Handwritten Indic Digit Recognition using Deep Hybrid Capsule Network

Mohammad Reduanul Haque[1], Rubaiya Hafiz[1], Mohammad Zahidul Islam[2], Mohammad Shorif Uddin[3]

reduan.cse@diu.edu.bd, rubaiya.cse@diu.edu.bd, zahid.eng@diu.edu.bd, shorifuddin@gmail.com
[1]Department of Computer Science and Engineering, [2]Department of English

[1,2]Daffodil International University, Dhaka, Bangladesh, Department of Computer Science and Engineering, [3]Jahangirnagar University, Savar, Dhaka, Bangladesh

Summary
Indian subcontinent is a birthplace of multilingual people where documents such as job application form, passport, number plate identification, and so forth is composed of text contents written in different languages/scripts. These scripts may be in the form of different indic numerals in a single document page. Due to this reason, building a generic recognizer that is capable of recognizing handwritten indic digits written by diverse writers is needed. Also, a lot of work has been done for various non-Indic numerals particularly, in case of Roman, but, in case of Indic digits, the research is limited. Moreover, most of the research focuses with only on MNIST datasets or with only single datasets, either because of time restraints or because the model is tailored to a specific task. In this work, a hybrid model is proposed to recognize all available indic handwritten digit images using the existing benchmark datasets. The proposed method bridges the automatically learnt features of Capsule Network with hand crafted Bag of Feature (BoF) extraction method. Along the way, we analyze (1) the successes (2) explore whether this method will perform well on more difficult conditions i.e. noise, color, affine transformations, intra-class variation, natural scenes. Experimental results show that the hybrid method gives better accuracy in comparison with Capsule Network.

Keywords:
Indic Digits; Capsule Network; BoF; ANN

1. Introduction

Indian subcontinent (the geographic region surrounded by the Indian Ocean: the Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan and Sri Lanka [1]) is a home of several major languages, of which the most notable are: Hindi (551 million), English(125 Million), Bengali (91 million), Telugu (84 million), Tamil (67 million) [2] etc. Majority of people here prefer their mother tongue to read, write and talk with each other. According to a report published by KPMG in 2017 [3],the expected growth of Indian language internet users is about 18\% each year and as a result the total number of people would reach 536 million by 2021. 68\% of them prefer digital contents on respective local language than the global language. So the overall internet ecosystem of contents, applications, social media platforms etc. need to be more native language friendly. For these reasons, reading, printed or handwritten digits in any language and convert them to digital media is very crucial and time consuming task. This is why recognition of handwritten indic digits play an active role in their day to day life.

A lot of work has already done with English [4-6], Hindi [6-8], Bengali [10], Tamil [11] handwritten digit recognition. Leo et al. [12] proposed an artificial neural network and HOG features based system to recognize handwritten digit in various south Indian languages. They got a recognition accuracy of 83% for Malayalam, 84% for Devanagari, 83% for Hindi, 85% for Telugu and 82% for Kannada. The overall classification rate for the same languages was 83.4%. A multi-language novel structural features based handwritten digit recognition system was proposed by Alghazo et al. and it was tested on six different popular languages, including Arabic Western, Arabic Eastern, Persian, Urdu, Devanagari, and Bangla [13]. Total 65 local structural features were extracted and among several classifiers. Random Forest was found to achieve the best results with an average recognition of 96.73%. Prabhu et al. [14] proposed a Seed-Augment-Train/Transfer (SAT) framework and tested it on real world handwritten digit dataset of five languages. When a purely synthetic training dataset with 140,000 training samples were employed, they achieved an overall accuracy varying from 60% to 75% for five different languages. They also found that, if the training dataset is augmented with merely 20% of the real-world dataset, the accuracy shot up by a significant amount.

Alom et al. introduced a deep learning based handwritten bangla digit recognition (HBDR) system and evaluated its performance on publicly available Bangla numeral image database named CMATERdb 3.1.1 [15]. They achieved 98.78\% recognition rate using the proposed method: CNN with Gabor features, outperforms the state-of-the-art

Manuscript received February 5, 2024
Manuscript revised February 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.2.10

algorithms for handwritten bangla digit recognition. Another deep learning based model was proposed by Ashiquzzaman and Tushar [16]. Their proposed method employed for Arabic numeral recognition and was given 97.4 percent accuracy.

