Journal of Engineering Science and Technology
Vol. 15, No. 6 (2020) 3601 - 3614

INTER PERSON ACTIVITY RECOGNITION USING RGB-D DATA

M. M. SARDESHMUKH[1,*], M. T. KOLTE[1], V. M. SARDESHMUKH[2]

[1]JSPM Narhe Technical Campus, Narhe Pune,
India, Pimpari Chinchwad College of Engineering, Pune, India
[2]Sinhgad Academy of Engineering, Pune, India
*Corresponding Author: mmsardeshmukh2016@gmail.com

Abstract

The large amount of video data from various sources like CCTV is widely available. Automatic analysis of video for scene understanding is essential and useful in many video surveillances, applications like anomaly detection, activity recognition, patent monitoring. In this paper, we have presented an Activity Recognition System in varying illumination. Proper segmentation and selection of features and classifiers are crucial in such applications. The depth information from the RGB-D sensor and the color cue is used to segment the person and the background. The use of depth information reduced the complexity and improved accuracy in the segmentation. We have used the novel motion feature along with the GEI of the silhouette and person skeletons for describing various activities. KNN, NN, Naive Bayes Classifier, and SVM are used for activity classification. The dataset used for experimentation is prepared with the help of 11 persons for 10 activities in four illumination conditions. Our study shows that the use of the depth information from Kinect sensor reduces the computational complexity in segmentation and motion feature improves the recognition rate.

Keywords: Activity recognition, Classification, Feature extraction, Motion feature, RGB-D sensor.

1. Introduction

Now a days as large video data is easily available activity recognition is an important research area. Activity recognition has a key role in many applications like video surveillance, content-based video retrieval, patent monitoring, etc. [1]. Automatic understanding of the activity is difficult and challenging. This is because of many reasons such as large variation in performing the activity by an individual, varying background and illumination. The problem becomes worst in the case of crowed and group activity. The standard features to describe a particular activity are not available. The computation complexity is another problem in the implementation of such systems. The use of depth information for foreground-background separation can be a good option. Kinect sensor is a low - cost device for obtaining the depth information; additionally, the skeleton information is available from it [2]. This paper proposes the use of depth information for segmentation and a novel motion feature with the GEI. The database prepared for experimentation comprises ten activities performed by eleven individuals.

2. Background

Microsoft Kinect sensor is used for reliable recognition of construction workers and their activities by Escorcia et al. [3] used colour and depth data [3]. They extracted important visual features from different poses of workers to achieve accurate activity recognition.

Escorcia et al. [3] trained and tested the algorithm by using 80 videos that are taken from 4 workers. Experimental results have shown that the average precision of the method used is 85.28%. Fanello et al. [4] used features based on a 3D Histogram of flow (3DHOF) and Global Histogram of Oriented Gradient (GHOG) [4].

Fanello et al. [4] referred to activity recognition concerning Human-Machine Interaction (HMI) and focused on activities performed by a human. Hidden Markov Model (HMM), Coupled Hidden Semi Markov Model, or action graphs suggested as classifiers in the activity recognition [5-7].

These methods require an expensive offline training phase. The Kinect sensor released in 2010 and then actively used in motion captures and motion analysis applications. Kinect used for pose analysis of construction to classify awkward postures by ray and Teizer. Poppe [8] studied the other line of research that analyses human motion from image and video.

Poppe [8] adopted the hierarchy used by Moeslund et al. [9]. Robertson and Reid had not considered many contexts such as environment [10], interaction between people [11, 12] or objects [13, 14]. Poppe [8] considered only activities and full-body movements and avoid work on gesture recognition [15, 16].

Recently the activity and style recognition are the aim of many approaches [17-19]. Li [21] discussed a cost-effective device that uses a depth sensor along with an RGB camera [20]. The depth image is calculated internally by comparing the spacing of return dots with its values of specific depth.

