Gesture Recognition from One Example Using Depth Images

Bin Liang, Lihong Zheng

***Abstract*—By using depth images, this paper presents an approach capable of recognizing the gesture from only one example of each class. Background removal and denoising are performed on depth images firstly. Motion Energy Information (MEI) images are then obtained through calculating the differences between consecutive frames. Within each MEI image, we represent successive movements by time series using Histograms of Oriented Gradients (HOG) descriptor. Principle Component Analysis (PCA) reconstruction approach is applied on the descriptor to find a set of discriminantly informative principle components (PCs) from the corresponding training gesture. Next the descriptors extracted from test gestures are reconstructed back utilizing each set of PCs from training gestures. Finally the test gestures are recognized according to the set of PCs which produces the lowest reconstruction error. We evaluate our approach on the task of recognizing gestures from one example using depth images, and compare the performance of our approach with other methods, reaching a promising result.**

***Index Terms*—gesture recognition, depth images, MEI image, HOG, PCA reconstruction.**

I. Introduction

With the potential for many applications of interactive, intelligent computing, human-computer interaction (HCI) is an important area in the computer vision community. Human gestures may provide more meaningful information, so gesture recognition can be termed as a method in this direction. The problem of understanding human gestures is complicated for many issues, including the fact that gestures are dynamic and may happen at various timescales. Gesture recognition is further complicated by the variation between people and even between instances of a single person [1]. All these issues make gesture recognition a challenging problem.

Sensors used for gesture recognition include wearable sensors and external sensors such as video cameras [2]. Compared with the extensive calibration, restricted natural movement and high cost of the wearable sensors, video-based gesture recognition addresses these issues. With the launch of Kinect [3], the task of gesture recognition has been greatly simplified by the introduction of a sequence of depth images collected by a Kinect camera. In general, a depth image contains depth information as 8-bit gray level for each pixel, which presents the distance between the captured object and the camera. Different from RGB images collected using ordinary camera, depth images significantly reduce the huge color and texture variability induced by clothing, hair, skin and background, as the images shown in Fig. 1.

(a) RGB image (b) Depth image

Fig. 1. RGB image and depth image.

Humans can recognize new gestures after seeing just one example, but for computers, recognizing even well-defined gestures, such as sign language, is much more challenging and has traditionally required thousands of training examples to learn. Our work focuses on the gesture recognition from a single training example. In the case of insufficient training examples, the standard tools of statistical machine learning would be very likely to fail because they will suffer from overfitting problem.

Gestures can be static or dynamic. Some gestures also have both static and dynamic elements, as in sign languages. A static gesture is a gesture in which a single posture is held for certain duration, while a dynamic gesture consists of a sequence of postures, which may be repetitive or not, and in which the posture order and the timing of the sequence may be critical [4]. There have been various tools to handle gesture recognition based on the approaches ranging from statistical modeling, computer vision and pattern recognition to image processing. Statistical modeling has been employed to address most of the problems, such as hidden Markov model (HMM) [5][6]. HMM-based approaches and related statistical models are well-known to analyze motion for interpreting and recognizing gestures. However, in the traditional HMM framework, the hidden states are typically coupled with the training examples, which could cause unexpected problems in the learning procedure. Besides, insufficient training examples would be likely to lead to the failure. Many gesture recognition systems have successfully used computer vision and pattern recognition techniques, including feature extraction, object detection, clustering, and classification. Spatio-temporal features [7][8] are considered by many researchers. However, the computational complexity and time-consuming in computation limit their applicability in real-time applications. Image processing techniques, involving analysis and detection of shape, texture, color, motion, optical flow, image enhancement, segmentation, and contour modeling, are also effective for gesture recognition. The major limitation of the traditional image-processing

Manuscript received April 15, 2013.

B. Liang is with the School of Computing and Mathematics, Charles Sturt University, NSW 2678 Australia (e-mail: bliang@ csu.edu.au).

