World Applied Sciences Journal 35 (8): 1325-1330, 2017
ISSN 1818-4952
© IDOSI Publications, 2017
DOI: 10.5829/idosi.wasj.2017.1325.1330

# Adaptive Log-Covariance Quaternions-Based Human Action Recognition System Using RGB-D Sensor

*[1]Huong Yong Ting, [2]Kok Swee Sim and [2]Fazly Salleh Abas*

[1]School of Computing, University College of Technology Sarawak,
868 Persiaran Brooke, 96000 Sibu, Sarawak, Malaysia
[2]Faculty of Engineering and Technology, Multimedia University,
Jalan Ayer Keroh Lama, 75450 Bukit Beruang, Melaka, Malaysia

**Abstract:** In this paper, we extended our previous non-adaptive human action recognition framework by incorporating an adaptive model in order to exhibit minimal supervision system where the intelligent system will handle the process of human action detection, labelling, training and recognition in automated manner. Action descriptor that is formed in non-adaptive model framework is compared with the stored action descriptors in order to determine a new human action. Then, the collected new action descriptors are clustered using agglomerative hierarchical clustering algorithm. Subsequently, all the clustered action descriptors are assigned with new labels and the multi-class SVM classifier is updated automatically. Additionally, we evaluated the adaptive model on public depth dataset namely MSR-Action 3D dataset. The experimental results reveal that the proposed adaptive framework is comparable with state of the arts methods which were developed in supervised manner. In future, conventional RGB information can be fused with the depth map data to produce algorithms with better recognition accuracy and robustness for both frameworks.

**Key words:** Kinect · Database · Agglomerative Hierarchical Clustering · Action Classification

## INTRODUCTION

Vision-based human action recognition is a computerized method to label image sequences with action labels automatically. The domain has been actively researched and applied in numerous real-world application, such as video surveillance systems, video analysis, content-based video retrieval, human-computer interaction, health care, etc. Basically, vision-based human action recognition consists of four main stages, namely model initialization, tracking, pose estimation and action classification (Figure 1) [1]. Nonetheless, achieving high recognition accuracy is still a challenging task in this domain. There are two major factors which contribute to the human recognition accuracy. The first factor is the input video sensor while the second factor is the human action modelling which may be ambiguous and dynamic.

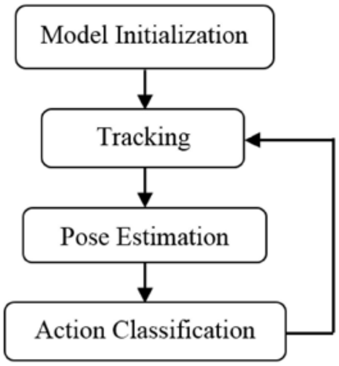


Fig. 1: Four main stages in vision-based human action recognition.

Several researches have been conducted on human action recognition using wearable sensors [2-3]. However, wearable sensors might result in uncomfortableness to the subject and is less preferable as compared to vision-based. A comprehensive survey regarding vision-based

**Corresponding Author:** Huong Yong Ting, School of Computing, University College of Technology Sarawak, 868 Persiaran Brooke, 96000 Sibu, Sarawak, Malaysia.

human action recognition has been covered by Aggarwal and Ryoo [4] where most of the human action recognition methods only adopted human body movement in 2-dimensional spatial (*x-y*) and temporal (*t*) information due to the high cost and low availability of depth sensors. In this case, RGB camera is usually used to capture human motion without depth information and thus leads to discriminative performance degradation. In the real world, human actions are of 4-dimensional, *x-y-z-t*.

