Real Time Multimodal Emotion Recognition System using Facial Landmarks and Hand over Face Gestures

Mahesh Krishnananda Prabhu and Dinesh Babu Jayagopi

***Abstract*—Over the last few years, emotional intelligent systems have changed the way humans interact with machines. The main intention of these systems is not only to interpret human affective states but also to respond in real time during assistive human to device interactions. In this paper we propose a method for building a Multimodal Emotion Recognition System (MERS), which combine mainly face cues and hand over face gestures which work in near real time with an average frame rate of 14 Fps. Although there are many state of the art emotion recognition systems using facial landmarks, we claim that our proposed system is one of the very few which also include hand over face gestures, which are commonly expressed during emotional interactions.**

***Index Terms*—Hand-over-face gesture, facial landmark, histogram of oriented gradient, space-time interest points.**

I. Introduction

We need Emotional AI because our emotional imperfections dwarf our ability to take decisions, to manage work life balance and to repair messed up relationships. Although we are surrounded by artificial Intelligent Systems and technologies with massive cognitive and autonomous abilities, all these systems have Intelligence Quotient (IQ) but not *Emotional Quotient* (EQ). The way we are interacting with machines is changing; it's becoming a lot more relational, intimate. When people are faced with problem even severe or life threatening ones they often reach for their smartphones or personal assistants for help and in most of the cases these devices might not understand very well. They may respond saying "I am sorry I don't know what you mean by that and at best, they will refer you to a call center. What people are looking for is a trustworthy companion, advisor. And this can have real value in terms of motivating behavioral change, such as improvement in the quality of Interpersonal relationship, minimizing distress, enhancing personal effectiveness etc. So the Emotional Intelligence is the key factor for socially engaging the user with devices.

People express emotions through multiple modalities i.e. through Speech, Facial Expression and Body Pose and also through their hands. Although the verbal aspect of the interactions have been there from long time Nonverbal communication plays a central role in how humans communicate and empathize with each other. There have been lot of emotion recognition solutions based on facial landmarks, but very few are based on multiple modalities with face, along with gestures using hands. In this paper we try to explore non-verbal cue, hand over face gestures and build a system which can respond to natural human behavior in real time. We consider this problem as two separate tasks: first being detection of emotions through facial landmarks; and second recognition through hand over face gestures.

The main objectives of this paper are as follows:

1) A facial landmark based emotion state recognition
2) Automatically code and classify hand over face gestures
3) Fuse the above two methods in a novel way in order to achieve a real time emotion recognition system

In Section II we present the related work, in Section III we describe about the overall system and method applied. In Section IV we talk about the experimental evaluation and final system in Section V. Lastly about the conclusion and Future work.

II. Related Work

Out of different modalities of emotion recognition afore mentioned the face has received the most attention from both psychologists and affective computing researchers [1]. It is not very surprising as faces are the most visible social part of the human body. They reveal emotions [2], communicate intent, and help regulate social interaction [3]. Body language also is an important method used to communicate affect [4]. Early research on adaptor style body language in [5] presented the importance of leaning, head pose and the overall openness of the body in identifying human affect. More recent research presented in [6] has shown that emotions displayed through static body poses are recognized at the same frequency as emotions displayed with facial expressions. One of the main factors that limit the accuracy of facial analysis systems is hand occlusion. As the face becomes occluded, facial features are lost, corrupted, or erroneously detected. However, there is empirical evidence that some of these hand-over-face gestures serve as cues for recognition of cognitive mental states [7]. Although there are large numbers of methods which can measure the emotion recognition through face, none of them include hand over face gestures. So the focus of this paper is mainly on combining the facial landmarks and hand over face gestures in building the system.

Manuscript received March 6, 2017, revised April 10, 2017.

Mahesh Krishnananda Prabhu is with Samsung R&D Institute, Bagmane Constellation Business Park, Phoenix Building, Outer Ring Road, Doddanekkundi, Bengaluru, Karnataka 560037, India (e-mail: mahesh.kp@iiitb.org).

