ITCS 6156 - Project Proposal

Keywords Identification from Handwritten Doctor's Prescription

Team Members

Team	Mavericks	
Name	Login	800#
Akanksha Bhalla	abhalla	801077998
Dhananjay Arora	darora2	801077164
Naman Manocha	nmanocha	801077765
Shubham Gupta	sgupta37	801081963

Introduction

Problem

As we all know that doctors have illegible handwriting and it is difficult for individuals from non-medical background to understand the disease and the medicines mentioned in the report. Most of the times, even pharmacists find it difficult to decipher the handwriting written in medical reports. This is the problem we have been observing from decades and many have suffered due to this problem.

In the modern era, few doctors have started to provide digitized prescriptions to maintain records, but most doctors still provide traditionally handwritten prescriptions on their printed letterhead. This is one of the main reasons we want to work on this problem.

Moreover, it is difficult to store and access physical documents in an efficient manner, search through them efficiently and to share them with others. Thus, a lot of important data gets lost or does not get reviewed because documents(prescription) never get transferred to digital format. This is another reason for us to work on this problem.

There are many existing models which convert handwritten images to digital text, but our aim is to go an extra mile and identify the keywords or medical terms from the prescriptions such as the disease, medicines prescribed, medical tests suggested, etc.

Motivation

Many people face a problem in identifying or understanding keywords written by doctors in their prescriptions and end up purchasing wrong medicines. One of us also faced a similar problem. Naman got a prescription from the doctor when he was suffering from pain in his abdominal area. Even though he got the medication he still didn't know what it was for. On his next visit to the doctor, he questioned why he was still having pain. The doctor said that he had mentioned in his prescription that he was suffering from a kidney stone and that might take more time for him to recover. Similarly, many people face such problems, and by building this model we are trying to overcome such issues.

Dataset

We have got the IAM Handwriting Dataset for the initial training of our model. This dataset is very large, and it consists of images of handwritten text segregated by words and sentences which we will be using. Apart from this, we build our own dataset using our Family & Friends Contact. Medical Dataset is very sensitive and kept confidential. So, we built our own dataset and plan not to share it with anyone in the future or upload to any repository.

Approach

Here is the brief high-level approach to solving this problem. We have discussed our approach for each of the below algorithm/models in detail in the "**Methods**" section of this report.

- Build dataset.
- Apply Image processing techniques such as:
  - Rescaling
  - Skew Correction
  - Slant Correction
  - Erosion and Dilation
  - Image Thresholding
- Page Segmentation - Segregate the handwritten text from the printed text
- Line Segmentation - Segregate each line from the segmented handwritten text
- Handwritten Text Recognition - Use Deep CNN and bi-LSTM to convert image text to characters.

- Keyword Spotting - Using the predefined python repositories of medical terms, we will be spotting the medical keywords from the medical prescriptions

Literature Survey

Paper: Handwritten Text Recognition using Deep Learning [4]

This paper classifies handwritten word and converts them to digital text. They try to accomplish the task using two approaches, one with classifying words directly, for which Convolutional Neural Network (CNN) was used and another approach based on character segmentation for which Long Short-Term memory was used to construct boundary boxes for each character.

They use Hidden Markov Model for the task of OCR which results in high accuracy for the OCR. A CNN with two convolution layers and two average pooling layers and the fully connected layer were used to classify. The architecture followed in this paper assumes that input images will be of the same size. To achieve this task, they pad the image with whitespace to the maximum width and height of the dataset.

Another important aspect mentioned by them was that some images of the words in the dataset was tilted. They decided to rotate the image to the right slightly with random probability before training. They then tried for the character level classification and further run various experiments for both approaches and with a different combination of hyperparameters.

Finally, they conclude that character level classification was more successful and mentioned that due to lack of enough data for their problem and their model suffers from segmenting cursive characters.

Paper: Offline Handwritten English Character Recognition based on Convolutional Neural Network [5]

This paper applied LeNet-5 CNN model with special settings for several neurons in each layer on UNIPEN dataset. The author of this paper has described his self-developed algorithm for training based on reinforcement of error samples, which they called Error-Samples-Reinforcement-Learning (ESRL) algorithm.

The CNN Model used in this paper is modified LeNet-5 with various modifications to tradeoff time-cost and recognition performance. CNN with 32 X 32 and 28 X 28 normalization lose information more quickly than 20 X 20 normalization. For preprocessing the UNIPEN Dataset, they used the DDA method and anti-aliasing. LeNet-5 with ESRL algorithm had the ability to reject illegal samples for printed character recognition. Hamming Distance was used to compare with rejecting distance.

Paper: Support Vector Machine for English handwritten character recognition [3]

This paper proposed the Freeman Chain Code as the representation technique for image character recognition. This code gives the boundary of a character image and the location of the next pixel. After that, the feature vector is built. Finally, the Support Vector Machine (SVM) is used to classify the text. Dataset was taken from NIST (National Institute of Standards and Technology).

The proposed model starts with pre-processing. It involves process like thinning which produces a skeleton of a character. This character is fed to the next stage which is feature extraction. FCC is generated from characters which are used as features for classification. It traverses through each pixel of the character. The heuristic function is used to generate FCC to represent characters correctly and in the final step, SVM classifiers help to build the recognition model. Test results showed high accuracy at that time of research.

