Depth Reconstruction Based Visual SLAM Using ORB Feature Extraction

Yatharth Ahuja, Tushar Nitharwal, Utkarsh Sundriyal, Sreedevi Indu, and Anup K. Mandpura
Delhi Technological University, New Delhi, India
Email: {yatharth.rd, tusharnitharwal, utkarshsundriyal.15}@gmail.com, {s.indu, kanup}@dtu.ac.in

***Abstract*—Mobile platforms are now computationally able to implement SLAM extending the scope of SLAM applications, which has been limited due to the requirement of elaborate sensor support to work. Visual SLAM methods help in tackling sensor limitations on such devices by exploiting information rich data from cameras. The proposed method aims to exploit such visual information from a monocular RGB input to derive depth information of contents of same using DenseNet-169 based Encoder-Decoder architecture. Thus, obtained depth map was combined with the keypoints extracted from monocular input to be processed by ORB SLAM. Further, analysis was done to evaluate the usage of various feature extractors vis-a-vis Oriented Fast and Rotated BRIEF (ORB), SIFT, BRISK. The map was generated from input visual trajectory and pipeline developed was able to implement RGB-D SLAM from only monocular input. The proposed system, thus, helps in executing an efficient SLAM algorithm using only the monocular RGB input.**

***Index Terms*—visual SLAM, depth reconstruction, encoder, decoder, ORB**

I. Introduction

Simultaneous Localization and Mapping (SLAM) is a process in which a robot can build the map of its environment and simultaneously compute its location as well. In the past decade there has been rapid development in solving the SLAM problem with various implementations of SLAM methods. With the advancements in mobile technology and corresponding camera technologies, the application of complex tools can be effectively carried out on the said platforms. SLAM offers to provide solution for various applications from surveying to surveillance [1]. Simultaneous Localisation and Mapping (SLAM) has been an active area of research and has gained tremendous progress in recent years. It has been deployed for the additional feature of mapping. It helps in constructing the map with systems perception data and consequent stream of perceived motion. Map generated by SLAM, generally helps in two manners [2]: It provides a platform to execute other features and creates a reliable prospective base to explore. And that it eliminates the error of estimation in a known environment for the user, thus creating a reliable application.

Manuscript received June 17, 2022; revised September 20, 2022.

Penetration of mobile phone technology also contributes to the scope of problems that can be approached through it. While considering adoption of SLAM in mobile phones, the camera becomes focal point of research in SLAM due to the ubiquitous nature and sophistication of its technology due to which cameras today are able to capture information rich visual data. Traditional SLAM methods [3] require additional sensor data which are either not easily found in mobile phones and are difficult to operate in remote scenarios.

Visual SLAM, as the name suggests, based on solely visual input from cameras provides a viable solution to the application of SLAM to mobile platforms. However, the challenge arises to extract information to compensate for the otherwise concomitant sensors. Depth is a critical information which is extracted from RGB-D or LiDAR instruments. Accurate depth estimation from images is a fundamental task in many applications including scene understanding and reconstruction in SLAM. With the rapid development of deep neural networks, monocular dense depth estimation based on deep learning has been widely studied recently and achieved promising performance in accuracy. For our supporting features, mapping based on robust feature extractions were looked into. Feature extraction is a method of extracting intelligible information from the image which is generally represented as pixel level data. Scale Invariant Feature Transform (SIFT) [4] is a well-known and deployed feature extraction algorithm which relies on keypoint generation. Binary Robust Invariant Scale Keypoint (BRISK) [5] and Oriented Fast and Rotated BRIEF (ORB) [6] are 2 other well-known techniques. Comparative studies have been carried out earlier as well [7].

In [8], the authors have tried to deploy the SIFT [4] based SLAM method for a Global Localisation using the mobile platform. And subsequently a Random K-D based optimisation approach to establish a trajectory of keyframes. Similarly, the study in [9] deploys a multi-sensor method combining inertial data with the visual input to create a robust SLAM application. However, the reliance on a multi-sensor suite is not accessible and scalable in multiple applications.