Dinesh Babu Jayagopi is with Multi Modal Perception Lab, International Institute of Information Technology Bangalore (IIITB), 26/C, Hosur Rd, Electronics City Phase 1, Electronic City, Bengaluru, Karnataka 560100, India (e-mail: jdinesh@iiitb.ac.in).

doi: 10.18178/ijmlc.2017.7.2.615

III. Overall System

We divide our problem statement into two parts. First find gesture through hand over face gestures and second through Facial Landmark points. For hand over face gesture first we find out if there is a hand occlusion or not by using some of the coding descriptors mentioned in [7] and then classify the gestures based on certain hypothesis. In Second part we extract and train the facial landmark region to classify different emotions. We recognize totally two gestures using hand over face gestures and four using facial landmark points. The overall system diagram is mentioned in Fig. 1.

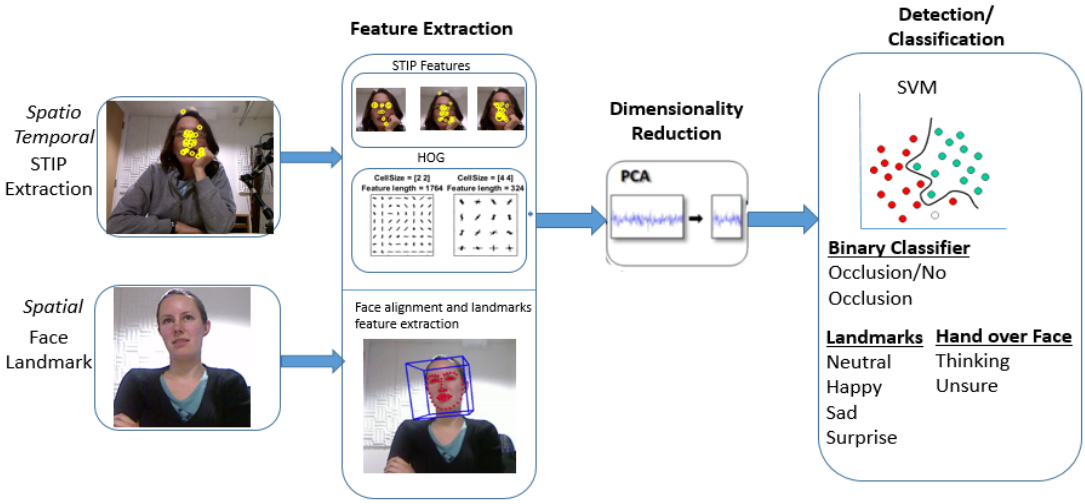

Fig. 1. Overall System showing different steps in the processing of the system.

A. Emotion Recognition using Hand over Face Cues

1) Coding descriptors

For classifying hand over face gestures we have used [8], 3D multi-modal corpus of naturally evoked complex mental states, which has been labeled using crowd-sourcing. It has a collection and annotation methodology of a 3D multimodal corpus of 80 audio/video segments of natural complex mental states. The corpus includes spontaneous facial expressions and hand gestures labeled using crowd-sourcing. Out of 80 videos 25 videos had hand over face gestures. We used similar coding descriptors mentioned in [7] but due to unbalanced dataset keeping real time system in mind we combined some of these descriptors. These videos were manually labeled into following categories

- Hand Action – coded as one label for entire video either static or dynamic. It could be touching, stroking or tapping
- Hand Occlusion – coded as one label for entire video, whether hand occlusion present or not. The occlusion could be any region on the face , forehead, chin, cheeks, lips and hair

The data set had around 80 videos which had 12 mental states. As per table mentioned by Mahmoud in [8] although people use their hand over face gestures for mental states for Happy, bored, Interested and others but majority of the people use for two states Thinking and Unsure. This is marked in blue in Fig. 2. Hence for our system we consider detecting only these two states.

2) Feature extraction for hand over face gestures

For feature extraction we wanted to choose those which would aptly represent the above mentioned descriptors. For hand action we considered Space Time Interest Points (STIP) that combined spatial and temporal features. For spatial features we used Histogram of Gradient (HOG). After feature extraction we used Principal Component Analysis (PCA) to obtain a compact representation.