Paper: Keyword spotting in doctor's handwriting on medical prescriptions [1]

There are many significant impacts in this piece of proposed work, for example, this work can be used in developing expert diagnostic systems, in extracting information from patient history, in detecting wrong medication and in making different statistical analyses of the medicines prescribed by the doctors. In order to extract the information from such document images, domain-specific knowledge of doctors was firstly extracted by identifying department names from the printed text that appears in the letterhead part. Then the specialty/expertise of doctors was understood by this letterhead text which helped to search only relevant prescription documents for word spotting in the handwritten portion.

The word spotting in letterhead part as well as in handwritten part has been performed using the Hidden Markov Model in this paper. The objective of this paper was aimed at the automatic transcription of the handwritten prescription with the help of keyword spotting.

Firstly, this paper proposes an approach by incorporating domain knowledge into the text retrieval process. From the printed letterhead part of the prescription, attributes are retrieved, regarding the doctor specifications, with the help of word spotting in printed text part. The doctor’s domain thus extracted is indexed with the search for disease/medicine keywords.

It utilizes a handwritten word spotting methodology to search the queried keywords in the written text portion of the relevant prescription. A Tandem-HMM system is proposed for word spotting. A PHOG feature was used for an efficient word recognition system. The proposed system in this paper attempts to eradicate the problems that a patient faces while reading a prescription because of poor handwriting of doctors.

Paper: A New Approach to Information Retrieval based on Keyword Spotting from Handwritten Medical Prescriptions [1]

This paper mainly focuses on spotting words in the handwritten portion of the prescription using the Hidden Markov Model (HMM). They performed experiments on 500 different types of prescriptions. This paper also proposed medical knowledge for text recognition.

Their first part was to get details about doctors like department and designation from the printed part of doctor's letterhead with the help of OCR. Thus, extracted details from the printed part mapped with disease and medicine prescribed by the doctor. And for the offline word spotting in the prescription is done with the help of the Hidden Markov Model. HMM is used to model sequential dependencies. Also, the separate Gaussian Mixture Model (GMM) is defined for each state of the model. This paper has used word level classification using the HMM approach.

Method

To get an idea on how to work with an image dataset, we first tried the below steps:

1. We have tried to predict the handwritten text with the help of the Tesseract [11] OCR library in Python, but we figured out that it works well with printed words but not with the handwritten text. So, we explored other ways to do so (Code uploaded our GitHub).

2. After that, we successfully ran the model built by Harald, Simple HTR [6] on the **DSBA Hadoop Cluster** due to computational complexities of our individual machines. In this algorithm we mainly use cv2, NumPy and TensorFlow imports. This model consists of 5 CNN layers, followed by 2 RNN (LSTM) layers and a CTC layer (for loss and decoding). The input is a gray value image of size 128x32, which is passed down to the CNN layers which are of size 32x256. The CNN layers match feature sequences to the image. In the LSTM

layers the feature sequence is matched to a matrix of size 32x80. The matrix stores the scores that are assigned to each of the 80 characters after they've passed through the first two steps. In the CTC layer if we are training the data then it returns the loss value and ground truth value. If we are validating the data then the CTC layer returns the decoded output, that is, it gives the final text either by using best path decoder or beam search decoder. The batch size that we have kept is 50, so that the beam width of word beam search can conform to the beam width of vanilla beam search. We first create a class called class FilePaths in the main.py program. In this class we start by reading the data. The variables we declare and define are CharList (a text file of 80 unique characters), fnAccuracy (this consists of the validation character error rate of the saved model), fnTrain (training data), fnInfer (test image), fnCorpus (dictionary containing all IAM dataset words). We then define four methods. The first method train () takes two parameters that is model and loader. This function starts with initializing variables epoch, bestCharErrorRate (which is the best validation character's error rate) and noImprovementSince (this is the number of epochs where there was no improvement in the charErrorRate) and are initialized with the value zero. The variable earlyStopping represents the number of epochs where we should stop if there is no improvement in the error rate and is set to 5. The method starts with increasing the epoch value by 1. It checks if there is still data in the training set, if it is then it calculates the loss of that batch and prints the batch information along with the loss. Then it validates the data by calling the validate () method and storing that value in a variable called charErrorRate. It then checks if charErrorRate is less than besCharErrorRate or not, if yes then it saves the model if not then it prints that there was no improvement and the noImprovementSince is increased by 1. The training stops when the value of noImprovementSince is greater than equal to 5 (i.e, the value of earlyStopping). In the validate () method we take in parameters model and loader. We start iterating over the data in first batch and find if any character has been recognized. If it is recognized, then we set that value as the ground truth value. Next, we find validation error and store it in charErrorRate and this value is returned by the method. We then define the infer () method which returns the probability of the word being recognized correctly. In the main () function we first choose which decoding method we want to use out of vanilla beam search or word beam search. Then we get to choose if we want to train or validate our data. If we choose to train, then we call the train method otherwise we call the validate function else we call infer function on the test image.

