IEIE Transactions on Smart Processing and Computing, vol. 4, no. 4, August 2015
http://dx.doi.org/10.5573/IEIESPC.2015.4.4.281

A Survey of Human Action Recognition Approaches that use an RGB-D Sensor

Adnan Farooq and Chee Sun Won*

Department of Electronics and Electrical Engineering, Dongguk University -Seoul, South Korea
{aadnan, cswon}@dongguk.edu

***** Corresponding Author: Chee Sun Won

Received June 20, 2015; Revised July 15, 2015; Accepted August 24, 2015; Published August 31, 2015

** Regular Paper: This paper reviews the recent progress possibly including previous works in a particular research topic, and has been accepted by the editorial board through the regular reviewing process.*

***Abstract*:** Human action recognition from a video scene has remained a challenging problem in the area of computer vision and pattern recognition. The development of the low-cost RGB depth camera (RGB-D) allows new opportunities to solve the problem of human action recognition. In this paper, we present a comprehensive review of recent approaches to human action recognition based on depth maps, skeleton joints, and other hybrid approaches. In particular, we focus on the advantages and limitations of the existing approaches and on future directions.

***Keywords*:** Human action recognition, Depth maps, Skeleton joints, Kinect

1. Introduction

The study of human action recognition introduces various new methods for understanding actions and activities from video data. The main concern in human action recognition systems is how to identify the type of action from a set of video sequences. Different systems like consumer interactive entertainment, gaming, surveillance systems, smart homes, and life-care systems include several feasible applications [1, 2], which have become the utmost inspiration for researchers, who have hence developed algorithms for human action recognition.

Previously, RGB cameras have always been a focal point of studies into identifying actions from image sequences taken by these cameras [3, 4]. Various constraints relating to 2D cameras are responsiveness to illumination changes, surrounding clutter, and disorder [3, 4]. It has been a tough and difficult task to precisely recognize human actions. However, with the development of cost-effective RGB depth (RGB-D) camera sensors (e.g., the Microsoft Kinect), the results from action recognition have improved, and they have become a point of consideration for many researchers [5]. Depth camera sensors provide more discriminating and clear information by giving a 3D structural view from which to recognize action, compared to visible light cameras. Furthermore, depth sensors also help lessen and ease the low-level complications found in RGB images, such as background subtraction and light variations. Also, depth cameras can be beneficial for the entire range of day-to-day work, even at night, like patient monitoring systems. Depth images enable us to view and assess human skeleton joints in a 3D coordinate system. These 3D skeleton joints provide additional information to examine for recognition of action, which in turn increases the accuracy of the human–computer interface [5]. Depth sensors, like the Kinect, usually provide three types of data: depth images, 3D skeleton joints, and color (RGB) images. So, it has been a big challenge to utilize these data, together or independently, to present human behavior and to improve the accuracy of action recognition.

Human movement is classified into four levels (motion, action, activity, and behavior), where motion is a small movement of a body part for a very short time span. However, motion is a key factor in actions, which helps to identify other movement, such as the following [6].

An action is a collection of recurring different motions, which show what a person is doing, like running, sitting, etc., or interaction of the person with certain things. The duration of the action lasts no more than a few seconds.

An activity is also an assortment of various actions that help in perceiving and understanding human behavior while performing designated tasks, like cooking, cleaning, studying, etc., which are activities that can continue for

Fig. 1. Flow of an action recognition system.

much longer times.

Behavior is extremely meaningful in understanding human motion that can last for hours (or even days) and that can be considered either normal or abnormal.

Action, activity and behavior can be differentiated on the basis of supportive dissimilar features concerning time scales. In this study, we focus on shorter and medium time period actions, such as raising a hand or sitting down.

Human action recognition has been mainly focused on three leading applications: 1) surveillance systems, 2) entertainment environments, and 3) healthcare systems, which comprise systems to track or follow individuals automatically [2, 7-14]. In a surveillance system, the authorities need to monitor and detect all kinds of criminal and suspicious activities [2, 7, 8]. Most surveillance systems, equipped with several cameras, require well-trained staff to monitor human actions on screen. However, using automatic human action recognition algorithms, it is possible to reduce the number of staff and immediately create a security alert in order to prevent dangerous situations. Furthermore, human action recognition systems can also be used to identify entertainment actions, including sports, dance and gaming. For entertainment actions, response time to interact with a game is very important. Thus, a number of techniques have been developed to address this issue using depth sensors [9, 10]. In healthcare systems, it is important to monitor the activities of a patient [11, 12]. The aim of using such healthcare systems is to assist the health workers to care for, treat, and diagnose patients, hence, improving the reliability of diagnosis. These medical healthcare systems can also help decrease the work load on medical staff and provide better facilities to patients.

Generally, human action–recognition approaches involve several steps, as shown in Fig. 1, where feature extraction is one of the important blocks, which performs a vital role in the action recognition system. The performance of feature extraction methods for an action recognition system is evaluated on the basis of classification accuracy. Several available datasets, recorded from depth sensors, are widely available and accessible for developing an innovative recognition system.

