Future Computing and Informatics Journal

Volume 3
Issue 1 (2018) *Volume 3, Issue 1, 2018*

Article 5

2018

Depth-based human activity recognition: A comparative perspective study on feature extraction

Heba Hamdy Ali
Beni-Suef University, Cairo, Egypt, heba.h.ali@fcis.bsu.edu.eg

Hossam M. Moftah
Faculty of Computers and Information, Beni Suef University, Egypt, hossamm@gmail.com

Aliaa A.A. Youssif
Helwan University, Cairo, Egypt, aliaay@fci.helwan.edu.eg

Available online at www.sciencedirect.com

ScienceDirect

Future Computing and Informatics Journal 3 (2018) 51–67
http://www.journals.elsevier.com/future-computing-and-informatics-journal/

Depth-based human activity recognition: A comparative perspective study on feature extraction

Heba Hamdy Ali [a,*], Hossam M. Moftah [a], Aliaa A.A. Youssif [b]

[a] Beni-Suef University, Cairo, Egypt
[b] Helwan University, Cairo, Egypt

Received 5 September 2017; revised 18 November 2017; accepted 26 November 2017
Available online 21 December 2017

Abstract

Depth Maps-based Human Activity Recognition is the process of categorizing depth sequences with a particular activity. In this problem, some applications represent robust solutions in domains such as surveillance system, computer vision applications, and video retrieval systems. The task is challenging due to variations inside one class and distinguishes between activities of various classes and video recording settings. In this study, we introduce a detailed study of current advances in the depth maps-based image representations and feature extraction process. Moreover, we discuss the state of art datasets and subsequent classification procedure. Also, a comparative study of some of the more popular depth-map approaches has provided in greater detail. The proposed methods are evaluated on three depth-based datasets "MSR Action 3D", "MSR Hand Gesture", and "MSR Daily Activity 3D". Experimental results achieved 100%, 95.83%, and 96.55% respectively. While combining depth and color features on "RGBD-HuDaAct" Dataset, achieved 89.1%.

Keywords: Activity recognition; Depth; Feature extraction; Video; Human body detection; Hand gesture

1. Introduction

Human activity recognition has an incredible significance in computer vision field. The human activity recognition objective is to examine and characterization progressing activities automatically from an unknown video. The advantages of recognizing human activities from videos are efficient in several critical applications. Such as automated surveillance systems [1] in public places, such as metro stations and air terminals require detection and of abnormal and normal activities.

There are different sorts of human activities [2] according to their complexity; these activities are divided into four types: "Gestures", "Actions", "Interactions", and "Group Activities". "Gestures" are simple motion of a part of a body. For example, "raising an arm and moving a leg." Actions will be exercises performed by one individual that might be made out of various gestures organized in a time order, "strolling", "waving", and "punching" are examples of "Actions". "Interactions" are human activities that involve at least two individuals or objects. As an example, "two persons checking hands" is an interaction between two individuals and "somebody pushing table" is an interaction includes one persons and an object. Finally, "Group activities" are that activities played by group composed of individuals or objects: "A group of people playing football" and two groups fighting" are typical examples.

Monitoring of changes in an actor's behavior is an important process in activity recognition. This task is in charge of acquiring applicable relevant data for activity recognition systems to recognize an activity. The two main activity recognition approaches are "vision-based" and "depth map-based".

* Corresponding author.
E-mail addresses: heba.h.ali@fcis.bsu.edu.eg (H.H. Ali), hossamm@gmail.com (H.M. Moftah), aliaay@fci.helwan.edu.eg (A.A.A. Youssif).
Peer review under responsibility of Faculty of Computers and Information Technology, Future University in Egypt.

https://doi.org/10.1016/j.fcij.2017.11.002

"Vision-based" activity recognition utilizes computer vision methodologies to analyze visual observations for activity recognition using visual sensing facilities, e.g., camera, and infra-red sensor, to capture activity [3–5] There has been significant work made on vision-based activity recognition [6], however, because of the multifaceted nature of true settings. These methodologies experience from issues related to reusability and scalability, such as highly variation of activities in the natural environment.

"Depth Maps-based" activity recognition depends mostly on features, either local or global, extricated from depth map images [6]. Depth maps give metric estimations of the geometry while visual information gives projective one that is invariant to lighting. Moreover, depth sequence representations for action recognition have a few difficulties. Above all else, depth map images may contain occlusions, which make the global features unsettled. Additionally, contrasted with color images, the depth images do not have texture but it difficult to apply local differential operators like gradients on because they are generally too noisy in both spatial and temporal cases.

The majority of vision-based systems are developed to work on normal visual information. There have been incredible reviews research [7–9]. There are inherent limits to the sort of image acquisition source. It is delicate to color, shading and illumination light variations, occlusions, and background clutters. In spite of great effort, the accuracy of recognizing actions is still a challenging research point. Due to the financially cost-effective "Kinect", depth cameras have gotten significant consideration from researchers in the vision and robotics community. The depth camera has two main benefits. Firstly, the depth sensor supplies information about 3D structure of the image to recover postures and recognize the activity. Secondly, the depth sensor can sense in darkness. This benefit is used for animal monitoring systems. These advantages are utilized in interesting research points like skeleton human detection from a depth map [10]. The skeletons measured from depth maps are precise and bring advantages to numerous applications including action and gesture recognition. Depth-map human activity recognition can be considered in its simplest form as a sequence of image representation, feature extraction process, and recognition of these activities.

The paper is organizes as, we first illustrate related work and discuss the key features and challenges of the human activity recognition as these motivate the different methodologies that detailed in the literature. We discuss images representation and feature extraction in Section 2. Many works will be depicted and examined in more detail in Section 3. In Section 4, we introduce the most well-known datasets. Then, we talk about impediments of state of the art approaches and outline future directions.

2. Feature extraction approaches

In this section, we debate various feature extraction techniques from depth map sequences. Ideally, these should be general over little varieties in appearance, background, perspective, and activity performing. In the meantime, the descriptor must be adequately generous to take into consideration powerful characterization of the activity. The temporal order is important in real life action performance. Several of the image representations approaches expressly consider the temporal order; others extract only image features for each image of the sequence. In this situation, the temporal variations are needed to be handled in the recognition phase.

2.1. 3D points (BOPs) features

Interest points provide an image content representation by depicting local parts of the image thus consider robustly solution to clutter, occlusions, and intra-class variations [11]. "Interest points" extracted from 2D-images can be employed for applications like image retrieval, and video classification. The extraction of the points on the outlines of the planar projections of the 3D depth-map is the simplest way to sample 3D point's representation. Regards to projection to projection planes number utilized, however, the number of points can still be significant. To address the issue; the idea of "bag of point" [12] is utilized. A sampling task which comprises of "projection", "contour sampling", and finally "retrieval" of the 3D points that are close to the sampled 2D points as shown in Fig. 1. BOPs features encode the activities in the expandable graphical model framework [13]. A static posture is represented as node in action graph which depicted by a little arrangement of sampled 3D points.

One limitation of 3D points (BOPs) features approach is the missing of space features between the interest points. In addition, it may not be reliable because of noise and occlusions in the depth maps, the silhouettes perspectives from the top, and the side views. It is hard to sample the interest points robustly given the geometry and movement varieties over various people. To address these issues, authors presented in [14] feature representation, defined as "Space Time Occupancy Patterns: STOP". The depth map sequence introduces in a "4D space-time grid". Saturation method is applied to improve the roles of the silhouettes points of the body parts movements. An action of Forwarding Kick depth sequence formed in space-time cells shown in Fig. 2. The sequence divided into three segments, and every portion contains about twenty depth-frames. The empty cells are not displayed, and the points in red color are the cells that have pointes greater than a defined level of points. A "STOP" feature vector is very scanty, that is, the most of its data are zero-elements. "Orthogonal Class Learning (OCL)" [15] is a modified version of "Principal Component Analysis (PCA)" [16] to perform a dimensionality reduction. OCL obtained for every "STOP" feature vector. A small feature vector is generated by reduce dimensions that called "PCA-STOP".