3. **Performed Beam Search on Word Data** [6] **-** Beam Search is a heuristic search algorithm that explores a graph by expanding the most promising node in a limited set. Beam search is an optimization of best-first search that reduces its memory requirements.

```
[darora2@mba-i2 src]$ python main.py –beamsearch

Validation character error rate of saved model: 11.334136%
Python: 3.6.5 |Anaconda, Inc.| (default, Apr 29, 2018, 16:14:56)
[GCC 7.2.0]
TensorFlow: 1.10.1
2019-04-03 13:37:40.602257: I
tensorflow/core/platform/cpu_feature_guard.cc:141] Your CPU
supports instructions that this TensorFlow binary was not
compiled to use: AVX2 FMA
Init with stored values from ../model/snapshot-24
Recognized: "little"
Probability: 0.69998664
```

Fig 1: Result of Word Beam Search

Note: *Harald's SimpleHTR Model is built on the IAM Dataset that consisted only of words, but our problem is much more complex. We are going build a model for Medical Documents and take entire scanned image dataset instead of a dataset that consisted of well pre-processed individual words. In addition to this, once we identify the text in documents, we plan to spot the medical jargons and terminologies that can be helpful in real-time.*

Maverick's Approach:

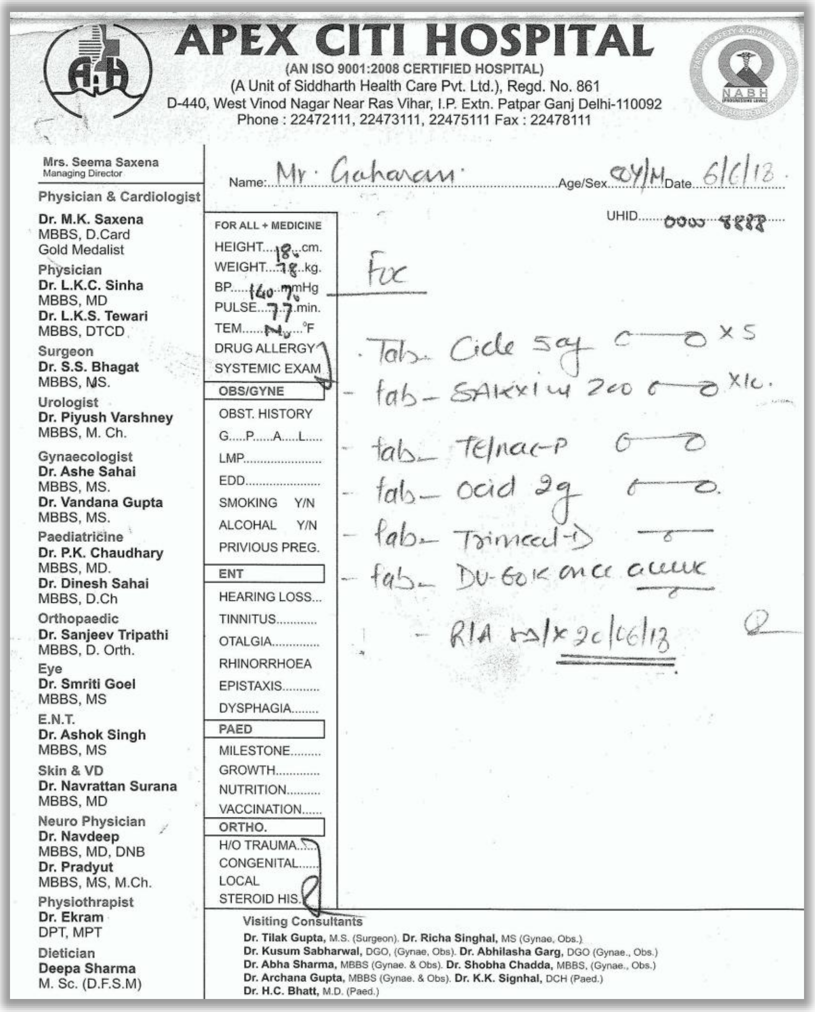
Here is our detailed approach for this project after Literature Survey & reviewing existing implementations. We have classified it into 5 major blocks, and we will be discussing each one of them in detail now:

A. Image Preprocessing:

In this section, we tried and tested out various image preprocessing techniques that were required for our dataset. Here is a summary of what we did:

1. **Rescaling**: Inter cubic interpolation in CV to scale the image to a larger size.
2. **Grayscale:** Converting the image to grayscale.
3. **Erosion and Dilation:** To reduce the noise we applied both the techniques with a various combination of kernel sizes.
4. **Blurring**: In order to reduce the noise, out of 3 methods i.e. Average, Gaussian, Median Blurring and we found that Gaussian, Median Blurring works fine with our image data.

5. **Image Thresholding**: Segmented the image in foreground and background. We tried various thresholding techniques and found **Adaptive Gaussian** Thresholding from CV2 provides a good image for our dataset.
6. Since the documents were manually scanned by inexperienced people, most of them were skewed. Hence, image skew correction was necessary. Skew correction is one of the first operations to be applied to scanned documents or images when converting data to a digital format. Its aim is to align an image before processing because text segmentation and recognition methods require properly aligned text data.
   To skew an image containing a rotated block of text at an unknown angle, we corrected the text skew by:
   i. Detecting the block of text in the image.
   ii. Computing the angle of the rotated text.
   iii. Rotating the image to correct for the skew.



APEX CITI HOSPITAL
(AN ISO 9001:2008 CERTIFIED HOSPITAL)
(A Unit of Siddharth Health Care Pvt. Ltd.), Regd. No. 861
D-440, West Vinod Nagar Near Ras Vihar, I.P. Extn. Patpar Ganj Delhi-110092
Phone : 22472111, 22473111, 22475111 Fax : 22478111

NABH

Mrs. Seema Saxena
Managing Director

Physician & Cardiologist
Dr. M.K. Saxena
MBBS, D.Card
Gold Medalist

Physician
Dr. L.K.C. Sinha
MBBS, MD
Dr. L.K.S. Tewari
MBBS, DTCD

Surgeon
Dr. S.S. Bhagat
MBBS, MS.

Urologist
Dr. Piyush Varshney
MBBS, M. Ch.