Every dataset includes different actions and activities performed by different volunteer subjects, and each dataset is designed to resolve a particular challenge. Table 1 provides a summary of the most popular datasets. Most of the methods reviewed in this paper are evaluated on one or more of these datasets. In this survey, we review human action recognition systems that have been proposed to recognize human actions. This review paper is organized as follows. In Section 2, we review human action systems based on depth maps, skeleton joints, and hybrid methods (i.e., depth and color, depth and skeleton). A summary of all the reviewed work is presented in Section 3, which includes the advantages and disadvantages of each reviewed method. The conclusion is presented in Section 4.

Table 1. Publicly available RGB-D datasets for evaluating action recognition systems.

Datasets	Size	Remarks
Microsoft Research Action3D [13]	10 subjects/ 20 actions/ 2-3 repetitions	There are a total of 567 depth map sequences with a resolution of 320x240. The dataset was recorded using the Kinect sensor. All are interactions with game consoles (i.e. draw a circle, two-hand wave, forward kick, etc.).
Microsoft Research Daily Activity 3D [14]	10 subjects/ 16 activities/ 2 repetitions	16 indoor activities were done by 10 subjects. Each subject performed each activity once in a standing position and once in a sitting position. Three channels are recorded using the Kinect sensor: (i) depth maps, (ii) RGB video, (iii) skeleton joint positions.
UT-Kinect Action [15]	10 subjects/ 10 actions/ 2 repetitions	In the UT-Kinect Action dataset, there are 10 different actions with three channels: (i) RGB, (ii) depth, and (iii) 3D skeleton joints.
UCF-Kinect [16]	16 subjects/ 16 activities/ 5 repetitions	The UCF-Kinect dataset is a long-sequence dataset that is used to test latency.
Kitchen scene action [17]	9 activities	Different activities in the kitchen have been performed to recognize cooking motions.

2. Human Action Recognition

2.1 Human Action Recognition Using Depth Maps

Li et al. [18] introduced a method that recognizes human actions from depth sequences. The motivation of this work was to develop a method that does not require joint tracking. It also uses 3D contour points instead of 2D points. Depth maps are projected on three orthogonal Cartesian planes, and a specified number of points along the contours of all three projections are sampled for each frame. These sampled points are then used as a “bag-of-points” to illustrate a set of salient postures that correspond to the nodes of an action graph used to model the dynamics of the actions. The authors used their own dataset for the experiments, which later became known as the Microsoft Research (MSR) Action3D dataset, and they achieved a

74.4% recognition accuracy. The limitation of this approach is that the sampling of 3D points from the whole body requires a large dataset. Also, due to noise and occlusion in the depth maps, YZ and XZ views may not be reliable for extracting 3D points.

To overcome some of the issues [13] in Table 1, Vieira et al. [18] proposed space–time occupancy patterns (STOP) to represent the sequence of depth maps, where the space and time axes are divided into multiple segments so that each action sequence is embedded in a multiple 4D grid. In order to enhance the role of spare cells, a saturation scheme was proposed, which typically consists of points on a silhouette or moving parts of the body. To recognize the actions, a nearest neighbor classifier based on cosine distance was employed. Experimental results on the MSR Action3D dataset show that STOP features for action classification yield better accuracy than that of Rougier et al. [12]. The major drawback to this approach is that they empirically set the parameter for dividing sequences into cells.

A method that addresses the noise and occlusion issues in action recognition systems using depth images was proposed by Wang et al. [19]. The authors considered a 3D action sequence as a 4D shape and proposed random occupancy pattern (ROP) features extracted from randomly sampled 4D sub-volumes of different sizes and at different locations using a weighted sampling scheme. An elastic-net regularized classification is then modeled to further select the most discriminative features, which are robust to noise and less sensitive to occlusions. Finally, support vector machine (SVM) is used to recognize the actions. Experimental results on the MSR Action3D dataset show that the proposed method outperforms previous methods by Li et al. [13] and Vieira et al. [18].

An action recognition system that is capable of extracting additional shape and motion information using 3D depth maps was proposed by Yang et al. [20]. In this system, each 3D depth map is first projected onto three orthogonal Cartesian planes. Each projected view is generated by thresholding the difference of consecutive depth frames and stacks to obtain a depth motion map (DMM) for each projected view. A histogram of oriented gradients (HOG) [21] is then applied to each 2D projected view to extract the features. Furthermore, the features from all three views are then concatenated to form a DMM-HOG descriptor. An SVM classifier is used to recognize the actions. Steps for extracting the HOG from the DMM are shown in Fig. 2. The drawback of this system is that their approach does not show the direction of the variation. Also, the authors explored the number of frames required to generate satisfactory results, which showed that roughly 35 frames are enough to generate acceptable results. Nonetheless, it cannot be applied to complex actions to get satisfactory results.

