RECOGNITION OF HANDWRITTEN MEITEI MAYEK SCRIPT BASED ON TEXTURE FEATURE

Sanasam Inunganbi and Prakash Choudhary

National Institute of Technology Manipur, Langol-795004, INDIA

ABSTRACT

Recognition of Manipuri Script called Meitei Mayek is still in the infant stage due to its complex structure. In this paper, an attempt has been made to develop an offline Meitei Mayek handwritten character recognition model by exploiting the texture feature, Local Binary Pattern (LBP). The system has been developed and evaluated on a large dataset consisting of 3,780 characters which are collected from different people of varying age group. The highest recognition rate achieved by the proposed method is 93.33% using Support Vector Machine (SVM). So, the contribution of this paper is bi-fold: firstly, a collection of a large handwritten corpus of Meitei Mayek Script and secondly developing character recognition model on the collected dataset.

KEYWORDS

Handwritten character recognition, Meitei Mayek script, Local Binary pattern, Support Vector Machine.

1. INTRODUCTION

Optical Character Recognition is one of the interesting and challenging topics in Pattern Recognition and Computer Vision. It may be defined as the procedure follow to recognized printed or written text, which can be acquired either off-line or online. Recognition of printed script is relatively easy because of its standard size and style. Recognition of handwritten script is the real challenge faced by research community because of massive variation in the writing style of different individuals. The general procedure followed by an OCR algorithm, irrespective of nature of script are image acquisition, pre-processing, feature extraction and recognition. Handwritten character recognition has multiple applications such as digitization of documents, automation in mailing system and bank data processing, etc. The result of OCR systems can act as an intermediate step in translation or speech synthesis. Meitei Mayek is already encoded into Unicode [1]. Therefore, the OCR output of text as Unicode would help to maintain a universal standard for further processing.

Meitei Mayek is the official language of Manipur, a state in north-eastern India. It belongs to Tibeto-Burman branch of the Sino-Tibetan language family and is also spoken in Bangladesh and Myanmar. The script has been reinstated recently, and that is why only a few researchers have been accomplished. The current script is the construction of the ancient Meitei Mayek script. The script consists of 27 alphabets (18 original alphabets called Eeyek Eepee and 9 additional letters called Lom Eeyek, derived from the original alphabets), 8 letters with short

ending (lonsum Eeyek), 8 vowel signs (Heitap Eeyek), 10 digits (Cheising Eeyek), 4 punctuation marks (Khudam Eeyek). Interestingly the alphabets of Meitei Mayek Scripts are named after parts of human body. The 27 alphabets of the Meitei Mayek Scripts are given in Figure 1 with the meaning of their names.