2.2. Spatial–temporal cuboid descriptors

The extension of interest points from 2D images into 3D is Space-Time Interest Points (STIP) [17] which is mostly used for action or activity recognition. The popular "STIP" descriptors

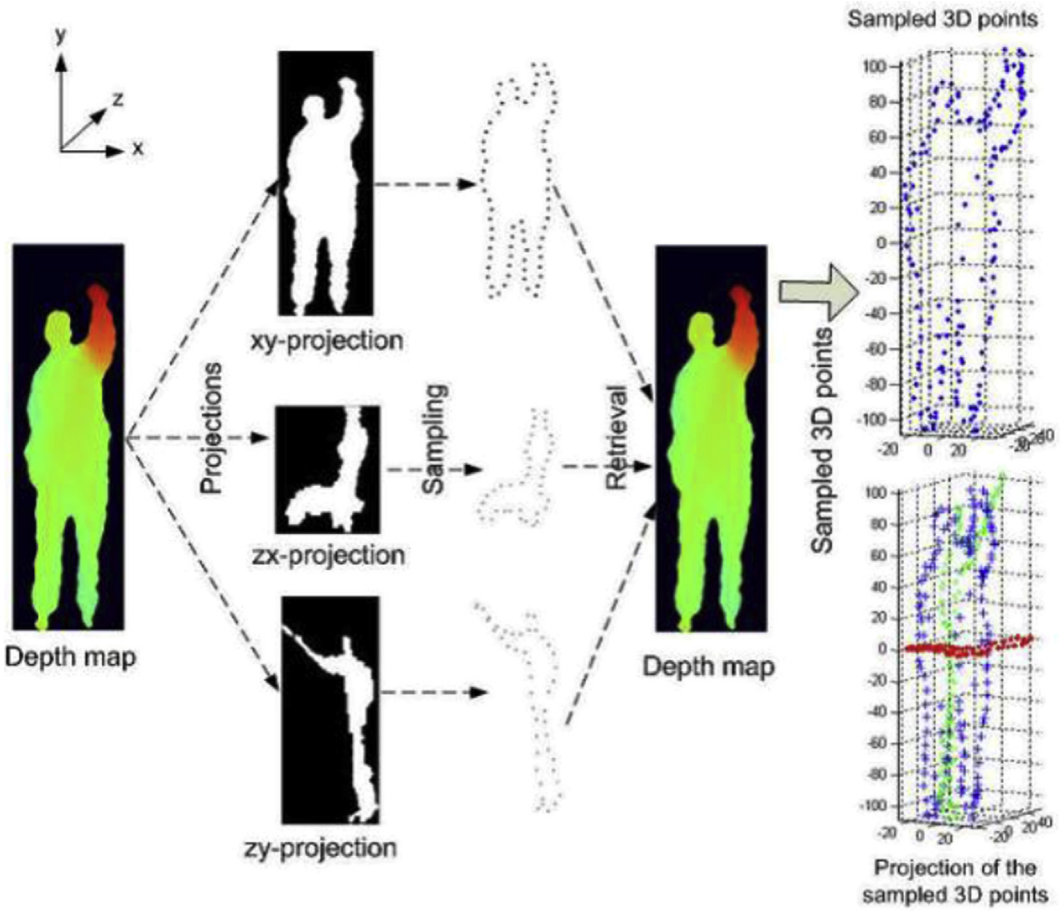

Fig. 1. Sampling process of 3D points representative from a depth image [12].

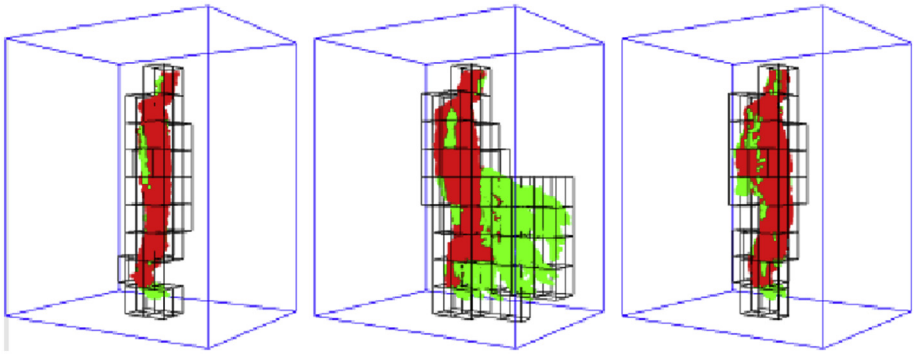

Fig. 2. Depth sequence Space-time cells for 'Forward Kick' action [14].

include the "cuboid detector" [18], "HOGHOF" [19], "HOG3D" [20] and "extended SURF" [21].

The filtering method is presented in [11] to extricate "STIPs" from depth recordings named as "DSTIP" that adequately suppress the noise estimations. Extra, new "depth cuboid similarity feature (DCSF)" to represent the local 3D depth cuboid around the DSTIPs" with adjustable size is built. The "cuboid codebook" is produced using K-means algorithm to cluster the "DCSF" as an outline in Fig. 3.

Another descriptor of the depth action analysis presented in [22] called Comparative Coding Descriptor (CCD). Small cuboids can be generated from the spatiotemporal of depth map with centers. That reference points (center) can be selected similarly as the corners of spatiotemporal or salient points for action representation. Cuboid with the side size of three is used as cuboid, dependent upon which "CCD" components is concentrated. The value of the center is compared with that of the other 26 points respectively, and the differences are coded. Fig. 4 illustrates the creation of "CCD" feature descriptor. Colored slices display depth frames in time, and the red vertex indicates the reference point.

2.3. Random Occupancy Pattern (ROP) features

The authors in [23] also studied activity distinguish issue from depth sequences acquired by a one depth sensor. They

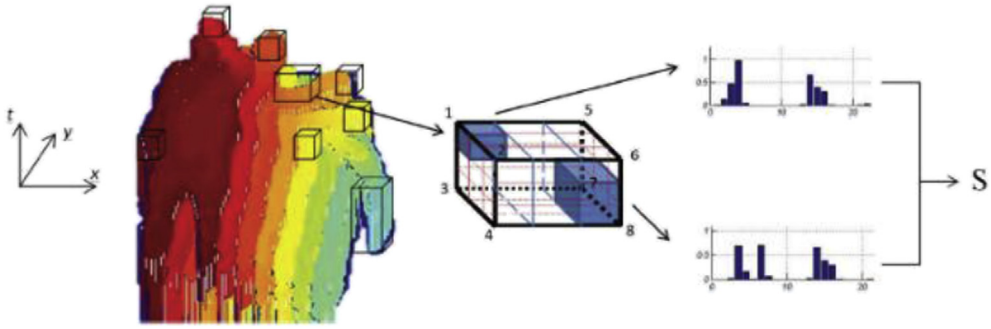

Fig. 3. Extracting DCSF from depth video [11].

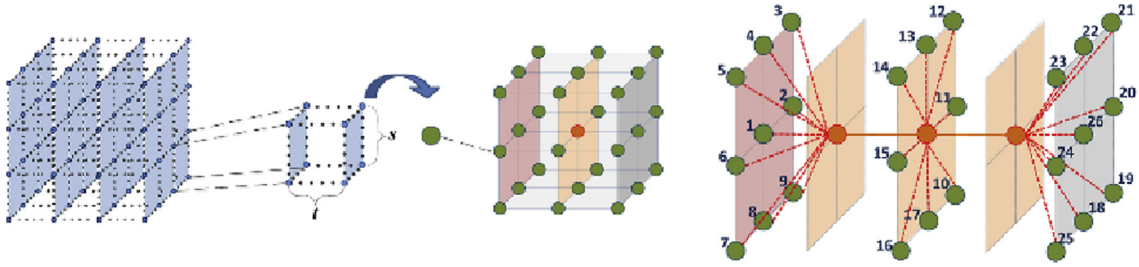

Fig. 4. "CCD" extraction process [22].

introduced "Random Occupancy Pattern (ROP)" features, which extracted from sampled 4D sub-volumes with various sizes and at various locations in Fig. 5.To handle the issues of noise and occlusion, depth-maps are represented in 4D sub-stitute of a three-dimensional movement succession. The ROP features are robust to noise when they are extricated at a bigger scale. Meanwhile, since they encode majority of the data starting with the regions that are most discriminative for those provided for an action, they are less conscious to occlusion. They also introduced a sampling method to represent the large space of sample in efficient way. Sparse coding [24] is applied further to enhance the proposed technique.

2.4. Depth silhouette

Motivated by the large success of silhouette that provides the shape information of human activities. "Depth silhouettes" demonstrate discernable parts besides to the shape information while "Binary silhouettes" contain less information because of its pixel intensity values distribution over the human body just shape information is available as demonstrated in Fig. 6a. While depth silhouette images for sample of rushing activity shown in Fig. 6b.

Fig. 7 shows method developed in [25] for utilizing "Depth silhouettes" to generate feature descriptors. The main concept is to employee "R transform" [26] on the depth silhouette to obtain compact shape representation rejecting time-sequential problems. In R transformation, a 2-D directional shape feature is calculated first through Radon transform of every depth silhouette, and then a 1-D feature profile, that is translation and scaling invariant, gets computed through R transform.

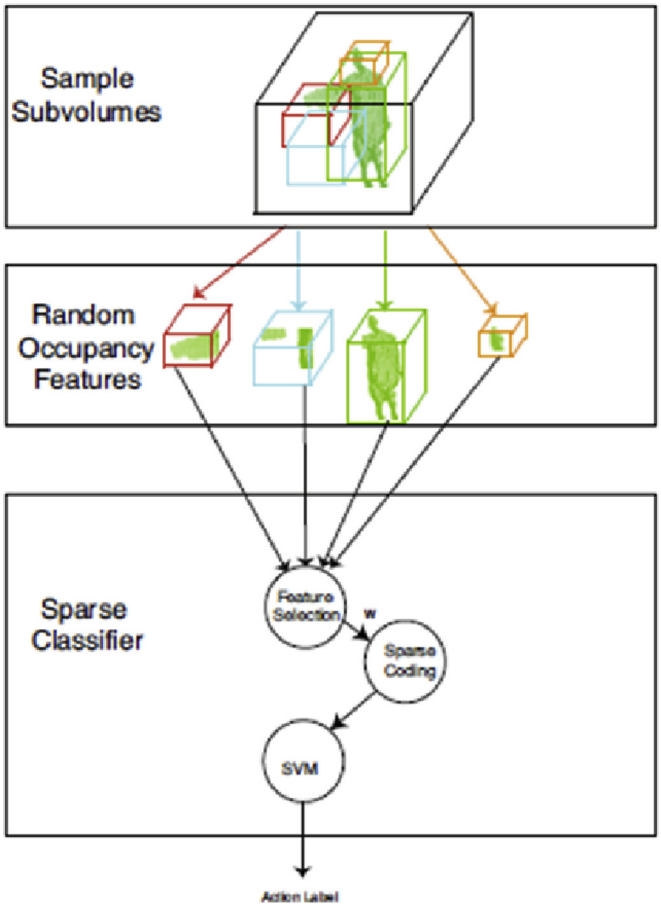

Fig. 5. Occupancy Pattern framework proposed in [23].