Ahmad et al. [22] employed an R transform [23] to compute a 2D angular projection map of an activity silhouette via Radon transform and to compare the proposed method with other feature extraction methods (i.e. PCA and ICA) [24, 25]. The authors argue that PCA and ICA are sensitive to scale and translation using depth silhouettes. Therefore, a 2D Radon transform converts into

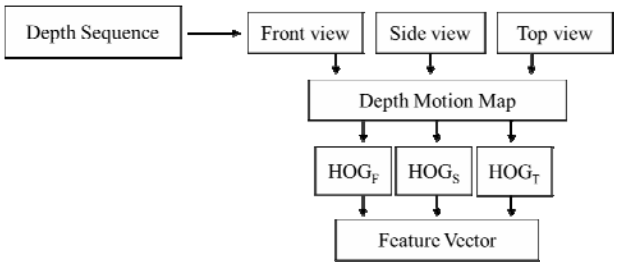

Fig. 2. Histogram of oriented gradients descriptor on motion maps.

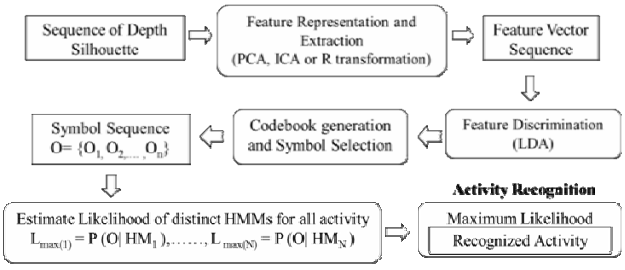

Fig. 3. Framework of the human activity recognition system using R transform [22].

a 1D R transform profile to provide a highly compact shape representation for each depth silhouette. That is, to extract suitable features from the 1D R transformed profiles of depth silhouettes, linear discriminant analysis (LDA) is used to make the features more discriminative. Finally, the features are trained and tested using hidden Markov models (HMMs) [26] on the codebook of vectors generated using the Linde-Buzo-Gray (LBG) clustering algorithm [27] for recognition. Fig. 3 shows the overall flow of the proposed method. Experimental results show that their feature extraction method is robust on the 10 human activities collected by the authors. Using this dataset, they achieved an accuracy of 96.55%. The limitation to this system is that the proposed method is view-dependent.

Using depth sequences, a new feature descriptor named histogram of oriented 4D surface normal (HON4D) was proposed by Oreifej and Liu [28]. The proposed feature descriptor is more discriminative than the average 4D occupancy [18] and is robust against noise and occlusion [18]. HON4D features consider the 3D depth sequences as a surface in 4D spatio-temporal space–time, depth and spatial coordinates. In order to construct HON4D, the 4D space is quantized using the 120 vertices of a 600-cell polychoron. Then, the quantization is refined using a discriminative density measure by inducing additional projectors in the direction, where the 4D normal is denser and more discriminative. An SVM classifier is used to recognize the actions. Experimental results show that HON4D achieves high accuracy compared to state-of-the-art methods. The limitation to this system is that HON4D can only roughly characterize the local spatial shape around each joint to represent human–object interaction. Also, differential operation on depth images can enhance noise.

Xia et al. [29] proposed an algorithm for extracting local spatio-temporal interest points (STIPs) from depth videos (DSTIPs) and described a local 3D depth cuboid using the depth cuboid similarity feature (DCSF). The DSTIPs deal with the noise in the depth videos, and DCSF was presented as a descriptor for the 3D local cuboid in depth videos. The authors claimed that the DSTIPs+DCSF pipeline recognizes activities from the depth videos without depending on skeleton joint information, motion segmentation, tracking or de-noising procedures. The experimental results reported for the MSR Daily Activity 3D dataset show that it is possible to recognize human activities using the DSTIPs and DCSF with an accuracy of 88.20% by using 12 out of 16 activities. Four activities that have less motion (i.e., sitting still, reading a book, writing on paper, and using a laptop) were removed from the experiments because most of the recognition errors come from these relatively motionless activities. Furthermore, the accuracy of the proposed system is highly dependent on the noise level of the depth images.

Recent work by Song et al. [30] focuses on the use of depth information to describe human actions in videos that seem to be of essential concern and can greatly influence the performance of human action recognition. The 3D point cloud is exercised because it holds points in the 3D real-world coordinate system to symbolize the human body's outer surface. An attribute named body surface context (BSC) is presented to explain the sharing of relative locations of neighbors for a reference point in the point cloud. Tests using the Kinect Human Action Dataset resulted in an accuracy of 91.32%. Using the BSC feature, experiments on the MSR Action3D dataset yielded an average accuracy of 90.36% and an accuracy of 77.8% with the MSR Daily Activity 3D dataset. Experimentation showed that superior performance is attained with the tested feature and it performed robustly when observing variations (i.e. translation and rotation).

