to it which is induced by the finest path, that is a non-text region, a word or a sequence of words

In this section, the basic working principle of character recognition is described followed by a detailed literature survey. Handwritten recognition is divided into six phases which are image acquisition, pre-processing (denoising, super-resolution segmentation), feature extraction, classification and post processing.

Its block diagram is shown below in Figure 2.

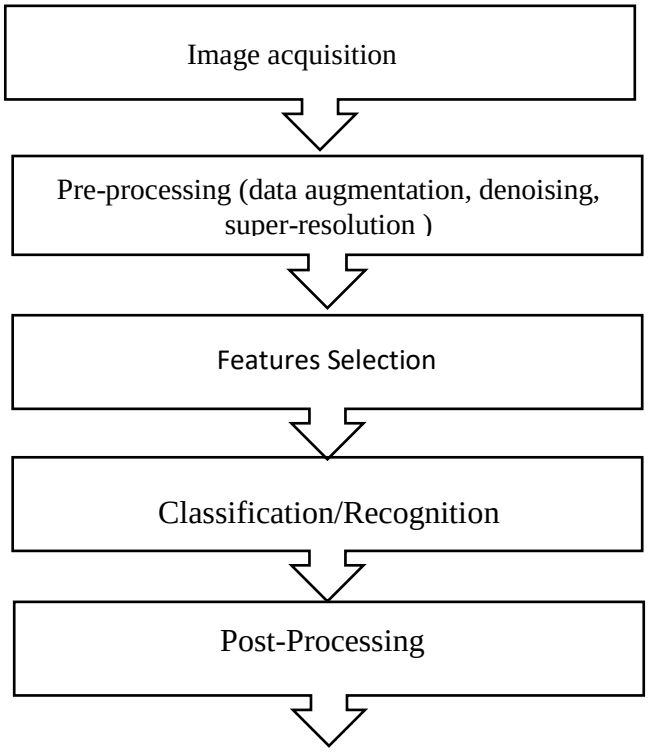

Figure 1.2: Block Diagram of Character Recognition

1.3 Motivation

In [1] Manoj Sharma, Anupama Ray et al, they have recognized the text from the noisy low-resolution (LR) images. By using the property of super-resolution that it leads to the spatial correlation in the noise, thus it has resulted from a noisy low-resolution text image result to a noisy high-resolution text image. This can further be de-noised appropriately. By using the conventional super resolution on the noisy image after the pre-processing of the image or it can be said that before application of the super-resolution denoising is done. But due to the denoising it leads to the loss of details present at the high frequency. Further it leads to the loss of information that is of the texture and edge in the output high resolution images. A non-resilient approach for text images and text recognition is proposed in their paper is done using the deep BLSTM network. This deep BLSTM network is trained on the high-resolution images. This proposed deep learning method which is based on end to end robust noise resilient framework for text image performs multiple tasks simultaneously that is image denoising as well as the preservation of

the loss of details. In the paper mentioned learning of the Stacked Sparse Denoising Auto-Encoder (SSDA) is done and for the text image super-resolution learning of the deep convolutional auto-encoder (CDCA) is performed. In their proposed framework of the end-to-end deep learning-based architecture, at the time of fine tuning the initial weights are served by the the pretrained weights.

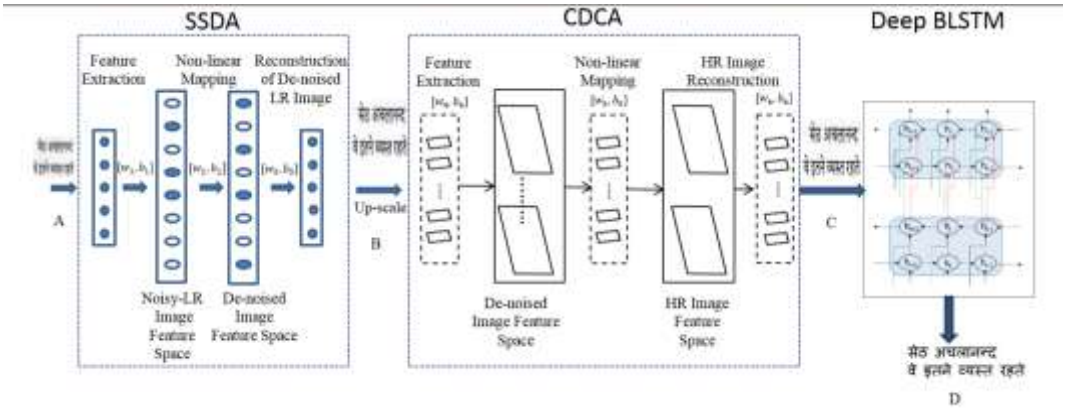

Figure 1.3. Block diagram of architecture. A: Noisy LR text image; B: Denoised Text image; C: HR text image; D: Recognized Text output

For both of the activities the network is regularized and optimized jointly. This proposed framework was tested on various Indian language datasets which produced at par results with that of the noisy images. The drawbacks of approach mentioned in their paper is that for each dataset, the LR images and the bicubic interpolated version suffered a lot in terms of recognition accuracy and PSNR. SSDA-CDCA had jointly trained denoising and super-resolution and then uses the super-resolved output as input to a BLSTM based OCR. Although this cascade network worked better than SSDA+CDCA, where each piece was trained separately, because individually trained models when cascaded together carry their errors, thus effecting the final recognition rates. At that point even a very robust OCR system would fail due to issues in the image enhancement parts.

This problem is solved by integrating four architecture together along with data augmentation to recognise large scale of different types of text in document images. First few layer of network perform data augmentation by adapting Cycle GAN architecture then next few layers based on REDNET perform denoising and next

few layers take those denoised images and perform super-resolution with underlying architecture of DBPN with skip connection then these super-resolved images are forwarded to next layer of bi directional LSTM empowered with CTC loss .And at the end we obtain a digitized version of text that was extracted by the scanned document image .It shows very good accuracy in recognising text along with preservation of its texture and information present at high frequency independent of document which are contaminated with different kind of noises either induced by scanning procedure or real time noises and also independent of its quality whether it is of low resolution or high resolution.

Following figure shows the block diagram of our proposed architecture .

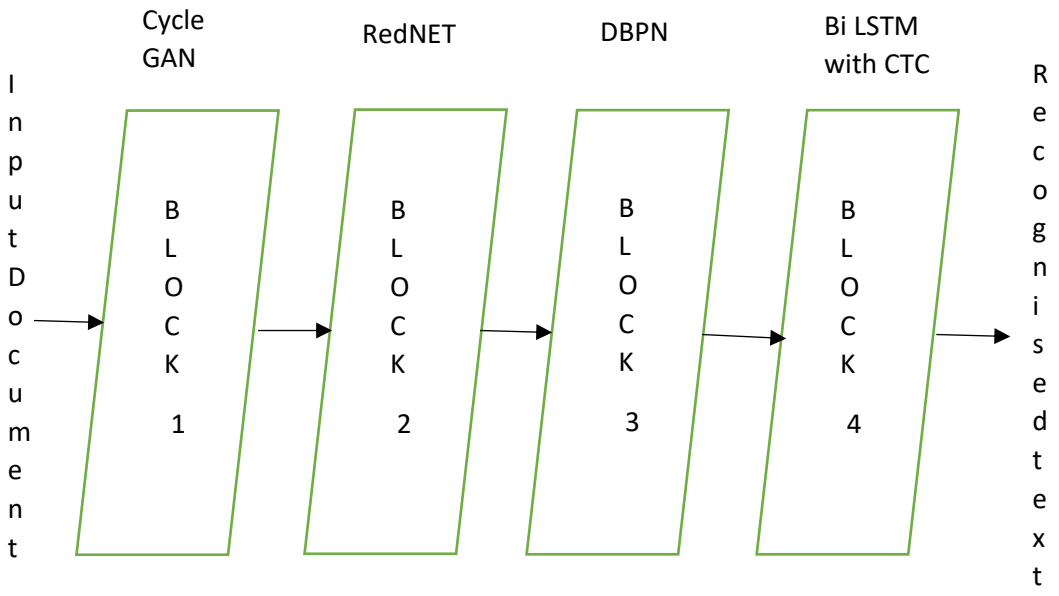

Figure 1.4: Block is for data augmentation block 2 for denoising block 3 for super-resolution and block 4 for text recognition

1.4 Thesis Organisation

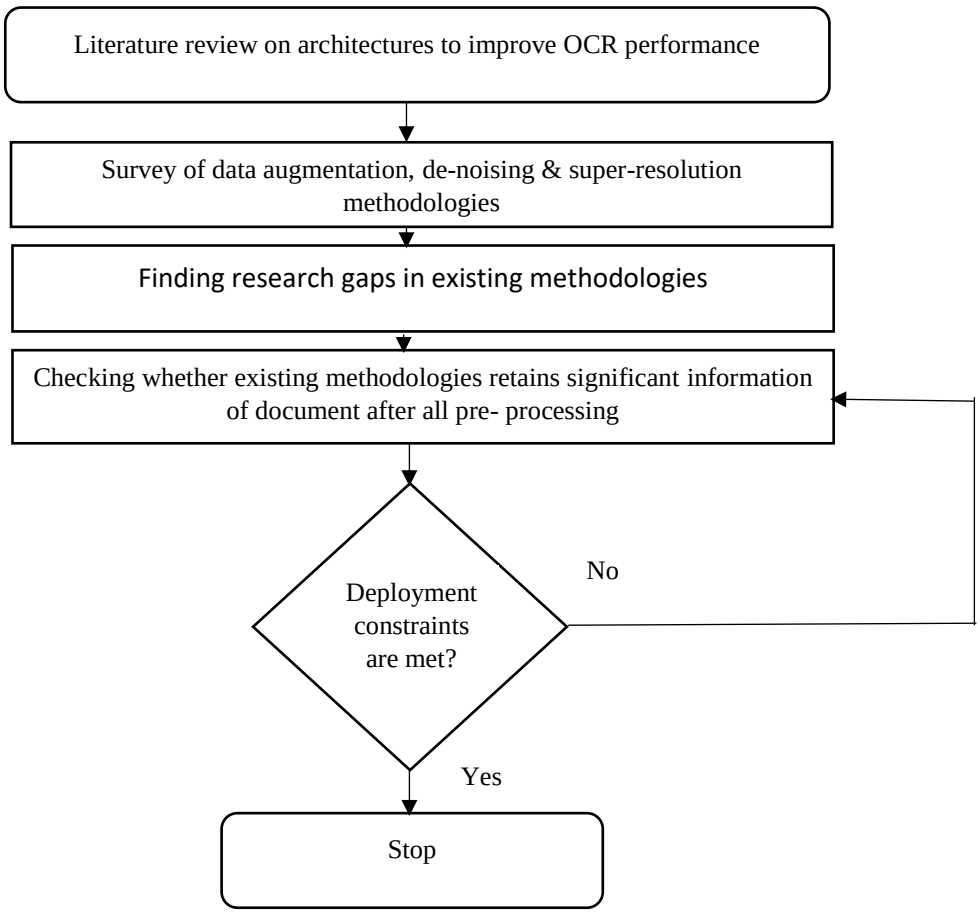

Figure 1.5: Thesis organisation.