Recently, Capsule Network (referred to as CapsNet), introduced by Geoffrey Hinton that encodes spatial information into features while using dynamic routing [17]. CapsNets has achieved state of the art only on MNIST dataset- a standard dataset of English handwritten digits [18].

From the above literature, it is clear that very few works have been reported for the digit recognition written in Indic scripts. The main reason for this slow progress could be attributed to the diverse shapes, ambiguous handwritten digits and disproportionate cursive handwritings. In addition, most of the above recognition systems fail to meet the desired accuracy when exposed to multinumerals scenario. Hence, it would be necessary to develop a method which is independent of script and yields good recognition accuracy. This has motivated us to introduce a numeral invariant handwritten digit recognition system for identifying digits written in five popular scripts, namely, English, Bangla, Devanagari, Tamil, and Telugu. In this paper, we propose a hybrid method and compare the accuracy on top five Indian sub-continent digit datasets as well as explore how this architecture performs on these numerals that are marginally harder in specific ways. Our paper puts forward the following contributions:

a. A hybrid model is developed combining simple artificial neural network with capsule network.
b. Analyze the success and explore whether this method will perform well in more difficult conditions such as noise, color, and transformations.
c. Compare the accuracy of the method on top five Indian sub-continent digit datasets and explore how this architecture performs on these numerals.

The forthcoming parts of this paper is fabricated as follows: step II gives the general approach of our suggested scheme. Section III contains a brief discussion of the overall experimental results and eventually, we have concluded the paper in section IV.

2. Methodology

There are different approaches for handwritten digit recognition which may be broadly classified into two categories: classical approaches (e.g. BoF and support vector machine) and neural-based methods (e.g. Simple neural network, deep convolutional neural network, Transfer Learning and Capsule Network). A brief description of these techniques is given below.

2.1 Color Based Bag of Features

Bag of features (BoF) is a well-established computer vision approach and it is applicable in image classification, object recognition, image retrieval and even visual localization [19]. Both color and SURF based BoF method was employed in our system for feature extraction and detection. In SURF, Hessian matrix based blob detection approach is imposed for the detection of interest point using the following equation:

$$H(X,\sigma) = \begin{bmatrix} L_{xx}(X,\sigma) & L_{xy}(X,\sigma) \\ L_{xy}(X,\sigma) & L_{yy}(X,\sigma) \end{bmatrix} \qquad (1)$$

SURF itself cannot be able to classify similar shaped objects accurately. This is why we imposed both 64 dimensional SURF descriptor and RGB color features to train out classifier.

2.2 Support Vector Machine

Basically, Support vector machine (SVM) [20-22] is a supervised classification technique used for binary classification. It uses a kernel trick to construct linear separating hyperplane in higher dimensional vector spaces. However, it can be extended to multiclass classification that works on more than two classes. Given a group of labeled training data the algorithm gives an optimal hyperplane that categorizes new examples.

2.3 CNN and Deep CNN

Convolutional Neural Network (CNN) [23] is a deep appearance of traditional Artificial Neural Network (ANN) architecture. A basic CNN architecture consisting of two main parts: feature extractor and classifier. The feature extraction unit consists of convolution layer, an activation function and pooling layer, where the output from the previous layer is served as the input to the next layer. The classification unit consists of a fully-connected layer.

In deep CNN, there are many layers (number of layers ≥ 3) [24] instead of a single convolutional-layer and a pooling-layer.

2.4 Transfer Learning

Transfer learning gained popularity recently due to its effectiveness that take pre-existing models for training and then transfer the knowledge learned from a similar task.