Gynaecologist
Dr. Ashe Sahai
MBBS, MS.
Dr. Vandana Gupta
MBBS, MS.

Paediatricine
Dr. P.K. Chaudhary
MBBS, MD.
Dr. Dinesh Sahai
MBBS, D.Ch

Orthopaedic
Dr. Sanjeev Tripathi
MBBS, D. Orth.

Eye
Dr. Smriti Goel
MBBS, MS

E.N.T.
Dr. Ashok Singh
MBBS, MS

Skin & VD
Dr. Navrattan Surana
MBBS, MD

Neuro Physician
Dr. Navdeep
MBBS, MD, DNB
Dr. Pradyut
MBBS, MS, M.Ch.

Physiothrapist
Dr. Ekram
DPT, MPT

Dietician
Deepa Sharma
M. Sc. (D.F.S.M)

Name: Mr · Gaharam · Age/Sex 60Y/M Date 6/6/13 ·

UHID 0000 4888

FOR ALL + MEDICINE
HEIGHT 186 cm.
WEIGHT 78 kg.
BP 140 76 mmHg
PULSE 77 min.
TEM N °F
DRUG ALLERGY
SYSTEMIC EXAM

OBS/GYNE
OBST. HISTORY
G.....P.....A.....L.....
LMP....................
EDD....................
SMOKING Y/N
ALCOHAL Y/N
PRIVIOUS PREG.

ENT
HEARING LOSS...
TINNITUS...........
OTALGIA.............
RHINORRHOEA
EPISTAXIS..........
DYSPHAGIA........

PAED
MILESTONE.........
GROWTH.............
NUTRITION..........
VACCINATION......

ORTHO.
H/O TRAUMA......
CONGENITAL......
LOCAL
STEROID HIS.

Fuc

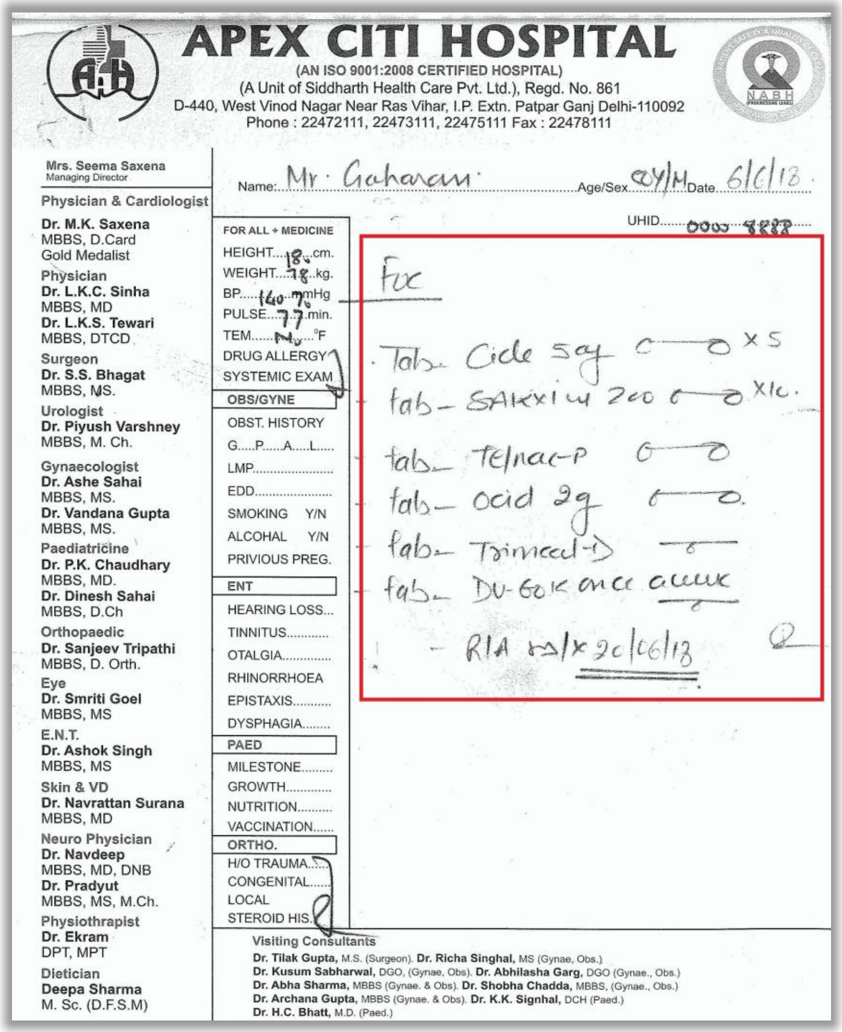
· Tab. Cicle 5g o—o ×5
- fab – SAKXIM 200 o—o ×10.
- fab – Telnac-P o—o
- fab – acid 2g o—o.
- fab – Trimead-D —o—
- fab – DU-60K once week —o—
- RIA r/x 20/06/13

Visiting Consultants
Dr. Tilak Gupta, M.S. (Surgeon). **Dr. Richa Singhal,** MS (Gynae, Obs.)
Dr. Kusum Sabharwal, DGO, (Gynae, Obs). **Dr. Abhilasha Garg,** DGO (Gynae., Obs.)
Dr. Abha Sharma, MBBS (Gynae. & Obs). **Dr. Shobha Chadda,** MBBS, (Gynae., Obs.)
Dr. Archana Gupta, MBBS (Gynae. & Obs). **Dr. K.K. Signhal,** DCH (Paed.)
Dr. H.C. Bhatt, M.D. (Paed.)