Chapter 2

LITERATURE REVIEW

For the domain and language, text detection from document images and handwritten character processing systems are very specific. It can be compared to the construction of a generic system where it is able to process all the different kind of and written text or scripts in the document image. Many efforts have been done on the Chinese Arabic and European languages for the documentation process. But the text detection for the domestic languages which includes Hindi, Punjabi, Bengali, Tamil has not been explored. The reason is because it is used very less as compared to English which is the most often used language. For the text recognition from the segmented sections of the documented images the earlier system processed the handcrafted features. Various new techniques have been demonstrated by the researchers in the recent years that demonstrated the cooperative nature of the various text which was taken at different configurations, documents with noisy image, etc. These new techniques have come around because of the advancement in in the area of machine learning and regularization method. This included an unsupervised feature learning, semi supervised learning, variational autoencoder, recurrent neural network, Long Short Term Memory (LSTM), convolutional neural networks (CNN), deformable part-based models (DPMs) , belief propagation and conditional random fields (CRF). The input used for the image acquisition is the digitized or the digital image. The most commonly used device for this purpose is the digitizer and the electronic tablet. Digital Pen is used buy this devices. Other methods can also be used for taking the input image for the handwritten characters. These methods include photograph, scanner for by using a stylus for writing the computer directly.

2.1. Data augmentation

Large amount of data required by the deep learning networks. But for the data sets such as historical documents it is not possible to acquire large data so for such purpose, we can modify the original data so that the data augmentation is

accomplished. In case of the recognition of the image augmentation process is applied by performing the basic transformations which includes flipping the image horizontally, sampling the sub Windows of an image or scaling [6]. In the case of handwriting transformation used is the affine transformation. The disadvantage of affine transformation is that in the word images it is not able to capture the variations which is due to the character level that is variations of slant and size. In [7] after applying the gaussian smoothening, Simard et al. created random displacement field which was able to show the development as compared to the case of affine transformation which was done by random elastic distortions over the images which consisted of single character. The method described above was applied on a small 28x28 handwritten digit images which was taken only for the recognition of single character. The distortion which occurred were dependant on the two parameters which included standard deviation that is σ and α.[8] formulated by Krishnan and Jawahar et al, they proposed a method in which the original characters were replaced by the normalized character. Further they were concatenated by providing them with dissimilar spacing stroke width, background distribution and alignment. By this new line images are formed for the recognition of line using the corpus of segmented handwritten characters.

But the drawback with this model is that they can be effective only when the handwriting matches with the fonts. This is not possible when we consider the case of historical documents. In [9] Curtis Wigington and Seth Stewart et al used the CNN-LSTM based architecture so that it can be used for the recognition of text. Two data augmentation and normalization techniques were introduced by them and they depicted reduction in the word error rate and character error rate. The reason for the normalisation was that in case of handwritten text it adjusted the variation in the scale while that of augmentation was that it helped in modelling the variation in natural character-to character scenario. On both of the images that is the training and the testing image normalization and augmentation are performed. First step was the augmentation random warp grid distortion (RWGD). For the alignment towards the baseline control points were placed on the regular grid. And then each of the control points which was in the x and y direction were perturbed randomly by random sampling of their normal distribution. At last the images are warped with accordance to the perturbed control points.

The IAM Database

The IAM database is used which consists of the handwritten text lines in this reference paper. There are more than 1500 scanned handwritten documents in the version 3.0 of this database. This database had more than 650 different writers reproducing more than 13000 transcribed handwritten lines, without any restrictions on the style of writing or the instruments used. The extraction of the sentences have been done from the Lancaster-Oslo/Bergen (LOB) text corpus [10]. The type of datasets that consists are the old and noisy with degraded documents along with the presence of the newer scanned and the printed books, thus it has diverse set of the recognition challenges.

The IAM datasets used in this work constitutes:

1. Training lines: 6161 (from 283 writers)
2. Validation lines: 920 (56 writers)
3. Test lines: 2781 (161 writers)

The datasets given above are disjoint and each writer has not contributed for more than one sets. The partitions used here are :

1. Instances: 87967
2. Different words occurring in the union: 11320
   (Training, Validation and Test sets)

The modelling of the lexicon is done with characters, which includes lowercase letters, uppercase letters, digits, punctuation marks, the space, and a character for garbage symbols.

1. Characters: 78
2. Lowercase letters: 26
3. Uppercase letters: 26
4. Digits: 10
5. Punctuation marks: 14

2.2. DENOISING AND SUPER-RESOLUTION

For reducing the variation in the handwritten texts including the style of the fonts, multiple steps for the pre-processing is performed. All these processes are performed by the automatic text recognition systems, thus the information that is pertinent for recognition is preserved. For achieving the good rate of character recognition pre-processing is a very important step. The pre-processing step is mainly done for the scanned documents or handwritten text lines, slope and slant correction and normalization of the size of the characters. For the alignment of the lower baseline with respect to the image's horizontal axis, the horizontal rotation is done which corrects the slope. In [11]For removing the slope and slant in the image for the alignment of the lower baseline with respect to the image's horizontal axis a new method is presented in case of the handwritten text lines. Also, for the normalization of the normalization of the size of text image by the use of Neural Networks (NNs), a new method is presented. For the accurate measurement/estimation of the slope as well as the horizontal alignment local extrema from the text image is considered which belongs to the classification of the lower baseline by a Multilayer Perceptron (MLP). For the removal of the slant a non-uniform method is used that is by ANNs. The final step is the computation of the reference lines of the slopes by using another MLP thus the text and the lines are aligned horizontally without any error.

The main objective involved in the pre-processing step are as follows:

1. De-noising
2. Normalizing strokes
3. Elimination of the variations
4. Super-resolve
5. Segmentation

If these steps are not performed then it would lead to the complicated recognition and further reduction in the recognition rate. The conventional process applied for the process of the pre-processing of the image by the de-noising algorithm and then super-resolution (SR) has drawback that is with the noise the high frequency details that is the texture detail specially is lost invariably. This step of denoising leads to

the restriction in the subsequent super-resolution (SR) step. So, we have the challenge of synthesizing such texture details.

Different steps involved in the pre-processing step are as follows:

1. Size normalization
2. Centring
3. Interpolating missing points
4. Smoothing
5. Slant correction
6. Resampling of points

Here stacked sparse de-noising auto-encoder (SSDA) is one of the state-of-the-art methods for the de-noising of the natural images.

In [12] by Abhishek Singh and Fatih Porikli et al, patch similarity-based SR algorithm was used for obtaining the high-resolution version of both the noisy as well as the de-noised images. By considering a convex combination of orientation and frequency selective bands of both the noisy and the de-noised images a high-resolution image can be obtained wherein:

1. Some of the high frequency detailed textures that was lost during the de-noising step can be effectively recovered in the high-resolution domain.
2. By varying the parameters of the convex combination some of the additional parameter can be synthesized easily.

In the above paper much importance has been given to the restoration of the originality of the documents. For the purpose of image denoising and super-resolution some of the conventional methods which includes total variation, BM3D algorithm and dictionary learning based methods have shown good results. Regularization method has been essential for the ill-posed problem of image restoration.

DNN based methods for the purpose of the image restoration is more promising as compared to the traditional based methods. The most popular DNN model for the purpose of image restoration is the stacked denoising auto-encoder [10]. For the low-level vision tasks and the denoising a combination of the sparse coding and DNN pre-trained with denoising auto-encoder was used by Xie et al. [14]. Some of the alternate methods for such as multi-layer perceptron [15] and CNN [16] for

image denoising, as well as DNN for image or video super-resolution [13] and compression artifacts reduction have also been studied in the recent years.

A patch based algorithm was proposed by Burger et al. [12] which was made to learn by a plain multi-layer perceptron. It was also concluded in the paper that with very large networks and training data the neural networks will surely achieve state-of-the-art image denoising performance. A fully convolutional CNN was prosed for the purpose of denoising. It was observed by them that the CNN was able to achieve more superior performance as compared to the wavelet and Markov Random Field (MRF) methods. A collaborative local auto-encoder was used for the purpose of super-resolution in a layer by layer fashion after applying it with a non-local self-similarity (NLSS) based search on the input image by Cui et al. Among the low/high-resolution images Dong et al introduced a method for directly learning the end-to-end mapping. It was argued that the conventional sparse coding's domain expertise can be integrated for achieving better results. One of the best advantage by the use of DNN method is that these are purely data driven and further no assumptions had to be made for the noise distributions.

Different author has different handwriting and among the same author the handwritings can change with time. As the modern methods have shown good results for the purpose of handwriting recognition but still it is not enough for capturing the variation that occur in the handwriting. In the recent trends CNN have shown very low error rate for the large, multi-author handwritten word datasets. These networks have made very less use of the feature representations with deep feature embedding and augmented training for performing the recognition and the spotting of the words. In the recently organized competition on the German handwriting recognition Recurrent Neural Networks (RNNs) was applied effectively to HWR. It produced top results in that competition. Two novel data augmentation and normalization techniques was introduced by Curtis Wigington and Seth Stewart et al for the improvement in the state-of-the-art of the handwriting recognition based on the neural networks. This novel techniques can be applied to any of the HWR neural network. In the context of both the word and the line level it achieved great results which was more precise than the current best model, that is:

1. Normalization of the profile to both the line and the word images.
2. Distortion of existing words using random perturbations on a regular grid aligned to the baseline.