2.2 Human Action Recognition Using Skeleton Joints

Xia et al. [31] showed the advantages of using 3D skeleton joints and represented 3D human postures using a histogram of 3D joint locations (HOJ3D). In their representation, 3D space is partitioned into bins using a modified spherical coordinate system. That is, 12 manually selected joints were used to build a compact representation of the human posture. To make the representation more robust, votes of 3D skeleton joints were cast into neighboring bins using a Gaussian weight function. To extract most dominant and discriminative features, LDA was applied to reduce the dimensionality. These discriminative features were then clustered into a fixed number of posture vocabularies which represent the prototypical poses of actions. Finally, these visual words were trained and tested using a discrete HMM. According to reported experimental results on the MSR Action3D dataset, and by using their own proposed dataset, their proposed method has the salient advantage of using 3D skeleton joints of the human posture. However, the drawback to their method is its reliance on the hip joint, which might potentially compromise recognition accuracy due to the noise embedded in the estimation of hip joint location.

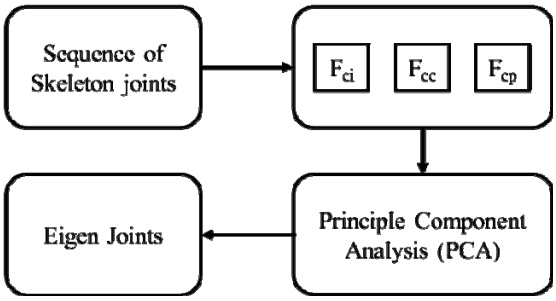

Fig. 4. Steps for computing eigen-joint features proposed by Yang et al. [32].

In a similar way, Yang et al. [32] illustrated that skeleton joints are computationally inexpensive, more compact, and distinctive compared to depth maps. Based on that, the authors proposed an eigen joints–based action recognition system, which extracts three different kinds of features using skeleton joints. These features include posture (Fcc), motion features (Fcp) that encode spatial and temporal characteristics of skeleton joints, and offset features (Fci), which calculate the difference between a current pose and the initial one. Then, applying PCA to these joint differences to obtain eigen joints by reducing the redundancy and noise, the Naive-Bayes-Nearest-Neighbor (NBNN) classifier [33] is used to recognize multiple action categories. Fig. 4 shows the process of extracting eigen joints. Also, they further explore the number of frames that are sufficient to recognize the action for their system. Experimental results on the MSR Action3D dataset show that a short sequence of 15-20 frames is sufficient for action recognition. The drawback to this approach is, while calculating the offset feature, the authors assume that the initial skeleton pose is neutral, which is not always correct.

Using the advantages of 3D joints, Yang et al. [32] proposed a compact but effective local skeleton descriptor that creates a pose representation invariant to any similarity conversion, which is, hence, view-invariant. The new skeletal feature, which is called skeletal quad [34], locally encodes the relation of joint quadruples so that 3D similarity invariance is assured. Experimental results of the proposed method verify its state-of-the-art performance in human action recognition using 3D joint positions. The proposed action recognition method was tested on broadly used datasets, such as the MSR Action3D dataset and HDM05. Experimental results with MSR Action3D using skeleton joints showed an average accuracy of 89.86%, and showed 93.89% accuracy with HDM05.

2.3 Human Action Recognition Using Hybrid Methods

The work done by Wang et al. [14] utilizes the advantages of both skeleton joints and point cloud information. Most of the actions differ mainly due to the objects in interactions, whereas in such cases, using only skeleton information is not sufficient. Moreover, to capture

the intra-class variance via occupancy information, the authors proposed a novel actionlet ensemble model. An important observation made by them in terms of skeleton joints is that the pairwise relative positions of joints are more discriminative than the joint positions themselves. Interaction between human and environmental objects is characterized by a local occupancy pattern (LOP) at each joint. Furthermore, the proposed method is evaluated using the CMU MoCap dataset, the MSR Action3D dataset, and a new dataset called the MSR Daily Activity 3D dataset. Experimental results showed that their method has superior performance compared to previous methods. The drawback of their system is that it relies on skeleton localization, which is unreliable for posture with self-occlusion.

Lei at al. [35] combined depth and color features to recognize kitchen activities. Their method successfully demonstrated tracking the interactions between hands and objects during kitchen activities, such as mixing flour with water and chopping vegetables. For object recognition, the reported system uses a gradient kernel descriptor on both color and depth data. The global features are extracted by applying PCA on the gradient of the hand trajectories, which are extracted by tracking the skin characteristics, and local features are defined using a bag-of-words for snippets of trajectory gradients. All the features are then fed into an SVM classifier for training. The overall reported accuracy is 82% for combined action and object recognition. This work shows the initial concept of recognizing the object and actions in a real-world kitchen environment. However, using such system in real time requires a large dataset to train the system.