Fig 2: Sample medical prescription in India

B. Paragraph Segmentation:

Once we processed the input image, we tried to segment the handwritten text from the image. A medical prescription consists of printed and handwritten text both. Our aim is to focus on Handwritten text. So, we used Deep CNN network to segment the paragraphs of text from the image. We found out the height and width of the handwritten text and used the bounding box to label the Handwritten data.



APEX CITI HOSPITAL

(AN ISO 9001:2008 CERTIFIED HOSPITAL)
(A Unit of Siddharth Health Care Pvt. Ltd.), Regd. No. 861
D-440, West Vinod Nagar Near Ras Vihar, I.P. Extn. Patpar Ganj Delhi-110092
Phone : 22472111, 22473111, 22475111 Fax : 22478111

Mrs. Seema Saxena
Managing Director

Physician & Cardiologist
Dr. M.K. Saxena
MBBS, D.Card
Gold Medalist

Physician
Dr. L.K.C. Sinha
MBBS, MD
Dr. L.K.S. Tewari
MBBS, DTCD

Surgeon
Dr. S.S. Bhagat
MBBS, MS.

Urologist
Dr. Piyush Varshney
MBBS, M. Ch.

Gynaecologist
Dr. Ashe Sahai
MBBS, MS.
Dr. Vandana Gupta
MBBS, MS.

Paediatricine
Dr. P.K. Chaudhary
MBBS, MD.
Dr. Dinesh Sahai
MBBS, D.Ch

Orthopaedic
Dr. Sanjeev Tripathi
MBBS, D. Orth.

Eye
Dr. Smriti Goel
MBBS, MS

E.N.T.
Dr. Ashok Singh
MBBS, MS

Skin & VD
Dr. Navrattan Surana
MBBS, MD

Neuro Physician
Dr. Navdeep
MBBS, MD, DNB
Dr. Pradyut
MBBS, MS, M.Ch.

Physiothrapist
Dr. Ekram
DPT, MPT

Dietician
Deepa Sharma
M. Sc. (D.F.S.M)

Name: Mr. Gaharam. Age/Sex 80Y/M Date 6/6/13.

UHID 0003 8882

FOR ALL + MEDICINE
HEIGHT 186 cm.
WEIGHT 78 kg.
BP 140/76 mmHg
PULSE 77 min.
TEM N °F
DRUG ALLERGY
SYSTEMIC EXAM

OBS/GYNE
OBST. HISTORY
G.....P.....A.....L.....
LMP......................
EDD......................
SMOKING Y/N
ALCOHAL Y/N
PRIVIOUS PREG.

ENT
HEARING LOSS...
TINNITUS............
OTALGIA.............
RHINORRHOEA
EPISTAXIS...........
DYSPHAGIA........

PAED
MILESTONE.........
GROWTH............
NUTRITION..........
VACCINATION......

ORTHO.
H/O TRAUMA......
CONGENITAL......
LOCAL
STEROID HIS.

Rx

Tab- Cicle 5g o—o ×5
tab- SAKXIM 200 o—o ×10.
tab- Telnac-P o—o
tab- ocid 20 o—o.
tab- Trimcd-D —o—
tab- DU-60K once week —o—
— R/A s/x 20/06/13

Visiting Consultants
Dr. Tilak Gupta, M.S. (Surgeon). Dr. Richa Singhal, MS (Gynae, Obs.)
Dr. Kusum Sabharwal, DGO, (Gynae, Obs). Dr. Abhilasha Garg, DGO (Gynae., Obs.)
Dr. Abha Sharma, MBBS (Gynae. & Obs). Dr. Shobha Chadda, MBBS, (Gynae., Obs.)
Dr. Archana Gupta, MBBS (Gynae. & Obs). Dr. K.K. Signhal, DCH (Paed.)
Dr. H.C. Bhatt, M.D. (Paed.)