To both the training and the test images normalization and augmentation was applied on. With the help of CNN-LSTM architecture [4] they evaluated the techniques of augmentation and normalization for handwriting recognition.

2.4. Classification

The features of the input image are extracted when it is given to the HCR system, and these features acts as an input to the trained classifier which can include the artificial neural network or support vector machine. The classification is done by comparing the features that are stored to that with the input feature. The class which matches closely to the input feature is labelled to that particular class. Connectionist Temporal Classification which labelled the Unsegmented Sequence Data with Recurrent Neural Networks was proposed by A G Alex, S Fandez [17]. In many of the real-world sequence learning tasks, it requires the estimation of the sequences from the noisy and the unsegmented input data. In the case of the speech recognition, the acoustic signal is changed into the words or sub-word units. Thus according to the need Recurrent neural networks (RNNs) are best suited as they are powerful sequence learners. But its use is limited as it requires the pre-segmented training data as well as the post-processing for transforming their outputs into label sequences. A novel technique which is used for the training of the Recurrent neural networks (RNNs) is presented which labels the unsegmented sequences directly, hence it removes both of the drawbacks. Its advantages are demonstrated by performing the experiment on the TIMIT speech corpus which reveals its advantages as compared to that of baseline HMM and a hybrid HMM-RNN.

A Graves, M Liwicki, Santiago Fern´Roman Bertolami, Horst Bunke, J¨urgen Schmi dhuber et al proposed [18] A Novel Connectionist System for Unconstrained Handwri ting Recognition. Recognising lines of unconstrained handwritten text is a challengin g task. The difficulty of segmenting cursive or overlapping characters, combined with

the need to exploit surrounding context, has led to low recognition rates for even the be st current recognisers. Most recent progress in the field has been made either through i mproved pre-
processing, or through advances in language modelling. Relatively little work has bee n done on the basic recognition algorithms. Indeed, most systems rely on the same hid den Markov models that have been used for decades in speech and handwriting recogn ition, despite their well-
known shortcomings. Their paper proposes an alternative approach based on a novel t ype of recurrent neural network, specifically designed for sequence labelling tasks wh ere the data is hard to segment and contains long range, bidirectional interdependencie s. They demonstrated the network's robustness to lexicon size, measure the individual influence of its hidden layers, and analyses its use of context.

Among several systems that have been found to perform well on UNIPEN among all d ata set is a writer independent method based on hidden Markov models [19]; a fusion te chnique called cluster generative statistical dynamic time warping , associations with dynamic time warping with HMMs and embeds clustering and statistical sequence mo delling in a single feature space; and a support vector machine with a novel Gaussian d ynamic time warping kernel . Emblematic error rates on UNIPEN range from 3 for dig it recognition, to about 10 for lower case character recognition.
Parallelly techniques can be used to classify isolated words, and this has given good re sults for small vocabularies (for example a writer dependent word error rate of about 4. 5 for 32 words). However an obvious downside of whole word classification is that it d oes not measure up to large vocabularies.
For large vocabulary recognition tasks, the only viable approach is to recognise discre te characters and record them onto comprehensive words using a dictionary. Naively, t his could be finished by pre-
segmenting words into characters and classifying each section. However, segmentati on is difficult for cursive or unconstrained text, unless the words have previously been recognised. This creates a circular reliance between segmentation and recognition tha t is sometimes referred to as Sayre's paradox .
One solution to Sayre's paradox is to simply overlook it, and carry out segmentation b efore recognition. A more promising approach to Sayre's paradox is to segment and re

cognise at the same time. Hidden Markov models (HMMs) are able to do this, which is one reason for their popularity in unconstrained handwriting recognition . The idea of applying HMMs to handwriting recognition was originally motivated by their success in speech recognition , where a similar conflict exists between recognition and segme ntation. Over the years, numerous refinements of the basic HMM approach have been proposed, such as the writer independent system considered , which combines point or iented and stroke oriented input features.
However, HMMs have several well-
known drawbacks. One of these is that they assume the probability of each observatio n depends only on the current state, which makes contextual effects difficult to model. Another is that HMMs are generative, while discriminative models generally give bett er performance in labelling and classification tasks.

Recurrent neural networks (RNNs) do not suffer from these limitations, and would the refore seem a promising alternative to HMMs. However the application of RNNs alon e to handwriting recognition have so far been limited to isolated character recognition. The main reason for this is that traditional neural network objective functions require a separate training signal for every point in the input sequence, which in turn requires p re-
segmented data.A more successful use of neural networks for handwriting recognitio n has been to combine them with HMMs in the so-
called hybrid approach. A variety of network architectures have been tried for hybrid h andwriting recognition, including multilayer perceptron's time delay neural networks and RNNs . However, although hybrid models alleviate the difficulty of introducing c ontext to HMMs, they still suffer from many of the drawbacks of HMMs, and they do n ot realise the full potential of RNNs for sequence modelling.

This paper proposes an alternative approach, in which a single RNN is trained directly for sequence labelling. The network uses the connectionist temporal classification (C TC) output layer, first applied to speech recognition. CTC uses the network to map dire ctly from the complete input sequence to the sequence of output labels, obviating the n eed for pre-
segmented data. It extends the original formulation of CTC by combining it with a dict

ionary and language model to obtain word recognition scores that can be compared dir ectly with other systems. Although CTC can be used with any type of RNN, best result s are given by networks able to incorporate as much context as possible. For this reason we chose the bidirectional Long Short-
Term Memory (BLSTM) architecture, which provides access to long range context al ong both input directions. Therefore we propose a novel robust noise resilient end to e nd generic architecture for noisy low resolution handwritten text detection and recogn ition. In this paper we overcame these problem by training a network to jointly remove the noise and side by side super-
resolving handwritten text along with data augmentation using synthetically produce d dataset. This architecture equip de-noising process to de-
noise along with high frequency details and texture preservation of hand written text. Moreover data augmentation helps to recognise hand written text degraded with any r eal noise and any writing style, the more will be the data the more will be the generaliza tion for handwritten text recognition. To overcome the limitation of generalisation cap ability with different styles of writing (handwritten or printed), type of fonts, and scrip t specific features etc. in frameworks, this paper uses Generative Adversarial Network s (GAN) for data augmentation, enhancement module to get noise resilient SR. These super-resolved features are further used to train few layers of Bidirectional Long-
Short-Term-
Memory (BLSTM) cells with Connectionist Temporal Classification (CTC) for text r ecognition. All modules (data augmentation with GAN, enhancement with GAN, and BLSTM based recognition) are trained end-to-
end to jointly optimize enhancement and recognition.

The proposed network being end-to-
end trainable boosts the recognition accuracies on degraded document images. The G AN based enhancement module is responsible for the de-noising and super-
resolution of degraded document images, while the BLSTM layers with its ability to le arn sequential data recognizes each sentence at one time.

CHAPTER –3

DATA AUGMENTATION

Cycle GAN

As deep learning-based application are based on data driven so there is a need to explore ways to better train document analysis systems via increasing number of data. It is definitely, in the instances where it is to be deal with single type (or domain) of document, such as scans, images or PDFs which contains printed or handwritten text: The instances where there is only presence of unsoiled document images (such as PDFs) to a model to be trained, such kind of model does not perform as good as on scanned document images. Such a realm incongruity often happens for datasets of document, as most of the time we regrettably depend on minute, unbroken, and cloistered datasets. So, we need to train a model that can encode images in one area such that it looks like one of the other stated domains. The model that can be fittingly itemized for the marked domain and which can also utilise more datasets within any other given domain. This is called field adaptation. There are all classes of reworking techniques to practice, but individual that gained an growing amount of consideration over the previous years is the Generative Adversarial Network (GAN) by Goodfellow et al. (2014). The elementary clue is that two adversarial representations, called the generator and the discriminator, are played counter to each other in a mini-max game setting. The generator generates pictures from some haphazard noise: G: z → x. The discriminator attempts to categorize pictures as dishonest (approaching from the generator) or true (approaching from the true data distribution): D: x → {real, fake}, by maximising log(D(x)). The goal line of the generator is to dupe the discriminator by decreasing log(D(G(z)). This system, ends up at a generator that has been cultured to generate faithful pictures. This model comprises two mapping methods G : X → Y and F : Y → X, and accompanied with adversarial discriminators D_Y and D_X. D_Y inspires G to decipher X into outputs vague from domain Y , and vice versa for D_X, F , and X. To additionally standardize the mappings, two “cycle consistency losses” are introduced that seizure the perception that if it is translated from one realm to the

other and back again then it should attain where it was in progress: headfirst cycle-consistency loss: x → G(x) → F (G(x)) ≈ x, and (c) backward cycle-consistency loss: y → F (y) → G(F (y)) ≈ y.

The goal of the technique is to understand the mapping among the two domains which are X and Y for the training samples that is $\{x_i\}_i^N{}_{=1} \in X$ and $\{y_j\}_j^M{}_{=1} \in Y$.