Due to the emergence of inexpensive, reliable and robust algorithms to capture the depth information, human action recognition using RGB-Depth sensor, such as Microsoft Kinect sensor is becoming more prevalent. Microsoft Kinect sensor, originally released with the intention to improve human computer interaction in gaming for the Xbox 360 game console. Despite being targeted mainly for the entertainment market, the sensor has gained enormous interests within the vision and robotics researchers for its broad applications [5]. With such sensor, the human action recognition task becomes relative easier as compared to the RGB imagery. Furthermore, a better view invariance and faster speed performance can be expected. Besides, supervised classifiers are inferring a decision rule from labelled training data. These approaches have demonstrated promising results [6]. However, acquiring of various and sufficient labelled data for general human actions might be difficult and computationally expensive [7]. On the other hand, unsupervised classification methods are of particular interest to analyze large amounts of unlabeled data.

In vision-based unsupervised human action recognition, Xiang and Gong [8] proposed a normalized affinity matrix to detect human behavior and abnormality without interference from human being. Moreover, Niebles *et al.* [9] demonstrated unsupervised learning for human action recognition using probabilistic Latent Semantic Analysis and Latent Dirichlet Allocation methods. In depth imagery context, however, there are only a few researchers explored unsupervised method. Chen *et al.* [10] randomly sample the sub volumes from RGB and depth video data as input to the bottom of Independent Subspace Analysis (ISA). After that, the learned ISA at the bottom layer is transferred to the top ISA layer. Eventually, the actions from different modalities are classified by the multi-class multi kernel learning algorithm. Besides, Ong *et al.* [11] utilize K-means method to cluster unlabeled features data from skeleton model and recognize the actions subsequently.

In our previous work [12], we developed a supervised log-covariance quaternions-based human action recognition system. Microsoft Kinect sensor is employed as input sensor to provide depth map sequences. The orientation of the bone is extracted and the human action is described by using log-covariance quaternions matrix. Subsequently, multi-class support vector machine (SVM) is used to train and classify human action. In this paper, our previous method is incorporated with an adaptive model to achieve minimal supervision human action recognition system. This hybrid model will enable an intelligent system to handle the process of human action detection, labelling, training and recognition in automated manner. In this research, a public depth map dataset, namely MSR-Action3D dataset [13] is utilized in order to evaluate effectiveness of the proposed method.

**Proposed Methodl:** Log-covariance quaternions action recognition framework which is initially developed in supervised classification manner is fused with an adaptive model in order to demonstrate intelligence capability to detect, label, train and recognize human action automatically. Figure 2 exhibits the overview of adaptive log-covariance quaternions action recognition framework.

**Detection of New Action:** In adaptive model, database plays an important role. The database is served as a storage for log-covariance quaternions action descriptors. The mean of log-covariance quaternions action descriptor is stored in the database as shown in Equation 1.

$$\overline{C_{\log}} = \frac{1}{M}\sum_{m=1}^{M} C_{\log}, m \qquad (1)$$

where $M$ is total number of quaternions action descriptor for a single action.

The newly formed of log-covariance quaternions action descriptor is compared with the stored action descriptors using Manhattan distance function as exhibited in Equation 2.

$$D = \sum_{i=1}^{J} | C_i - S_i | \qquad (2)$$

where $R$ is the stored log-covariance quaternions action descriptor, $S$ is the incoming log-covariance quaternions action descriptor and $J$ is the total elements in the vector. A threshold value, $T$ is employed in order to detect a new action. If the incoming log-covariance quaternions action