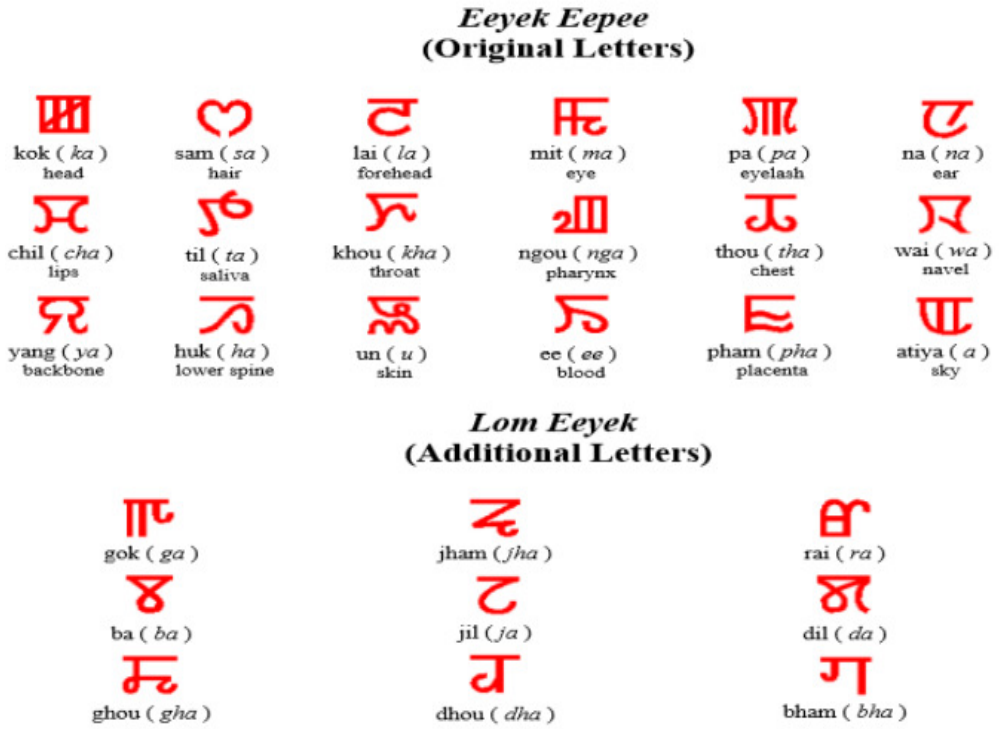

Figure 1: Meitei Mayek Alphabets

Various Character Recognition System has been developed for different languages in the world. However, very few research papers existed in literature for recognition based on Meitei Mayek Script. In 2010, identification of the handwritten character of Meitei Mayek script using Artificial Neural Network was proposed by [2]. The authors have used 594 samples images, out of which 459 have been trained, and the remaining 135 had been used for testing. Probabilistic and fuzzy feature and their combination had been used as the feature vector in their work. In 2013, an OCR system of Meitei Mayek Script on printed documents was proposed using Support Vector Machine (SVM) in [3]. They had scanned 100 pages at 300 dpi from textbook, magazine, and newspaper and stored in ``tiff" or ``png" format. Chain code and directional distribution had been used as local features, aspect ratio and longest vertical run had been used as the global feature for classification. In 2014, another OCR system for Meitei Mayek Script using Neural Network (NN) with back propagation had been proposed in [4]. The authors had used 1000 samples of images, 500 images were used for training while the remaining 500 were used for testing. They had used the binary pattern as the feature for their model. Recognition of Meitei Mayek numerals or digits had also been carried out using SVM in [5-6] and NN in [7]. Various work performed on Meitei Mayek script has been summarized in Table 1.

Table 1: Various work performed on Meitei Mayek Script

Database	Pre-processing Methods	Recognition Methods
594 sample images [2]	Connected Component analysis performed on dilated and filled character images for calculating area, centroid and bounding box to extract characters	Artificial Neural Network with probabilistic and fuzzy features and their combination
100 pages printed document [3]	Horizontal and vertical projection were used for line and word segmentation	SVM using Chain code and directional feature as local feature, aspect ratio and longest vertical run as global feature.
1000 samples [4]	Morphological operation was used for line, word and characters segmentation	Neural Network with back propagation
1000 samples [7]	Morphological operation was used for line, word and digit segmentation	Neural Network using pixel density as feature
800 samples [6]	Explicit character segmentation technique had been applied	Gabor filter with SVM

The remaining part of the paper is organized as Section 2 presents the methodology of the proposed scheme which is subdivided into Image Acquisition, Pre-processing, Feature Extraction, and Recognition. Section 3 gives the Experimental Results and Section 4 concludes the paper by providing Conclusions and Future Work.

2. PROPOSED METHODOLOGY FOR CHARACTER RECOGNITION

Four essential stages have been covered to accomplish the recognition problem of Meitei Mayek Script. The steps that have been considered in recognizing handwritten characters of this script are Image Acquisition, Pre-processing, Feature Extraction, and Recognition. This paper focuses on the recognition of the 27 alphabets. A schematic diagram of this recognition system has been illustrated in Figure 2.