3) Space TIME interest point (STIP)

Local space-time features [9], [10] are popularly used feature for detection of action recognition [11]. Recently, Song *et al*. [12] used them to encode facial and body micro expressions for emotion detection. They reflect interesting events that can be used for a compact representation of video data as well as for its interpretation. We used the approach proposed by Song *et al*. [12]. STIP capture salient visual patterns in a space-time image volume by extending the local spatial image descriptor to the space-time domain. Obtaining local space-time features has two steps: spatio temporal interest point (STIP) detection followed by feature extraction. Mahmoud [7] used Harris3D interest point detector followed by a combination of the Histograms of Oriented Gradient (HOG) and the Histogram of Optical Flow (HOF) feature descriptors. Keeping real time scenario in mind in our approach we used Harris interest points with Local jet features rather than HOG. We saw that with this the feature detection process speeded up by 15-20 times compared to HOG.

4) Histogram of GRADIENTS (HOG)

Histograms of Oriented Gradients (HOG) are very popularly used for pedestrian detection [13], and facial landmark detection [14] amongst others. HOG technique counts occurrences of gradient orientation in localized portions of an image. These occurrences are represented as a histogram for each cell normalized in a larger block area. HOG features capture both appearance and shape information making them suitable for a number of computer vision tasks.

B. Emotion Recognition Using FACIAL Landmarks (Coding Descriptors)

For emotion recognition using landmark detectors we used Cohn-Kanade dataset [15] which has emotions and AU labels,

along with the extended image data and tracked landmarks. This database contains Image sequences in which the subject's emotion changes from a neutral expression to a peak expression. The data base was taken and then manually separated out based on the Emotions. We have considered 4 Emotion states viz. Happy, Sad, Surprise and Neutral.

1) Feature extraction for FACIAL landmark detection

There has been a flurry of activity in the area of facial feature detection and tracking libraries in the last 5 years. Part of the reason for this activity is the availability of large annotated datasets like LFPW and Helen. We chose the one implemented in dlib since it had a Real time pose estimation solution [16]. The Faces are detected using HOG features implementation

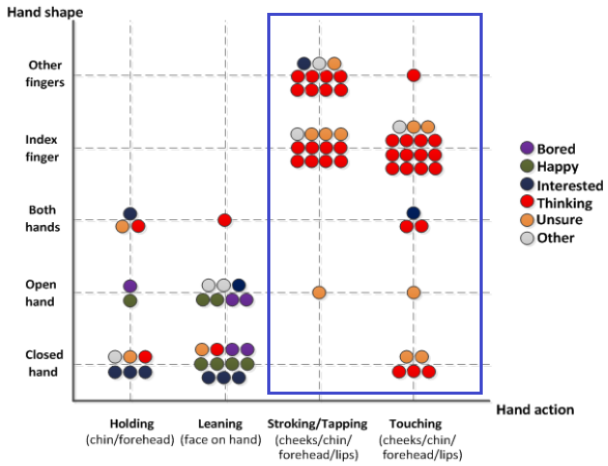

Fig. 2. Heat map of thinking and unsure mental states (Reference [8]).

IV. Experimental Evaluation

For our classification tasks, we used the labeled subset of Cam3D described in Section III.A to evaluate our approach. Below table mentions about the data set we considered

TABLE I: Data Set Consideration For Training

Hand Action	Static or Dynamic	Whole Video	25 videos(4191 frames)
Hand Occlusion	Present or Absent	Whole Video	25 videos(4625 frames)

As a part of preprocessing step we did a face alignment for all our videos and then performed scaling, the final resolution of the image was 160×120. Space time features were extracted at the original video frame rate (30 frames per second) as mentioned in Section 3.3. We removed the features not in the facial region by using the results from the landmark detector. For HOG features we used very similar approach presented in [7]. We extracted HOG features from a normalized 160×120 pixel image of a face. We used 8×8 pixel cells with 18 gradient orientations and block size of 2×2 cells. 9576 dimension HOG Vectors were reduced to 1035 through PCA. Window frame of 10 was considered and aggregated. We used uni-modal and multimodal fusion approaches and standard Liblinear [17] library for SVM.