Recently, Althloothi et al. [36] proposed a human activity recognition system using multi-features and multiple kernel learning (MKL) [37]. In order to recognize human actions from a sequence of RGB-D data, their method utilizes surface representation and a kinematics structure of the human body. It extracts shape features from a depth map using a spherical harmonics representation that describes the 3D silhouette structure, whereas the motion features are extracted using 3D joints that describe the movement of the human body. The author believes that segments such as forearms and the shin provide sufficient and compact information to recognize human activities. Therefore, each distal limb segment is described by orientation and translation with respect to the initial frame to create temporal features. Then, both feature sets are combined using an MKL technique to produce an optimally combined kernel matrix within the SVM for activity classification. The drawback to their system is that the shape features extracted using spherical harmonics are large in size. Also, at the beginning and at the end of each depth sequence in the MSR Action3D and MSR Daily Activity 3D datasets, the subject is in a stand-still position with small body movements. However, while generating the motion characteristics of an action, these small movements at the beginning and at the end generate large pixel values, which ultimately contribute to large reconstruction error.

3. Summary

The advantages and disadvantages of the above reviewed methods, based on depth maps, skeleton joints, and hybrid approaches, are presented in Table 2. Although

Table 2. Advantages and disadvantages of the existing methods.

Feature Extraction Methods	General comments	Pros	Cons
3D sampled points [13]	Using depth silhouettes, 3D points have been extracted on the contour of the depth map.	They extend RGB approaches to extract contour points on depth images. However, their method can recognize the action performed by single or multiple parts of the human body without tracking the skeleton joints.	Due to noise and occlusion, contours of multiple views are not reliable, and the current sampling scheme is view-dependent.
STOP: Space–Time Occupancy Patterns [18]	Space–time occupancy patterns are presented by dividing the depth sequence into a 4D space–time grid. All the cells in the grid have the same size.	Spatial and temporal contextual information has been used to recognize the actions, which is robust against noise and occlusion.	There is no method defined to set the parameter for dividing the sequence into cells.
Random Occupancy Patterns (ROP) [19]	ROP features are extracted from randomly sampled 4D sub-volumes with different sizes and different volumes. Then, all the points in the sub-volumes are accumulated and normalized with a sigmoid function.	The proposed feature extraction method is robust to noise and less sensitive to occlusion.	Feature patterns are highly complex and need more time during processing.
Motion maps [20]	Motion maps provide shape as well as motion information. However, HOG has been used to extract local appearance and shape of motion maps.	They are computationally efficient action recognition systems based on depth maps for extraction of additional shape and motion information.	Motion maps do not provide directional velocity information between the frames.

R Transform [22]	R transform has been used to extract features from depth silhouettes, comparing the proposed method with PCA and ICA.	R transform–based translation and scale-invariant feature extraction methods can be used for human activity recognition systems.	The R transform–based feature extraction method is not view-invariant.
HON4D [24]	Captures histogram distribution of the surface normal orientation in the 4D volume of time, depth and spatial coordinates.	The proposed feature extraction method is robust against noise and occlusion and more discriminative than other 4D occupancy methods. Also, it captures the distribution of changing shape and motion cues together.	This method can roughly characterize the local spatial shape around each joint. Differential operation on a depth image can enhance noise.
DCSF [29]	Extracting STIP from depth videos and describing local 3D DCSF around interest points can be efficiently used to recognize actions.	Uses DSTIPs and DCSF to recognize the activities from depth videos without depending on skeleton joints, motion segmentation and tracking or de-noising procedures.	It is difficult to analyze the method for full activities, and most of the recognition errors come from those activities.
Body surface context (BSC) [30]	3D point clouds have been used to represent the 3D surface of the body, which contains rich information to recognize human actions.	3D point clouds of the body's surface can avoid perspective distortion in depth images.	It is based on different combinations of features for each dataset, but it is not feasible for an automatic system to select the combination for high accuracy.
HOJ3D [31]	Twelve manually selected skeleton joints are converted to a spherical coordinate system to make a compact representation of the human posture.	Skeleton joints are more informative and can achieve high accuracy with a smaller number of joints.	Relying only on the hip joint might potentially compromise recognition accuracy.
Eigen joints [32]	This is an action recognition system that extracts spatiotemporal change between the joints. Then, PCA is used to obtain eigen joints by reducing redundancy and noise.	It is a skeleton joint–based feature extraction method that extracts features in both spatial and temporal domains. It is more accurate and informative than trajectory-based methods.	Offset feature computation depends on the assumption that the initial skeleton pose is neutral, which is not correct.
Quadruples [34]	A skeleton joint–based feature extraction method called skeletal quad ensures 3D similarity invariance of joint quadruples by local encoding using a Fisher kernel.	A view-invariant descriptor using joint quadruples encodes Fisher kernel representations.	It is not a good choice to completely rely on skeleton joints, because these 3D joints are noisy and fail when there are occlusions.
Hybrid method (3D point cloud + skeleton) [14]	Local occupancy pattern (LOP) features are calculated from depth maps around the joints' locations.	A highly discriminative and translation invariant feature extraction method captures relations between the human body parts and the env ironmental objects in the interaction. Also, it represents the temporal structure of an individual joint.	Heavily relying on skeleton localization becomes unreliable for postures with self-occlusion.
Kitchen activities (depth + RGB) [35]	Fine-grained kitchen activities are recognized using depth and color cues.	It is an efficient feature extraction method taking advantage of both RGB and depth images to recognize objects and fine-grained kitchen activities.	Requires a large dataset to train the system.
Multi-feature (3D point cloud and skeleton joints) [36]	This human activity recognition system combines spherical harmonics features from depth maps and motion features using 3D joints.	It is a view-invariant feature extraction method based on shape representation and the kinematics structure of the human body. That is, both features are fused using MKL to produce an optimal combined kernel matrix.	Shape features are large in size, which may be unreliable for postures with self-occlusion, whereas it extracts motion features on the assumption that the initial pose is in a neutral state, which is not the case.