Formulation

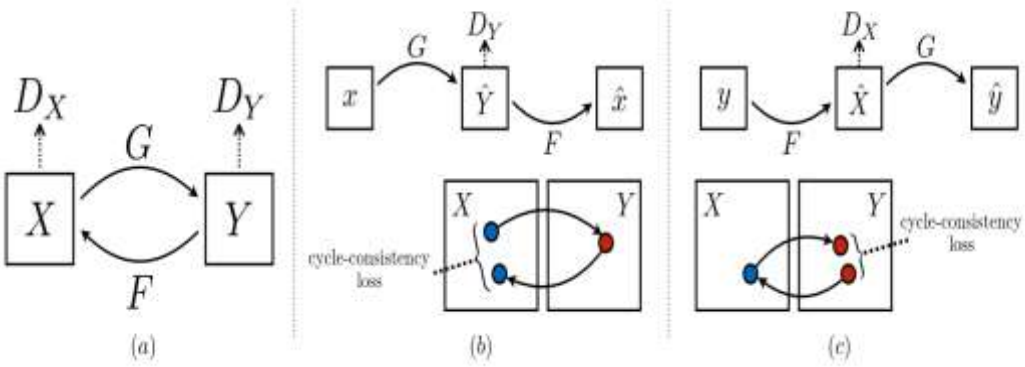

Figure 3.1: Mapping between generators and discriminators

Its shown in the Figure 3.1(a), that the proposed model consists of two mapping mappings G : X → Y and F : Y → X. Also,we have included the two discriminators that is D_X and D_Y. The aim of D_X is to differentiate among the image {x} and the translated image that is {F (y)}; in a similar manner D_Y is used to aim and differentiate between the image {y} and the translated image that is {G (y)}.

This framework considers two types of standings: adversarial losses for matching the pattern of generated images to the data pattern in the aimed distribution; and a cycle consistency loss to avoid the learned mappings G and F from opposing individually one another.

3.1. Adversarial Loss

The adversarial losses are applied to equally recording functions. For the recording function G: X →Y and its discriminator D_Y , the objective is expressed as:

$$L_{GAN}(G, D_Y , X, Y) = E_{y\sim p_{data}(y)}[\log D_Y (y)] + E_{x\sim p_{data}(x)}[\log(1 - D_Y (G(x))] \quad \dots (3.1)$$

where G attempts to produce pictures G(x) that appear alike as pictures from field Y , whereas D_Y intentions to differentiate amongst deciphered examples G(x) and real examples y. A similar adversarial loss intended for the mapping function F : Y →X and aits discriminator D_X as well: i.e. $L_{GAN}(F, D_X, Y, X)$.

3.2. Cycle Consistency Loss

In theory, mapping of G and F can be learned in Adversarial training which can produce outputs Y and X in target region with similar distribution accordingly(this condition is fulfilled when G and F are stochastic function) .A network can equate over a same set of images given as input to any of randomly arrangement of images in the target region, with very large capacity, where output distribution can be induced that matches the target region with the help of learned mappings. To reduce the domain of function which can be mapped possibly, the cyclic consistency of learned mapping functions is necessary. The image transformation cycle must be able to bring x back toathe original image, for each and every image x from domain X, i.e.a x →G(x) →F (G(x)) ≈x. This is called as forwarda cycle consistency. Similarly, from domain Y , for each image y, backward cycle consistency should be satisfied with G and F: y →F (y) →G(F (y)) ≈y. This behaviour can be incentivize using a cycle consistency loss:

$$L_{cyc}(G, F) = E_{x\sim p_{data}(x)}[kF(G(x)) - xk_1] + E_{y\sim p_{data}(y)}[kG(F(y)) - yk_1]. \quad \ldots (3.2)$$

An adversarial loss between F (G(x)) and x can be used to replace with The L1 norm in this loss with and between G(F (y)) and y. But improved performance is not registered with it.

$$L(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + \lambda L_{cyc}(G, F), \quad \ldots(3.3)$$

The relative importance of the two objectives is controlled by the lamda, λ. The aim

is to solve following equation.

$$G^*, F^* = \arg min_{G,F}\, max_{D_x, D_y}\, l(G, F, D_x, D_y).$$

… (3.4)

This model can be noticed as if two autoencoder are being trained, one autoencoder $G \circ F : Y \rightarrow Y$ is jointly trained with another autoencoder i.e $F \circ G : X \rightarrow X$. However, a special internal structure is possessed by each autoencoder: with the help of an intermediate representation that is a transformation of the image into another field, a special internal structure is drawn. It can be seen as a special case of "adversarial autoencoders" with such setup where matching of an arbitrary target distribution is done by training adversarial loss to the bottleneck layer of an autoencoder to match an arbitrary target distribution.

CHAPTER -4

DENOISING VIA REDNET

Nature for image restoration or reconditioning, this entirely convolutional auto-encoder network is very deep in it is an encoding and decoding framework which has symmetric convolutional and deconvolutional layers.

It can be said that the network [20] consists of many layers of the convolution and the deconvolution operators. It learns the end-to-end mapping from the tarnished images to the original images. The convolutional layers extract the concept of the image and simultaneously it removes any of the corrupted patches. The image can be recovered by the up-sampling of the feature maps by the use of the deconvolutional layers. For dealing with the problem that deeper networks are more difficult to train, symmetrically link convolutional and deconvolutional layers is used for the purpose of avoiding it with the presence of the skip layers. This leads to the convergence of the training rate much fast and also achieves better results. In RedNet,athe residual block is used as the building module to avoid the model degradation problem [21]. This allows the performance of networks to improve as the structure goes deeper. Moreover, application of fusion structure to incorporate depth information into the network and use of skip-architecture to bypass the spatial information from encoder to decoder. Further, inspired by the training scheme in [22], the pyramid supervision is used that apply supervised learning for better optimization and regularisation over different layers on the decoder.

4.1 Architecture

This architecture is fully convolutional and deconvolutional. After every deconvolutio n and convolution rectification layers are added. The preservation of the primary com ponents of objects in the image and elimination of the corruptions is done by the conv olutional layer which side by side also work as a feature extractor. The deconvolution al layers are then combined to recover the details of image contents. The deconvoluti onal layers produces the "clean" version of the input image as the output image. More over, skip connections are also used in this framework.

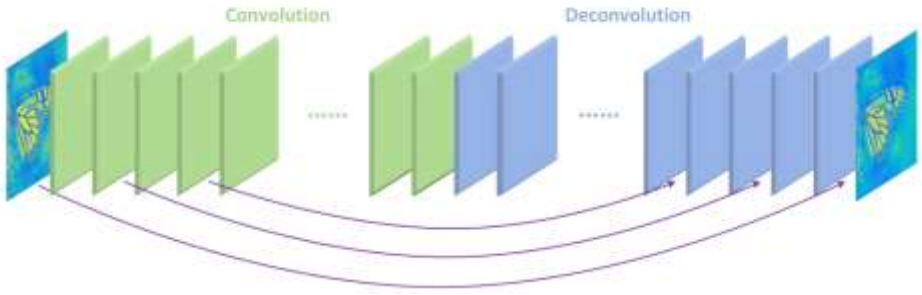

Figure 4.1: The overall architecture of our proposed network. The network contains l ayers of symmetric convolution (encoder) and deconvolution (decoder). Skip-layer connections are connected every a few (in our experiments, two) layers.

The Skip connections which are added from a convolutional layer to its corresponding mirrored deconvolutional layer. The convolutional feature maps are passed to and sum med with the deconvolutional feature maps elementwise, and passed to the next layer af ter rectification.

Neither pooling nor unpooling is preferred in the network for low resolution image or lowlevel image restoration problems, as usually useful details of the image those are e ssential for these tasks are discarded by pooling layer. The kernel size for convolution and deconvolution is set to 3×3, motivated by the VGG model , excellent image recog nition performance was shown by this size setup. The network which has been taken do es prediction based on the pixels, hence the size that can be taken of the network is rand om. The size of the image produced as the output and which is considered as input iswx hxc in which the variables are, w as width, h as height and c as the total number of chann els. The value of c that is chosen here is 1 but we can take any value of c which is greater t han 1. Higher is the value of feature maps, better is the performance but here 64 feature maps were considered as it gives optimum result for the deconvolutional and convoluti onal layers. As from the abovementioned architecture only two network experiments were conducted that is with 20 and with 30 layer respectively.

4.1.1 Deconvolution decoder

Recently for the semantic segmentation combination of the convolution and deconvol ution is proposed. As we know in convolutional layers multiple input activation are fu

sed in the same filter window for the output of a single activation. Deconvolution laye rs is one to many that is one input to multiple output function. For learnable upsamplin g layers, deconvolution is mostly used. Now in place of deconvolution we can replace it with convolution thus resulting in the architecture which is close to the latest very deep f ull convolutional neural networks. But there is a difference between the full convoluti on system and the model proposed here. In the case of the full convolution system, aft er each of the steps there is a reduction in the noise that is the noises are demarcated on e by one. And these processes leads to no loss of the data in the image. In the model pro posed the convolution is used for preserving the main content of the image. After that the use of deconvolution is for the compensation of the details. With this network pro posed, very deep full convolutional neural networks five and ten layers are used.

Padding and upsampling of the input is used for fully convolutional networks, to mak e the output and input of the same size. The 5 layers in the starting of the network are co nvolutional and the next 5 layers of the network are deconvolutional. The remaining pa rameters for training are kept similar. Use of deconvolution provides somewhat muc h better results counterpart in terms of the Peak SignaltoNoise Ratio (PSNR) than the fully convolutional layered framework.

Moreover, it is needed to speedup the testing phase so that application of deep learnin g models on devices with limited computing power such as mobile phones can be don e. In such conditions, to quickly complete the testing phase the use of the down-sampling in convolutional layers comes in handy which reduces the size of the feature maps. Deconvolution is used to upsample the feature maps in the symmetric deconvol utional layers in order to obtain an output of the same size as that of input. The efficien cy of testing is shown to improve well with performance degradation of negligible qua ntity.

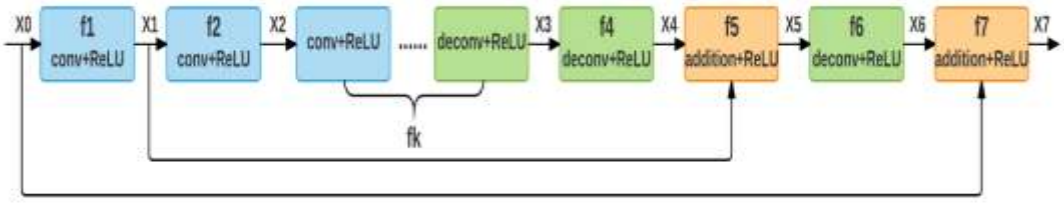

Figure 4.2: An example of a building block in the proposed framework. For ease of visualization, only two skip connections are shown in this example, and the ones in layers represented by f_k are omitted.