A. Hand Occlusion Detection

We manually labeled corpus videos as occluded or not and then trained using linear SVM classifier using single modalities and feature-level fusion. Table 1 shows the classification accuracy results of uni-modal features and multi-modal fusion. We found that the best performance is obtained from the multi-modal linear classifier.

B. Hand Action Detection

For Hand action, the data was labeled as one label per video, describing the hand action as static or dynamic in the majority of the video frames. Therefore, we aggregated the features to obtain one feature set per video. We used a binary classification approach to categorize the hand action as dynamic or static. Table II shows the accuracy

TABLE II: Hand Over Face Detection

Method	STIP	HOG	Fusion
Hand Occlusion (1754 Frames)	44.4%	66.7%	**70%**
Hand Action (937 frames)	83.3%	66.7%	**83.3%**

C. Emotions through Facial Landmarks

As explained in Section III.B the database was manually separated based on emotions. Firstly faces were detected using dlib library and then landmark Points were aligned on the detected faces. HOG features were generated based on these landmarks. There were around 68 landmark points. The length and slope of line segment from every point to every other point is used as feature. Therefore there were 4556 features for a single face. SVM was used because it is computationally less demanding than Artificial Neural Network and Logistic Regression. Columns were separated to Features and labels. Rows were randomized and Columns of features normalized. A multiclass approach was used for training using [18]. Both One-versus-One (OVO) and One-versus-All (OVA) methods were implemented. OVA is easier to get the probability values corresponding to each Emotion whereas OVO is less effected by problems of imbalanced datasets but computationally expensive. Considering real time situation demand we used OVA. We used cross validation to find optimal parameters for SVM. 3 Fold cross validation was applied to find optimal parameters (v=0.0498789 and γ = 1.4641e-05). After the training classifiers were stored, OVO had 6 classifiers and OVA had 4 classifiers. For Emotion detection faces are detected and re-checked in same way as annotation. All the faces in a given image are Cropped and written to disk and numbered. Landmark points are aligned and features are generated similar to annotation. Features are generated from each image and stored in the form of a vector. The feature is passed to the classifier and emotion is detected.

V. Final System

Our final aim was to combine the 3 methods mentioned in the previous section and build a near real time system to detect emotions through face and hand over face gestures. From the data set we could infer that the hand over face gestures was mostly used for Unsure and Thinking emotions. Hence we followed the approach mentioned in Fig. 3 to

differentiate them. Our final system could detect around 6 emotional states in total with fairly near real time as showed in Fig. 4. In our system the execution time for calculating the emotion recognition based on landmarks was a bit on higher side compared to hand over face which affected the final FPS the most. Hence in order to achieve real time system we skipped certain frames under the assumption that the emotion might not change in five frames i.e. in one sixth of a second. For Facial Landmark detection we used MUG [19] database, for benchmarking frame rate we used AMFED [20] which had around 242 videos captured in real world scenarios. MUG database consists of image sequences of 86 subjects performing facial expressions. We have compared our system against [21] in terms of emotions with landmark solution which uses Gabor transforms. We found that our system was better in terms of accuracy and also in time. Table III shows the comparison results.