According to Malima's principal image, the circle base descriptor applies to the depth [22]. The use of a sequence of whole-body silhouette overtime for the covariance matrix approach is given in [23]. The hand silhouette feature vector is also enough in many applications instead of full-body [24]. Kulsheshth et al. [25] used centroid distance Fourier descriptor to perform gesture recognition.

3. Methodology

The proposed algorithm uses Gait Energy Image (GEI) extracted from the human body along with the depth information. Human activity recognition is done with two-stage classification and motion features.

3.1. Segmentation

The Segmentation of human/object from the background is one of the challenges for researchers due to its complexity. The colour and depth information contain complementary information. If both are used together, then the segmentation becomes less complicated and more accurate [1]. The image is converted into binary after pre-processing using the following equation

$$Bi(x,y) = \begin{Bmatrix} 1\ if\ Pi(x,y) > 1 \\ 0\ if\ Pi(x,y), 1 \end{Bmatrix} \tag{1}$$

where *Pi* = pre-processed, *Bi* = binary image

Sum of column and row gives the location of the object (person) in the image as the other points are zero except object (person)

$$Location_{\ person\ Y\ axis} = p\ if\ Sum\ (p)_{horizontal} > T_H \tag{2}$$

The first and last value in the location person Y-axis gives us the start (L_1) and end (L_2) of the object (person) in the image which decides the limit of the bounding box. Similar process can carry out to find the X-axis limits.

$$Location_{\ person\ X\ axis} = p\ if\ Sum\ (p)_{vertical} > T_H \tag{3}$$

This gives the X-axis limits L_3 and L_4 respectively. From this X and, Y-axis limits bounding boxes are applied to separate the object (image) from the background.

$$Width_{bounding\ box\ 1} = L_2 - L_1 \tag{4}$$

$$Width_{bounding\ box\ 2} = L_4 - L_3 \tag{5}$$

3.2. Feature extraction

In this paper, we are considering 4 types of features. From these image features, we get the silhouette bound locations. The features are ‘GEI image with single person’, ‘GEI image with two persons’, ‘GEI image with single person skeleton’, and ‘GEI image with the two-person skeleton.

3.2.1. GEI (Gait Energy Image)

Gait Energy Image (GEI) is obtained from the silhouette of the object (person) in all the images of the video [26].

3.2.2. Motion features

Motion features are calculated from 10% of the initial and final frames. The difference between the average of the initial and final frame gives the motion feature. This is given by the equation:

$$D = \left(\sum_{j=ly-n}^{ly} Po(x)\right)\Big/ n - \left(\sum_{i=1}^{n} Po(x)\right)\Big/ n \tag{6}$$

where $Po(x)$ = Position of the image on X-axis, lv = Length of the video file and n = 10% of the length of the video file

$$D = \begin{cases} +1 \; if \; D > n \\ -1 \; if \; D \leq n \; (or) \\ 0 \; Elsewhere \end{cases} \tag{7}$$

Case 1: If D = +1, then the person is moving in a forward direction.

Case 2: If D = -1, the person is moving in a backward direction.

Case3: If D = 0 then the person is standing constant and not moving to any direction

Classification uses the motion feature along with the GEI of the silhouette and GEI of the skeleton.

3.3. Classification

A novel two-stage classification method is used to recognize the human activity. Figure 1 shows the classification flow.

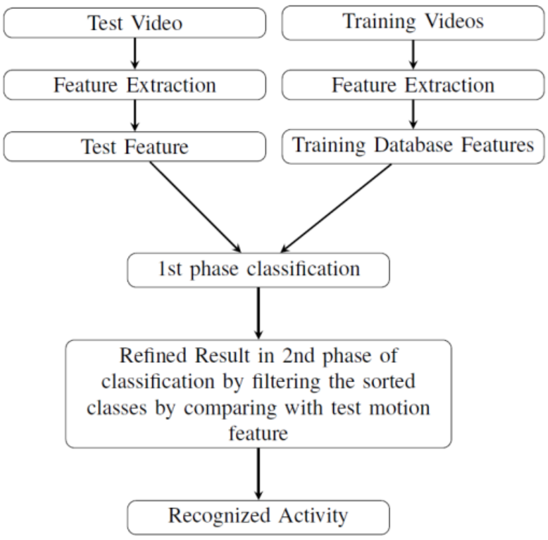


Fig. 1. A novel two-stage classification.