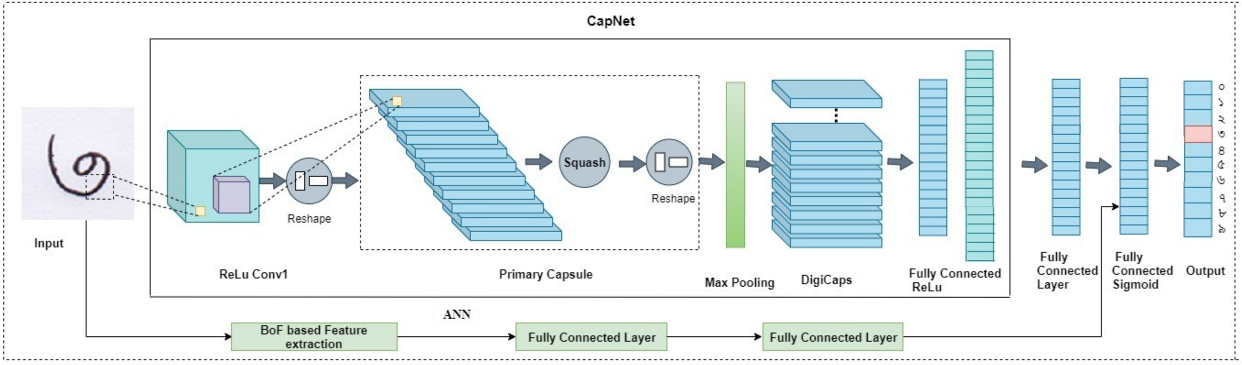


Fig. 1: General Structure of our proposed hybrid capsule network

As high recognition accuracy of CNNs demands huge labelled data to train its deep architecture and it is expensive and exhausting to create labelled data, so it is impossible to get huge number of training samples. For this reason, transfer learning is a promising alternative to train CNNs with scarce dataset. In case of transfer learning, it is easier and much faster to fine-tune a network than do the training from scratch. In the structure of CNNs, early layers contain generic features that can be re-used for various tasks. In contrast, the final layers are more specific to the applications. Based on this property, the initial layers are well-preserved while the final ones are adjusted to train with the new dataset of interest [25-26].

2.5 Capsule Network

In a capsule network [17], the network learns to render an image inversely; that is by looking at the image it tries to predict the instantiation parameters for that image. Initially, the input image is converted into a block of activations by employing convolution layer and supplied as an input into the primary capsule layer. Dynamic routing between primary capsules are calculated to generate the values of digit capsules. C_{ij} is the coupling coefficients and are used to combine the individual digit capsules and form the final digit capsule as follows.

$$C_{ij} = \frac{exp(W_{\text{ij}}^{\text{DC}})}{\sum_k exp(W_{\text{ij}}^{\text{DC}})} \qquad (2)$$

Total S_j number of input vectors are processed by j-th capsule to produce an activation vector v_j.

$$S_j = \sum_i C_{ij}\hat{U}_{j|i} \qquad (3)$$

The resultant squashed combined digit capsule V_j is given by

$$V_j = \frac{\|S_j\|^2}{1+\|S_j\|^2}\cdot\frac{S_j}{\|S_j\|} \qquad (4)$$

To produce the resultant digit capsule, equations 2 to 4 are repeatedly performed.

2.5 Proposed Methodology

The hybrid model consists of two parts depicted in Fig. 1. The first part of the model is a Capsule Network. And the second part is a simple artificial neural network with two hidden layers. The input to this network is the BoF feature vector extracted from an image. The output of the two parts are combined into a larger feature vector, which is fed as input to another network comprising of two fully connected layers, the last layer is a softmax activation function that provides the probability distributions of class predictions.

Let m1 (.) and m2(.) be the independent slices (slice 1 and slice 2) obtained from BoF ANN and Capsule Network respectively. Then the two evidence sources m1 (:) and m2 (:) can be combined to feed the third components of the hybrid model (here, this component performed as a softmax classifier according to the following combination or orthogonal sum).

$$m_{12}(C) = m_1(C) \oplus m_2(C) = \frac{\sum_{A\cap B=C} m_1(A)m_2(B)}{1-\sum_{A\cap B=\varnothing} m_1(A)m_2(B)} \qquad (5)$$

3. Experimental Results and Discussions

We have experimented our method with five different datasets each of which represents individual language. Some sample digit images of our datasets are shown in Table I.

TABLE I: Digits of different languages

English	0	1	2	3	4	5	6	7	8	9
Bangla	০	১	২	৩	৪	৫	৬	৭	৮	৯
Tamil	0	௧	௨	௩	௪	௫	௬	௭	௮	௯
Hindi	०	१	२	३	४	५	६	७	८	९
Telugu	౦	౧	౨	౩	౪	౫	౬	౭	౮	౯

3.1 Dataset Description

a. MNIST: It is the standard set of normalized and centered 28 x 28 black and white images of handwritten digits (0-9). It contains 60,000 training and 10,000 testing images.

b. NumtaDB: This is a diverse dataset consisting of more than 85,000 images of hand-written Bengali digits. 85% of them are considered as training and remaining 15% are used as test images. This dataset is very difficult to work with because of highly unprocessed and augmented images.