L. Zheng is with the School of Computing and Mathematics, Charles Sturt University, NSW 2678 Australia (e-mail: lzheng@ csu.edu.au).

based method is that it is very easily affected by the conditions of gesture sequences, such as illumination variations, dynamic background, occlusion and skin color, etc.

In this paper, we propose an alternative method to overcome some of the drawbacks from these previous approaches. Human silhouettes can be easily segmented due to the unique property of the depth image, which is highly effective for the gesture recognition in the following procedure. And then, the Histogram of Oriented Gradients (HOG*)* descriptor is performed on Motion Energy Information (MEI*)* images to represent the corresponding gesture. Because a test gesture could be better reconstructed with a set of principle components (PCs) that has been obtained from the same or similar gesture, the most relevant gesture will be recognized according to the lowest reconstruction error. In our experiment, we explore the two components, MEI and HOG in proposed approach, affecting the performance, and compare proposed method with other methods. We demonstrate the proficiency of our approach on Chalearn gesture dataset [4].

II. GESTURE REPRESENTATION

If all the motion information in a gesture sequence can be represented exactly, recognition would simply be a matter of determining whether the movement representation in a given gesture meets the information. Gesture representation must quantify the variance and how the object moves over the course of the gesture.

A. Preprocessing

Background removal is a very important initial step for gesture recognition with the aim of extracting the moving regions from the gesture sequences. Usually, background removal can be completed by calculating the differences between the current frame and the background image for each pixel and then applying threshold to detect the regions of gestures. However, in the case of having no prior background image, it is a challenging task for traditional methods. In depth images, the values of pixels belonging to background have a great difference from those belonging to the object. Utilizing this property, the gesture region in a sequence can be easily segmented from the background by using OTSU [9] bilevel threshold method to classify the pixels.

The depth images also have some drawbacks, one of which is the noise at the edge of objects. With missing bits and a pretty serious flickering issue, noise in depth images resembles a type of salt and pepper noise. According to [10], a 5×5 median filter is adopted for spatial filtering, which can replace the pixel value with the median value of the sub-image. Therefore, the random noise is removed in this way and the original depth image is smoothed.

The original depth images and pre-processed images are shown in Fig. 2. After background removal and denoising for the original depth images, the gesture representation is less prone to negative effect of the original depth images. Experimental results demonstrate that preprocessing operation can reduce the average error rate as much as 6%. Thus these operations are highly effective for the recognition in the following procedure.

(a) Original depth images

(b) Preprocessed images after background removal and denoising

Fig. 2. Original depth images and preprocessed images

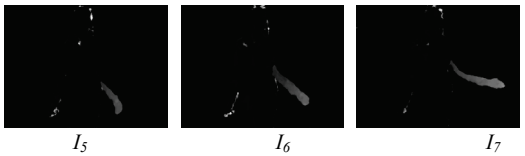
B. Motion Energy Information

In the proposed approach, the variance and movements of the gesture are quantified by a series of Motion Energy Information (MEI) images, which represents where motion has occurred in a gesture sequence. Each gesture corresponds to a video sequence containing consecutive video frames, where each frame is a depth image. G denotes a gesture sequence composed of N frames, $G=\{F_1,\ldots,F_N\}$, in which F_i is the i^{th} frame. Let $V=\{G_1,\ldots,G_K\}$ be a set of gesture sequences corresponding to each gesture vocabulary, where G_j is the j^{th} gesture sequence of the gesture vocabulary in the dataset.

A gesture sequence can be represented by a set of MEI images $I_i\in\{I_1,\ldots,I_{N-1}\}$ according to the differences between consecutive frames in gesture sequence G. MEI image can be obtained by subtracting consecutive frames in the whole gesture sequence:

$$I_i=F_{i+1}-F_i,\ i=1,\ldots N-1 \qquad (1)$$

In Fig. 3, we display the MEI images obtained from a sample gesture. According to the MEI images, the motion of the gesture could be illustrated well using the difference between consecutive frames.