After reducing the feature dimension, the extracted features are fed to the different classifiers by using either Principal Component Analysis (PCA) or Linear Discriminate Analysis (LDA). The use of PCA and LDA reduces the feature dimension, in turn, the computational complexity. The following classifiers are used to recognize human activity in the first phase [27-30].

i) K-Nearest Neighbour (KNN)

ii) Support Vector Machine (SVM)

iii) Neural Network (NN)

iv) Naive Bays (NB) Classifier .

In second stage of classification the sorted class in first phase filtered using the motion feature. In activities like approach and depart the motion is exact opposite so using this motion feature the previous classification done by first phase classifier is redefined. In some activities like handshake there is no motion so motion feature is 0 which help in redefining the classification in second stage.

4. Experimental Results

4.1. Dataset

The experimentation is carried out on a dataset, which has ten possible activities (approach, depart, punch, handshake, push, kick, pat, point, lift, salaam). All these activities are recorded by using the Microsoft Kinect sensor, which provides colour and depth information. These activities are performed by eleven subjects under four illumination conditions. The four illumination conditions produced by controlling the light source and curtains on the window. The details are shown in Table 1. The dataset consists of a total of one hundred and ten videos.

Table 1. Illumination conditions.

Illumination	Light Source 1	Light source 2	Curtains
L0	ON	OFF	Closed
L1	ON	ON	Closed
L2	OFF	OFF	Opened
L3	OFF	OFF	Opened

4.2. Training

A fully supervised approach is used for training using all the four classifiers namely KNN, NN, SVM, and Naive Bayes classifier. Training is done using four different feature sets of seven subjects.

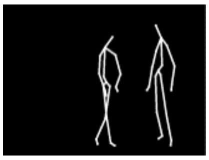
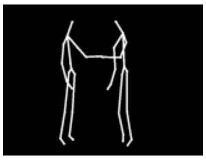
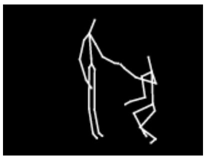
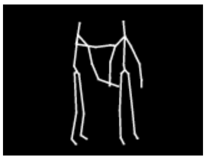
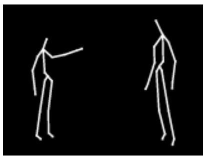
4.3. Activity recognition

Of the ten activities, it is challenging to discriminate between handshake and punch, pat, and point based on GEI features due to no or small amount of inter-class similarity. In this work, four sets of features used are with and without motion feature and four classifiers along with two subspace representation techniques.

Table 2 shows the dataset images for various activities. In Table 3, segmentation results in different illumination conditions are given. From these results, it can be observed that segmentation has no effect of varying illumination, because we have used the depth information. Table 4 shows the silhouette extracted for single and two persons.

Figures 2 to 5 show the Confusion matrix for K Nearest neighbour, Neural Network, Support Vector Machine, and Naive Bayes classifier, respectively.

Table 2. Dataset images.

Action Type	Color	Depth	Skeleton
Approach			
Depart			
Handshake			
Kick			
Lifting			
Patting			
Pointing			

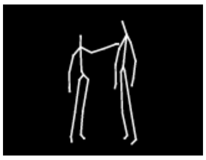
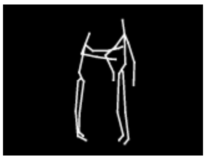
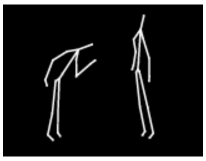
Punch			
Push			
Salam			

Table 3. Images under illumination conditions.