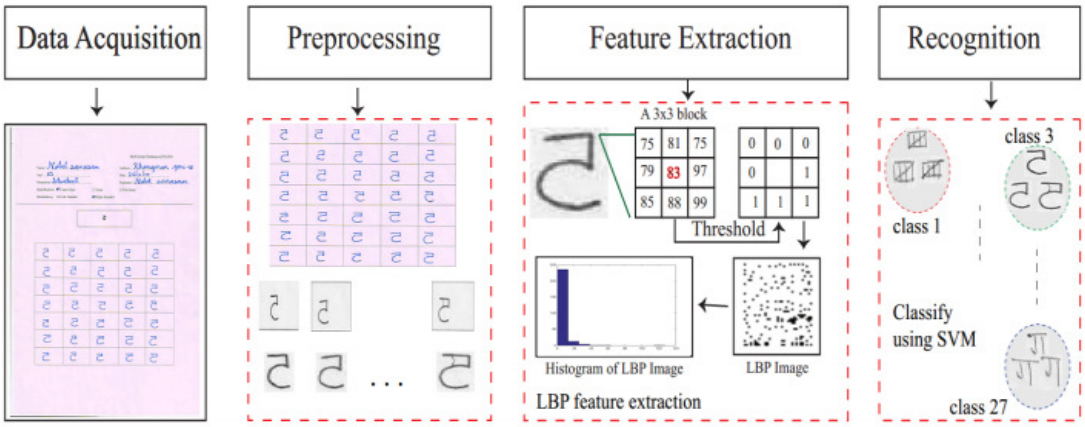

Figure 2: Schematic Diagram of the Proposed Method

2.1 IMAGE ACQUISITION

The first and foremost important step in developing an OCR system is to design and acquire databases for evaluation and validation of the proposed algorithms or techniques. So, for development and evaluation of an efficient character recognition system of Meitei Mayek, we have collected 4 sets of the 27 alphabets in 108 pages. Each alphabet has 35 samples collected per page so that one set comprises of 945 (27 x 35) isolated characters. So, in total there are 3,780 samples of the alphabets written by 70 different people of various age group. The image acquisition is associated with demographic information of the writer such as name, address, occupation, qualification signature, etc. so that this information can be made available for other application like signature verification as well. The collected image has been appropriately scanned using a scanner at 300 dpi and stored in ``jpg" file format for further processing..

2.2 PRE-PROCESSING

Pre-processing is vital for further processing and crucial for maximizing recognition accuracy. It is the procedure performed to enhance image properties and suppress undesirable noise. The use of pre-processing technique improved document image and prepared it for next stage in Character recognition system. So, it is essential to have an active pre-processing step to achieve higher recognition rate. Besides, it makes an OCR system more robust through accurate enhancement, noise removal, image thresholding, character segmentation, character normalization, additional space removal around the character and morphological technique. Segmentation is the main pre-processing perform in our methodology

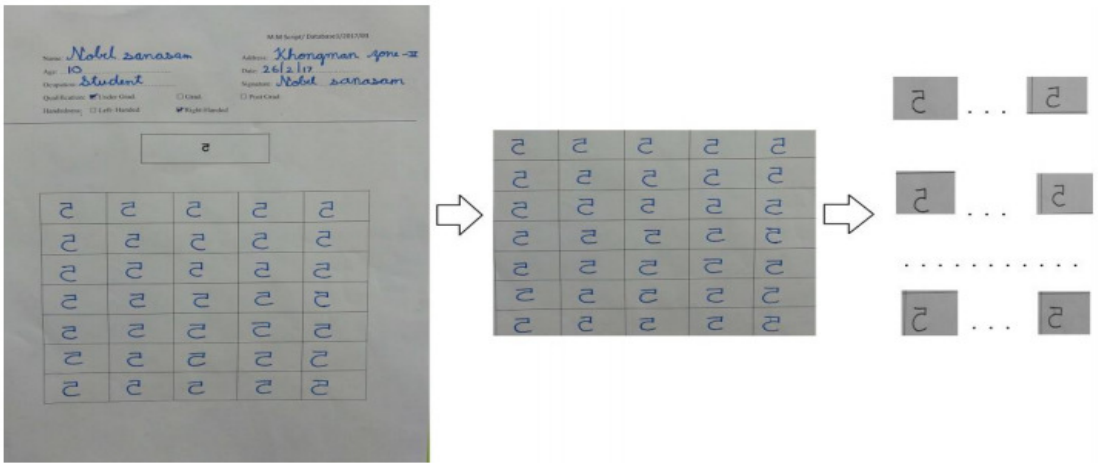

Figure 3: Pre-processing perform in our character recognition System

The characters have been collected in tabular form where each character occupies one slot, and every character needs to be isolated for further processing. Firstly, the character table (the section where writer replicates characters in their writing style) has been separated from the rest of the part by manual cropping as illustrated in Figure 3. Then, every character has been isolated by diving the table into an appropriate number of rows and columns since the table has been appropriately aligned, and further processing has been performed to eliminate extra space other than the alphabets.