In the developed system, we alleviate such dependence by estimating the depth of monocular visual input using DenseNet-169 [10] based Encoder-Decoder neural network. The RGB data, along with extracted depth estimation, is then processed by ORB features based

doi: 10.18178/joig.10.4.172-177

SLAM algorithm to generate the map by evaluating the trajectory represented in the visual input sequence.

II. Methodology

As depicted in Fig. 1, the developed system uses only the visual input data in the form of RGB image or video frame from camera. It then processes this data to get a depth-map from this input. The RGB images are then provided to the feature extraction model which identifies the landmarks in an area and this data is finally fed to the SLAM algorithm which helps autonomous agents in building maps and carrying out corresponding operations.

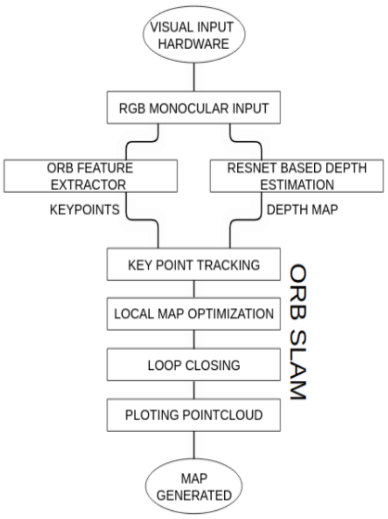


Figure 1. High level system representation.

A. Depth Reconstruction

In order to provide RGB-D input to mapping algorithm, we need to incorporate the depth component for the image as well. For that purpose, we need an RGB camera and a depth sensor. Instead of using an additional sensor, we have achieved the same result by deploying Depth Reconstruction which takes as input an RGB image and converts it into a depth map where the intensity of each pixel corresponds to the distance of that point from the camera. We deployed an Encoder-Decoder architecture based on DenseNet-169 [10] for Encoder-Decoder, which is a sequential neural network, with the architecture as represented in Fig. 2. The role of the encoder is to encode an image, i.e. reduce it in number of features and thus reduce its size and make it a dense vector which represents the features of the image. The height and width of the image is reduced in this process. The role of the decoder is to decode this dense vector based on the conditions that we specify to achieve our objective. Both the input and output of an Encoder-Decoder model is an image.

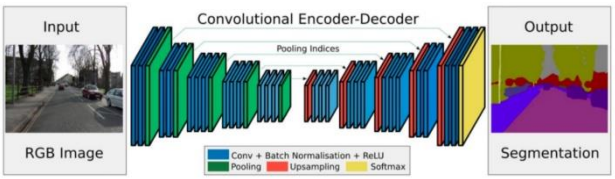


Figure 2. Depth prediction pipeline [11].

For our application, we have used the Encoder-Decoder architecture to generate a depth map. The input is an RGB image from the camera, which is passed to the model and the output is obtained as an image where the intensity of each pixel corresponds to the distance of that point from the camera. We have trained the model on the NYU-V2 Depth dataset [12] using the loss function [1] deploying the weighted sum of gradient loss, depth loss and SSIM loss [13]. The encoder down samples into a feature dense vector using convolutions. Decoder up samples using deconvolution. Data augmentation was also performed to enhance the quality of the dataset. For relating the depth of obtained image to distance in real time, the following equation is used:

$$depth = \frac{baseline \times focal\ Length}{disparity} \quad (1)$$

B. Feature Extraction

Feature extraction in images is the process in which an image is reduced to a smaller number of features, i.e. the essential features of the image like corners, edges etc. are extracted so that the processing of these images becomes easier and manageable, thereby reducing the number of computational resources required. Feature extraction has two parts: First is the process in which an input image is taken and important features are identified, such as corners, edges, blobs etc. This part is called feature detection. For our case we are taking input from a video which means consecutive frames from a video, as represented in Fig. 3.