Table 3. Summary of feature selection, classification and recognition methods

Paper	Extracted Features	Feature Selection/ Dimension reduction	Clustering	Classification
[13]	3D points at the contour of a depth map			Action graph
[18]	Depth values	PCA	K-means	HMM
[19]	3D point cloud	LDA	Elastic net regularized classifier	SVM
[20]	Histogram of gradients			SVM
[22]	Depth values	PCA, LDA	LBG	HMM
[28]	Histogram of surface normal			SVM
[29]	Histogram of depth pixels	PCA	K-means	DS-SRC
[30]	3D point cloud	PCA	K-means	SVM
[31]	Histogram of 3D joints in spherical coordinates	LDA	K-means	HMM
[32]	Skeleton joints	PCA		NBNN
[34]	Gradient values			SVM
[14]	Low-frequency Fourier coefficients	Actionlet ensemble		SVM
[35]	Gradient values			SVM
[36]	3D point cloud and skeleton joints			SVM

Table 4. Recognition accuracies of reviewed action recognition systems on benchmark datasets.

Paper	MSR Action3D	MSR Daily Activity 3D	UCF Kinect dataset	Kitchen scene action
[13]	74.7%			
[18]	84.80%			
[19]	86.50%			
[20]	91.63%			
[22]				
[28]	88.89%	80%		
[29]	89.3%	83.6%		
[30]	90.36%	77.8%		
[31]	78.97%			
[32]	82.33%		97.1%	
[34]	89.86%			
[14]	88.2%	85.75%		
[35]				82%
[36]	79.7%	93.1%		

all the above methods are capable of dealing with the actions and activities of daily life, there are also drawbacks and limitations to using depth map–based, skeleton joint–based and hybrid methods for action recognition systems. Depth maps fail to recognize human actions when fine-grained motion is required, whereas extracting 3D points at the contours may incur loss of inner information from the depth silhouettes. Furthermore, shape-based features do not provide any information for calculating the directional velocity of the action between the frames, and it is an important parameter for differentiating the actions. Hence, depth-based features are neither very efficient for, nor sufficient for, certain applications, such as entertainment, human–computer interaction, and smart healthcare systems. The 3D skeleton joints estimated using the depth maps are often noisy and may have large errors when there are occlusions (e.g., legs or hands crossing over each other). Moreover, motion information extracted using 3D joints alone is not sufficient to differentiate similar activities, such as drinking water and eating. Therefore, there is a need to include extra information in the feature vector to improve classification performance. Thus, a hybrid method

can be helpful by taking full advantage of using depth maps and 3D skeleton joints to enhance the classification performance of human action recognition.

A summary of all the feature-selection, clustering and recognition methods used in the above reviewed papers is in Table 3. Because most of the studied action recognition systems select dominant and discriminative features using LDA, these features are then represented by the codebook, which is generated using a k-means algorithm. Finally, after training the system, it recognizes the learned actions via the trained SVM.

The recognition accuracy of the reviewed methods on the datasets mentioned in Table 1 is summarized in Table 4. The assessment method adopted by the mentioned works for the MSR Action3D dataset is a cross-subject test. This method was originally proposed by Li et al. [13] by dividing the 20 actions into three subsets, with each subset containing eight actions. For the MSR Daily Activity 3D dataset, all the authors verified the performance of their method using a leave-one-subject-out (LOSO) test. For the UCF Kinect dataset, 70% of the actions were used for training and 30% for testing. Jalal et al. [22] proposed their own human activity dataset and evaluated the performance of their proposed method using 30% video clips for training and 70% for testing.