c. UJTDchar: The UJTDchar dataset contains 100 labelled image samples in JPG format for each character in Tamil language.

d. Devanagari: Devanagari is an Indic script and forms a basis for over 100 languages spoken in India and Nepal including Hindi, Marathi, Sanskrit, and Maithili. This dataset comprises of grayscale images of 47 primary alphabets, 14 vowels, and 33 consonants, and 10 digits in png format. All the characters are centered within 28 $\times$ 28 pixels.

e. CMATERdb 3.4.1: We have collected handwritten Telugu numerals from CMATERdb database repository [27], [28].

3.2 Distortions

To find out the robustness of our method on all above-mentioned datasets, we have imposed some deformations. As a result, we would be able to figure out the extent to which the model deformation invariant for recognition.

For all datasets, an alternate, deformed dataset is generated by applying a random affine deformation consisting of:

a. Rotation: Rotated image by a uniformly sampled angle within [-20°, 20°].
b. Shear: Sheared along x and y axes by uniformly sampling shear parameters within [-0.2, 0.2]. (Shear parameters are numbers added to the cross-terms in the 2 x 3 matrix describing an affine transformation.)
c. Translation: Translated along x and y axis by a uniformly sampled displacement parameters within [-1, +1]. (Displacement parameters are numbers added to the constant terms in the 2 x 3 matrix.)
d. Scale: Always scaled image 150%.

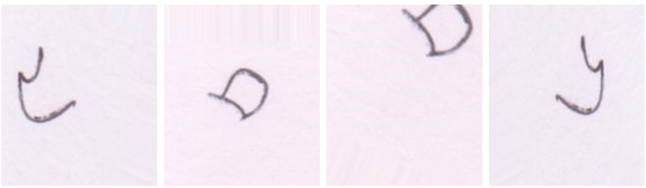
Tables II and III show the comparative accuracy results of top five Indian sub-continent digit dataset on different conditions, such as with and without affine transformations as well as random rotation. We can see that when hybrid method is imposed, it gives better accuracy in almost every cases. Few images from the NumtaDB Dataset are shown in Fig. 2. Also Fig. 3 shows some misrecognized images from different datasets.

TABLE II: Overall Classification Accuracy Across 5 Datasets After 100 Epochs. With and Without Affine Transformations.

Datasets	Normal		Affine	
	CapsNet	Hybrid Method	CapsNet	Hybrid Method
MNIST	99.2	99.6	74.91	82.5
Tamil	91.8	95.5	72.9	80.05
Telegu	94.2	96.2	76.4	80.2
Hindi	94.8	96.1	75.3	82.9
Bangla	96.2	99.3	74.4	88.9

TABLE III: Recognition accuracy after random rotation (-30 to +30)

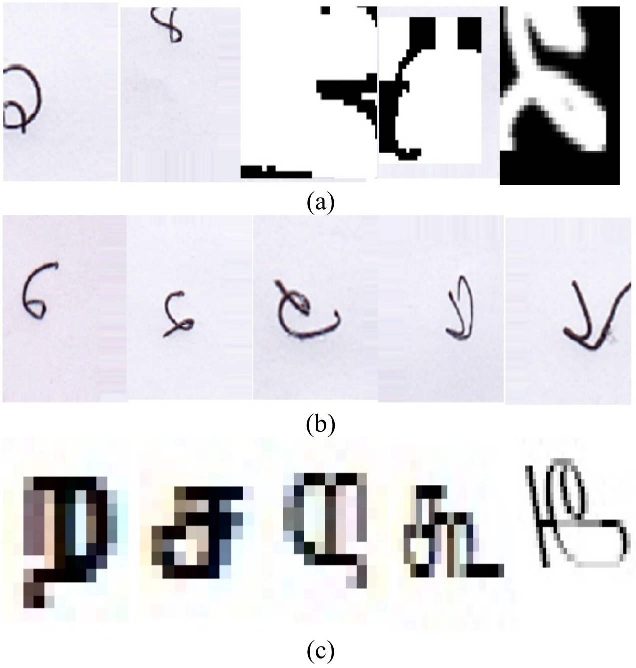
Dataset	CapsNet	Hybrid Method
MNIST	77.68	85.53
Tamil	75.8	86.5
Telegu	73.2	80.9
Hindi	77.8	84.1
Bangla	76.2	88.3



(a) (b) (c) (d)

Fig. 2: Examples of images from the NumtaDB Dataset. Digit images of (a), (d) are recognized by all hybrid network correctly; However, failed to recognize (b), and (c) due to their distortions

(rotation and occlusion).