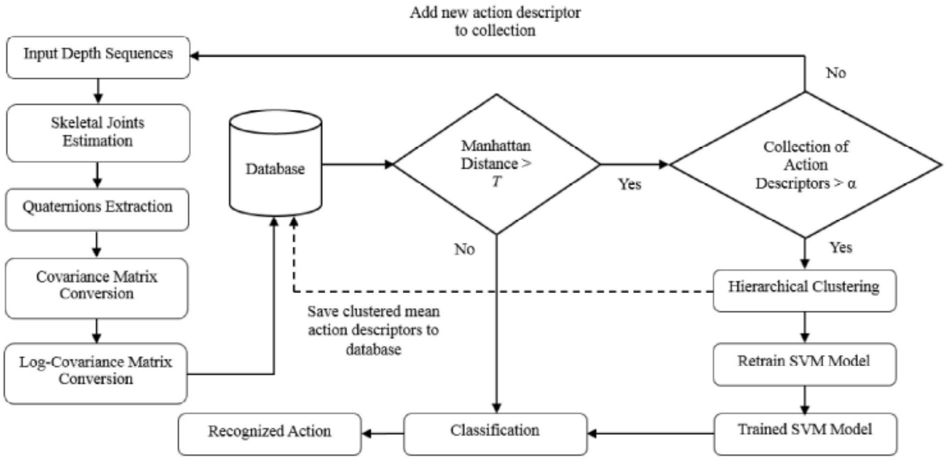

Add new action descriptor
to collection
Input Depth Sequences
No
Skeletal Joints Estimation
Database
Manhattan Distance > T
Yes
Collection of Action Descriptors > α
Quaternions Extraction
No
Yes
Covariance Matrix Conversion
Hierarchical Clustering
Save clustered mean action descriptors to database
Log-Covariance Matrix Conversion
Retrain SVM Model
Recognized Action
Classification
Trained SVM Model

Fig. 2: Overview of adaptive log-covariance quaternions action recognition framework.

descriptor produces a threshold value which is more than *T*, the algorithm will proceed to examine the amount of action descriptors in action descriptor collection pool. In this research, the *T* value was set as 250 after obtaining the minimum intra-class variation value from same action that performed by ten different individuals. The action descriptor collection pool functions as a container to temporary store the action descriptors and serves as an input to the clustering algorithm when the collection pool exceeds limit *α*.

**Hierarchical Clustering:** Agglomerative hierarchical clustering [14] method is adopted to group similar action descriptors in this research. Generally, hierarchical clustering depends on linkage criterion to group objects into clusters. The commonly used linkage criteria between two sets of observations *A* and *B* are [15]:

Single linkage clustering (Equation 3),

$$D(A,B) = \min_{\alpha \in A, b \in B} d(a,b) \tag{3}$$

Complete linkage clustering (Equation 4) and

$$D(A,B) = \max_{\alpha \in A, b \in B} d(a,b) \tag{4}$$

Average linkage clustering (Equation 5)

$$D(A,B) = \frac{1}{|A|\,B|} \sum_{a \in A} \sum_{b \in B} d(a,b) \tag{5}$$

where *d*(*a*, *b*) is the distance between two elements *a* and *b*.

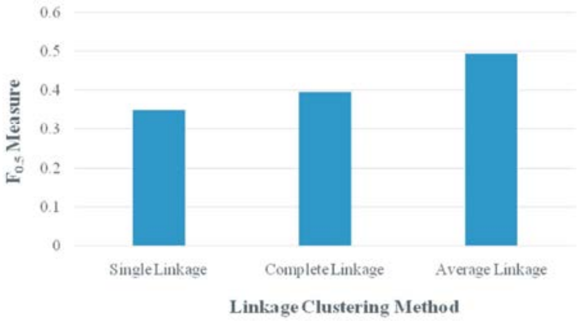


Fig. 3: Comparison of different linkage methods.

In this research, we tested all three linkage clustering methods and found that Average linkage clustering performs the best as compared to other clustering methods. Figure 3 shows a comparison of different linkage methods. In the testing, three actions from MSR-Action3D dataset [13] which are *draw circle*, *draw tick* and *draw x* were chosen. These actions were picked due to the small inter-class variation value. We cut in the hierarchy that yielded three clusters and examine the clustering accuracy. The mean of precision and recall measures were generated from three actions (five samples from different individuals for each action) in order to compute $F_{0.5}$ measure.