Fig. 6. Depth sequences of (a) "binary silhouette" and (b) "depth silhouettes" [25].

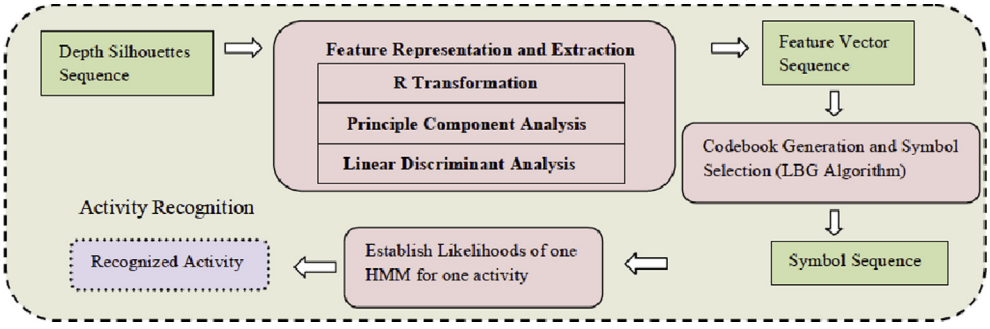

Fig. 7. The Flow of Depth Silhouettes and ℜ Transformation method proposed by [25].

Finally PCA is used for dimension reduction for a set of the "R transform" profiles of different activities, and then it applies "Linear Discriminant Analysis" [27] to extract prominent activity feature descriptor which are more minimized and robust. Finally, the features are feed into "Hidden Markov Model (HMM)" for activity classification. "Linear Discriminant Analysis" is modified to get discriminate vectors as in [28] like HMM, the most used sequential methods for visual data. "R-transform" extracts scale, periodic, and translation invariant features from the group of activities also used in [29]. The authors use the "kernel discriminate analysis (KDA)" [30] to improve the high postures resemblance of various actions. KDA is significant increases discrimination among the various categories of actions by using non-linear techniques.

In an example of real-time systems proposed [31] in for dynamic hand gesture recognition. The authors developed two sorts of visual features: cell occupancy features and silhouette features as demonstrated in Fig. 8 respectively. With the large dimensionality of both shape descriptors, PCA to reduce the dimensionality. This approach based on action graph, which similar to standard HMM with their robust properties but by allowing states share among different gestures, they require less training data. To deal with hand orientations, the authors have implemented a new method for hand segmentation and orientation normalization.

2.5. Surface normals features

Another depth-based descriptor introduced in [32], the authors has used a histogram to capture the distribution of the normal surface in the 4D volume of depth to represent the depth video sequence, time, and spatial coordinates. To obtain the remarked structure, "Histogram of oriented 4D surface normals (HON4D)" computed for depth sequence. A quantization process is applied in 4D space normal polychoron to construct HON4D, Afterward, refining the quantization to become more discriminative Fig. 9 outlines the different steps engaged with computing the HON4D descriptor.

Rather than using depth maps only, 4d local Spatio–temporal features proposed in [33] used to represent human activities. They utilize a weighted straight of a visual and geometric features combinations. The approach at that point concatenate the elements and their gradients using a spatial–temporal window into one vector is about more than 105 element features. K-means clustering [34] is implied on all vectors to decrease the high dimensionality. Feature vectors are grouped from a training subset with 600 vocabularies they utilized six movement classifications. "Latent Dirichlet Allocation (LDA)" [35] model used to anticipate activities from input recordings; this methodology address this issue with respect to six activities classes are considered as "topics" and feature computed from 4D feature space are considered as "words". Because of the efficiency of this sampling schema, it applied for approximate estimation.

Some work based on hypersurface normals presented in [36], by cluster a depth maps to generate the "polynormal" which is used jointly to represent both movements and shape. To extract the spatial features and temporary orders, the Adaptive "Spatio-temporal pyramid" is implied to a depth sequence to subdivide into a set of "space-time grids" as shown in Fig. 10. A new method of gathering the low level "polynormals" into the one "supernormal vector (SNV)" which considered as a modified Fisher kernel descriptor [37].

Another using of normals is the "polynormal" which presented in [37] to distinguish human activities from video depth sequences. It gathers hypersurface normals of local neighboring from a depth video to generate the "polynormal" that jointly represent shape cues and local motion. "Polynormal" Fisher Vector is the aggregation of the low-level "polynormals" using Fisher vector. The "Spatio-temporal pyramid" subdivides

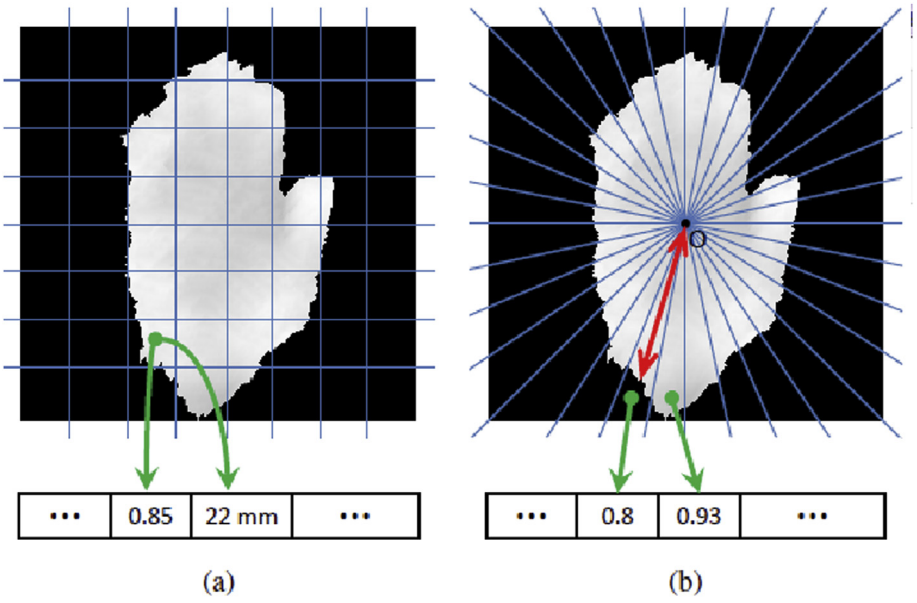

Fig. 8. Feature extraction (a) Occupied area of each cell, for cell occupancy features, (b)Fan-like sectors are divided, For silhouette feature [31].

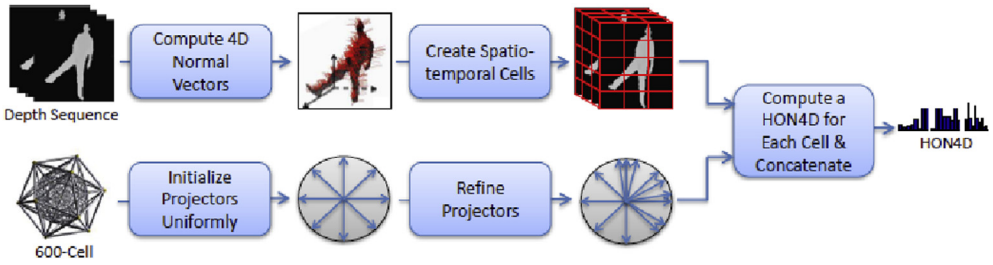

Fig. 9. The HON4D descriptor computing steps [32].

a depth video into a set of space-time cells to extract the spatial information and temporal order; "Polynormal" Fisher Vectors from these cells are aggregated as the one feature descriptor of a depth-map sequence illustrated in Fig. 11.

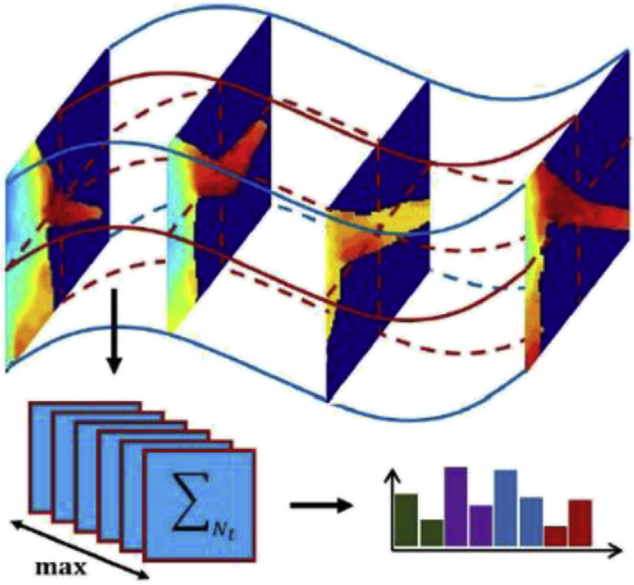

Fig. 10. The joint trajectory volume, "polynormals" proposed in [36].