(a)

(b)

(c)

Fig. 3: Some misclassified images from different datasets dueto various distortions/deformations.

4. Conclusions

Accurate recognition of handwritten digits in real-world scenarios is a challenging task that has attracted considerable attention over the last few years. In this work, a hybrid model is proposed that combined Capsule Network and ANN with BoF. Experimental results show that the hybrid method gives better accuracy (i.e. most robust) in comparison with other methods in terms of shear, scaling and rotational situations.

Funding

No funding was received for this research.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] A. Roy, "Indian shield: Pristine shape, size and tectonic framework," in *Geological Evolution of the Precambrian Indian Shield*, pp. 1–15, Springer, 2019**.**

[2] B. E. Sawe, *What Language Is Spoken in India? - WorldAtlas.com*, July 10, 2018 (Accessed July 3, 2019). https://www.worldatlas.com/articles/the-most-widely-spoken-languages-in-india.html.

[3] *Indian Languages–Defining India's Internet - KPMG International Cooperative [NL]*, April 25, 2017 (Accessed July 4, 2019). https://assets.kpmg/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf.

[4] S. Pratt, A. Ochoa, M. Yadav, A. Sheta, and M. Eldefrawy, "Handwritten digits recognition using convolution neural networks," *The Journal of Computing Sciences in Colleges*, p. 40, 2019.

[5] B. Lopez, M. A. Nguyen, and A. Walia, "Modified mnist," 2019.

[6] S. Majumder, C. von der Malsburg, A. Richhariya, and S. Bhanot, "Handwritten digit recognition by elastic matching," *arXiv preprint arXiv:1807.09324*, 2018.

[7] B. N. Dhannoon and H. H. Al, "Handwritten hindi numerals recognition," *International Journal of Innovation and Applied Studies*, 05 2013.

[8] M. Chaudhary, M. H. Mirja, and N. Mittal, "Hindi numeral recognition using neural network," *Int. J. Sci. Eng. Res.*, vol. 5, no. 6, pp. 260–268, 2014.

[9] G. Singh and S. Lehri, "Recognition of handwritten hindi characters us- ing backpropagation neural network," *International Journal of Computer Science and Information Technologies*, vol. 3, no. 4, pp. 4892–4895, 2012.

[10] R. Noor, K. M. Islam, and M. J. Rahimi, "Handwritten bangla numeral recognition using ensembling of convolutional neural network," in *2018 21st International Conference of Computer and Information Technology (ICCIT)*, pp. 1–6, IEEE, 2018.

[11] M. Kumar, M. Jindal, R. Sharma, and S. R. Jindal, "Performance eval- uation of classifiers for the recognition of offline handwritten gurmukhi characters and numerals: a study," *Artificial Intelligence Review*, pp. 1– 23, 2019.

[12] Pauly, R. D. Raj, and B. Paul, "Hand written digit recognition system for south indian languages using artificial neural networks," in *2015 Eighth International Conference on Contemporary Computing (IC3),* pp. 122–126, IEEE, 2015.

[13] J. M. Alghazo, G. Latif, L. Alzubaidi, and A. Elhassan, "Multi-language handwritten digits recognition based on novel structural features," *Journal of Imaging Science and Technology*, vol. 63, no. 2, pp. 20502–1, 2019

[14] V. U. Prabhu, S. Han, D. A. Yap, M. Douhaniaris, P. Seshadri, and J. Whaley, "Fonts-2-handwriting: A seed-augment-train framework for universal digit classification," *arXiv preprint arXiv:1905.08633*, 2019.

[15] M. Z. Alom, P. Sidike, T. M. Taha, and V. K. Asari, "Handwritten bangla digit recognition using deep learning," *arXiv preprint arXiv:1705.02680*, 2017.

[16] M. Z. Alom, P. Sidike, T. M. Taha, and V. K. Asari, "Handwritten bangla digit recognition using deep learning," *arXiv preprint arXiv:1705.02680*, 2017.

[17] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between cap- sules," in *Advances in neural information processing systems*, pp. 3856– 3866, 2017.