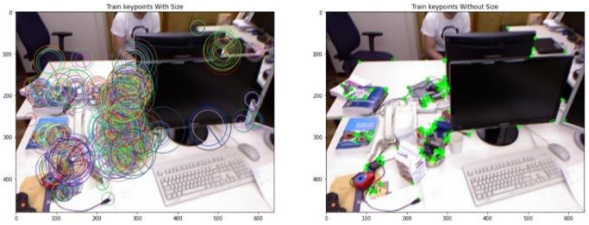


Figure 3. Feature detection with & without keypoint size.

The second part is when we receive the next frame. Once again, we detect important features in this image. Now we use the previous frame and the current frame (2 consecutive frames) and match the feature points within these 2 frames based on some specified tolerable limit. This part is called feature matching.

This is therefore an integral step in visual SLAM. Various different methods are used to detect interest points (corners, blobs, edges) in a frame. These interest points are then matched over continuous frames for data association and landmark extraction which give information about the robot's motion.

For our work, we have compared the speed and accuracy of SIFT, BRISK and ORB detectors and descriptors as they are known to perform well for SLAM based applications [7] due to their speed and accuracy.

C. ORB Feature Extraction Based SLAM

Generation of a map based on Oriented Fast and Rotated BRIEF (ORB) feature extraction, as discussed above is

done. RGBD version of ORB-SLAM2 [14], a feature-based Mapping and Localisation algorithm is used which exploits the previously obtained depth frames. Main advantages of ORB feature-based SLAM method is the robustness with respect to rotation and the speed of operation [6]. It has been achieved in primarily the following three major parallel threads:

1) *Tracking:* Each visual frame obtained is optimised for its pose with respect to the re-projection error [14]. First the incoming frames are processed with a feature extractor which gives us keypoint descriptors. These keypoints are compared with the preceding frame and an optimised pose is determined based on motion-only bundle adjustment. A local map is them maintained with keyframes sharing visibility of the map in terms of common view points. Subsequently, the pose with respect to keyframe under consideration is optimised through the re-projection error [15] of common map viewpoints. Based on the optimisation results, keyframe is inserted into the discerned trajectory.
2) *Local Mapping:* After the decision and insertion of keyframe, it's linked with the keyframe with most strong correspondence in the covisibility local map. Consequent map points are calculated based on ORB triangulation between subsequent keyframes in the covisibility. Using local bundle adjustment, the local map, comprising of current keyframe. Keyframes exceeding the co-visibility and similarity threshold with other keyframes are discarded for reducing complexity.
3) *Loop Closing:* Loop closings are done, if obtained, for greatly reducing the accumulated error in map representation and thus carries immense importance for an efficient system. In order to detect possible loop closures, the current keyframe and the ones in co-visbility are compared with the vocabulary of keyframes in the system. Number of consecutive keyframe matches is directly proportional to the strength of candidature for loop closure. RANSAC [16] iterations are performed on the candidate frames to identify inliers. Consequently, the mapping is optimised by pose graph optmisation [17] over the identified loop closures.

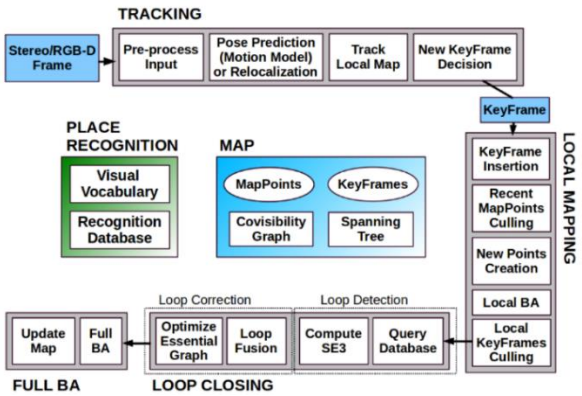


Figure 4. ORB-SLAM schematic [14].

These threads can be represented in functional form as in Fig. 4. The system has embedded a Place Recognition module based on DBoW2 [18] for relocalization. The TUM RGB-D [19] dataset was used for trajectory tracking testing.