Illuminatio n Condition	Color	Depth	Silhouette
L0			
L1			
L2			
L3			

Table 4. Single and two-person colour and silhouette images.

Action Type	Col_Two_person	Sil_Two_person	Sil_Single_person
Approach			
Handshake			
Patting			
Pointing			

Figures 6 to 9 give an overall recognition rate for all the four classifiers with three data representation sets and with and without motion features. The four feature sets used in classification are

1) GEI of acting person silhouette

2) GEI of all person silhouette

3) GEI of acting person skeleton and

4) GEI of all person skeleton.

These features are used with and without motion features for all the four classifiers. From the results, it is observed that the use of motion features increases the recognition rate in the case of all the classifiers using any feature set. It indicates the importance of understanding the motion characteristics in activity recognition.

The recognition rate of the KNN classifier for all person silhouette features with LDA is 90%. The same can be obtained using SVM and the original feature set. The confusion matrix is also obtained for all the classifiers which are an essential measure in evaluating the performance of the algorithm. In the case of KNN classifier 100% recognition is obtained for activities approach, depart, kick, lift, push, and salaam whereas handshake and punch, lift and pat, punch and push, point, and salaam sometime get misclassified. Neural network classifier has a 100% recognition ratio for the activities approach, depart, kick, lift, push, and salaam whereas this classifier fails in classifying punch, pat and push. Also, it had confusion in understanding handshake, punch and push.

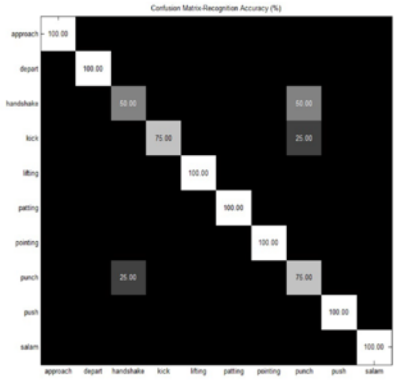


Fig. 2. Confusion Matrix of KNN.

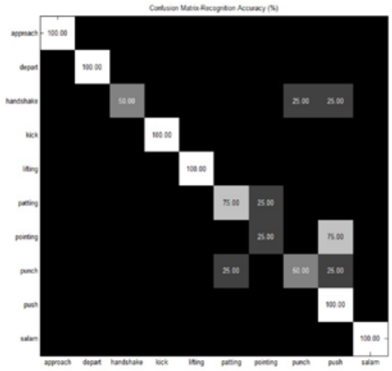


Fig. 3. Confusion Matrix of KNN.

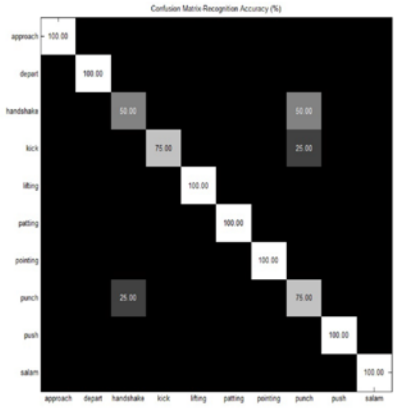


Fig. 4. Confusion Matrix of SVM.

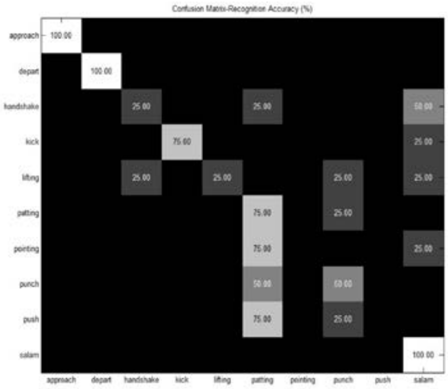


Fig. 5. Confusion Matrix of NB.