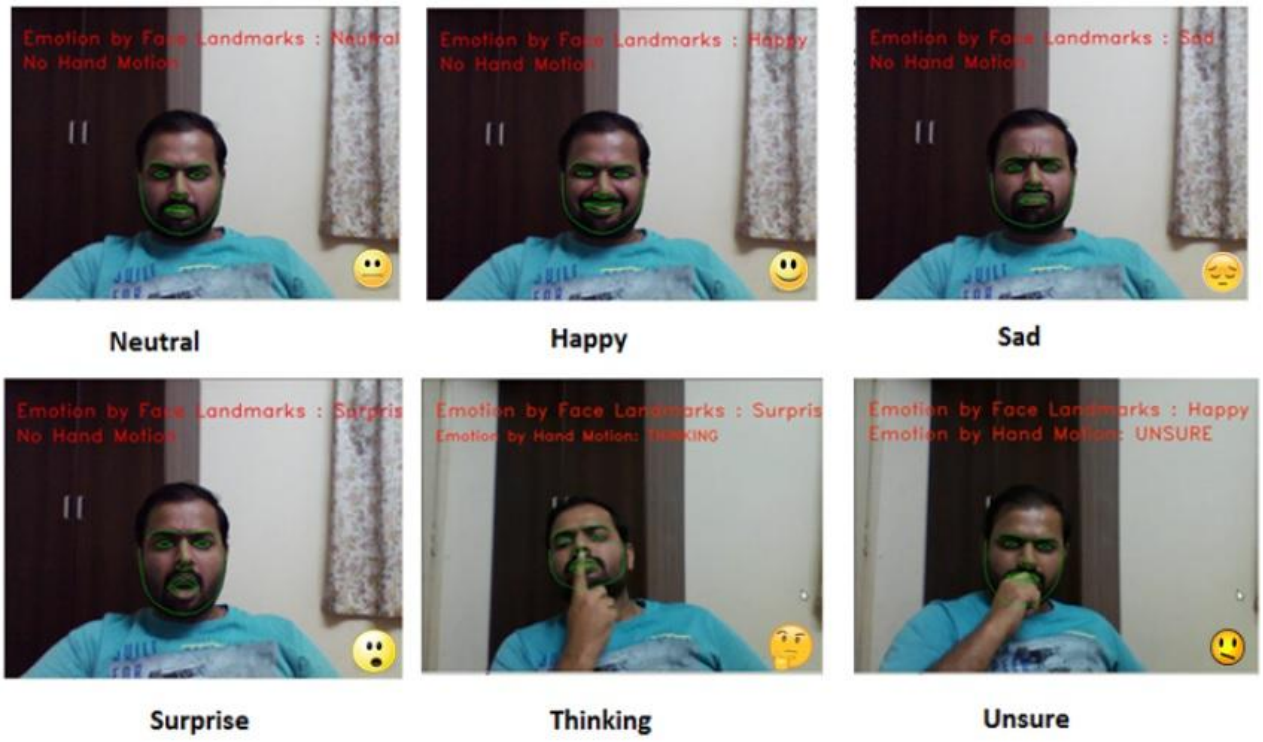

Fig. 3. Our final system showing 6 emotional states.

TABLE III: FPS DETAILS OF OUR SYSTEM

Video Resolution	AMFED DB	Modality	FPS (Average)
320 × 240	Little Wort [18]	Emotions through Facial Landmarks	**12.2fps**
320×240	Our Method	Facial Landmarks and Hand over Face Gestures	**14.1fps**

TABLE IV: ACCURACY DETAILS OF OUR SYSTEM

Emotion (MUG DB [20]	LittleWort [21]	Our Method
Happy	82.62	**99.78**
Sad	26.8	**58.6**
Neutral	80.04	**85.18**
Surprise	24.31	**91.78**

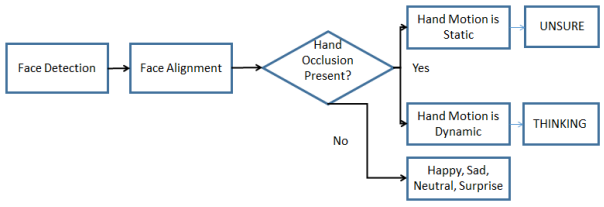

Fig. 4. Flow diagram of our final system.

VI. CONCLUSION AND FUTURE WORK

In this work we showed how hand over face gestures and facial landmarks can be effectively used to build a multimodal emotion recognition system. Also we depicted how we could make it run in near real time. Going forward more descriptors can be added to make accuracy of the system better and also add more mental states. And also there could ways to improve the emotion prediction system by bringing in the audio modality. This requires joint learning of the both audio and video parameters. Exploiting co-relations between these two dimensions will be one of the important challenges. Our future work would be to tune this system to work in un-constrained system conditions.