4.2 Skip connections

The recovery of image details from the image abstraction can be done with ability of deconvolution layer. Deconvolution has the ability to recover the image abstraction with only a few layers of convolution present in shallow networks. However, deconvolution does not work so well while the network uses operations such as max pooling or when it goes deeper, possibly due to the loss of too much image detail has already done by the deeper convolutional layers. The performance gain is also reduced along with the depth of layers in the deeper convolutional network. Moreover. vanishing gradient is also seen in deeper network which makes the training of deeper network very hard problem.

The above two problems can be addressed inspired by deep residual networks and the highway networks, skip connections are added between two corresponding deconvolutional and convolutional layers. For using such connections there are two reasons. First, image details can be lost when the network goes deeper, which makes recovery of deconvolution weaker. However, skip connections carry forward much details of the image. Second, the benefits on backpropagating the gradient to bottom layers can also be achieved by the skip connections, which makes it easy to train the much deeper network. Note that these skip layer connections are5very different from the approach that has only its concern over the regularization side. In this scenario the information related to the convolutional feature maps is passed to the equivalent deconvolutional layers.

The network is used to fit the residual of the problem instead of direct learning of the mappings from output Y to the input X, which is described by the equation

$Y - X = F(X)$. In inner blocks of the encodingdecoding network, such learning strateg y is applied to style training more operative. To every two convolutional layers , the ski p connections are passed to their mirrored deconvolutional layers. This configuration w orks very well although other kind of configurations are also possible. It becomes very e asier to be train the network by the use of such skip connections and by increasing the de pth of network the gain in performance of restoration can be observed.

Th feedforward LSTMs without gates are the CNN layers of deep residual networks an d the feedforward long shortterm memory (LSTMs) with forget gates are essentially de ep highway networks. The format of standard feedforward LSTMs are not in general sa me as the Deep residual networks.

CHAPTER-5

SUPER-RESOLUTION

For performing the Superresolution of the document images which are emerging from the denoising block are passed to the DBPN with Skip Connection to fulfil the task. Th e representations of the nonlinear mapping from those to highresolution output and lo wresolution input are learned by the feedforward framework of recently proposed dee p super-resolution architectures. However, the mutual dependencies of low- and high-resolution images cannot be fully addressed by such kind of approach. Deep BackProj ection Networks (DBPN), providing an error feedback apparatus for error projection at every stage, iterative up and down sampling layers can be exploited. Representation of different types of image degradation and highresolution components, can be done b y mutual connection of each up and downsampling stages. This idea can be extended a llowing up and down sampling stages concatenation of features across all the layers of Dense DBPN allowing improvement in further superresolution or reconstruction, lau nching new state of the art outputs for large scaling factors such as 8x across multiple d ata sets and in particular yielding superior outputs.

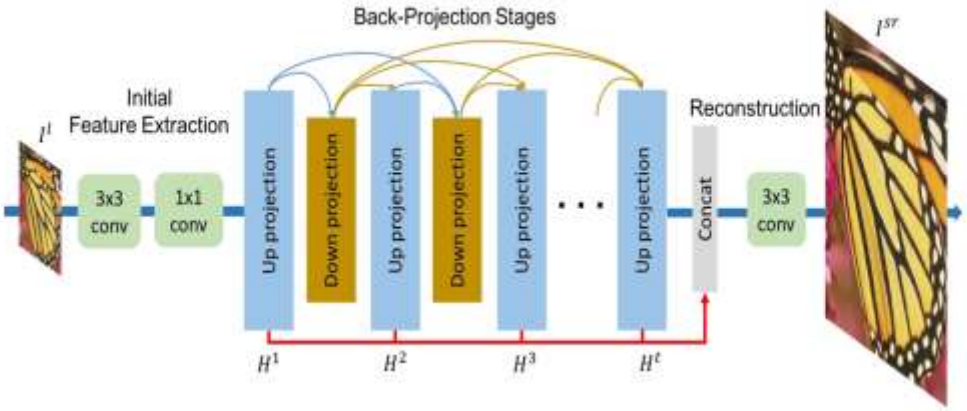

Figure 5.1: DBPN exploits densely connected projection unit to encourage feature reu se.

5.1 Deep Back-Projection Networks

Let I^l and I^h be LR and HR image with $(M' \times N')$ and $(M \times N)$ respectively, where $N' < N$ and $M' < M$. The projection unit is the main building block of proposed DBPN architecture, which is trained to map either an LR feature map to an HR map (upprojection), or an HR map to an LR map (downprojection) as part of the endtoend training of the SR system.

5.2 Projection units

The up-projection unit can be define as follows:

Scale down: $l_0^t = (h_0^t * g_t) \downarrow_s$,

Scale up: $h_0^t = (l^{t-1} * p_t) \uparrow_s$,

Scale residual up: $h_1^t = (e_t^l * q_t) \uparrow_s$,

Residual: $e_t^l = l_0^t - l^{t-1}$,

Output feature map: $h^t = h_0^t + h_1^t$

… (5.1)

where $\uparrow_s$ and $\downarrow_s$ are, respectively, the upaand down sampling operator with scaling factors, p_t, g_t, q_t are deconvolutional layers at stage t and is the spatial convolution operator.

This projection unit takes the previously computed LR feature map l^{t-1} as input, and maps it to an (intermediate) HR map H_0^t; then it attempts to map it back to LR map l_0^t ("backproject"). The residual (difference) e_t^l between the observed LR map l^{t-1} and the reconstructed l_0^t is mapped to HR again, producing a new intermediate(residual) map h_1^t; the final output of the unit, the HR map h^t, is obtained by summing the two intermediate HR maps.

The down projection unit is defined very similarly, but now its job is to map its input HR map H^t to the LR map l^t.

scale down: $L_0^t = (H^t * g_t') \downarrow_s$,

scale up: $H_0^t = (L_0^t * p_t') \uparrow_s$,

residual: $e_t^h = H_0^t - H^t$,

scale residual down: $L_1^t = (e_t^h * g_t') \downarrow_s$,

output feature map: $L^t = L_0^t + L_1^t$

… (5.2)

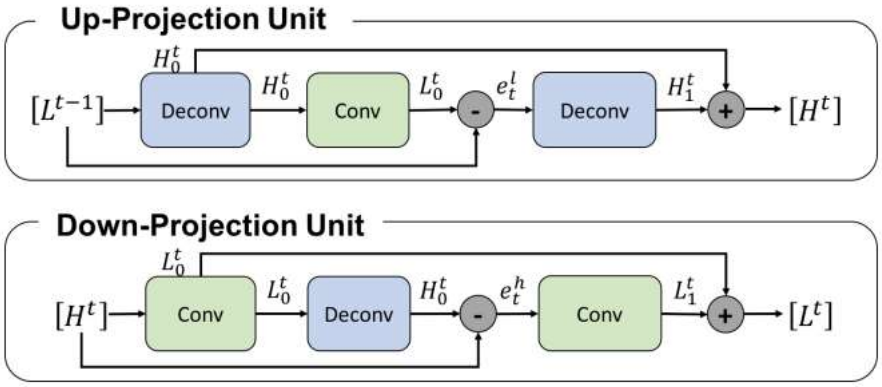

Figure 5.2: Proposed up- and down-projection unit in the DBPN.

Between H and L the projection units are organised alternatively in a successions of st ages. These projection units can be understood as a selfcorrecting procedure which fee ds a projection error to the sampling layer and iteratively changes the solution by feedi ng back the projection error.

As a projection unit huge sized filters such as 8×8 and 12 ×12 are handed down. In othe r existing networks, the use of largesized filter is avoided because it slows down the co nvergence speed and there is a chance that they produce suboptimum outcomes. Howe ver, network can suppress this disadvantage can show good performance on very large scale even with very few layered network on repeatedly using the projection unit.

5.2. Dense projection units

Feature reuse, production of improved feature, elimination of vanishing gradient prob lem can be achieved by the dense inter-layer connectivity pattern in DenseNets.

Dropout and batch normalization is avoided in these layers of deep DBPN which prov ed to be unsuitable for SR because their nature to remove range flexibility of attributes . Before entering the projection unit dimensional reduction and feature pooling via 1X

1 convolution layer is acheived.

In this DBPN architecture,output from earlier layers are concatenated to proved as a in put to each unit. Let the $h^{t\tilde{}}$ and $l^{t\tilde{}}$ be the input for dense down and up projection unit, res pectively. To merge all previous outputs from each unit, $h^{t\tilde{}}$ and $l^{t\tilde{}}$ are generated using c onv (1, N) . This improvement enables us to generate the feature maps effectively, as s hown in the experimental results.

5.3. Network architecture

The Network architecure can be divided into three portion: in the beginning oof archit ecture extraction of feature can be done, then next their projection is done , and last par t is reconstruction. Here, the convolutional layer be conv(fs, N) where N is the number of filters and where fs is the filter size.

1. In first step of feature extraction. Construction of low resolution is done.

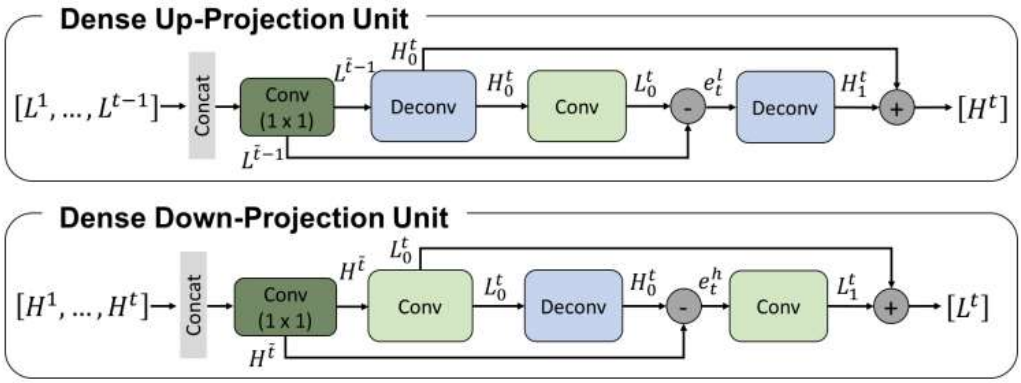

Figure 5.3: Proposed up and down-projection unit in the D-DBPN.