4. Conclusion

Over the last few years, there has been a lot of work by researchers in the field of human action recognition using the low-cost depth sensor. The success of these works is demonstrated in entertainment systems that estimate the body poses and recognize facial and hand gestures, by smart healthcare systems to care for patients and monitor their activities, and in the security systems that recognize suspicious activities and create an alert to prevent dangerous situations. Different databases have been used by the authors to test the performance of their algorithms. For the MSR Action3D dataset, Yang et al. [20] achieved 91.63% accuracy, whereas for the MSR Daily Activity 3D dataset, Althloothi et al. [36] achieved 93.1% accuracy. Moreover, Yang et al. [32] achieved 97.1% accuracy for the UCF Kinect dataset. Currently, human action systems focus only on extracting boundary information from depth silhouettes. However, using only skeleton information may not be feasible, because the skeleton joints are not always accurate. Furthermore, to overcome the limitations and drawbacks of the current human action recognition systems, it is necessary to extract valuable information from inside the depth silhouettes. Also, it is necessary to use the joint points with the depth silhouettes for an accurate and stable human action recognition system.

Acknowledgement

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2013R1A1A2005024).

References

[1] A. Veeraraghavan et al., “Matching shape sequences in video with applications in human movement analysis,” *Pattern Analysis and Machine Intelligence, IEEE Transactions*, pp.1896-1909, Jun. 2004. Article (CrossRef Link)

[2] W. Lin et al., "Human activity recognition for video surveillance,” *in Circuits and Systems, IEEE International Symposium on*, pp. 2737-2740, May. 2008. Article (CrossRef Link)

[3] H. S. Mojidra et al., “A Literature Survey on Human Activity Recognition via Hidden Markov Model,” *IJCA Proc. on International Conference on Recent Trends in Information Technology and Computer Science 2012 ICRTITCS*, pp. 1-5, Feb. 2013. Article (CrossRef Link)

[4] R. Gupta et al., “Human activities recognition using depth images,” *in Proc. of the 21st ACM international conference on Multimedia*, pp. 283-292, Oct. 2013. Article (CrossRef Link)

[5] Z. Zhang et al., “Microsoft kinect sensor and its effect.” *MultiMedia, IEEE*, Vol. 19, No. 2, pp. 4-10, Feb. 2012. Article (CrossRef Link)

[6] A. A. Chaaraoui, "Vision-based Recognition of Human Behaviour for Intelligent Environments," *Director: Florez Revuelta, Franciso*, Jan. 2014. Article (CrossRef Link)

[7] M. Valera et al., “Intelligent distributed surveillance systems: a review,” *Vision, Image and Signal Processing, IEE Proceedings*, Vol. 152, No. 2, pp. 192-204. Apr. 2005. Article (CrossRef Link)

[8] J. W. Hsieh et al., “Video-based human movement analysis and its application to surveillance systems,” *Multimedia, IEEE Transactions on*, Vol. 10, No. 3, pp. 372-384, Apr. 2008. Article (CrossRef Link)

[9] V. Bloom et al., “G3d: A gaming action dataset and real time action recognition evaluation framework,” *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012 IEEE Computer Society Conference on, pp. 7-12, Jul. 2012. Article (CrossRef Link)

[10] A. Fossati et al., “Consumer depth cameras for computer vision: research topics and applications,” *Springer Science & Business Media*, Article (CrossRef Link)

[11] M. Parajuli et al., “Senior health monitoring using Kinect,” *Communications and Electronics (ICCE), Fourth International Conference on*, pp. 309-312, Aug. 2012. Article (CrossRef Link)

[12] C. Rougier et al., “Fall detection from depth map video sequences,” *Toward Useful Services for Elderly and People with Disabilities*, Vol. 6719, pp. 121-128. Hun. 2011. Article (CrossRef Link)

[13] W. Li et al., “Action recognition based on a bag of 3d points,” *Computer Vision and Pattern Recognition Workshops (CVPRW) IEEE Computer Society Conference on*, pp. 9-14, Jun. 2010. Article (CrossRef Link)