2.6. Depth motion maps

The "Depth Motion Maps (DMM)" developed to capture the combined temporal movement energies. More spastically, the depth map is projected using "orthogonal Cartesian planes" and then normalized. By computing the difference between two successive frames and thresholding for every projected depth-map, a binary depth map is created. Then summed the binary maps to obtain the "DMM" for every projective view [39]. "Histogram of Oriented Gradients (HOG)" [40] is then applied to every perspective view to extracting the features. "DMM-HOG" descriptors generated by concatenating three aspects together as shown in Fig. 12.

An approach for human actions recognition is proposed in [41] using depth images. The motion of the object are computed by both depth image average and the depth difference image. They use hierarchical structure of the silhouette bounding box to get the features from space-time depth difference images. The temporal feature of the action represented using motion history of depth images. The author uses the scale, translation, and Hu moments to describe the features of the average depth image and the motion history image. Then using SVM to classify human actions.

Real-time action recognition presented in [42], DMMs from three projection views (front, side, and top) are used to

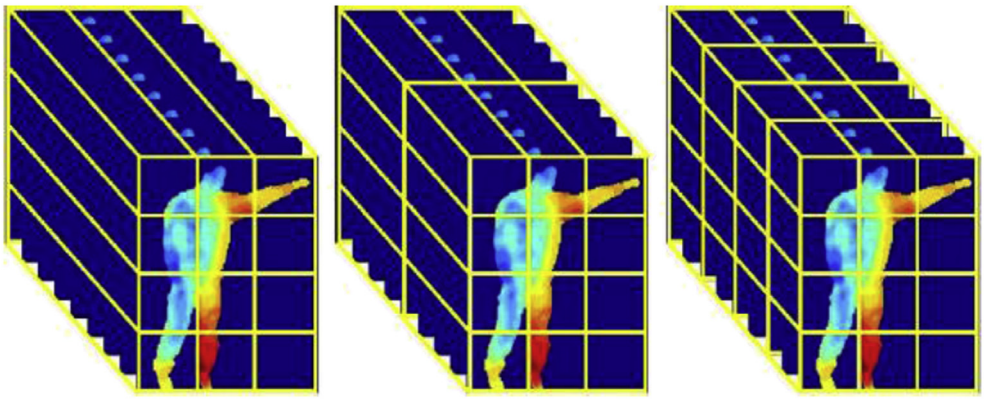

Fig. 11. Spatio-temporal example [38].

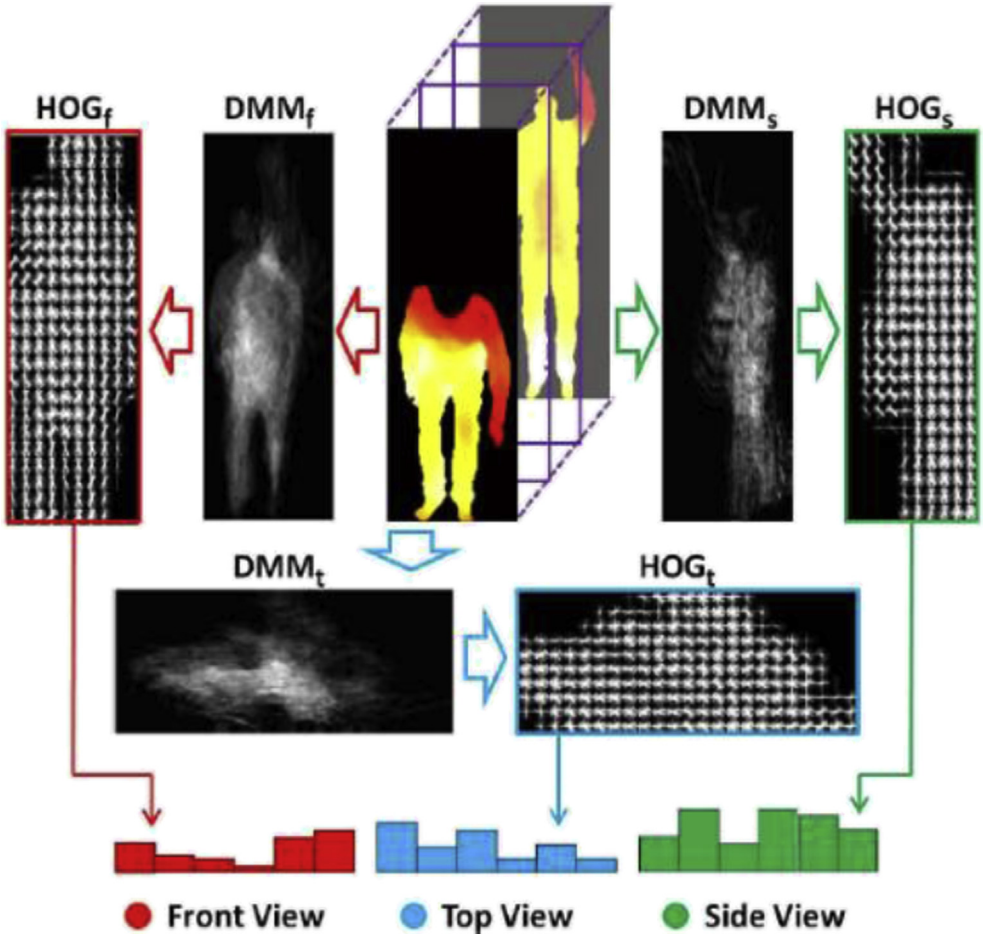

Fig. 12. Depth motion maps-based HOG framework proposed by [39].

describe motion and use PCA for dimensionality reduction as shown in Fig. 13. To recognition an action, an "l_2-regularized collaborative representation classifier" using a distance-weighted "Tikhonov" matrix is then utilized. The developed algorithm efficient computationally allowing it to run in real-time. To gain a compacted feature representation. The authors extend their work [43] and presented new methodology in [44]. Fig. 14 shows two sorts of fusions composed of feature-level decision-level fusions. In the feature level, "LBP features" are merged from three depth motion maps to gain a compact feature descriptor while in the decision level; a soft decision merged rule is used to aggregate the classification outputs.

A compact and discriminative action representation represented at [44]. The proposed feature extraction and action classification framework are shown in Fig. 15. Firstly, it

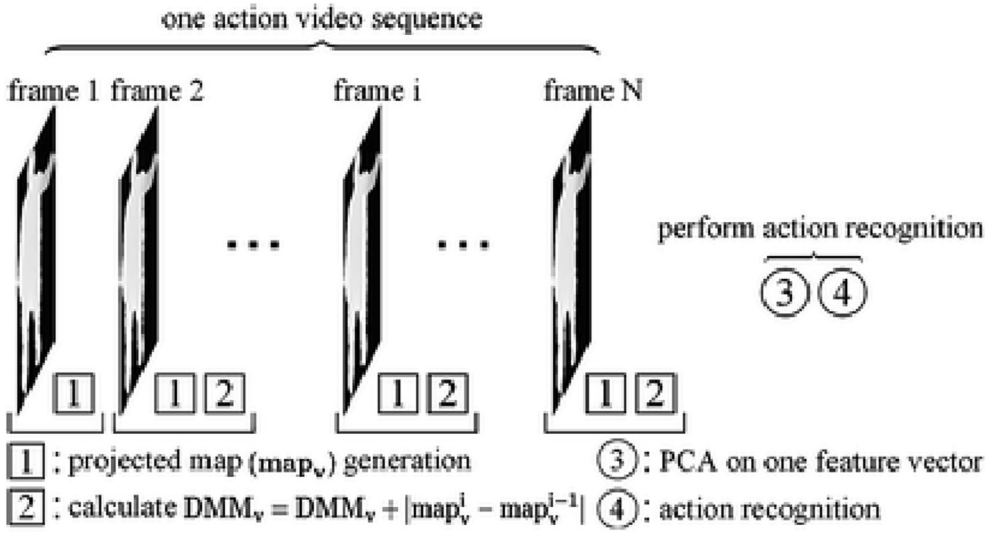

Fig. 13. Real-time action recognition [42].