The feature maps of all preceding units (i.e., $[L^1, ..., L^{t-1}]$ and $[H^1, ..., H^t]$ in up- and downprojections units, respectively) are concatenated and used as inputs, and its own feature maps are used as inputs into all subsequent units.

Featuremaps L^0 from the input using conv(3, n_0). Then conv(1, n_R) is used to reduce th e dimension from n_0 to n_R before entering projection step where n_0 is the number of filte rs used in the initial LR features extraction and n_R is the number of filters used in each p rojection unit.

2. Backprojection stages. Following initial feature extraction is a sequence of projecti on units, alternating between construction of LR and HR feature maps H^t, L^t; each unit

has access to the outputs of all previous units.

3. Reconstruction. Finally, the target HR image is reconstructed as $I^{sr} = f_{Rec}([H^1, H^2, ..., H^t])$, where f_{Rec} use conv(3, 3) as reconstruction and $[H^1, H^2, ..., H^t]$ refers to the concat enation of the feature-maps produced in each up-projection unit.

Due to the definitions of these building blocks, our network architecture is modular. W e can easily define and train networks with different numbers of stages, controlling the depth. For a network with T stages, we have the initial extraction stage (2 layers), and t hen T upprojection units and T −1 downprojection units, each with 3 layers, followed b y the reconstruction (one more layer). However, for the dense network, we add conv(1 , n_R) in each projection unit, except the first three units.

CHAPTER-6

RECOGNITION

BiLSTM CTC

6.1. Recurrent Neural Networks

Self connected connectionist model containing hidden forms the Recurrent neural net works (RNNs). To allow the use of past context recurrent connection is useful because the previous input to the network is memorized by its internal state. The modulation in the rate of change of internal state by recurrent weights is necessary as it helps in build ing robustness to localised distortions of the input data which can be achieved through recurrent connection. For handwriting recognition, context is important as illustrated i n Figure 6.1 .

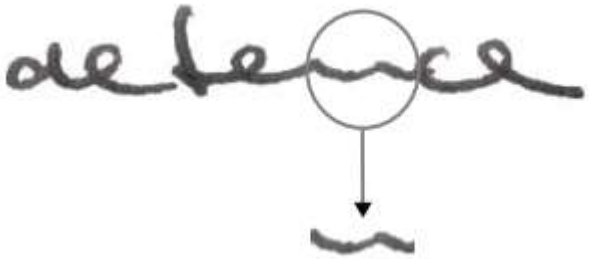

Figure 6.1: Importance of context in handwriting recognition. The word 'defence' is cl early legible, but the letter 'n' in isolation is ambiguous.

6.2 Long Short-Term Memory (LSTM)

Limited range of contextual information is provided by the Recurrent Neural Network . As the input iterates all over the network's recurrent connections, the effect of input o n hidden layer and corresponding effect on output layer decays or increases exponenti ally, which may lead to gradient vanishing and gradient explosion problem.

These problems of gradient vanishing and gradient explosion of RNN makes it difficu lt to fulfil the loss of information of more than 10 time steps between desired input and

required [23].

Long Short-

Term Memory (LSTM) [24], [25] is also the special case of RNN architecture which is particularly proposed to resolve the problem of vanishing gradient. Recurrently connected hidden layers forming the subnets connected recurrently are the units of LSTM network which are called as memory block. Each block contains a set of internal units or cell. Three multiplicative gates controls the the activation of every internal units. The input gate, forget gate and output gate are the three multiplicative gates.

From the units over a long periods of time the three gates are allowed to store and access information. As long as the input gate remains bolted the activation of the units will not be overwritten by the new inputs arriving in the network (i.e. has an activation close to 0). Similarly, when the output gate is open, the activation of units is individually permissible to the rest of the network.

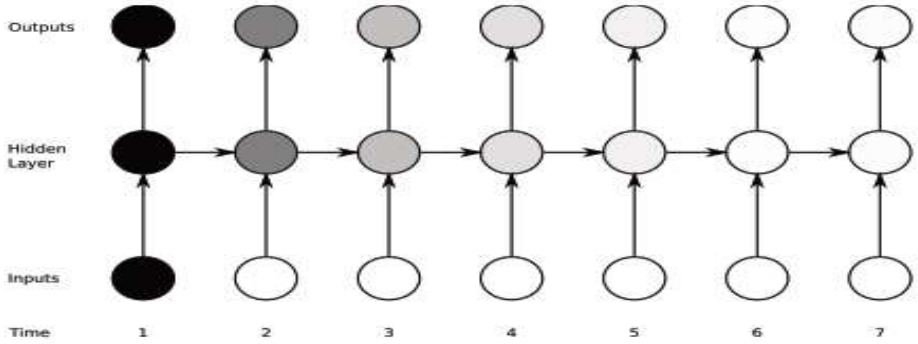

Figure 6.2: Illustration of the vanishing gradient problem.

A recurrent network unrolled in time is represented in the diagram. The elements are coloured bestowing to how subtle they are to the input at stretch 1 (where white is low and black is high). The impact of the first input decays exponentially over stretch.

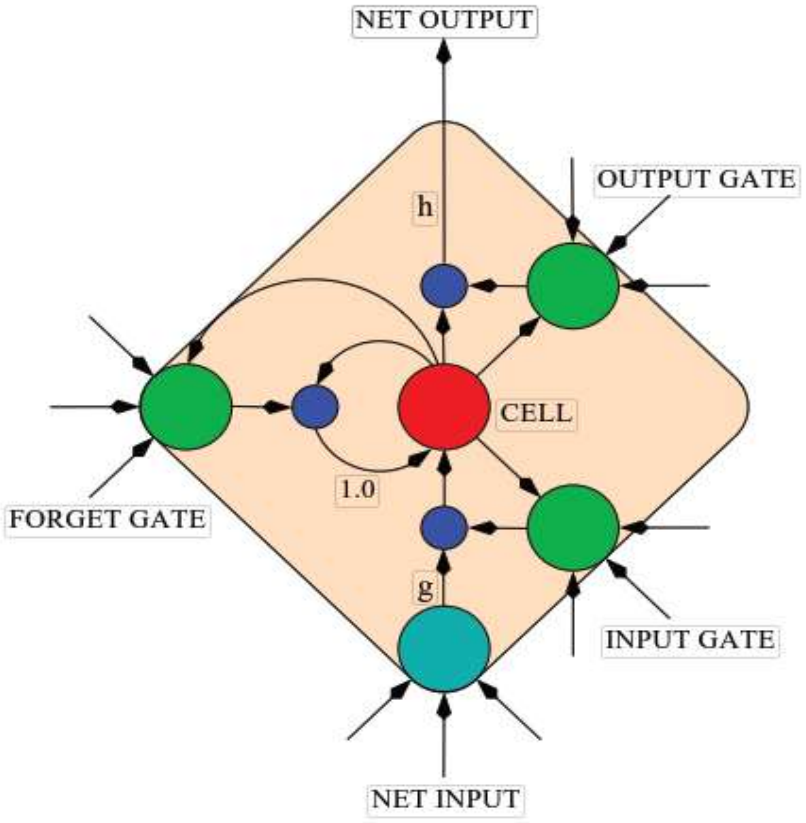

Figure 6.3: LSTM memory block with one cell.

The white recurrent connection with fixed weight 1.0 is low. From the rest of the netwo rk the three gates assembles input from the remaining network and also governs the cel l via multiplicative units (small circles). The input and output of the cell is scaled by the input and output gates while the recurrent connection of the cell is scanned by the inpu t and output gates .At the pointed positions the cell squashing functions (g and h) are ap plied . The internal connections from the cell to the gates are known as peephole weigh ts.

By the forget gate the recurrent connection of the unit is connected and disconnected. I t is to be noted that the dependency is 'carried' by the memory unit as long the input gat e is closed, as the forget gate is open and and that the output dependency can be switche d on and off by the output gate, without affecting the hidden cell.

6.3 Bidirectional Recurrent Neural Networks

It is essential to have information about past as well as future context for numerous use case. The recognition of a given letter is helped by knowing the letters both to the right and left of it in identification of text/ handwriting.

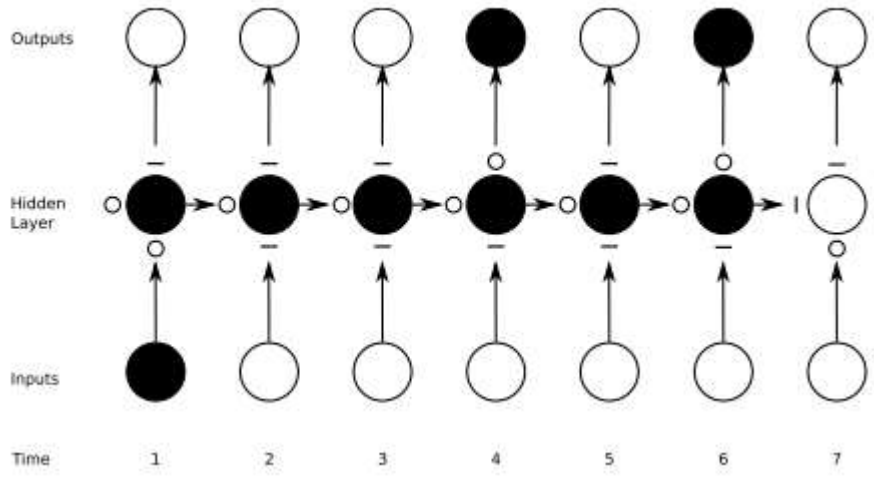

Figure 6.4: Preservation of gradient information by LSTM.

The diagram characterizes single hidden LSTM memory block of a network unrolled i n time. The input, forget, and output gate activations are respectively displayed below, to the left and above the memory block. As in Figure 13 , the shading of the units corres ponds to their sensitivity to the input at time 1. For reducing the complexity, the gates a re either entirely open ('O') or entirely closed ('—').