III. Testing and Results

The system has been evaluated for various intermediate inputs from subsystems. For evaluation of the model for depth prediction from monocular input, we use the Root-Mean-Square-Error (RMSE) score. The model gave improvements on existing models trained on the NYU-V2 dataset [12]. The corresponding results are provided in Table I.

TABLE I. Comparison of Our Work with Existing Models

Architecture	Loss Function	RMSE
PEM and EAM Combination [20]	Scale-invariant error	0.439
DEM [21]	Custom Loss Function [21]	0.497
Our Model	**SSIM**	**0.4025**

Further in the system, we evaluate the usage of various feature extractors from monocular inputs to be further processed for pose estimation and optimisation. ORB, SIFT and BRISK were evaluated on the basis of: the number of features or keypoints detected in the image as shown in Fig. 5, the number of features matched between two consecutive frames, and the total time of execution of feature detection on any particular frame. The metrics chosen are most relevant to the application of this system in real-time scenarios and thus provided us best perspective of optimal feature extractors. The corresponding results are provided in Table II, Table III, Table IV respectively.

TABLE II. Number of Features Detected

	Image 1	Image 2
SIFT	2011	2471
BRISK	2471	1876
ORB	500	500

TABLE III. Number of Features Matched

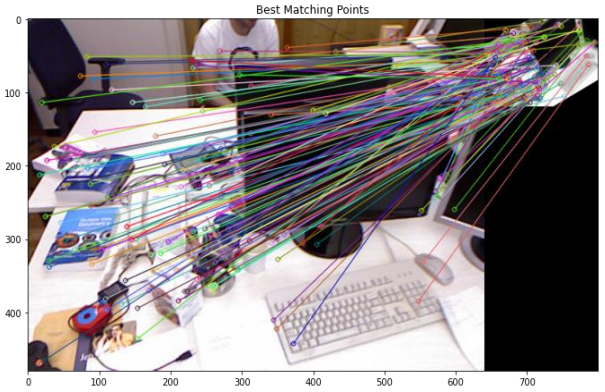
	Image 1	Image 2
SIFT	148	151
BRISK	97	92
ORB	84	85

TABLE IV. Execution Time

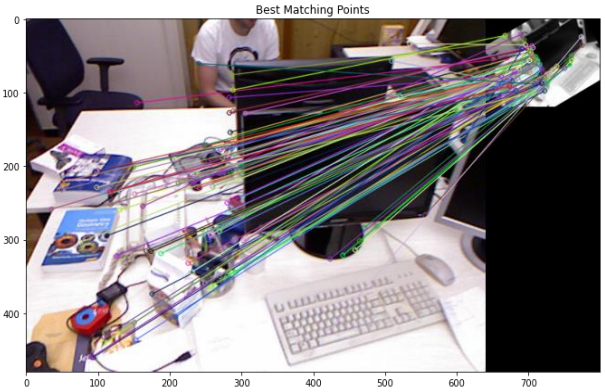
	Image 1	Image 2
SIFT	0.3721s	0.3799s
BRISK	0.3503s	0.3510s
ORB	0.3072s	0.2969s

From the above results we can see that: BRISK has the greatest number of features detected, followed by SIFT

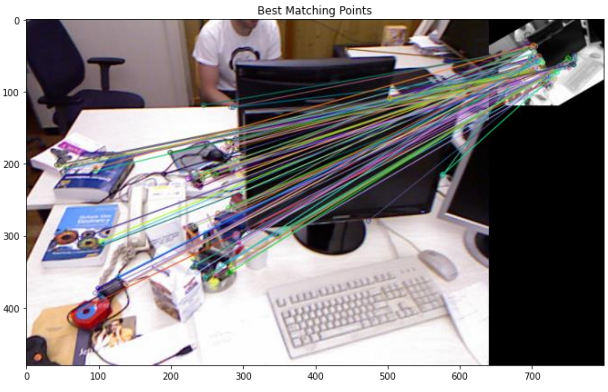
and then ORB. SIFT has the greatest number of features matched followed by BRISK and then ORB. Compared to the number of features detected, the ratio of features detected to features matched is highest in SIFT, followed by ORB and then BRISK. In terms of computation time, ORB is clearly the fastest, followed by BRISK and then ORB. Therefore, taking these factors into account, we decided to work with ORB.