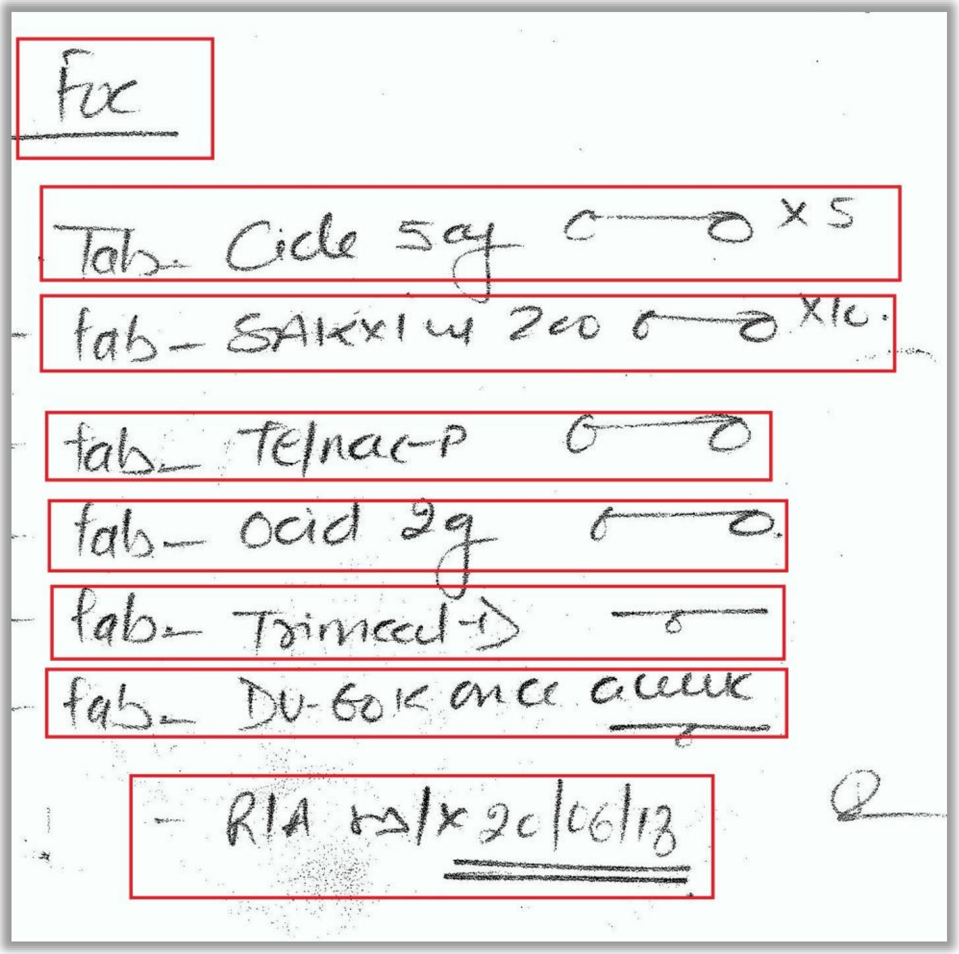
Fig 3: Paragraph Segmentation

C. Line Segmentation:

After segmentation of image to paragraphs, we will pass each paragraph bounded by boxes as an input for our Line Segmentation model and it will output the image marking each line bounded by boxes.

For the segmentation of paragraph into the lines, we will be using **Single Shot MultiBox Detector (SSD)** with horizontal anchor boxes to identify each line from the paragraph. SSD predicts the bounding boxes and class as it processes the image simultaneously.

Our plan for feature extraction is to create a network based on ResNet [8] because the main advantage of Resnet is that it solves the famous problem of **Vanishing Gradient.** With ResNet, gradients can back propagate by skipping some connections from last to front layer and the size of ResNet depends on how big our layer is and how many layers do we have.



Fxc

Tab- Cicle 5ay ○—○ x5

Tab- SAKXIN 200 ○—○ x10.

Tab- Telnac-P ○—○

Tab- ocid 2g ○—○.

Tab- Trimcad-D —○

Tab- DU-60K once aweek

R/A after 20/06/13

Fig 4: Line Segmentation

D. Handwriting Recognition:

We will build a model that will generate the image features from the input image through CNN, and we will feed those features to LSTM sequentially. We will use an encoding layer for extracting the image features from CNN. Then we will try to downsample so that the features of the image are reduced. We will also create a feature extraction network. Most of the previous work [7] was done with multi-dimensional LSTM to recognize the handwriting. But we will implement that using single dimensional LSTM as it is computationally much cheaper than multidimensional.

E. Keyword Spotting:

For keyword spotting, medical keyword repositories such as PyMedTermino[10], PubMed, and RISmed which are developed by National Institute of Health will be used to hunt for medical keywords in the recognized text.

Updated Plan

This was the initial high-level plan; we have marked the tasks that are completed successfully. Along with this, **as per the feedback received in the project proposal**, we have come up with a detailed timeline for our action plan which is followed right after this table.

Previous Plan:

Milestones/Tasks	Estimated Bandwidth	Task Status
Literature Survey	02/01 - 02/25	Done
Project Proposal	02/26 - 03/02	Done
Dataset Collection	03/01 - 03/06	Done
Preprocessing and Visualization	03/11 - 03/15	Done
Training & Testing the model with IAM dataset	03/15 - 03/22	Done

Training the medical data	03/23 - 03/29	Partially Done
Mid Term Report	03/30 - 04/01	Done
Building Various Models using TensorFlow	04/02 - 04/12	To be done
Testing the model performance and interpret results	04/12 - 04/15	To be done
Find ways to improve the model	04/16 - 04/25	To be done
Final Written Report and Poster preparation	04/26 - 05/03	To be done

Table 1: Previous Execution Plan

Updated Detailed Plan:

Milestones/Tasks	Estimated Bandwidth	Task Status/Comments
Explore Project Domain	1/15 – 1/22	Completed Successfully
Finalize the project topic of common interest of the group	1/22 – 1/24	Completed Successfully
Discussion among team members and mentor for the feasibility of the project	1/25 – 1/28	Completed Successfully
Project Team & Project Topic Submission	29-Jan	Completed Successfully
Literature Survey of the existing work done in Handwriting Recognition Domain	1/31 – 2/24	Completed Successfully
Discussion with the mentor about the Handwriting Domain	2/20 – 3/4	Completed Successfully
Find out the Medical Dataset having Patient's prescriptions	02/12	Completed Successfully
Figure out the enhancements/improvements that the idea can be integrated into Medical Domain	2/24 – 2/28	Completed Successfully
Draft Project Proposal with our ideas	2/28 – 3/1	Completed Successfully
Reach out to Research Paper authors to get the dataset for academic purpose	2/28 – 3/7	This was unsuccessful as the research authors denied sharing the dataset due to confidentiality.
Building the dataset on our own asking Friends/Relatives to share their Medical Data	3/2 – 3/10	Completed Successfully
Explore Image processing techniques for the dataset	3/11 – 3/12	Completed Successfully
Figure out what image processing a medical dataset requires	3/11 – 3/12	Completed Successfully