Bidirectional recurrent neural networks (BRNNs) [41], [42] are able to access context in both directions along the input sequence. BRNNs contain two separate hidden layer s, one of which processes the input sequence forwards, while the other processes it bac kwards. Both hidden layers are connected to the same output layer, providing it with ac cess to the past and future context of every point in the sequence. BRNNs have outperf ormed standard RNNs in several sequence learning tasks, notably protein structure pr ediction [27] and speech processing [26], [28].

Combining BRNNs and LSTM gives bidirectional LSTM (BLSTM). BLSTM has pre viously outperformed other network architectures, including standard LSTM, BRNN s and HMM-RNN hybrids, on phoneme recognition.

6.4 Deep BLSTM Recognition Engine

The network which is used as for spotting/recognising the text is proposed. 2D BLSTM's three layers are fixed together with each other or stacked. The connection are done with every hidden layer with two forget. As dimension of data is in synchronization with the architecture there is no need to normalize high weights. With the usage of the arrangement that is ground truth as the Unicode sequence, the raw binary pixels of image are provided as input to the LSTM. This network possess the maximum competence to pick up learning for recognizing the text at (i, j) that is supreme for any 2D data or the image, as two dimensional network is being used. The size of the mini batch used was 128 along with the mini batch stochastic gradient. The base learning rate was started as $1 * 10^{(-3)}$ and with the decay of 0.95. For preventing the classifier against overfitting early stopping was used. Connectionist Temporal Classification (CTC) was used as the output layer and by this use our learning of the many to many mapping among the input and output sequence could be done. Connectionist Temporal Classification is used for transforming the LSTM's output to the conditional probability distribution over all the probable sequence of the labels which is conditioned over the input sequence. For the most possible of the labelling with the modified forward backward algorithm prefix search decoding is used.

6.5 Connectionist Temporal Classification (CTC)

The conventional RNN's objective functions requires an input sequence which is presegmented. A separate target for every of the segment is used. The application the RNN has depleted which can be seen as in the cursive handwriting recognition, at the place where the segmentation is difficult to determine. Since, the output which is produced are an independent series, local classification, so different type of post processing is needed for the conversion into desirable label sequence.

The RNN output layer which is specifically designed for sequence labellingtask is called Connectionist temporal classification (CTC) .It does not require the data to be presegmented, and it directly outputs a probability distribution over label sequences. CTC h

as been shown to outperform both HMMs and RNN,HMM hybrids on a phoneme reco gnition task [31]. CTC can be used for any RNN architecture.

A CTC output layer contains as many units as there are labels in the task, plus an additi onal 'blank' or 'no label' unit. The output activations are normalised using the softma x function [46] so that they sum to 1 and are each in the range (0, 1):

$$y_k^t = \frac{e^{a_k^t}}{\sum_{k'} e^{a_{k'}^t}} \quad \ldots (6.1)$$

where a_k^t is the unsquashed activation of output unit k at time t, and y_k^t is the final output of the same unit.

The normalised outputs are used to estimate the conditional probabilities of observing label (or blank) k at time t in the input sequence x, i.e. $y_k^t = p(k, t|\mathrm{x})$ (from now on we wi ll use a bold font to denote sequences). Note that each output is conditioned on the entir e input sequence. For this reason, CTC is best used in conjunction with an architecture capable of incorporating long range context in both input directions. One such archite cture is bidirectional LSTM, as described in the previous section.

The conditional probability $p(\pi|\mathrm{x})$ of observing a particular *path* π through the lattice o f label observations is found by multiplying together the label and blank probabilities a t every time step:

$$p(\boldsymbol{\pi}|\boldsymbol{X}) = \prod_{t=1}^{T} p(\pi_t, t|\boldsymbol{X}) = \prod_{t=1}^{T} y_{\pi_t}^t \quad \ldots (6.2)$$

where π_t is the label observed at time t along path π. Paths are mapped onto label sequen ces by an operator B that removes first the repeated label, then the blanks. For example, both $B(a, -, a, b, -)$ and $B(-, a, a, -, -, a, b, b)$ yield the labelling (a,a,b). Since the path s are mutually exclusive, the conditional probability of some labelling l is the sum of th e probabilities of all the paths mapped onto it by B:

$$p(\mathbf{1}|X) = \sum_{\pi \epsilon \beta^{-1}(\mathbf{1})} p(\pi|X)$$

… (6.3)

The above step is what allows the network to be trained with unsegmented data.

6.6 CTC Forward Backward Algorithm

In the output paths to allow for blanks, modified label sequences l' are to be considered where spaces are added at the starting and the end of l, and between each pair of uninter rupted labels they are inserted. l' is therefore have the length equals to $2|l| + 1$. For the c alculation of the probabilities of prefixes of l' shifts amongst blank and nonblank label s is done , and transition is also done on any pair of distinct non-blank labels.

For a labelling l, define the *forward variable* α_s^t as the summed probability of all paths whose length *t* prefixes are mapped by *B* onto the length *s*/2 prefix of l, i.e.

$$\alpha_s^t = \sum_{\substack{\pi: \\ \beta(\pi_{1:t}) = l_{1:s/2}}} \prod_{t'=1}^{t} y_{\pi_{t'}}^{t'}$$

… (6.4)

where, for some sequence s, $s_{a:b}$ is the subsequence $(s_a, s_{a+1}, ..., s_{b-1}, s_b)$, and *s*/2 is roun ded down to an integer value. As, α_s^t can be calculated recursively.

Allowing all paths to start with0either a blank (b) or the first symbol in l (l1), the follo wing rules for initialisation are used:

$$\alpha_1^1 = y_b^1$$

$$\alpha_2^1 = y_{l_1}^1$$

$$\alpha_s^1 = 0, \forall s > 2$$

… (6.5)

and recursion:

$$\alpha_s^t = y_{l'_s}^t \begin{cases} \sum_{i=s-1}^{s} \alpha_i^{t-1} \; if \; l'_s = b \; or \; l_s = b \; or \; l'_{s-2} = l'_s \\ \sum_{i=s-2}^{s} \alpha_i^{t-1} \; otherwise \end{cases}$$

Note that

$$\alpha_s^t = 0 \, \forall s < |l'| - 2(T - t) - 1$$

… (6.6)

because2these variables correspond to states for which there are not enough time-steps left to complete the sequence.

The backward variables $\beta_s{}^t$ are defined as the summed probability of all path whose suffixes starting at t map onto the suffix of star ting at label $s/2$:

$$\beta_s^t = \sum_{\substack{\pi: \\ \beta(\pi_{t:T}) = l_s/2:|1|}} \prod_{t'=t+1}^{T} y_{\pi_{t'}}^{t'}$$

… (6.7)

The rules for initialisation and recursion of the backward variables are as follows

$$\beta_{|l'|}^T = 1$$

$$\beta_{|l'|-1}^T = 1$$

$$\beta_s^t = 0, \forall s < |l'|$$

$$\beta_s^t = \begin{cases} \sum_{i=s}^{s+1} \beta_i^{t+1} y_{l_i'}^{t+1} \; if \; l_s' = b \; or \; l_{s+2}' = l_s' \\ \sum_{i=s}^{s+2} \beta_i^{t+1} y_{l_i'}^{t+1} \; otherwise \end{cases}$$

… (6.8)

Note that

$$\beta_s^t = 0 \, \forall s > 2t$$

because these variables correspond to unreachable states.

Finally, the label sequence probability is given by the sum of the products of the forwar d and backward variables at any time:

$$p(l|x) = \sum_{s=1}^{|l'|} \alpha_s^t \, \beta_s^t$$

… (6.9)

6.7 CTC Objective Function

The CTC objective function is defined as the negative log probability of the network c orrectly labelling the entire training set. Let *S* be a training set, consisting of pairs of inp ut and target sequences (x, z). Then the objective function *O* can be expressed as

$$O = - \sum_{(x,z) \epsilon S} ln \, p(z|x)$$

… (6.10)

The network can be trained with gradient descent by first differentiating *O* with respec t to the outputs, then using backpropagation through time to find the derivatives with re spect to the network weights.

Noting that the same label (or blank) may be repeated several times for a single labellin

g l, defining the set of positions where label *k* occurs as

$$\boldsymbol{lab(l,k) = \{s: l'_s = k\}}$$

… (6.11)

which may be empty. Then set l = z and differentiate it with respect to the network outp uts to obtain

$$\frac{\partial p(z|x)}{\partial y_k^t} = \frac{1}{y_k^t} \sum_{s \in lab(z,k)} \alpha_s^t \beta_s^t$$

… (6.12)

Substituting this, we get

$$\frac{\partial O}{\partial y_k^t} = -\frac{1}{p(z|x) y_k^t} \sum_{s \in lab(z,k)} \alpha_s^t \beta_s^t$$

… (6.13)

To backpropagate the gradient through the output layer, the objective function derivat ives with respect to the outputs a_k^t before the activation function is applied. For the soft max function

$$\frac{\partial y_{k'}^t}{\partial \alpha_k^t} = y_{k'}^t \delta_{kk'} - y_{k'}^t y_{k'}^t$$

… (6.14)

and therefore

$$\frac{\partial O}{\partial u_k^t} = y_k^t - \frac{1}{p(z|x)} \sum_{s \in lab(z,k)} \alpha_s^t \beta_s^t$$

… (6.15)

RESULTS

Generated data via data augmentation makes a system very robust as the generated im ages could be both close to a handwritten or printed document with varied font types an d style of writingTable I shows the recognition accuracy and PSNR for IAM datasets. The results have been compared with endtoend trainable proposed framework with re sults from HR image, LR4x, LR3x and other cascaded networks of denoising, followe d by superresolution and OCR. In all cases the OCR engine used is the same Bi LSTM CTC network architecture. For the dataset, the LR images version suffered a lot in ter ms of recognition accuracy and PSNR. RedNet+DBPN had jointly trained denoising a nd superresolution and then uses the superresolved output as input to a BLSTM based OCR. This network although worked better than SSDA+CDCA, where each piece wa s trained separately. The proposed framework outperformed all the cascaded models with a significant margin and is at par with the OCR accuracy and PSNR of the SR4x v ersion of the same image. As explained earlier, individually trained models when casc aded together carry their errors, thus effecting the final recognition rates. At that point even a very robust OCR system would fail due to issues in the image enhancement part s.