Figure 4: Process for obtaining minimum bounding rectangle of a character

From the isolated image, the edge of the alphabet is determined using Sobel operator. To thicken the border, we have operated the morphological dilation [8] using structural element disk of size three. Image filling operation is performed on the binary image to fill any hole exists in the picture. Then, the connected component analysis is implemented on the filled image to obtain properties like area and bounding box. The bounding box corresponding to the most significant object (here the alphabet) is taken into account, and the image is cropped accordingly. The whole procedure can be depicted diagrammatically from Figure 4.

2.3 FEATURE EXTRACTION

An image can be represented by distinct numerical properties of various feature, which can be analyzed by a classifier for correct classification. Texture represents an essential property of an object that does not change with the change of illumination conditions. Texture can outline regularity, randomness, smoothness or coarseness. Texture gives the spatial arrangement or structure of color or pixel intensities over a local area or of the whole image. It defines how a pixel intensity correlates with their neighboring pixels.

The Local Binary Pattern (LBP), introduced by Ojala [9], has been widely used to analyse the texture of an image. It may be defined as an ordered set of binary numbers formed by thresholding between central and its neighbour pixels. The order set of binary number thus generated is then multiplied by the weighted power of two to be converted into decimal. On the other hand, histogram provides the global appearance of an image. So, the histogram of LBP image is considered as the feature vector of an image in our recognition system. Mathematically, LBP can be calculated using equation 1 where gc is the centre pixel of 3 x 3 block and gp are the neighbouring pixels. For our experiment, we have taken P = 8 and R = 1. After obtaining the LBP image LBP_I(x,y), the LBP histogram can be obtained using equation 2.

$$LBP_{P,R} = \sum_{i=0}^{P-1} s(g_i - g_c)\, 2^i \,, s(x) = \begin{cases} 1, if\ x \geq 0 \\ 0, otherwise \end{cases} \quad (1)$$

$$H_i = \sum_{x,y} I\{(LBP_{I(x,y)}) = i\}\ i = 0, 1, \ldots n-1 \quad (2)$$

where n is the number of different labels produced by the LBP operator, and I{A} is 1 if A is true and 0 otherwise. In this modality, we have obtained the value of n= 256.

LBP has many advantages. First, its computation is simple with a low dimensionality of feature space. Second, it provides highly discriminant features. It has proven to be highly tolerant to illumination changes. It is required to extract features for a large number of images for developing this system and hence simple computation yet efficient that takes less time is preferred.

The collected dataset of 3,780 characters are processed and 256 LBP features are extracted from each character which will be used for Recognition. The feature vector space is independent of the image size.

2.4 RECOGNITION

Recognition is an important step to evaluate the usefulness of database of the proposed method. It requires some decision to be taken to give labels to the test images of which categories they belong. This decision is made by a classifier based on the features being selected. Support Vector Machine (SVM) is an efficient and discriminant classifier. It was first introduced by Boser et al. [10]. The SVM is a supervised learning algorithm for classification and regression. The SVM has been considered for our experiment with linear kernel because it is robust to noise and it suffers less from over fitting. The feature size in our experiment is not very small, and hence there is no need to transform feature points to a higher dimensional space. That means nonlinear mapping does not improve the performance. Using linear kernel is good enough. Moreover, a thumb rule of SVM asserts to use linear kernel when the number of features is larger than the number of observations.

A guide to the proper understanding of SVM was given by A. Ben-Hur et al. in [11]. Training and testing phases are present in SVM like other classifiers. During the training phase, SVM constructs a large margin hyperplane by setting Lagrange multiplier (α) and biased (b).

3. EXPERIMENTAL RESULTS

For the technical validation and contribution of the dataset as a standard benchmark platform for linguistic research on Meitei Mayek Script, we have applied character recognition technique and presented the results of the experiments. The experiment has been carried out on the collected samples of characters. The database of 27 alphabets has been collected in sets of four. One set consists of 945 samples, having 35 samples for each alphabet. Out of 945 isolated characters, 540 (20 from each alphabet) has been used for training and 405 (15 each alphabet) for testing. The experiment has been carried out separately on each of the four sets using SVM classifier. Each set has been written by a group of different individuals and does have variation. In the collected dataset, no two individuals have written the same alphabet twice, and the overall variation found in the database is around 80\% and above. So, the proposed method is robust to various writing styles. The variance in writing has been depicted in Figure 5. The recognition rate achieved in each of the sets are 90.62%, 92.59%, 93.33% and 91.36% respectively. The highest accuracy achieved is 93.33% by the Set 3. The accuracies of each alphabet in Set 3 have been highlighted

in Table 2. It can be examined from the table that majority of the alphabets attain 100% of the recognition rate and only two classes obtained lesser accuracy of 53.33% and 20%.