Apply Image pre-processing techniques to the dataset	3/13- 3/20	Completed Successfully
• Image Rescaling	3/13- 3/20	Completed Successfully
• Grayscale	3/13- 3/20	Completed Successfully
• Dilation & Erosion	3/13- 3/20	Completed Successfully
• Gaussian Blurring	3/13- 3/20	Completed Successfully
• Adaptive Threshold	3/13- 3/20	Completed Successfully
• Skew Correction/Slant Correction	3/13- 3/20	Completed Successfully
Page Segmentation using Deep CNN	3/21 – 3/26	Completed Successfully
Line Segmentation	3/27 – 4/6	In Progress
• Trying out the Sliding Window Method	3/27 – 3/30	We didn’t achieve good results with Sliding Windows method and the text was segmented randomly and very less accurately.
• Trying out ResNet	3/31 – 4/6	In Progress
Project Mid Progress Report	4/1 – 4/4	Completed Successfully
Explore OCR Techniques for Handwriting Recognition	3/20 – 2/10	To Be Done
Apply Tesseract to find out how it works for Handwritten Text	3/22 – 3/24	Failed. Tesseract works well with printed text but not at all with Handwritten text.
Run & Understand Existing Implementation [5] trained on IAM Dataset	3/22 – 3/26	Completed Successfully
Use DCNN and LSTM to recognize the handwritten text from images	4/5 – 4/10	To Be Done
Comparison of our approach to the existing approaches	4/10 – 4/12	To Be Done
Finalize the model for Handwriting Recognition	4/10 – 4/14	To Be Done
Keyword Spotting using Python packages	4/15 – 4/18	To Be Done
Checking accuracy/error rate of identified medical keywords	4/18 – 4/24	To Be Done
Poster Preparation	4/20 – 4/25	To be Done
Final Report Preparation	4/25 – 5/5	To be Done
Poster Presentation	30-Apr	To Be Done
Final Report Submission	5-May	To Be Done

Table 2: Updated Execution Plan

Difficulties and Problems Encountered

Initially, we faced an issue collecting Prescriptions' dataset. We read several research papers and contacted the authors to get the dataset, but they denied providing the dataset due to privacy issues. We overcame that issue by collecting several images ourselves. Now, we have around 400 Prescriptions' images. Also, we have 1GB of another dataset from IAM Dataset for training purpose.

The medical practitioners have illegible handwriting. Sometimes, it is very hard for patients to understand that is written in their medical report, even when they know the context of their report. The models that we have tried so far works well with good calligraphy, however, we get poor results when we test it on our dataset. That is where we are focusing on and build an approach to detect keywords. We have described the approach that we are using in the "Method" section.

Response to the feedback

<Jake>: good to have a personal motivation. :)

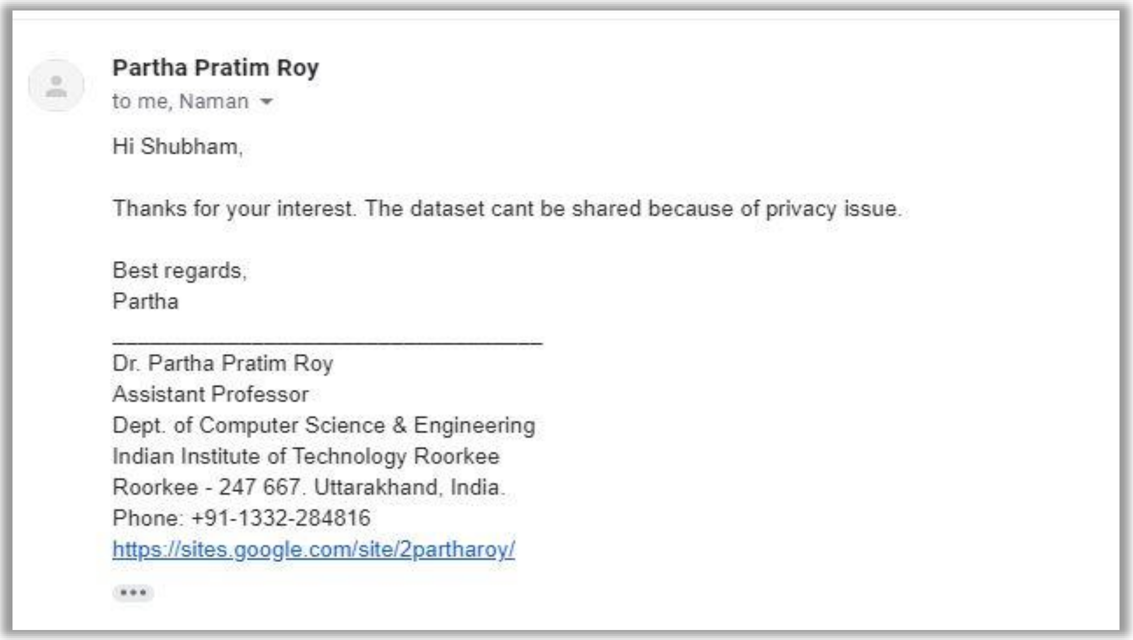
<Mavericks>: Thanks :)

<Jake>: do you have a plan B when you were not able to find a proper dataset?