I_5 I_6 I_7

I_8 I_9 I_{10}

Fig. 3. MEI images from a sample gesture sequence

C. HOG Descriptor

The Histograms of Oriented Gradients (HOG) descriptor is proposed for human detection by [11]. Apart from various human detection, HOG descriptor is also employed for gesture recognition by [12]-[15] due to its robust performance. The HOG descriptor can provide details of human gestures. The essence of HOG is to describe local edge structure or appearance of object by local distribution of gradients [11]. Thus, HOG descriptor gains robust representation by representing the motion features as directional patterns over coarser spatial regions.

In our approach, HOG is implemented on MEI images

obtained from difference between consecutive frames. The MEI image is divided into $M \times M$ non-overlapping spatial grids. We calculate the gradient vector of each pixel in the grid.

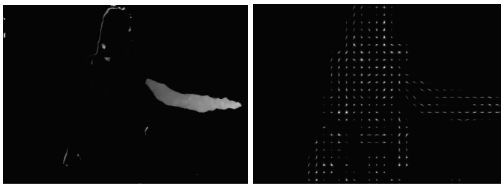
$$g(x,y) = [\mathrm{g}_x(x,y)\ \mathrm{g}_y(x,y)]^T \quad (2)$$

The magnitude and orientation of each gradient vector $g(x,y)$ are denoted as

$$m(x,y) = (\mathrm{g}_x(x,y)^2 + \mathrm{g}_y(x,y)^2)^{1/2} \quad (3)$$

$$\theta(x,y) = \tan^{-1}(\mathrm{g}_y(x,y) / \mathrm{g}_x(x,y)) \quad (4)$$

Each magnitude $m(x,y)$ is weighted in order to vote for the nearest local orientation bins of which the number is B and the adjacent grid histograms, respectively. Here, $\theta(x,y)$ is insensitive to the signs of contrasts, because color variations and background do not produce extra information for gesture recognition in our work. After accumulating the gradient histogram in each grid, L2-norm performs on the concatenated histogram vector. Hence, a single MEI image can be represented as a feature vector of $M \times M \times B$ dimensions. Generally, the more grids that one MEI image is divided into, the more descriptors are extracted. Meanwhile, more negative factors could be involved in. In our experiment, the number of bins B is 9, and three grid sizes were tested, 3×3, 8×8 and 16×16. When the grid size was chosen as 3×3, the average error rate is the lowest, 22.41%, while for 8×8 is 24.94% and for 16×16 is 27.92%. The HOG descriptor in one MEI image is illustrated in Fig. 4.



(a) I_8 (b) HOG descriptor of I_8

Fig. 4 MEI image and corresponding HOG descriptors

III. Gesture Recognition

In our task, only one training example of each class is provided. Traditional generative methods, *e.g.* HMM [5][6], would be very likely to fail due to the lack of training data and cause the overfitting problem. Additionally, for other discriminative models, the one-example condition also limits the discriminative power. Only one example cannot effectively define the hyperplane for discriminating multi-class when using SVM [16]. Adaboost [17] also needs certain quantity of positive and negative examples for training the weak classifiers. The decision trees methods, *e.g.* Random Forest [18], require hundreds of thousands of training examples to avoid overfitting problem [19].

In the case of one training example, we employ PCA reconstruction approach for gesture recognition. This is inspired by the one-class classification task [20] and email classification [21], where the reconstruction error through PCA is used to identify the outliers. The idea of this approach is that PCA can only perform a good reconstruction of the data that was used to compute the PCA basis, and that for other kinds of data the result of reconstruction is poor [21]. In our work, the underlying hypothesis is that a test gesture sequence can be better reconstructed with a set of principle components (PCs) which is obtained from another sequence containing the same or similar gesture.