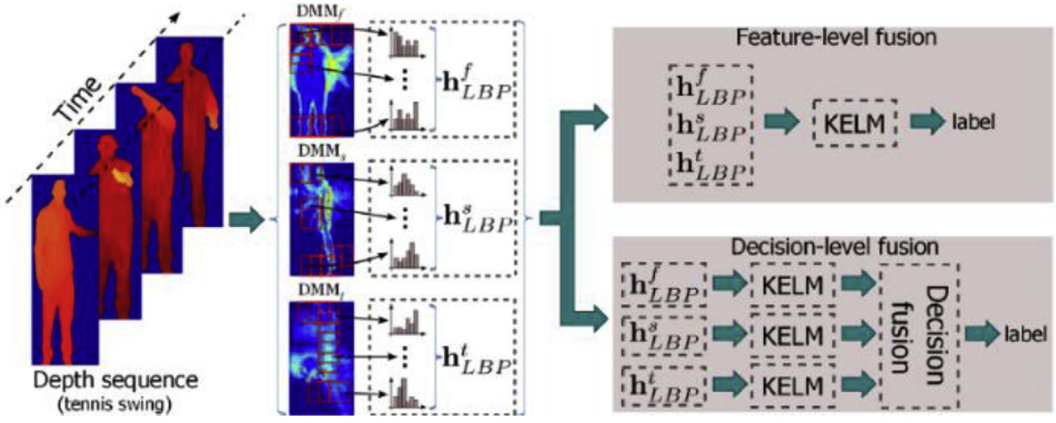

Fig. 14. The Pipeline of the developed action recognition method [44].

creates a side-viewed depth map from the input front-viewed depth map to capture additional information. Then by accumulating a series of depth, the framework create Depth Motion Appearance (DMA) and an extended version of motion history image called "Depth Motion History (DMH)". "DMH" has dynamic information of the sequence of the movement. Finally, "DMAs" and "DMHs" are merged into one single HOG descriptor. The linear SVM categorizes the "HOG" feature vector which yields the action class of the testing video.

The prior depth map-based approach does not consider dynamic movement of the body. On the other hand, the method in [45] combines both appearance and temporal features that extracted by an extended version of motion history image.

An another framework for recognize human activity based on depth maps proposed in [46]. It employs the "local gradient auto-correlations (GLAC)" [47] to extract shift-invariant image features from DMMs of depth map images. The "GLAC" descriptor is relying on the 2nd order of gradients. It can extract rich information from images. This work based on the "extreme learning machine (ELM)" [46,48] is introduced to concatenate the GLAC features from DMMs to recognize human actions. "ELM" is "single hidden layer feed-forward neural networks (SLFNs)". It has been effectively utilized in different applications [49,50]. Although DMMs obtained using all depth image sequence can represent the motion and shape of a depth sequence, the temporal information could not be captured in a subdivision of depth images. Therefore, the authors introduce a new framework [49,51] based 2D and 3D auto-correlation of gradients features to extend their work in [46]. Fig. 16 summarizes the proposed action recognition method. They use another feature extraction method named "space-time autocorrelation of gradients (STACOG) [52]. The "STACOG" feature is an adopted version of "GLAC" in 3D space and was developed for RGB video. Finally, a weighted fusion combines the "GLAC-STACOG" features based on "ELM" in order to recognize actions.

Depth motion maps (DMMs) have demonstrated viability in human action recognition; but, they lose the temporal

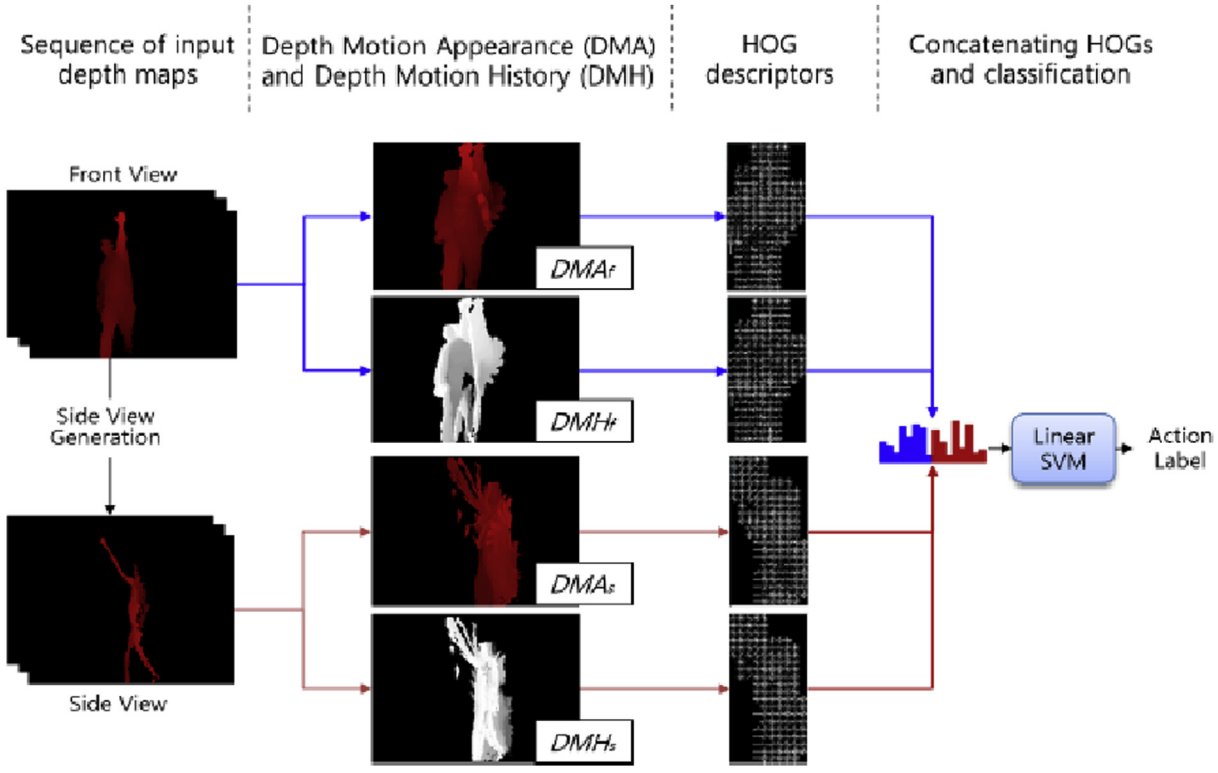

Fig. 15. The proposed framework in [45].

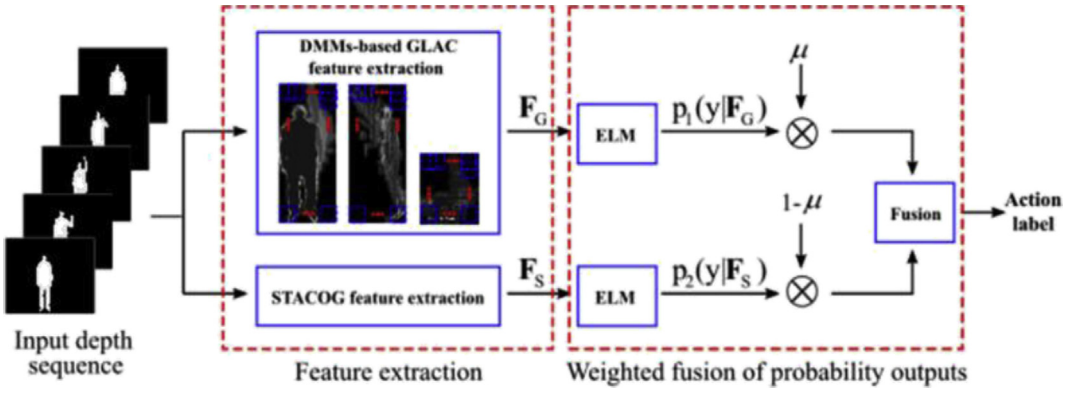

Fig. 16. Action recognition method based gradients features [51].

information and suffer from intra-class varieties caused by movements speed varieties. To address these difficulties, for human action recognition using depth map sequences, a framework introduced in called "Hierarchical Depth Motion Maps (HDMM)" and "Convolutional Neural Networks (3ConvNets)" [53]. They rotate the original depth data in 3D point clouds to mimic the rotation of cameras so that it can deal with variation cases. Next, to extract effectively the body shape and movement information weighted "DMM" is generated at several temporal scales which referred as "HDMM". Then, three channels of ConvNets are trained on the "HDMMs" from three projected orthogonal planes independently.

The most recent descriptor for human action recognition proposed in [54], it called Adaptive Hierarchical Depth Motion Maps (AH-DMMs). Fig. 17 is a specific example of generating AHDMMs. The AH-DMMs are calculated over multi-size temporal hierarchical windows of a video sequence; therefore they encode more details of motion and shape information which lost in DMMs. Meanwhile, by using motion energy based segmentation strategy, adaptive windows and steps are generated, making the AH-DMMs robust to action speed variations. Then, Gabor features encoding the texture information of AH-DMMs are extracted to improve the discriminative ability of the descriptors further. Third, after reducing dimensions by PCA, the final representations are classified by l2-regularized CRC. Compared with DMMs, the AH-DMMs encode temporal information of action sequences, more details of motion and more discriminative shape clues can be involved.