Comparison of some of the existing work in literature for recognition of Meitei Mayek alphabet has been performed with the experimental results of our proposed work, and the results have been summarized in Table 3. It can be observed from the table that our proposed model for recognition has achieved higher recognition rate as compared to the one existing in the literature. However, the accuracy in [3] is higher as they had considered printed documents for their experiment and ours had been developed and evaluated on the handwritten script.

Table 2: Recognition Rate (in %) for each alphabet in Set 3

Alphabets	Recognition Rate	Alphabets	Recognition Rate
ꯀ	100	ꯘ	100
ꯁ	100	ꯙ	100
ꯂ	100	ꯚ	100
ꯃ	100	ꯛ	100
ꯄ	100	ꯜ	100
ꯅ	20	ꯝ	100
ꯆ	86.6	ꯞ	100
ꯇ	100	ꯟ	53.33
ꯈ	80	ꯠ	100
ꯉ	100	ꯡ	100
ꯊ	100	ꯢ	100
ꯋ	100	ꯣ	80
ꯌ	100	ꯤ	100
ꯍ	100		

Table 3: Comparison of Recognition Rate (RR) with the existing methods in literature.

Paper	Database	Recognition
[2]	594(training-459, testing-135)	90.3 %
[3]	100 pages of printed documents	96%
[4]	1000 samples (500 each for training and testing)	80%
Set 1	945(training-540 testing- 405)	90.62%
Set 2	945(training-540 testing- 405)	92.59%
Set 3	945(training-540 testing- 405)	93.33%
Set 4	945(training-540 testing- 405)	91.36%