(a) SIFT matching results



(b) BRISK matching results



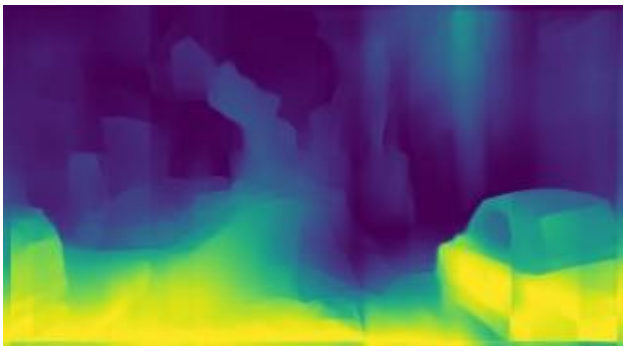
(c) ORB matching results

Figure 5. Matching result from (a) SIFT, (b) BRISK, (c) ORB on sample from the TUM-RGBD dataset [19].

Thus, concluded ORB feature-based mapping algorithm was tested on sequence 01 of the KITTI dataset [22] and corresponding map is presented in Fig. 6 where the path was generated along the detected trajectory and depth points perceived along it. The intermediate feature and depth map as obtained for one of the frames in the sequence are represented in Fig. 7.

Figure 6. ORB-SLAM output map for KITTI-01 sequence.

(a) ORB features



(b) Depth map

Figure 7. Intermediate processing inferences from a KITTI 01 Sequence frame: (a) ORB features and (b) depth information extracted.

IV. Conclusion

The system thus developed provides a pipeline for utilising a depth prediction model to compensate for absence of depth sensors to perform RGB-D input-based SLAM, which provides better performance than Monocular input-based SLAM. For the purpose of depth prediction, we have trained a ResNet based Encoder-Decoder architecture on the NYU-V2 depth dataset [12] and deployed it on the monocular input stream to provide more information to pose estimation and mapping stages in subsequent sub-system. The performance of deployed

architecture could be improved with more suitable data as per specific application and improvements in loss function, which is left for future work. The study done over various feature extractors affirms the usage of ORB feature extractor due to it's greater speed and matching capabilities. Consequently, deployed ORB SLAM algorithm provides us with a final map from various points detected in the visual trajectory with corresponding depth information. Loop closing could be improved with a better vocabulary and is left for future iterations. Finally, the system developed enables a wider usage of mobile computational platforms in applications related to vision-based mapping and/or localisation.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Yatharth Ahuja conceptualized, designed and executed majority of experiments vis-à-vis the encoder-decoder architecture for depth estimation, considering various feature extractors, executing feature-based SLAM and incorporating depth data in SLAM algorithm. He also contributed heavily in finalizing manuscript; Tushar Nitharwal carried on the experiments on various feature extractors and also contributed in depth estimation architecture's deployment; Utkarsh Sundriyal helped in compiling and completing the manuscript; Prof S. Indu and Dr. Anup K. Mandpura provided invaluable inputs and feedbacks throughout the project with respect to concepts, experiments, executionary details and manuscript structure. All authors had approved the final version.

Acknowledgment

The authors would like to thank Dr. Nathan Silberman of New York University for providing a great dataset, NYU-V2 depth dataset, to carry out this research.