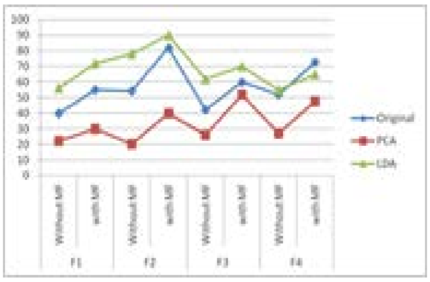


Fig. 6. Recognition Rate KNN.

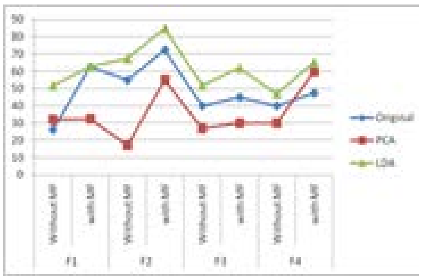


Fig. 7. Recognition Rate NN.

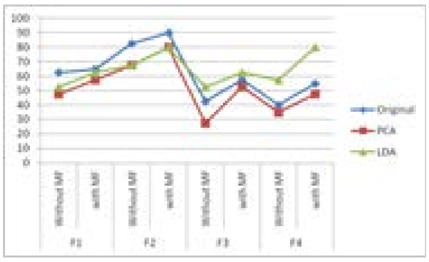


Fig. 8. Recognition Rate SVM.

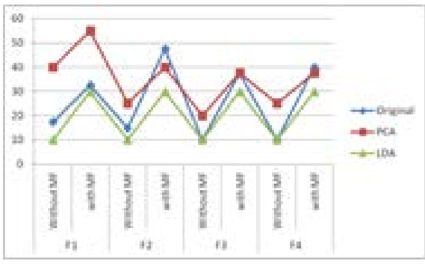


Fig. 9. Recognition Rate NB.

5. Discussion

The recognition rate obtained using different classifier and set of feature vectors presented in Table 5.

Table 5. Recognition ratio of different classifiers.

Classifier	Data Compression	Recognition Ratio: Acting Person Silhouette		All Persons Silhouette		Acting Person Skeleton		All Person Skeleton	
		without motion	With Motion	without motion	with motion	without motion	with motion	without motion	With Motion
KNN	Original	40	55	54	82	42	60	52	72.5
	PCA	22	30	20	40	26	52	27	47.5
	LDA	56	72	78	90	62	70	55	65
NN	Original	26	62.5	55	72.5	40	45	40	47.5
	PCA	32	32.5	17	55	27	30	30	60
	LDA	52	62.5	67.5	85	52	62	47	65
SVM	Original	62.5	65	82.5	90	42.5	57.5	40	55
	PCA	47.5	57.5	67.5	80	27.5	52.5	35	47.5
	LDA	52.5	62.5	67.5	80	52.5	62.5	57.5	80
NB	Original	17.5	32.5	15	47.5	10	37.5	10	40
	PCA	40	55	25	40	20	37.5	25	37.5
	LDA	10	30	10	30	10	30	10	30

From the results, it is observed that use of motion feature increases the recognition rate in case of all the classifiers using any of features set. It indicates the importance of understanding the motion characteristics in activity recognition.

The recognition rate of KNN classifier for all person silhouette features with LDA, is 90%. The same can be obtained using SVM and original feature set. By our experiments, we have found that using motion features we can boost the performance of the system.

We have also checked the effectiveness of the motion feature with four classifiers and found that it can boost the performance of all the classifiers. Also due to a use of depth data for segmentation we have found that activity recognition process becomes independent of illumination.

Apart from the classifiers experiment we have also performed experiments with four features type and found that it can boost the performance regardless of features and classifiers.

By our experiments, we have found that using motion feature we can boost the performance of the system. We have also checked the effectiveness of the motion feature with four classifiers and found that it can boost the performance of all the classifiers. Also, due to the use of depth data for segmentation, we have found that the activity recognition process becomes independent of illumination. Apart from the classifiers experiment, we have also performed experiments with four features types and found that it can boost the performance regardless of features and classifiers.