[18]] Y. LeCun, C. Cortes, and C. Burges, "Mnist handwritten digit database. 2010," *URL http://yann. lecun. com/exdb/mnist*, vol. 3, no. 1, 2010.

[19] S. O'Hara and B. A. Draper, "Introduction to the bag of fea- tures paradigm for image classification and retrieval," *arXiv preprint arXiv:1101.3354*, 2011.

[20] E. Mayoraz and E. Alpaydin, "Support vector machines for multi-class classification," in *International Work-Conference on Artificial Neural Networks*, pp. 833–842, Springer, 1999.

[21] J. P. Jones and L. A. Palmer, "An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex," *Journal of neurophysiology*, vol. 58, no. 6, pp. 1233–1258, 1987.

[22] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[23] LeCun *et al.*, "Lenet-5, convolutional neural networks," *URL: http://yann. lecun. com/exdb/lenet*, vol. 20, p. 5, 2015.

[24] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning

algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[25] Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Advances in neural information processing systems*, pp. 3320–3328, 2014

[26] K. Nogueira, O. A. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognition*, vol. 61, pp. 539–556, 2017.

[27] N. Das, J. M. Reddy, R. Sarkar, S. Basu, M. Kundu, M. Nasipuri, and D. K. Basu, "A statistical–topological feature combination for recognition of handwritten numerals," *Applied Soft Computing*, vol. 12, no. 8, pp. 2486–2495, 2012

[28] N. Das, R. Sarkar, S. Basu, M. Kundu, M. Nasipuri, and D. K. Basu, "A genetic algorithm based region sampling for selection of local features in handwritten digit recognition application," *Applied Soft Computing*, vol. 12, no. 5, pp. 1592–1606, 2012.

MOHAMMAD REDUANUL HAQUE received his Master of Science and Bachelor of Science degrees in Computer Science and Engineering from Jahangirnagar University, Savar, Dhaka in 2012 and 2011, respectively. He is currently served as a Senior Lecturer in the Department of Computer Science and Engineering at Daffodil International University, Dhaka, Bangladesh. His research interests includes Computer Vision, Deep Learning and Image Processing.

RUBAIYA HAFIZ received her Master of Science and Bachelor of Science degrees in Computer Science and Engineering from Jahangirnagar University, Savar, Dhaka. She is currently served as Senior Lecturer in the Department of Computer Science and Engineering at Daffodil Interenational University, Savar, Dhaka, Bangladesh. Her research interests includes Computer Vision, Deep Learning, Image Processing, Artificial Intelligence etc.

MOHAMMAD ZAHIDUL ISLAM received her Master of Science and Bachelor of Science degrees in English from Jahangirnagar University, Savar, Dhaka. He is currently served as a Lecturer in the Department of English at at Daffodil International University, Dhaka, Bangladesh.

MOHAMMAD SHORIF UDDIN (M'13–SM'15) received his Doctor of Engineering degree in information Science from Kyoto Institute of Technology in 2002, Japan, Master of Technology Education degree from Shiga University, Japan in 1999, Bachelor of Electrical and Electronic Engineering degree from Bangladesh University of Engineering and Technology in 1991 and also MBA in from Jahangirnagar University in 2013. He began his teaching career as a Lecturer in 1991 at the Bangladesh Institute of Technology, Chittagong (Renamed as CUET). He joined in the Department of Computer Science and Engineering of Jahangirnagar University in 1992 and currently, he is a Professor of this department. He undertook postdoctoral researches at Bioinformatics Institute, Singapore, Toyota Technological Institute, Japan and Kyoto Institute of Technology, Japan, Chiba University, Japan, Bonn University, Germany, Institute of Automation, Chinese Academy of Sciences, China. His research is motivated by applications in the fields of imaging informatics and computer vision. Mohammad Uddin is an IEEE Senior Member and also a Fellow of Bangladesh Computer Society and The Institution of Engineers Bangladesh. He wrote more than 100 journal and received the Best Paper award in the International Conference on Informatics, Electronics & Vision (ICIEV2013), Dhaka, Bangladesh and Best Presenter Award from the International Conference on Computer Vision and Graphics (ICCVG 2004), Warsaw, Poland. He holds two patents for his scientific inventions. Currently, he is the Chair of IEEE CS Bangladesh Chapter and an Associate Editor of IEEE Access.