References

[1] P. Ackland, S. Resnikoff, and R. Bourne, "World blindness and visual impairment: despite many successes, the problem is growing," *Community Eye Health*, vol. 30, no. 100, pp. 71-73, 2017.
[2] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Towards the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309-1332, 2016.
[3] M. Bousbia-Salah, A. Redjati, M. Fezari, and M. Bettayeb, "An ultrasonic navigation system for blind people," in *Proc. IEEE International Conference on Signal Processing and Communications*, 2007, pp. 1003-1006.
[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.
[5] M. T. N. Truong and S. Kim, "A review on image feature detection and description," in *Proc. the Korea Information Processing Society Conference*, 2016, pp. 677-680.
[6] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. International Conference on Computer Vision*, 2011.
[7] Z. Chen, X. Liu, M. Kojima, Q. Huang, and T. Arai, "A wearable navigation device for visually impaired people based on the real-time semantic visual SLAM system," *Sensors*, vol. 21, no. 4, p. 1536, 2021.
[8] J. Ventura, C. Arth, G. Reitmayr, and D. Schmalstieg, "Global localization from monocular SLAM on a mobile phone," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 4, pp. 531-539, April 2014.
[9] J. Li, B. Yang, D. Chen, N. Wang, G. Zhang, and H. Bao, "Survey and evaluation of monocular visual-inertial SLAM algorithms for augmented reality," *Virtual Reality & Intelligent Hardware*, vol. 1, no. 4, 2019.
[10] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2261-2269.
[11] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5162-5170.
[12] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Lecture Notes in Computer Science*, Berlin, Heidelberg: Springer, 2012, vol. 7576.
[13] J. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600-612, 2004.
[14] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147-1163, 2015.
[15] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2003.
[16] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381-395, June 1981.
[17] E. Olson, J. Leonard, and S. Teller, "Fast iterative alignment of pose graphs with poor initial estimates," in *Proc. IEEE International Conference on Robotics and Automation*, 2006, pp. 2262-2269.
[18] D. Galvez-Lopez and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Trans. Robot.*, vol. 28, no. 5, pp. 1188-1197, 2012.
[19] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, 2012, pp. 573-580.
[20] M. Lee, S. Hwang, C. Park, and S. Lee, "EdgeConv with attention module for monocular depth estimation," ArXiv, abs/2106.08615, 2021.
[21] X. Tu, *et al.*, "Learning depth for scene reconstruction using an encoder-decoder model," *IEEE Access*, vol. 8, p. 1, 2020.
[22] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *The International Journal of Robotics Research*, vol. 32, pp. 1231-1237, 2013.

Yatharth Ahuja is expected to receive his Bachelor of Technology degree in Electrical Engineering from the Delhi Technological University, New Delhi, India in 2022. His research interests are mainly the perception, learning and control design of intelligent and robotic systems. He is a member of the IEEE student community.

Tushar Nitharwal is expected to receive his Bachelor of Technology in Electrical Engineering from Delhi Technological University, New Delhi, India. His current interests include deep learning and computer vision and their applications. He has worked on several projects in the same field such as Facial Emotion Recognition, Line following robot etc.

Utkarsh Sundriyal, a data science enthusiast, is currently expected to receive his Bachelor of Technology degree in Electrical Engineering from the Delhi Technological University, New Delhi, India in 2022. His current interests include quantitative analysis and computer vision. He has participated in various competitions of martial arts. He is actively involved in social work as well.

Sreedevi Indu received the B.Tech. and M.Tech. degrees from University of Kerala, India and Ph.D. degree in Electronics and Communication engineering from Delhi University, India. She is currently a professor at the Department of Electronics and Communication Engineering, Delhi Technological University, India. Her current research interests include computer vision, image processing, wireless sensor networks and their applications. She was the branch counsellor of IEEE Student branch and received awards for her contributions.

Anup K. Mandpura received the B.Tech. degree from BHU, India, M.Tech. degree from IIT Guwahati, India and Ph.D. degree from IIT Delhi, India. He is currently an assistant professor at the Department of Electrical Engineering, Delhi Technological University, India. His current research interests include Performance Analysis of Wireless Energy Harvesting Systems, Machine Learning with application to Communication systems and PV systems, Networked Control Systems.