REFERENCES

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[2] P. Ekman and E. L. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*, Oxford University Press, USA, 1997.

[3] J. F. Cohn, "Human facial expressions as adaptations: Evolutionary questions in facial expression research Karen L," Schmidt Departments of Psychology and anthropology.

[4] C. Shan, S. Gong, and P. W. McOwan, "Beyond facial expressions: Learning human emotion from body gestures," In *BMVC*, pp. 1–10, 2007.

[5] W. T. James, "A study of the expression of bodily posture," *The Journal of General Psychology*, vol. 7, no. 2, pp. 405–437, 1932.
[6] K. L. Walters and R. D. Walk, *Perception of Emotion from Body Posture*, 1986.
[7] M. M. Mahmoud, T. Baltrušaitis, and P. Robinson, "Automatic detection of naturalistic hand-over-face gesture descriptors," in *Proc. the 16th International Conference on Multimodal Interaction*, 2014, pp. 319–326.
[8] M. Mahmoud and P. Robinson, "Interpreting hand-over-face gestures," in *Proc. International Conference on Affective Computing and Intelligent Interaction*, pp. 248–255, Springer, 2011.
[9] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.
[10] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.
[11] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.
[12] Y. Song, L.-P. Morency, and R. Davis, "Learning a sparse codebook of facial and body microexpressions for emotion recognition," in *Proc. the 15th ACM on International conference on multimodal interaction*, pp. 237–244, ACM, 2013.
[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR 2005. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893.
[14] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2879–2886, 2012.
[15] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Proc. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 94–101, 2010.
[16] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1867–1874, 2014.
[17] C.-C.g Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
[18] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes," *Pattern Recognition*, vol. 44, no. 8, pp. 1761–1776, 2011.
[19] N. Aifanti, C. Papachristou, and A. Delopoulos, "The mug facial expression database," in *Proc. 2010 11th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)* , 2010, pp. 1–4.
[20] D. McDuff, R. Kaliouby, T. Senechal, M. Amr, J. Cohn, and R. Picard, "Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 881–888.
[21] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Fully automatic facial action recognition in spontaneous behavior," in *Proc. 7th International Conference on Automatic Face and Gesture Recognition*, 2006, pp. 223–230.

Mahesh Krishnananda Prabhu is pursuing his masters at IIITB, Bangalore in information technology. He obtained bachelor of engineering degree from Bangalore Institute of Information Technology, VTU in 2005. His research interest includes Image Processing, Machine Learning and Human Robot Interaction and has been part of Multimodal Perception Lab at IIIT-B from past 2 years.

He is currently working in Samsung Research Institute India Bangalore and has close to 13 years on Software Industry experience in computer vision algorithms, machine learning, cloud computing and mobile software development in multimedia domain.

Dinesh Babu Jayagopi obtained his doctorate from Ecole Polytechnic Federale Lausanne (EPFL), Switzerland in 2011. His research interests are in audio-visual signal processing, machine learning, and social computing. He is currently working at IIITB as an assistant professor, he was a postdoc at the Social Computing Lab, Idiap Research Institute (EPFL) for 2.5 years. Prior to his PhD, he worked as a senior research engineer at Mercedes-Benz Research and Technology, Bangalore for 3 years. He completed his M.Sc.(engg) from I.I.Sc, Bangalore in 2003, specializing in system science and signal processing; and B.Tech in electronics from Madras Institute of Technology in 2001.

He heads Multimodal perception Lab at IIITB. The Multimodal Perception lab focuses on human-centered sensing and multimodal signal processing methods to observe, measure, and model human behavior. These methods are used in applications that facilitate behavioral training, and surveillance; and enable human-robot interactions (HRI). The focus is mainly on vision and audio modalities. Probabilistic graphical models form the backbone of the underlying formalism.