In gesture representation, descriptors are obtained through performing HOG on MEI images. For each gesture, feature vector D_i is the HOG descriptors of the corresponding MEI image I_i, $1 \le i \le N-1$. In this way, each gesture can be represented using a descriptor matrix $\mathbf{D}$, where each row vector denotes one feature vector D_i. Then PCA is applied on each of the matrices $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_K$, each of which represents the corresponding gesture sequence G_j in the gesture vocabulary V. We store the top t PCs for each gesture, which consist of eigenvalues $\mathbf{W}$ and the corresponding eigenvectors $\mathbf{U}$. Therefore, each gesture sequence G_j in the vocabulary can be represented by a set of PCs using the pattern of $(\mathbf{W},\mathbf{U})_j, 1 \le j \le K$. Identifying the correct number of PCs in PCA is a difficult and computationally expensive problem [22]. If we select too many PCs, the noise would be included from the sampling fluctuations in the analysis. If we choose too few PCs, relevant information would be lost and the analysis is incomplete. Unfortunately, there is not an indisputable approach for the determination of the number of PCs. In our approach, we determined the number of PCs by experiments, and the average error rate is the lowest when the number of PCs is 13.

In the dataset provided by Chalearn [4], each test video contains 1 to 5 gestures. Note that in order to verify the performance of our approach, we use the temporal segmentation annotation provided by [4]. Thus, a test video is firstly segmented so that we can get several single gesture sequences from a test video. Each test gesture sequence is processed using the same method as for the training gesture and hence it is represented by a matrix $\mathbf{T}$. Matrix $\mathbf{T}$ is then projected into each of the K- spaces derived from the training set, and projections are reconstructed back. $\mathbf{R}_1, \ldots, \mathbf{R}_K$ denotes the reconstructions of $\mathbf{T}$ through PCA models $(\mathbf{W},\mathbf{U})_j, 1 \le j \le K$, from the training gesture sequences, respectively. The reconstruction error e of each test gesture is measured as follows

$$e(h) = (1/n) \times \sum_{i=1}^{n} (\sum_{j=1}^{m} (\mathbf{R}_{i,j} - \mathbf{T}_{i,j})^2)^{1/2}, h = 1, \ldots, K \quad (5)$$

where n and m are the number of rows and columns of $\mathbf{T}$. Then we measure reconstruction error over the descriptor matrix corresponding to the test video. Finally, the gesture is recognized associated to the PCA model which produces the lowest reconstruction error, that is $\arg\min_h(e(h))$.

PCA is a process to reveal the internal structure of the data that are being analyzed, by means of seeking the set of PCs that best describes the discrimination or the distribution of such data. Thus, the PCs preserve better the information of the gestures on which the PCA was applied, or of those that are similar.

IV. Experiments and Results

In the following, we present our experimental results. We firstly explore the two components, MEI and HOG, affecting

the performance of the proposed approach. And then we compare the results with other three methods, baseline method provided by [4], principle motion method [23], and Extended-MHI method [24] on Chalearn gesture dataset [4]. In order to study the performance of the proposed approach, we conducted the experiments on 10 different data batches from the dataset, with each data batch presenting a gesture vocabulary.

A. Dataset

The data in Chalearn gesture dataset [4] are recorded using a Kinect camera including both hand and arm gestures. The data are organized in batches of 100 gestures pertaining to a small gesture vocabulary of 8 to 12 gestures, recorded by the same user. Our experiments are performed on the first 10 data batches of the dataset, including 10 different gesture vocabularies: Canada Aviation Ground Circulation 1, Referee Wrestling Signals 1, Gang Hand Signals 1, Diving Signals 2, Gestuno Disaster, Diving Signals 3, Referee Volleyball Signals 1, Gestuno Topography, Referee Volleyball Signals1, and Surgeon Signals. Each gesture vocabulary is made of 47 gesture sequences and split into a training set and a test set. The gesture sequences were recorded at approximately 10 frames per second. For most of the batches the depth accuracy is in the range of 50 to 800 levels. The depth value was normalized and then mapped to 8-bit integers. The image sizes are 320×240.

B. Evaluation Method

In our experiments, we quantify the recognition error rate by computing the Levenshtein distance [25] between the list of predicted gesture labels R and the corresponding list of true gesture labels T. The Levenshtein distance is the minimum number of edit operations that one has to perform to go from R to T (or vice versa). Accordingly, the final recognition error rate can be obtained through the sum of the generalized Levenshtein distances for all the lists of the result compared to the corresponding lists of the truth value.