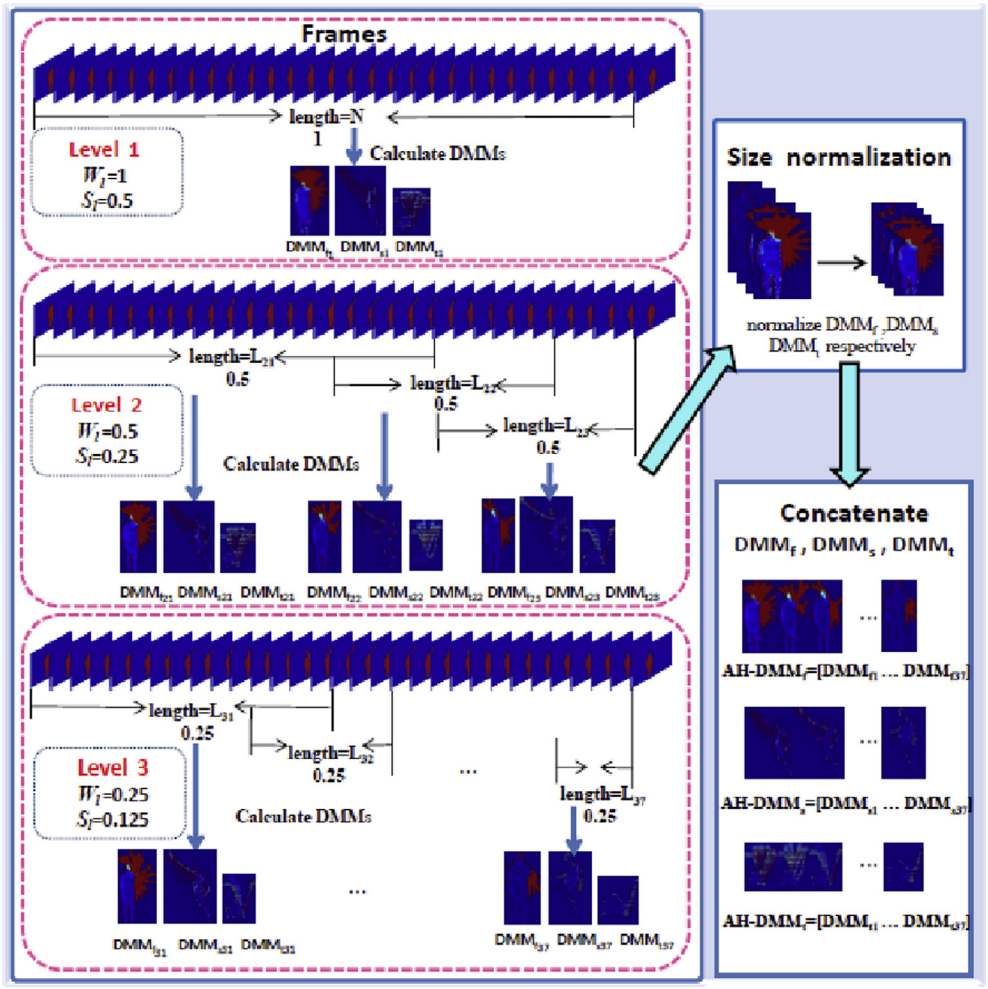

Fig. 17. The generation progress of AH-DMMs with three levels from depth sequence [54].

2.7. Depth and color features

Color data provide the visual appearance of actions, while, depth data supply the structural information. Work demonstrated in [55] combine depth and color data. It presents a successful model for tracking the association among a human hand and equipment in kitchen, for example, blending with water and cleaning vegetables. It studies both object and action recognition using object tracking techniques. The framework utilizes the "SIFT feature" from color and depth images. These features are considered as input to train SVM. They use PCA on the gradients of 3D hand trajectories to extract the global features. A hand is tracked using skin color, local feature is represented as "bag-of-words" of gradients.

Authors in [56] also, uses both depth and color images they have used various extracting interest points methods and made accuracy comparison of it. Finally, their work showed that the best results achieved when combining interest points extracted from the RGB channel and depth map features as illustrated in Fig. 18.

In [57], proposed a home activities benchmark dataset named as "RGBD-HuDaAct", using both a color video camera and a depth sensor. Two state-of-the-art image representation methodologies for action recognition are combined. "Spatio-temporal interest points (STIPs)" color image and "motion history images (MHIs)" are extracted from color and depth images respectively shown in Fig. 19.

An adaptive learning approach [58] automatically extract "Spatio-temporal features," and also combines the RGB and depth features, from RGBD video. The outline of this method illustrated in Fig. 20. "Graph-based genetic programming (RGGP)" methodology is presented, a set of primitive 3D operators is first randomly constructed as combinations and then grew generation by generation by assessing on a collection of RGBD video sequences.

Most recently framework combines depth and color cues exhibited in [59] and calculation to break down RGBD recordings caught from the robot interfacing with people. Four unique descriptors that have appeared to perform well in movement in activity recognition tasks: "3D optical flow", "Spatio−temporal interesting points" in RGB data, "depth data", and "body posture descriptors". By combining those features aims at generating a mechanism like when humans experience identified activities. Then a "Bag-of-Words" histogram for each

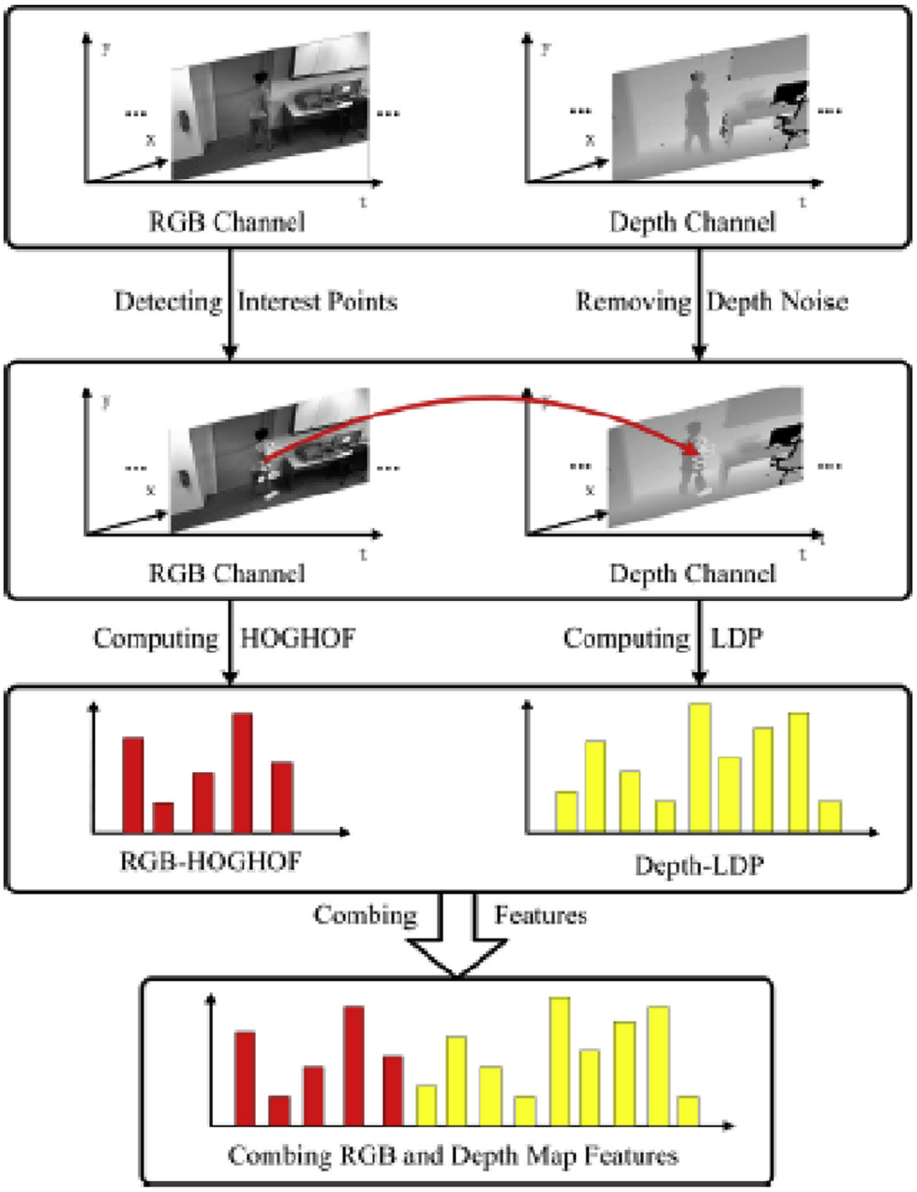

Fig. 18. The framework of combining both the RGB- and depth-map descriptor [56].

sort of feature has been created. SVM classifier is intended to incorporate every one of the descriptors effectively, in particular conditions, to privilege a descriptor concerning another one. A hopeful approach intending to assign different weights to different types of features is the "multi-channel kernel".

Recent work in [60,61] have applied the deep learning concept. An extensive scale dataset for RGB + Depth actions with more than 56 thousand video tests are presented [60] in. Their dataset contains 60 distinctive activity classes. A temporal features for each body part is modeled using recurrent neural network, and achieved better classification results.

3. Experiments and discussions

In this study, we aimed at showing a comparative perspective between different feature extraction techniques as shown in Table 1 that utilized in-depth map based activity recognition. These techniques are local interest points, occupancy patterns, depth silhouette, surface normal and depth motion maps.