## Experimental Results and Discussions

**Experimental Setup:** The MSR-Action3D dataset is a public action dataset which was acquired by a depth camera. Generally, this dataset consists of 20 actions performed by ten subjects where each action was repeated two or three times. The 20 actions are: *high arm wave*, *horizontal arm wave*, *hammer*, *hand catch*,

Table 1: The three action subsets employed in our experiments.

| Action Set 1 (AS1) | Action Set 2 (AS2) | Action Set 3 (AS3) |
|---|---|---|
| Horizontal wave | High wave | High throw |
| Hammer | Hand catch | Forward kick |
| Forward punch | Draw x | Side kick |
| High throw | Draw tick | Jogging |
| Hand clap | Draw circle | Tennis swing |
| Bend | Hand wave | Tennis serve |
| Tennis serve | Forward kick | Golf swing |
| Pickup and throw | Side boxing | Pickup and throw |

*forward punch*, *high throw*, *draw x*, *draw tick*, *draw circle*, *hand clap*, *two hand wave*, *side-boxing*, *pick and throw*, *bend*, *forward kick*, *side kick*, *jogging*, *tennis swing*, *tennis serve* and *golf swing*. These actions involve numerous movement of arms, legs, torso and combinations of all. The spatial resolution of the dataset is 640 × 480 and the total depth map frames is 23797 for 402 action samples. In order to perform a fair comparison, we followed the cross-subject (CrSub) test setup [13] where the total samples were divided into half where half of the subjects were used as training data while the rest of the subjects were used as testing data. Table 1 exhibits the same experimental setting to split 20 actions into three subsets.

Below is the experimental setup for the proposed adaptive model framework:

- The system was preloaded with three actions (*walking*, *jumping* and *boxing*) so that system can perform comparison in order to detect new action.
- Training subjects in each action subset was clustered by using hierarchical clustering algorithm.
- Hierarchy was cut for eight clusters for each action subset.
- Multi-class SVM classifier was retrained for each action subset and used to perform action recognition on testing subjects AS1, AS2 and AS3 respectively.
- Evaluation will be done based on clustering and recognition accuracies on each action subset.

Table 2: $F_{0.5}$ Scores for three activities subset after clustering.

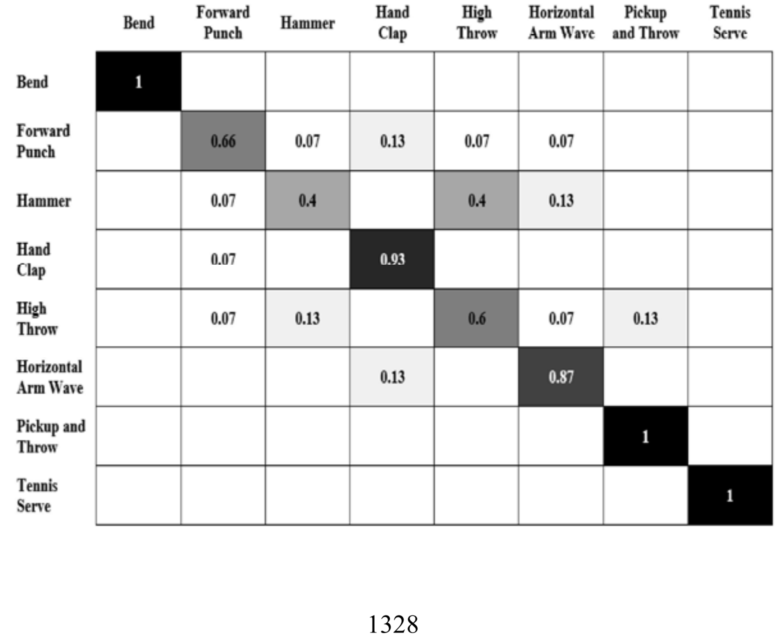
| Action Subset | $F_{0.5}$ Score |
|---|---|
| AS1 | 0.8779 |
| AS2 | 0.7473 |
| AS3 | 0.9785 |

Table 3: Comparison of the proposed adaptive framework with state of the art methods.