<Mavericks>: Yes, we had it before the first proposal. We were looking for the pre-processed datasets. However, we later realized that medical data is very sensitive and not easily available. So, all the four group members used their contacts in Friends and Family to collate the dataset keeping the confidentiality. We have even discussed not to upload it anywhere and will be used only for this project work.
As far as the dataset is concerned, we were able to gather more than 400 medical prescriptions. In addition to this, we will be using IAM Dataset (which is the most famous and easily available resource online [5]) to train our model. Out of the 400 medical prescriptions, we will be using a small subset of it for testing the performance of our model.

<Jake>: did you already get a response from the authors?

<Mavericks>: Yes, we received, but unluckily a No! But this didn't stop us from continuing our work. As already mentioned, we have a good dataset now to build this project. Here is the snippet of the response that we received:



Partha Pratim Roy
to me, Naman

Hi Shubham,

Thanks for your interest. The dataset cant be shared because of privacy issue.

Best regards,
Partha

Dr. Partha Pratim Roy
Assistant Professor
Dept. of Computer Science & Engineering
Indian Institute of Technology Roorkee
Roorkee - 247 667. Uttarakhand, India.
Phone: +91-1332-284816
https://sites.google.com/site/2partharoy/

<Jake>: ok. currently, you do not have any idea yet. Hope you have one soon.

<Mavericks>: Yes, we were able to develop our approach to build and train our model. We have already explained it in the Method section.

<Jake>: without clear direction, the timeline is high-level without any detail.

<Mavericks>: Our bad!! Apologies for that. We were not aware of the details that we were supposed to mention in the timeline. This time we have provided a very detailed timeline right from the word "GO" and whatever we have done so far and what we plan to do for the rest of the project. However, it is prone to little changes here and there (if we find/develop new ideas while implementing the models)

<Jake>: Take a direction toward deep learning approach with additional design to better handle medical hand-writing. what is the difference in the data from normal handwriting? This may be the starting point to develop ideas.

<Mavericks>: Ok, this was a good clue to start with! Thanks!!
Normal handwriting is a plain document with text written in it. But when we try to implement this approach with Medical Handwriting, there is a lot of difference. The most challenging one being the segregation of Printed and Handwritten text in the medical reports. (An example prescription which we already demonstrated in method section). We started with image pre-processing as the documents were scanned with different scanners and different conditions and had a lot of noise. We implemented image pre-processing techniques like rescaling, dilation, erosion, skew correction, slant correction, page segmentation, and line segmentation. Once we segment the lines, we will apply handwriting recognition models to identify the characters and spot the keywords.

Difference/Novelty (Updated)

Most of the literature we read about the handwritten data to digital text conversion were having good handwriting. There is not much research done in the medical domain, especially with handwritten medical reports. The research work that we found in this domain was mainly done using Hidden Markov model (HMM) and Optical Character Recognition (OCR).

We used CNN and ResNet for feature extraction which is a unique approach.

Moreover, most of the previous work [7] was done with multi-dimensional LSTM to recognize the handwriting. But we will implement that using single dimensional LSTM as it is computationally much cheaper than multidimensional.

We aim to solve this problem using above mentioned techniques and will aim to achieve more accuracy. The proposed system attempts to eradicate the problems that a patient faces while reading a prescription because of poor handwriting of doctors.

References

[1] A. Mukherjee, A. Halder, S. Nath, S. Sarkar. "A New Approach to Information Retrieval based on Keyword Spotting from Handwritten Medical Prescriptions", Advances in Industrial Engineering and Management, American Scientific Publishers, Vol. 6, No. 2 (2017),90-96, https://pdfs.semanticscholar.org/0af8/ec347b87569e3f68da04891fafce0c3f1fc8.pdf

[2] P. Roya, A. Bhunia, A. Das, P. Dhar, U. Pal "Expert Systems With Applications", Elsevier, Volume 76 (2017), Pages 113-128, https://doi.org/10.1016/j.eswa.2017.01.027

[3] D. Nasien, H. Haron, S. Yuhaniz "Support Vector Machine (SVM) for English Handwritten Character Recognition", 2010 Second International Conference on Computer Engineering and Applications, IEEE COMPUTER SOC (2010), https://ieeexplore.ieee.org/document/5445830

[4] B. Balci, D. Saadati, D. Shiferaw "Handwritten Text Recognition using Deep Learning" (2017), http://cs231n.stanford.edu/reports/2017/pdfs/810.pdf

[5] A. Yuan, G. Bai, L. Jiao, Y. Liu "Offline handwritten English character recognition based on convolutional neural network", IEEE COMPUTER SOC (2012), 10th IAPR International Workshop on Document Analysis Systems, https://ieeexplore.ieee.org/document/6195348

[6] Harald Scheidl [@githubharald], SimpleHTR, (2013), GitHub repository, https://github.com/githubharald/SimpleHTR

[7] Joan Puigcerver, "Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition?", 9-15 Nov. 2017, http://www.jpuigcerver.net/pubs/jpuigcerver_icdar2017.pdf

[8] Pablo Ruiz Ruiz, "Understanding and visualizing ResNets", Oct 8, 2018, Towards Data Science, https://towardsdatascience.com/understanding-and-visualizing-resnets-442284831be8

[9] U. Marti and H. Bunke. The IAM-database: An English Sentence Database for Off-line Handwriting Recognition. Int. Journal on Document Analysis and Recognition, Volume 5, pages 39 - 46, 2002.

[10] Lamy JB, Venot A, Duclos C. PyMedTermino: an open-source generic API for advanced terminology services. Stud Health Technol Inform 2015;210:924-928

[11] "Tesseract Open Source OCR Engine (main repository), GitHub repository, https://github.com/tesseract-ocr/tesseract