3.1. 3D points (BOPs) and spatial–temporal features

"Bag of 3D Points" [12]; to represent the 3D structure of every posture in a salient state, it utilizes few number of 3D points beside uses a graph to characterize the main postures in actions. This approach suffered from losing of spatial context information between interest points. Depth Cuboid Similarity Feature "DCSF" [11] where a filtering technique extricate "STIPs" from depth sequence (called "DSTIP") that suppresses the noise. It may not be to view the silhouettes from side-view and top-view reliable because of noise and occlusions in depth images. That makes it hard to sample the interest points given the geometry and movement varieties crosswise over various people.

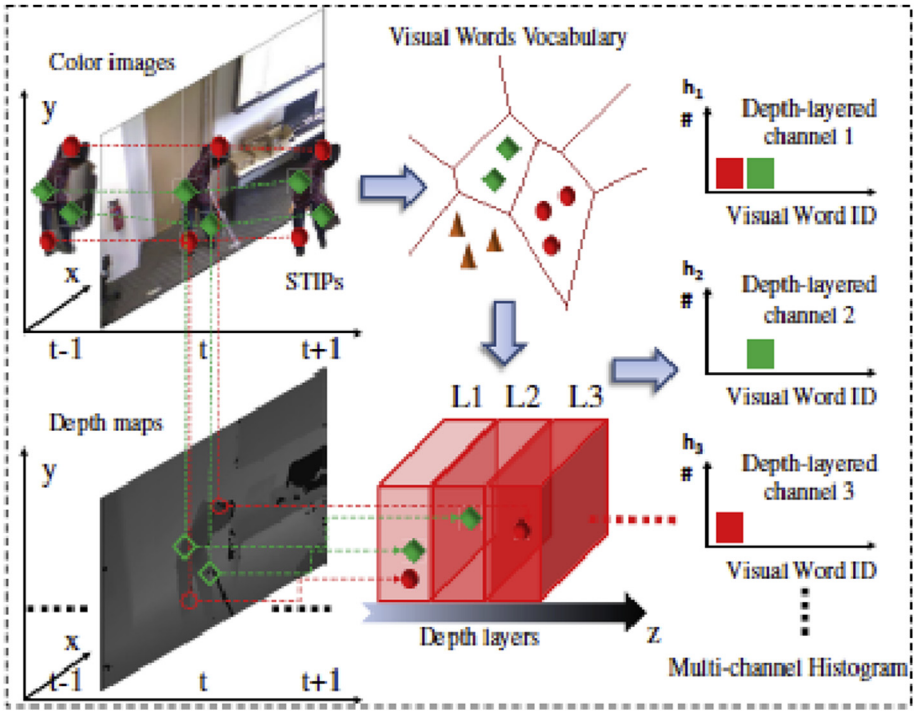

Fig. 19. The generation process of DLMC-STIP representation in [57].

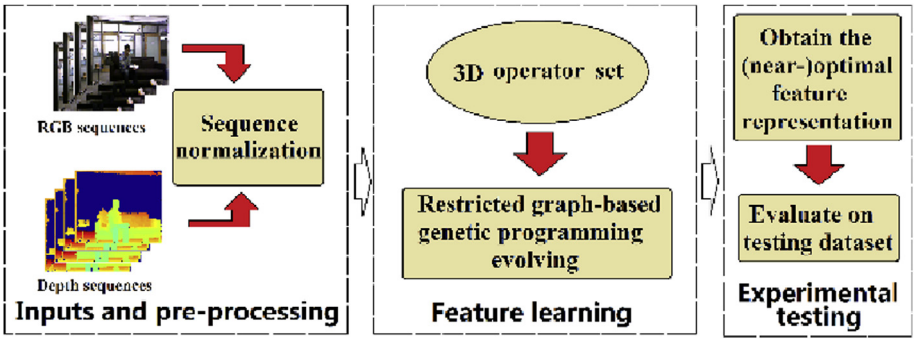

Fig. 20. The main flowchart for our proposed method [58].

3.2. Random Occupancy Pattern (ROP) features

"STOP" feature [14], it maintains both spatial and temporal contextual relation among space-time cells while being adaptable enough to handle intra-action varieties. Random Occupancy Patterns [23], where the depth maps are randomly sampled then the commonly characterized samples are picked and used as a descriptor. It also utilizes a sparse coding approach to encoding these features. The occupancy features are utilized in "a real-time system for a dynamic hand gesture" [31]. Although the silhouette features usually work better than cell occupancy features in hand gesture recognition because the most discriminative information about the hand shape is encoded in the outline of the hand.

3.3. Depth silhouette

Methods based depth silhouette [25,26,31]; for a binary human silhouette, "R transform" [26], is utilized to describe low-level features. The complexity benefit of the "R-transform" in both computation and geometric invariance is obvious. Although, the binary silhouettes only provide the shape information of actions. 2D directional shape feature map is calculated first through Radon transform of every depth silhouette [25], and then a 1D feature profile, that is translation and scaling invariant, gets calculated using "R transform". Action Graph based on Silhouette [31] presented a real-time recognition system by a depth camera.

3.4. Surface normals features

Descriptors depending on the surface normal. Histogram of oriented 4D normal represented in [32], it describes the depth sequence based on the histogram to catch the distribution of the surface in "time, depth, and spatial coordinates" space. SNV [36] which cluster hypersurface normals to generate the "polynormal" which is used to represent the local motion and shape information jointly. To capture the global spatial and

Table 1
Comparison between feature extraction techniques.

Methods based on	Method	Features	Representation	Classifier
Interest points	DCSF [13]	Depth cuboid similarity "DCSF"	PCA + kmeans clustering	SVM
	DCSF + joint [11]	DCSF + joint position feature	PCA + kmeans clustering	SVM
	Bag of 3D points [12]	3D points	2d projection	Action graph
Occupancy features	"STOP" feature [14]	STOP	PCA	Action graph
	ROPs [23]	ROP	PCA	SVM
	ROPs [23]	ROP + Haar feature	PCA	SVM
	ROPs [23]	ROP	Sparse coding	SVM
	Action graph on occupancy features [31]	Occupancy features	PCA	Action graph
Silhouettes	Binary silhouettes [25]	LDA on PCA- R features	LBG algorithm	HMM
	Depth silhouettes [25]	PCA- R feature	LBG algorithm	HMM
Surface normals	HON4D + Ddisc [32]	HON4D	Histogram	SVM + discriminative density Ddisc
	HON4D [32]	HON4D	Histogram	SVM
	SNV [36]	A joint trajectory	Fisher kernel	Gaussian mixture model (GMM)
	Polynormals [38]	PFV	Polynormal Fisher vector	Gaussian mixture model
Depth motion maps 'DMMs'	DMM-HOG [39]	DMM	HOG	SVM
	DMM-l2-regularized [41]	DMM	PCA	l2-regularized CRC
	DMM-LBP-FF [44]	LBP	PCA	KELM classifier
	DMM-LBP-DF [44]	LBP	PCA	KELM classifier
	DMA + DMH + HOG [45]	DMA + DMH	HOG	SVM
	DMMs-based GLAC [46]	DMM	GLAC	ELM classifier
	DMM- STACOG + GLAC [51]	DMM	STACOG, GLAC	ELM classifier
	HDMM + 3CONVNETS [53]	DMM	Sampling + HDMM	ConvNets-Neural Network
	AH-DMMs + Gabor [54]	DMM	Gabor filter + PCA	l2-regularized CRC

temporal orders, an adaptive spatiotemporal pyramid is used to extract set of space-time grids by subdivided a depth video. PFV [38] also follows this direction. It assembles local neighboring hypersurface normals from a depth sequence to form the "polynormal" which jointly represents movement and shape features. Fisher vector is applied to combine the low-level "polynormals" into the one feature descriptor.

3.5. Depth motion maps

The "depth motion maps (DMMs)" generated by gathering motion energy of projected depth maps are used as feature descriptors. "DMMs" are 2D images that supply an encoding of movement attributes of an action. Depth Motion Maps [39,41,44], where motion maps are acquired by summation the subtraction result of the depth frames. This descriptor in [41] utilized "DMMs" to capture the motion cues of activities, whereas "LBP" histogram features were utilized to accomplish minimized representation of "DMMs". Both feature level and decision-level fusion approaches were considered which included "kernel-based extreme learning machine (KELM)" classification. "Hierarchical Depth Motion Maps (HDMM)" [52] is present a weighted depth motion maps at several temporal scales. "AH-DMMs" [53] can capture more details of motion and shape clues by preserving the temporal information of actions. Meanwhile, the "AH-DMMs" are adaptive to action speed variations for using energy-based hierarchical structure. Gabor filter is then adopted to encode texture information of AHDMMs and generates more compact action representations.

The "DMMs-based GLAC" [46] features are utilized to capture the rich surface data from the DMMs of a sequence of depth images. The "STACOG" [51], which is a 3D adopted version of the "GLAC" feature descriptor, describes the space-time motion shape of a depth sequence. It likewise bring more depth sequence temporal features that it has been lost in the "DMMs". A weighted combination method based on "ELM" was introduced in [50] to give greater adaptability in grouping the two sets of features.

4. State-of-art datasets results

In this section, the main benchmark depth sequence datasets for action, gesture and activity recognition are depicted. All of these datasets included here are a large and different repertoire of different actions or activities that can be applied to different contexts or situations.