| Method | Average CrSub Test Recognition Accuracy (%) |
|---|---|
| Wang *et al*. [16] | 86.50 |
| Vieira *et al*. [17] | 84.80 |
| Yang and Tian [18] | 82.33 |
| The Proposed Adaptive Framework | 81.88 |
| Miranda *et al*. [19] | 80.30 |
| Xia *et al*. [20] | 78.97 |
| Li *et al*. [13] | 74.70 |

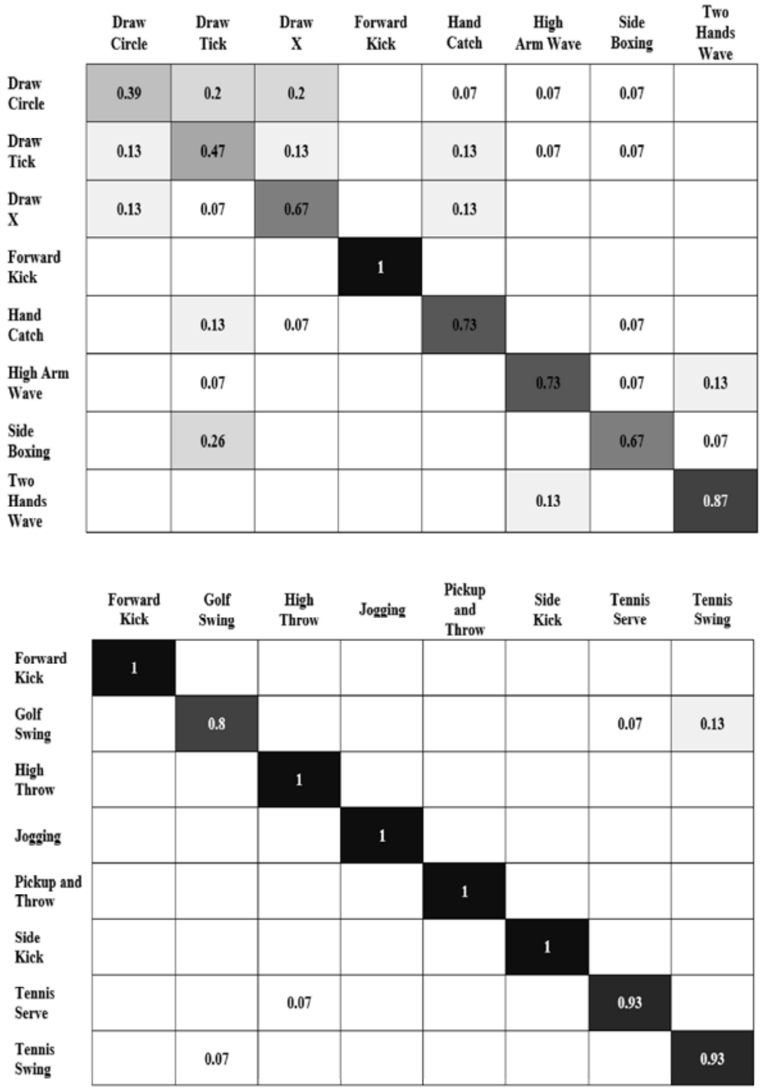
## RESULTS AND DISCUSSION

Table 2 exhibits the clustering accuracies ($F_{0.5}$ score) for three subset actions in adaptive model framework. From Table 2, the $F_{0.5}$ score is relatively lower in AS2 due to the inter-class action variations between *draw circle*, *draw tick* and *draw x* actions are not significant enough. Figure 4 presents the confusion matrices of the adaptive model under cross subject test after the multi-class SVM



| | Bend | Forward Punch | Hammer | Hand Clap | High Throw | Horizontal Arm Wave | Pickup and Throw | Tennis Serve |
|---|---|---|---|---|---|---|---|---|
| Bend | 1 | | | | | | | |
| Forward Punch | | 0.66 | 0.07 | 0.13 | 0.07 | 0.07 | | |
| Hammer | | 0.07 | 0.4 | | 0.4 | 0.13 | | |
| Hand Clap | | 0.07 | | 0.93 | | | | |
| High Throw | | 0.07 | 0.13 | | 0.6 | 0.07 | 0.13 | |
| Horizontal Arm Wave | | | | 0.13 | | 0.87 | | |
| Pickup and Throw | | | | | | | 1 | |
| Tennis Serve | | | | | | | | 1 |