C. Experiment 1: Effect of Components

In order to study the effect of two components in our proposed method: MEI images and HOG descriptor, we experiment extensively on the performance of the method only using MEI images (MEI) and method only using HOG descriptor (HOG). And then we compare the results to explore the effect of the two components in our approach (MEI+HOG). The experimental results shown in Table I prove the viability of our approach.

TABLE I: COMPARISONS OF TWO COMPONENTS IN PROPOSED APPROACH

Method	Average error rate (%)
MEI	31.02
HOG	23.73
MEI+HOG	22.41

MEI images are used in our approach to encode the information of motions in the gesture, and capture the temporal information of the motions in the corresponding sequence. Thus, MEI images emphasize movements of the gesture. However, MEI images are poor at representing the local shape of the motion while HOG descriptor provides the complementary ability to capture edge or gradient structure that is very characteristic of local shape. Therefore, the proposed method has the complementary property by combining MEI images and HOG descriptor for the representation of a gesture sequence.

D. Experiment 2: Comparison with Other Methods

Table II compares the average recognition error rate of our approach with results from other three methods, baseline method [4], principal motion method [23] and Extended-MHI [24] method. Compared to the existing approaches, our method shows a better performance. Our approach achieves 22.41% average error rate, which illustrates that the approach can be effectively adopted for gesture recognition from only one example.

TABLE II: COMPARISONS WITH OTHER METHODS

Method	Average error rate (%)
Baseline	62.80
Principle motion	37.42
Extended-MHI	26.00
Our approach	22.41

E. Results Analysis

Fig. 5 shows our approach's performance on the first 10 data batches. The results reveal that the proposed method performs well when there are large arm movements in a gesture sequence, *e.g.* data batch 1, 4, 5, 8 and 9. However, there are particular batches, *e.g.* data batch 3 and 10, where the main movements are hand movements and finger movements. The example frames of these two batches are show in Fig. 6. These subtle movements dominate the whole gesture causing it to be confused with other similar gestures. We reason the higher error rates in these batches that there is no hand detector in our approach to locate the hand position.

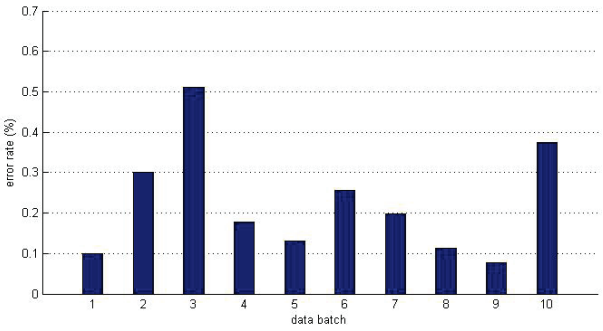


Fig. 5. Results on the 10 data batches

Fig. 6. Gesture sample frames from batch 3 and batch 10

Nevertheless, our approach is encouraging as we are capable to recognize the gesture from one example by using depth images. As for the subtle hand movements and finger movements, a reliable hand detector is expected.

V. CONCLUSIONS

In this paper, we have presented and evaluated an approach for gesture recognition from one example using the depth images from a Kinect camera. By using the unique property of the depth images, we conduct the preprocessing for the original images including background removal and denoising. With the combination of MEI images and HOG descriptor, the information of movements is encoded by MEI images, and then HOG descriptor is explored to represent the gesture sequence. Because a set of PCs enables to preserve the diversity and represents important class information in terms of variance of each gesture class, when it is used to reconstruct an unseen gesture to be recognized, the set of PCs with similar properties is able to reconstruct it with minor loss of information. According to the reconstruction error, the test gesture with lowest reconstruction error is close to the same or similar training example. The experimental results have shown that the proposed approach is effective for gesture recognition from only one example using depth images, and performs better than other methods included in the Chalearn gesture dataset.

In the future work we would like to further explore other properties in depth images, and employ more advanced appearance-based descriptor for more accurate subtle gesture recognition, which could conquer the ineffectiveness in discriminating hand and finger movements by introducing a reliable hand detector.