6. Conclusion

The use of depth information with colour reduces the computational complexity in segmentation as continuous updating of background is not required. This makes segmentation accurate in varying illumination condition which makes the algorithm robust. From observation, it can be seen that the use of motion features increases the recognition rate in the case of all the classifiers. So, we can say that motion features are helpful in activity recognition along with the traditional feature set. It is also observed that there is an increase in recognition rate due to two-phase classification and the support vector machine seems to be a good classifier in activity recognition. LDA and PCA help in reducing the dimensionality as well as computational complexity and improve the recognition rate. It is observed that the recognition rate is more in case we use LDA to represent the extracted feature data as compared to the original data-set.

7. Future Scope and Limitations

The use of RGBD sensor limits the use to indoor applications which is the major limitation of the work presented in the paper. This work can be extended with modification to recognize the complex activities specially the crowd analysis. It is useful in security applications at many places like Air Port Railway Stations etc. Useful in the elderly person monitoring who are alone at home.

Nomenclatures

Bi	Binary image
D	Motion feature value
L	Location of the person (object)
Ly	Length of video
N	10 % of the length of the video
Pi	Pre-processed image
Po(x)	Position of the image on the x-axis

Abbreviations

3DHOF	3D Histogram of Flow
CCTV	Closed Circuit Television

GEI	Gait Energy Image
GHOG	Global Histogram of Oriented Gradient
HMI	Human Machine Interaction
HMM	Hidden Markov Model
KNN	K Nearest Neighbour
LDA	Linear Discriminant Analysis
NB	Naive Bays
NN	Neural Network
PCA	Principal Component Analysis
RGB	Red Green Blue
RGB-D	Red Green Blue Depth
SVM	Support Vector Machine

References

1. Han, S.; Achar, M.; Lee, S.; and Pena-Mora, F. (2013). Empirical assessment of an RGB-D sensor on motion capture and action recognition for construction worker monitoring, *Visualization in Engineering*, 1, 1-13.
2. Sardeshmukh, M.M.; Kolte, M.T.; and Chaudahri, D.S. (2013). Activity recognition using multiple features, subspaces and classifiers, proceedings of swarm, evolutionary, and memetic computing. *Springer International Publishing*, 617- 624.
3. Escorcia, V.; Davila, M.A.; Golparvar-Fard, M.; and Niebles, J.C. (2012). Automated vision-based recognition of construction worker actions for building interior construction operations using RGBD cameras, *American Society of Civil Engineer*, 879-888.
4. Fanello, S.R.; Gori, I.; Metta, G.; and Odone, F. (2013). One-shot learning for real-time action recognition, *Pattern Recognition and Image Analysis.* Springer Berlin Heidelberg, 31-40.
5. Yamato, J.; Ohya, J.; and Ishii, K. (1992). Recognizing human action in time-sequential images using hidden Markov model. *Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Champaign, IL, USA.
6. Natarajan, P.; and Nevatia, R. (2007). Coupled hidden semi Markov models for activity recognition, Motion and Video Computing, *WMVC. 2007 IEEE Workshop on Motion and Video Computing*. Austin, TX, USA.
7. Li, W.; Zhang, Z.; and Liu, Z. (2010). Action recognition based on a bag of 3d points. 2010 IEEE *Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. San Francisco, CA, USA.
8. Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6), 976-990.
9. Moeslund, T.B.; Hilton, A.; and Kruger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding* , 104(2-3), 90-126.
10. Robertson, N.; and Reid, I. (2006). A general method for human activity recognition in the video. *Computer Vision and Image Understanding*, 104(2), 232-248.