(a)



(b)

| | Draw Circle | Draw Tick | Draw X | Forward Kick | Hand Catch | High Arm Wave | Side Boxing | Two Hands Wave |
|---|---|---|---|---|---|---|---|---|
| Draw Circle | 0.39 | 0.2 | 0.2 | | 0.07 | 0.07 | 0.07 | |
| Draw Tick | 0.13 | 0.47 | 0.13 | | 0.13 | 0.07 | 0.07 | |
| Draw X | 0.13 | 0.07 | 0.67 | | 0.13 | | | |
| Forward Kick | | | | 1 | | | | |
| Hand Catch | | 0.13 | 0.07 | | 0.73 | | 0.07 | |
| High Arm Wave | | 0.07 | | | | 0.73 | 0.07 | 0.13 |
| Side Boxing | | 0.26 | | | | | 0.67 | 0.07 |
| Two Hands Wave | | | | | | 0.13 | | 0.87 |

(c)

| | Forward Kick | Golf Swing | High Throw | Jogging | Pickup and Throw | Side Kick | Tennis Serve | Tennis Swing |
|---|---|---|---|---|---|---|---|---|
| Forward Kick | 1 | | | | | | | |
| Golf Swing | | 0.8 | | | | | 0.07 | 0.13 |
| High Throw | | | 1 | | | | | |
| Jogging | | | | 1 | | | | |
| Pickup and Throw | | | | | 1 | | | |
| Side Kick | | | | | | 1 | | |
| Tennis Serve | | | 0.07 | | | | 0.93 | |
| Tennis Swing | | 0.07 | | | | | | 0.93 |

Fig. 4: Recognition accuracy confusion matrices for the adaptive human action recognition framework. (a) Confusion matrix for AS1CrSub test, (b) confusion matrix for AS2CrSub test and (c) confusion matrix for AS3CrSub test.

classifier was retrained using the clustered inputs for each action subset. In Figure 4(a) and (b), the recognition rates for AS1CrSub and AS2CrSub are significantly lower as compared AS3CrSub (Figure 4(c)). This is mainly due to the errors of clustering and labelling in clustering stage. Besides, Table 3 summarizes the comparison results. The average recognition rate from the proposed adaptive model is comparable with other state of the art methods.

## CONCLUSIONS

In this paper, we have proposed an adaptive model framework based on log-covariance quaternions-based action recognition framework, which is also known as non-adaptive model framework. Basically, the fusion of both models will promote an intelligent system to handle the process of human action detection, labelling, training and recognition in automated manner. Database in adaptive model is employed to store human action descriptors in order to detect a new human action. Subsequently, the collected new action descriptors are clustered using agglomerative hierarchical clustering algorithm. The clustered action descriptors are assigned with labels and the multi-class SVM classifier is updated automatically. Besides, the experimental results on the MSR-Action 3D dataset exhibit the proposed adaptive model framework is comparable with state of the arts methods which were developed in supervised manner. In future, conventional RGB information can be fused with the depth map data to produce algorithms with better recognition accuracy and robustness for both frameworks.