[14] J. Wang et al., “Mining actionlet ensemble for action

recognition with depth cameras," *Computer Vision and Pattern Recognition (CVPR) IEEE Conference on*, pp. 1290-1297. Jun. 2012. Article (CrossRef Link)
[15] L. Xia et al., "View invariant human action recognition using histograms of 3d joints," *Computer Vision and Pattern Recognition Workshops (CVPRW) IEEE Computer Society Conference on*, pp. 20-27, Jun. 2012. Article (CrossRef Link)
[16] C. Ellis et al., "Exploring the trade-off between accuracy and observational latency in action recognition," *International Journal of Computer Vision*, Vol. 101, No. 3. Pp. 420-436. Aug. 2012. Article (CrossRef Link)
[17] A. Shimada et al., "Kitchen scene context based gesture recognition: A contest in ICPR2012," *Advances in Depth Image Analysis and Applications*, Vol. 7854, pp. 168-185, Nov. 2011. Article (CrossRef Link)
[18] A. W. Vieira et al., "Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences," *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, Vol. 7441, pp. 252-259. Sep. 2012. Article (CrossRef Link)
[19] J. Wang et al., "Robust 3d action recognition with random occupancy patterns." *12th European Conference on Computer Vision*, pp. 872-885. Oct. 2012. Article (CrossRef Link)
[20] X. Yang et al., "Recognizing actions using depth motion maps-based histograms of oriented gradients," *in Proc. of the 20th ACM international conference on Multimedia*, pp. 1057-1060, Nov. 2012. Article (CrossRef Link)
[21] N. Dalal et al., "Histograms of oriented gradients for human detection," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, Vol. 1, pp. 886-893, Jun. 2005. Article (CrossRef Link)
[22] A. Jalal et al., "Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home," *Consumer Electronics, IEEE Transactions on*, Vol. 58, No. 3, pp. 863-871, Aug. 2012. Article (CrossRef Link)
[23] Y. Wang, K. Huang, and T. Tan. "Human activity recognition based on r transform." In Computer Vision and Pattern Recognition, IEEE Conference on, pp. 1-8. Jun. 2007. Article (CrossRef Link)
[24] M. Z. Uddin et al., "Independent shape component-based human activity recognition via Hidden Markov Model," *Applied Intelligence*, Vol. 33, No. 2, pp. 193-206. Jan. 2010. Article (CrossRef Link)
[25] J. Han et al., "Human activity recognition in thermal infrared imagery," *Computer Vision and Pattern Recognition. IEEE Computer Society Conference on*, pp. 17, Jun. 2005. Article (CrossRef Link)
[26] H. Othman et al., "A separable low complexity 2D HMM with application to face recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 25. No. 10, pp. 1229 – 1238, Oct. 2003. Article (CrossRef Link)
[27] Y. Linde et al., "An algorithm for vector quantizer design," *Communications, IEEE Transactions on*, Vol. 28, No. 1, pp. 84–95, Jan. 1980. Article (CrossRef Link)
[28] O. Oreifej, & Z. Liu., "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013. Article (CrossRef Link)
[29] L. Xia et al., "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," *Computer Vision and Pattern Recognition, IEEE Conference on*, pp. 2834-2841. Jun. 2013. Article (CrossRef Link)
[30] Y. Song et al., "Body Surface Context: A New Robust Feature for Action Recognition from Depth Videos," *Circuits and Systems for Video Technology, IEEE Transactions on*, Vol. 24, No. 6, pp. 952-964, Jan. 2014. Article (CrossRef Link)
[31] L. Xia et al., "View invariant human action recognition using histograms of 3d joints," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, pp. 20-27. Jun. 2012. Article (CrossRef Link)
[32] X. Yang et al., "Effective 3d action recognition using eigenjoints," *Journal of Visual Communication and Image Representation*, Vol. 25, No. 1, pp. 2-11, Jan. 2014. Article (CrossRef Link)
[33] O. Boiman et al., "In defense of nearest-neighbor based image classification," *Computer Vision and Pattern Recognition, IEEE Conference on*, pp. 1-8, Jun. 2008. Article (CrossRef Link)
[34] G. Evangelidis et al., "Skeletal quads: Human action recognition using joint quadruples," *Pattern Recognition (ICPR), 22nd International Conference on*, pp. 4513-4518. Aug. 2014. Article (CrossRef Link)
[35] J. Lei et al., "Fine-grained kitchen activity recognition using rgb-d." *in Proc. of the ACM Conference on Ubiquitous Computing*, pp. 208-211. Sep. 2012. Article (CrossRef Link)
[36] S. Althloothi et al., "Human activity recognition using multi-features and multiple kernel learning," *Pattern Recognition*, Vol. 47. No. 5, pp. 1800-1812. May. 2014. Article (CrossRef Link)
[37] M. Gönen and E. Alpaydin, "Multiple kernel learning algorithms," *The Journal of Machine Learning Research*, Vol. 12, pp. 2211-2268. Jan. 2011. Article (CrossRef Link)

Adnan Farooq is a Ph.D. student in Department of Electrical and Electronics Engineering at Dongguk University, Seoul, South Korea. He received his B.S degree in Computer Engineering from COMSATS Institute of Science and Technology, Abbottabad, Pakistan and M.S. degree in Biomedical Engineering from Kyung Hee University, Republic of Korea. His research interest includes Image Processing, Computer vision.

Chee Sun Won received the B.S. degree in electronics engineering from Korea University, Seoul, in 1982, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Massachusetts, Amherst, in 1986 and 1990, respectively. From 1989 to 1992, he was a Senior Engineer with GoldStar Co., Ltd. (LG Electronics), Seoul, Korea. In 1992, he joined Dongguk University, Seoul, Korea, where he is currently a Professor in the Division of Electronics and Electrical Engineering. He was a Visiting Professor at Stanford University, Stanford, CA, and at McMaster University, Hamilton, ON, Canada. His research interests include MRF image modeling, image segmentation, robot vision, image retrieval, image/video compression, video condensation, stereoscopic 3D video signal processing, and image watermarking.