Table1

Comparison of Proposed Enhancement and Recognition Frameork On OCR Accuracy And PSNR.

Enhancement Methods	Noise	Handwritten text IAM Database OCR Accuracy	
		PSNR	
HR Image		93.04	
SR4x		91.23	28.14
RedNet+DBPN 4x	G 0.01 G 0.1 SnP0.01 SnP0.1	91.14 90.09 91.20 90.34	28.08 27.56 28.01 27.39
SSDA+CDCA 4x	G 0.01 G 0.1 SnP0.01 SnP0.1	80.52 76.97 80.16 75.99	20.98 18.97 20.59 19.04
SR3x		92.56	29.04
RedNet+DBPN 3x	G 0.01 G 0.1 SnP0.01 SnP0.1	92.39 92.07 92.48 91.86	28.86 28.29 28.89 28.19
SSDA+CDCA 3x	G 0.01 G 0.1 SnP0.01 SnP0.1	82.56 76.87 81.38 74.78	23.29 20.17 22.91 20.05

NOTE: Only in Proposed framework in all cases is BLSTMCTC. both enhancement module and OCR are trained in endtoend fashion in all other cases these modules are c ascaded.

Fig. 7.1: Illustrates the comparison between the OCR output before and after the enhan cement of degraded document image for printed text. It can be clearly seen that enhanc ement done by proposed framework has provided a significant boost to OCR accuracy . We observed that GAN have the capability to generate visually better images but not i n terms of PSNR.

Figure 7. OCR performance comparison between Degraded and Enhanced Document image

To demonstrate performance of the OCR part, we show in Table II the comparison bet ween endtoend trained proposed framework, cascaded modules of our framework and our enhancement module cascaded with Tesseract. The issues could be from the SR fr amework which might have increased some noise leading to merged characters or chal lenging text segmentation, or from denoising/inpainting which either loses actual text information or adds extra info (inpainting) leading adding its errors to SR module.

Table 2

Comparison Of Proposed Framework With Other Cascaded Frameworks For Robust 4x Enhancement And Recognition.

Module for Robust Recognition	**Noise**	**OCR Accuracy**
Proposed Framework	G 0.01 G 0.1	92.16 91.87
Our Enhancement Module + Google Tesseract	G 0.01 G 0.1	80.15 77.08

The outputs of proposed framework is clearly outperforming the other conventional a nd stateoftheart frameworks for degraded document image enhancement and recognit ion and in turn the proposed framework will aid in increasing the OCRs accuracy.

CONCLUSIONS

This paper hypothesizes and validates experimentally that an endtoend trainable netw ork that jointly optimizes image enhancement (denoising/inpainting and SR) and text recognition would make a highperforming robust OCR system for degraded lowresol ution document images. The data augmentation performed helped to recognise every kind of font style and degraded document images with real noises. Experimental result s show the robustness and efficacy of framework for printed and handwritten degrade d lowresolution document images with different types of noises, scripts, writing and f ontstyles. Stateoftheart recognition results have been achieved for both printed and ha ndwritten datasets for English. Exhaustive comparison has been shown with other stat e-of-the-art or conventional frameworks for robust text recognition.

REFERENCES

[1] Manoj Sharma, Anupama Ray, Santanu Chaudhury, Brejesh Lall "A No se- Res l ent Super-Resolut on framework to boost OCR performance", CDAR 2017.

[2] Mohammed Javed' · P. Nagabhushan, B dyut B. Chaudhur "A rev ew on document mage analys s techn ques d rectly n the compressed doma n" publ shed onl ne: 21 March 2017© Spr nger Sc ence+Bus ness Med a Dordrecht 2017.

[3] K. Dabov, A. Fo , V. Katkovn k, and K. O. Eg azar an. mage Deno s ng by sparse 3-d transform-doma n collaborat ve filter ng. EEE Trans. mage Process, 16(8):2080–2095, 2007.

[4] S. Gu, W. Zuo, Q. X e, D. Meng, X. Feng, and L. Zhang. Convolut onal sparse cod ng for mage super-resolut on. n Proc. EEE nt. Conf. Comp. V s., pages 1823–1831, 2015.

[5] P. Chatterjee and P. M lanfar. Cluster ng-based deno s ng w th locally learned d ct onar es. EEE Trans. mage Process., 18(7):1438–1451, 2009.

[6] K. Chatf eld, K. S monyan, A. Vedald , and A. Z sserman, "Return of the dev l n the deta ls: Delv ng deep nto convolut onal nets," Proc. BMVC, 2014.

[7] J. P. Patr ce Y. S mard, Dave Ste nkraus, "Best pract ces for convolut onal neural networks appl ed to v sual document analys s." nst tute of Electr cal and Electron cs Eng neers, nc., August 2003.

[8] P. Kr shnan and C. V. Jawahar, "Match ng handwr tten document mages," The 14th European Conference on Computer V s on (ECCV), 2016.

[9] Data Augmentat on for Recogn t on of Handwr tten Words and L nes us ng a CNN-LSTM Network.

[10] V. A. Lamme and P. R. Roelfsema. "The d st nct modes of v -s on offered by feedforward and recurrent process ng". Trends n neurosc ences, 23(11):571–579, 2000.

[11] " mprov ng Offl ne Handwr tten Text Recogn t on w th Hybr d HMM/ANN Models" EEE Transact ons On Pattern Analys s And Mach ne ntell gence, Vol. 33, No. 4, Apr l 2011.

[12] Abh shek S ngh "Super-Resolv ng No sy mages" Un vers ty of ll no s at Urbana-Champa gn, CVPR2014

[13] C. Dong, C. C. Loy, K. He, and X. Tang. mage super-resolut on us ng deep convolut onal networks. EEE Trans. Pattern Anal. Mach. ntell., 38(2):295–307,2016.

[14] J. X e, L. Xu, and E. Chen. mage deno s ng and npa nt ng w th deep neural networks. n Proc. Advances n Neural nf. Process. Syst., pages 350–358, 2012.

[15] H. C. Burger, C. J. Schuler, and S. Harmel ng. mage deno s ng: Can pla n neural networks compete w th BM3D? n Proc. EEE Conf. Comp. V s. Patt. Recogn., pages 2392–2399, 2012.

[16] V. Ja n and H. S. Seung. Natural mage deno s ng w th convolut onal networks. n Proc. Advances n Neural nf. Process. Syst., pages 769–776, 2008.

[17] Alex Graves, Faust no Gomez Schm dhuber, "Connect on st temporal class ficat on Labell ng unsegmented.

sequence data w th recurrent neural networks," n Proceed ngs of the 23rd nternat onal Conference on Mach ne Learn ng, 2006, CML '06.

[18] A. Graves, M. L w ck , S. Fernndez, R. Bertolam , H. Bunke, and J. Schm dhuber, "A novel connect on st system for unconstra ned handwr t ng recogn t on," EEE Transact ons on Pattern Analys s and Mach ne ntell gence,
2009.

[19] J. Hu, S. L m, and M. Brown, "Wr ter ndependent on-l ne handwr t ng recogn t on us ng an HMM approach," Pattern Recogn t on, vol. 33, no. 1, pp. 133–147, January 2000.

[20] X ao-J ao Mao, Chunhua Shen, and YuB n Yang, " mage restorat on us ng convolut onal auto-encoders w th symmetr c sk p connect ons," CoRR, vol. abs/1606.08921, 2016.

[21] He, K., Zhang, X., Ren, S., Sun, J.: Deep res dual learn ng for mage recogn t on. n: Proceed ngs of the EEE Conference on Computer V s on and Pattern Recogn t on. pp. 770–778 (2016).

[22] Szegedy, C., L u, W., J a, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rab nov ch, A.: Go ng deeper w th convolut ons. EEE Conference on Computer V s on and Pattern Recogn t on. pp. 1–9 (2015).

[23] S. Hochre ter, Y. Beng o, P. Frascon , and J. Schm dhuber, "Grad ent flow n recurrent nets: the d fficulty of learn ng long-term dependenc es," n A F eld Gu de to

Dynam cal Recurrent Neural Networks, S. C. Kremer and J. F. Kolen, Eds. EEE Press, 2001.

[24] S. Hochre ter and J. Schm dhuber, "Long Short-Term Memory," Neural Computat on, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] F. Gers, N. Schraudolph, and J. Schm dhuber, "Learn ng prec se t m ng w th LSTM recurrent networks," Journal of Mach ne Learn ng Research, vol. 3, pp. 115–143, 2002.

[26] M. Schuster and K. K. Pal wal, "B d rect onal recurrent neural networks," EEE Transact ons on S gnal Process ng, vol. 45, pp. 2673–2681, November 1997.

[27] P. Bald , S. Brunak, P. Frascon , G. Pollastr , and G. Soda, "B d rect onal dynam cs for prote n secondary structure pred ct on," Lecture Notes n Computer Sc ence, vol. 1828, pp. 80–104, 2001.

[28] T. Fukada, M. Schuster, and Y. Sag saka, "Phoneme boundary est mat on us ng b d rect onal recurrent neural networks and ts appl cat ons," Systems and Computers n Japan, vol. 30, no. 4, pp. 20–30, 1999.

[29] J. K m, J. Kwon Lee, and K. Mu Lee. Deeply-recurs ve convolut onal network for mage super-resolut on. n Proceed ngs of the EEE Conference on Computer V s on and Pattern Recogn t on, pages 1637–1645, 2016.

[30] J. Johnson, A. Alah , and L. Fe -Fe . Perceptual losses for real-t me style transfer and super-resolut on. n European Conference on Computer V s on, pages 694–711. Spr nger, 2016.

[31] C. Dong, C. C. Loy, K. He, and X. Tang. mage super-resolut on us ng deep convolut onal networks. EEE transact ons on pattern analys s and mach ne ntell gence, 38(2):295–307, 2016.

[32] C. Dong, C. C. Loy, and X. Tang. Accelerat ng the super-resolut on convolut onal neural network. n European Conference on Computer V s on, pages 391–407. Spr nger, 2016.