## REFERENCES

1. Moeslund, T.B., A. Hilton and V. Krueger, 2006. A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding, 104: 90-126.
2. Stickin, M., K. Van Laerhoven and B. Schiele, 2008. Exploring semi-supervised and active learning for activity recognition. In the Proceeding of International Symposium on Wearable Computers, pp: 81-88.
3. Huynh, T., M. Firtz andB. Schiele, 2008. Discovery of activity patterns using topic models. In the Proceeding of the 10th International Conference on Ubiquitous Computing, pp: 10-19.
4. Aggarwal, J.K. and M.S. Ryoo, 2010. Human activity analysis: a review. ACM Computing Survey, pp: 43.
5. Giles, J., 2010. Inside the race to hack the Kinect. The New Scientist, 208: 22-23.
6. Altun, K., B. Barshan and O. Tuncel, 2010. Comparative study on classifying human activities with miniature inertial and magnetic sensors. Pattern Recognition, 43: 3605-3620.
7. Cvetkovic, B., M. Lustrek and B. Kaluza, 2011. Gams M. Semi-supervised learning for adaptation of human activity recognition classifier to the user. In the Proceeding of International Joint Conference on Artificial Intelligence, pp: 24-29.
8. Xiang, T. and S.G. Gong, 2005. Video behaviour profiling and abnormality detection without manual labelling. In the Proceeding of IEEE International Conference of Computer Vision, pp: 1238-1245.
9. Niebles, J.C., H.C. Wang and L. Fei-Fei, 2008. Unsupervised learning of human action categories using spatio-temporal words. International Journal of Computer Vision, 79: 299-318.
10. Chen, G., F. Zhang, M. Giuliani, C. Buckl and A. Knoll, 2013. Unsupervised learning spatio-temporal features for human activity recognition from RGB-D video data. Lecture Notes in Computer Science, 8239: 341-350.
11. Ong, W.H., T. Koseki and L. Palafox, 2013. Unsupervised human activity detection with skeleton data from RGB-D Sensor. In the Proceeding of 2013 5th International Conference on Computational Intelligence, Communication Systems and Networks, pp: 30-35.
12. Ting, H.Y., K.S. Sim, F.S. Abas and R. Besar, 2014. Vision-based human gesture recognition using Kinect sensor. Lecture Notes in Electrical Engineering, 291: 239-244.
13. Li, W., Z.Y. Zhang and Z.C. Liu, 2010. Action recognition based on a bag of 3D points. In the Proceeding of Computer Vision and Pattern Recognition Workshops, pp: 9-14.
14. Johnson, S.C., 1967. Hierarchical clustering schemes. Psychometrika, 2: 241-254.
15. Szekely, G.J. and M.L. Rizzo, 2005. Hierarchical clustering via joint between-within distances: extending ward's minimum variance method. Journal of Classification, 22: 151-183.
16. Wang, J., Z. Liu, J. Chorowski, Z. Chen and Y. Wu, 2012. Robust 3D action recognition with random occupancy patterns. In the Proceeding of 12th European Conference on Computer Vision, pp: 872-885.
17. Vieira, A.W., E.R. Nascimento, G.L. Oliveira, Z. Liu and M. Campos, 2012. Stop: space-time occupancy patterns for 3D action recognition from depth map sequences. In the Proceeding of 17th Iberaamerica Congress on Pattern Recognition, pp: 252-259.
18. Yang, X. and Y. Tian, 2012. Eigenjoints-based action recognition using Naïve-Bayes-Nearest-Neighbour. In the Proceeding of Computer Vision and Pattern Recognition Workshops, pp: 14-19.
19. Miranda, L., T. Vieira, D. Martinez, T. Lewiner, A.W. Vieira and M.F.M. Campos, 2012. Real-time gesture recognition from depth data through key poses learning and decision forests. In the Proceeding of 25th SIBGRAPI Conference on Graphics, Patterns and Images, pp: 1182-1188.
20. Xia, L., C.C. Chen and J.K. Aggarwal, 2012. View invariant human action recognition using histograms of 3D joints. Proceeding of the 2nd International Workshop on Human Activity Understanding from 3D Data in conjunction with IEEE CVPR, pp: 20-27.