REFERENCES

[1] Y. M. Lui, and J. R. Beveridge, "Tangent bundle for human action recognition," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 2011, pp. 97-102.

[2] J. Suarez and R. R. Murphy, "Hand gesture recognition with depth images: A review," in *RO-MAN, 2012 IEEE*, 2012, pp. 411-417.

[3] Microsoft Corp. Redmond WA. Kinect for Xbox 360.

[4] Chalearn gesture dataset. *CGD2011, ChaLearn, California*, 2011.

[5] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-285, Feb. 1989.

[6] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," in *Computer vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on*, 1992, pp. 379-385.

[7] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. "Behavior recognition via sparse spatio-temporal features," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, 2005, pp. 65-72.

[8] I. Laptev and T. Lindeberg. "Space-time interest points," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on,* 2003, pp. 432-439.

[9] N. Otsu. "A threshold selection method from gray-level histograms", *Automatica*, 1975.

[10] S. Park, S. Yu, J. Kim, S. Kim, and S. Lee. "3D hand tracking using kalman filter in depth space," *EURASIP Journal on Advances in Signal Processing*, 2012.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, pp. 886-893.

[12] C.-C. Chen and J. Aggarwal, "Recognizing human action from a far field of view," in *Motion and Video Computing, 2009. WMVC'09. Workshop on*, 2009, pp. 1-7.

[13] S. Satkin and M. Hebert, "Modeling the temporal extent of actions," in *Computer Vision--ECCV 2010*, 2010, pp. 536-548.

[14] 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*, 2008, pp. 1-8.

[15] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Computer Vision, 2009 IEEE 12th International Conference on*, 2009, pp. 32-39.

[16] C. Cortes and V. Vapnik. "Support-vector networks," *Machine learning*, pp. 273-297, 1995.

[17] Y. Freund and R. E. Schapire. "Experiments with a new boosting algorithm," in *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE*-. 1996, pp. 148-156.

[18] L. Breiman. "Random forests," *Machine learning*, pp. 5-32, 2001.

[19] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, et al., "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, pp. 1297-1304, 2011.

[20] D. M. Tax, "One-class classification," Ph.D. dissertation, Delft University of Technology, 2001.

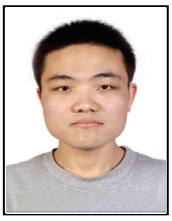
[21] J. C. Gomez and M.-F. Moens, "PCA document reconstruction for email classification," *Computational Statistics & Data Analysis*, 2011.

[22] D. M. Blei, A. Y. Ng, and M. I. Jordan. "Latent Dirichlet allocation," *the Journal of Machine Learning Research*, pp. 993-1022, 2003.

[23] H. J., Escalante and I. Guyon. Principal motion: PCA-based reconstruction of motion histograms. Technical report, ChaLearn Technical Memorandum, June 2012. http://www.causality.inf.ethz.ch/Gesture/principal_motion.pdf, 2012.

[24] D. Wu, F. Zhu, and L. Shao, "One shot learning gesture recognition from rgbd images," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, 2012, pp. 7-12.

[25] V. Levenshtein. "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet Physics Doklady*, 1966, pp. 707-710.



Bin Liang was born in China in 1987. In 2009 he received his bachelor's degree in computer science and technology from Taiyuan University of Technology, China. In 2012 he received his master degree in computer software and engineer from Taiyuan University of Technology. Now he is a PhD student in School of Computing and Mathematics, Charles Sturt University, Australia.

He has authored and co-authored several papers in the field of image processing and pattern recognition.

Lihong, Zheng received her PhD degree in Computing Sciences from University of Technology, Sydney in 2008. She is currently senior lecturer at School of Computing and Mathematics, Charles Sturt University. Her previous research interest was on automation, robotics and artificial intelligence. Her current research interest is on computer vision, image processing, pattern recognition and machine learning.

She had one patent and authored and co-authored more than forty papers in the field of artificial intelligence and image processing, machine learning and Robotics.