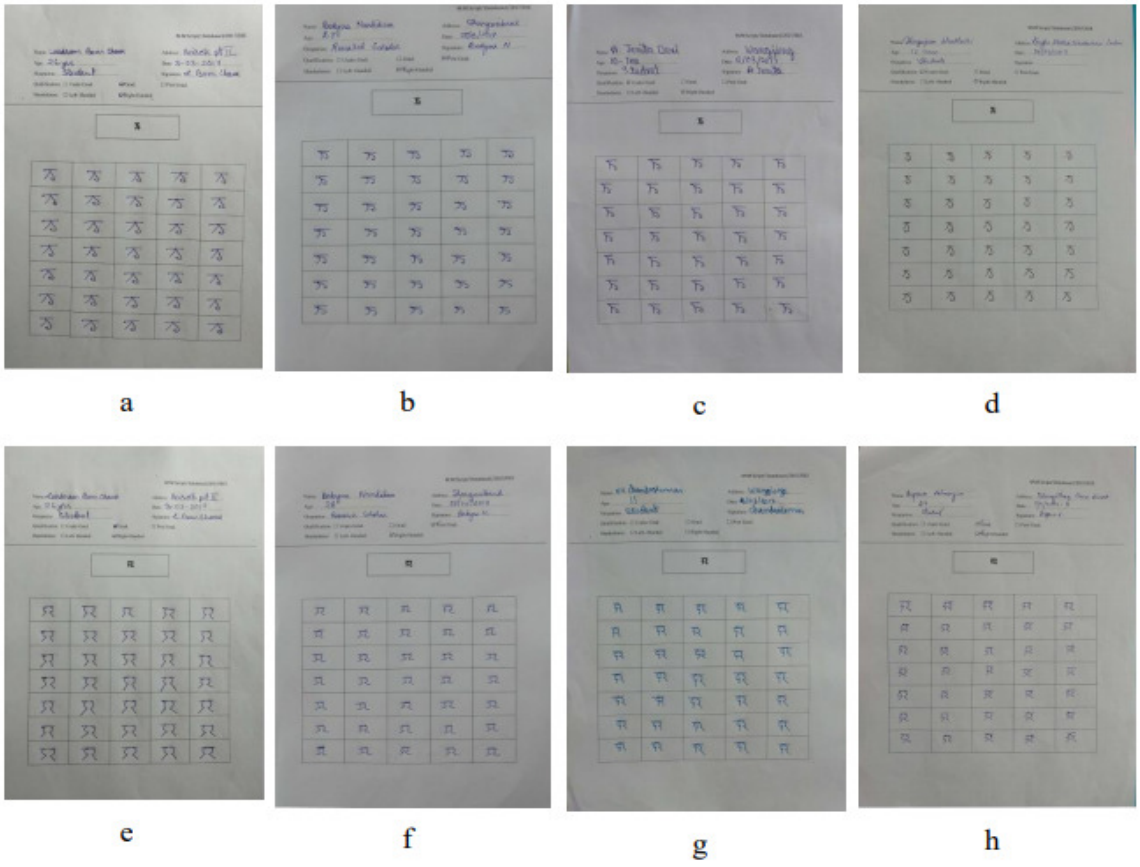

Figure 5: Sample of variation found among the same alphabets over the four sets (a,b,c,d) and (e,f,g,h)

4. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a character technique on a self-collected database of size 3,780 sample of Meitei Mayek alphabets. Till date, this is one of the largest available datasets for Meitei Mayek Script. The database has been collected in four sets, and four set of experiments have been conducted achieving varying accuracy. The variance in accuracy is because different individuals have written the database. Some individual has written with a pen while some, with a pencil. Moreover, each has their unique characteristic of writing. Hence, we have achieved varying accuracy. This issue can be resolved, and efficiency can be improved further in revision with new character recognition technique.

This paper only deals with the recognition of isolated Meitei Mayek alphabets. In future, identification of dataset having complete sentences of Meitei Mayek Script can be performed. In future, the focus can be made on the full-length dataset of both printed and handwritten, encoding the corpus, benchmarking of the dataset on different standards.

REFERENCES

[1] P. Daniels, "The unicode consortium: The unicode," Language: journal of the Linguistic Society, vol. 69(1), p. 225–225, 1993.

[2] T. Thokchom, P. Bansal, R. Vig and S. Bawa, "Recognition of handwritten character of Manipuri Script," JCP, vol. 5(10), pp. 1570-1574, 2010.

[3] S. Ghosh, U. Barman, P. Bora, T. . H. Singh and B. Chaudhuri, "An OCR system for the Meetei Mayek script," in Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2013.

[4] R. Laishram, P. . B. Singh, S. S. D. Thokchom , S. Anilkumar and A. U. Singh, "A neural network based handwritten Meitei Mayek alphabet optical character recognition system," in IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2014.

[5] C. J. Kumar and S. K. Kalita, "Recognition of handwritten numerals of manipuri script," International Journal of Computer Applications, vol. 84(17), 2013.

[6] K. A. Maring and R. Dhir, "Recognition of cheising iyek/eeyek-manipuri digits using support vector machines," Ijcsit, vol. 1(2), 2014.

[7] R. Laishram, A. U. Singh, N. C. Singh, A. Singh and H. James, "Simulation and Modeling of Handwritten Meitei Mayek Digits using Neural Network Approach," in Proc. of the Intl. Conf. on Advances in Electronics, Electrical and Computer Science Engineering-EEC, 2012.

[8] B. R. Masters, R. C. Gonzalez and R. Woods, "Digital image processing," Journal of biomedical optics, vol. 14(2), no. 029901, p. 029901, 2009.

[9] T. Ojala, M. P. ainen and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," Pattern, vol. 29 (1), pp. 51-59, 1996.

[10] B. E. Boser, I. M. Guyon and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in Proceedings of the fifth annual workshop on Computational learning theory, 1992.

[11] A. Ben-Hur and J. Weston, "A user's guide to support vector machines," Data mining techniques for the life sciences, pp. 223--239, 2010.

AUTHORS

Sanasam Inunganbi, is currently pursuing PhD in NIT Manipur. She completed her M.Tech from PDPM IIITDM Jabalpur in Computer Science and Engineering. Her area of interest includes Pattern Recognition, Image processing.

Prakash Choudhary, obtained his Ph.D. in Computer Science and Engineering from MNIT Jaipur. He received his M.Tech. in Computer Science and Engineering from VNIT Nagpur and B.E. from Rajasthan University in Information Technology. He is currently working as Assistant Professor in NIT Manipur since 2013. His area of research interests are Biomedical Image analysis, Pattern Recognition, Annotation and retrieval of Images.