11. Park, S.; and Trivedi, M.M. (2008). Understanding human interactions with track and body synergies (TBS) captured from multiple views. *Computer Vision and Image Understanding*, 111(1), 2-20.
12. Ryoo, M.S.; and Aggarwal, J.K. (2009). Semantic representation and recognition of continued and recursive human activities. *International Journal of Computer Vision*, 82 (1), 1-24.
13. Gupta, A.; Kembhavi, A.; and Davis, L.S. (2009). Observing human object interactions: using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10), 1775-1789.
14. Moore, D.J.; Essa, I.A.; and Hayes, M.H. (1999). Exploiting human actions and object context for recognition tasks. *Proceedings of the International Conference on Computer Vision*. Kerkyra, Greece.
15. Erol, A.; Bebis, G.; Nicolescu, M.; Boyle, R.D.; and Twombly, X. (2007). Vision-based hand pose estimation: a review. *Computer Vision and Image Understanding*, 108(12), 52-73.
16. Mitra, S.; and Acharya, T. (2007). Gesture recognition: a survey. *IEEE Transactions on Systems, Man, and Cybernetics*, *Part C: Applications and Reviews*, 37(3), 311-324.
17. Cuzzolin, F. (2006). Using bilinear models for view-invariant action and identity recognition. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. New York, USA.
18. Elgammal, A.; and Lee, C.S. (2004). Separating style and content on a nonlinear manifold. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Washington, DC, USA.
19. Wang, J.M.; Fleet, D.J.; and Hertzmann, A. (2007). Multifactor Gaussian process models for style-content separation. *Proceedings of 24th International Conference on Machine Learning*. 975-982.
20. Li, Y. (2012). Hand gesture recognition using Kinect. *2012 IEEE International Conference on Computer Science and Automation Engineering.* Beijing, China, 196-199.
21. Han, J.; Shao, L.; Xu, D.; and Shotton, J. (2013). Enhanced computer vision with Microsoft Kinect sensor: A review. *IEEE Transactions on Cybernetics,* 43(5), 1318-1334.
22. Malima.; Ozgur.; and Cetin. (2006). A fast algorithm for vision-based hand gesture recognition for robot control. *2006 IEEE 14th Signal Processing and Communications Applications.* Antalya, Turki, 1-4.
23. Guo, K.; Ishwar, P.; and Konrad, J. (2009). Action recognition in video by covariance matching of silhouette tunnels. *2009 XXII Brazilian Symposium on Computer Graphics and Image Processing*. Rio De Janiero, Brazil, 299-306.
24. Zhang, D.; and Lu, G. (2002). A Comparative study of Fourier descriptors for shape representation and retrieval. *Proceeding of 5th Asian Conference on Computer Vision.* 646-651.
25. Kulshreshth, A.; Zorn, C.; and Laviola, J.J. (2013). Poster: Real-time markerless Kinect based finger tracking and hand gesture recognition for HCI. *2013 IEEE Symposium on 3D User Interfaces* (3DUI). Orlando, USA, 187-188.

26. Ali, H.; Dargham, J.; Chekima, A.; and Moung, E. (2011). Gait recognition using gait energy image. *International Journal of Signal Processing, Image Processing, and Pattern Recognition*, 4, 141-152.
27. Arseneau, S.; and Cooperstock, J.R. (1999). Real-time image segmentation for action recognition. *1999 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM 1999). Conference Proceedings (Cat. No.99CH36368).* Victoria, Canada, 86-89.
28. Jieping Ye, Shuiwang Ji (2010). Discriminant analysis for dimensionality reduction: an overview of recent developments. *Biometrics: Theory, Methods, and Applications*. John Wiley and Sons, Inc.
29. Nanda, K.; and Leen, T. (1997). Dimension reduction by local principal component analysis. *Neural Computation*, 9(7), 1493-1516.
30. Draper, B.A.; Baek, K.; Bartlett, M.S.; and Beveridge, J.R. (2003). Recognizing faces with PCA and ICA. *Computer Vision and Image Understanding,* 91(1-2), 115-137.