4.1. MSR action 3D dataset

"MSR Action 3D" dataset [12] is an depth action dataset of acquired by a depth camera. The depth map images are well-segmented, there are no objects in the background, and the person appears at the same distance to the camera. It includes twenty actions are "horizontal arm wave", "high arm wave", "two hand wave", "hand catch", "hammer", "forward punch", "high throw", "draw x", "draw circle", "draw tick", "hand clap", "side-boxing", "bend", "sidekick", "forward kick", "jogging", "tennis swing", "tennis serve", "golf swing", "pick up" and "throw". Every action was played by ten persons for 3 times. The video resolution is 640 × 480, and the frame rate is 15 frames/second. Fig. 21 demonstrates the example of the dataset.

"Polynormal Fisher Vector" [38] achieved 92.73%, results show the recognition advantages from the global temporal context. The approaches relying on joints are exposed to joints

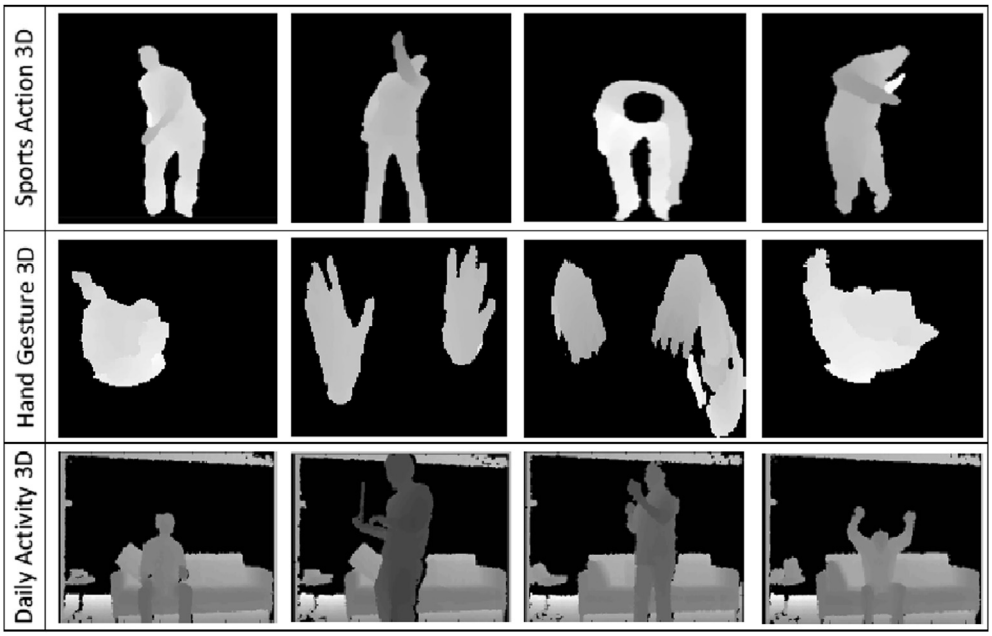

Fig. 21. Example depth frames from "MSRAction3D" dataset [12], "MSRH and Gesture" dataset [31], and "MSRDailyActivity3D" [33].

errors regarding to severe self-occlusions. The approaches in [14,33] still enhance the accuracy in [12] because cloud points are more stable and present extra shape characteristics. SNV [36] achieved an accuracy of 93.09% which significantly outperforms all methods that depend on surface normals. Although SNV and HON4D methods are based on hypersurface normals, SNV outperforms HON4D [32] by 4.20%. "DMM-HOG" [39] achieved an accuracy 94.6% which outperforms all previous methods. The accuracy goes down to 93% in Depth Motion Maps-based Local Binary Patterns [44]. It is easy to see that using both GLAC and STACOG features in [51] shows over 4% higher accuracy than using GLAC features in [46] only. Finally, deep convolution neural networks [51,52] achieved an accuracy 100%; it outperforms all the previous approaches; this is mostly because it can easily segment subject by thresholding the depth values, generating HDMM without much noise; also Pre-trained models can initialize the image-based deep neural networks well.

4.2. MSR gesture 3D dataset

The "MSR Gesture3D" dataset [31] is a depth sequences dataset for hand gestures. This dataset contains group of "American Sign Language (ASL)" gestures. There are twelve different gestures: "finish", "green", "milk", "hungry", "past", "blue", "pig", "store", "where", "letter j", "letter z", and "bathroom". Some depth sequences example are demonstrated in Fig. 21. Notice that despite the fact that this dataset includes the depth and color sequences, the deph images only are utilized as a part of the experiments. There are ten subjects, each playing out each motion 2–3 times. The dataset consists of 336 depth sequences totally. The self-occlusion is generally common with ASL dataset.

"Polynormal" Fisher Vector [38] achieved an accuracy of 95.83% which outperforms all compared approaches as demonstrated in Table 2. "DMM-GLAC-ELM" [46] also achieved high accuracy of 95.5%. The accuracy goes down to 94.74% in SNV [36]. PFV outperforms methods that depend on Occupancy features [23,31] by 7.33%, "DMM-LBP-DF" [44] by 1.23%. PFV [38] outperforms SNV by 1.09%. PFV and SNV hypersurface normals based methods achieve these results because "polynormals" obtain more discriminative local motion and shape information; in addition, Fisher vector is utilized to concatenate the low-level "polynormals" into the

Table 2
Accuracy of Compared feature extraction techniques.

Method	"MSR Action 3D dataset"	"MSR Hand gesture dataset"	"MSR Daily activity 3D dataset"
DCSF [13]	89.30%	×	83.60%
DCSF+Joint [11]	×	×	88.20%
Bag of 3D points [12]	74.70%	×	×
"STOP feature [14]	84.80%	×	×
ROPs [23]	85.92%	86.80%	×
ROPs [23]	86.50%	×	×
ROPs [23]	86.20%	88.50%	×
Action graphon occupancy features [31]	×	88.50%	×
Binary silhouettes [25]	×	×	85.75%
Depth silhouettes [25]	×	×	96.55%
HON4D + Ddisc [32]	88.89%	92.45%	×
HON4D [32]	85.85%	87.29%	×
SNV [36]	93.09%	94.74%	86.25%
Polynormals [38]	92.73%	95.83%	×
DMM-HOG [39]	94.60%	×	×
DMM-l2-regularized [41]	90.50%	×	×
DMM-LBP-FF [44]	91.90%	93.40%	×
DMM-LBP-DF [44]	93%	94.60%	×
DMA + DMH + HOG [45]	90.45%	×	×
DMMs-based GLAC [46]	90.48%	95.50%	×
DMM- STACOG + GLAC [51]	94.87%	98.50%	81.88%
HDMM + 3CONVNETS [53]	100%		×
AH-DMMs + Gabor [54]	94.18%	×	×

"Polynormal Fisher Vector". It is observed that, by using convolution neural network [53], the overall recognition accuracy beats all comparison methods, leading to almost 2.67% growth over the next best result 95.83% in PFV [38].

4.3. MSR daily activity 3D dataset

The "MSR Daily Activity 3D" dataset [30,33] is a depth sequences dataset contains daily activities. The dataset is surroundings objects, and humans show up at various distances to the camera. The most of activities is "human–object interaction". There are sixteen different activities: "drink", "eat", "read a book", "cell phone call", "write on paper", "use a laptop", "cheer up", "sit still", "toss paper", "play a game", "lie down on the sofa", "walking", "play guitar", "stand up" and "sit down". There are ten different persons, and each person plays each activity two times in two situations, "standing" and the "sitting" positions. The activity player in this dataset presents significant variations in spatial and scaling. Furthermore, most activities in this dataset include interactions with objects. Fig. 21 shows a sample of the dataset.

The accuracy result of Depth Silhouettes [25] shows 96.55% in the mean recognition rate over 10 typical home activities whereas using binary silhouettes the system achieved only 85.75%; binary silhouettes only provide the shape information of activities. The system should be useful as a smart HAR system for smart homes. "Spatio–temporal Depth Cuboid Similarity Feature" [11] makes accuracy 88.2%. It accomplishes an accuracy of 88.2% which is greater than SNV [36] archives 86.25% SNV descriptor describes both local movements and shape features in "polynormals" which in the high level encode the motion of hand and shape of the object.

4.4. RGBD-HuDa act dataset

"RGBD-HuDaAct" [57] is dataset for activities captured by a "Kinect" shown in Fig. 22. This database includes twelve activities: "stand up", "sit down", "make a phone call", "enter the room", "exit the room", "mop the floor to bed", "get up", "drink water", "eat a meal", "put on the jacket", and "take off the jacket". This dataset organized into 14 daily activities, 30 persons are performing. Every activity video is about 30–150 s repeated at maximum 4 times by every person. There are 1189 labeled video samples in this dataset.

In the depth map, more bright pixels means more depth values. As a result of surface reflections, some black regions cause depth measurement errors [55].

When gathering RGB and depth features [56], the various method performances of extracting interest points are compared. Also, it shows that the best accuracy is accomplished when extracting when combine the "RGB-based" interest points with the "depth-based" descriptor.

The accuracy comparison is in Table 3. .It shows various sorts of combinations of "RGB" and "depth-map" descriptors;

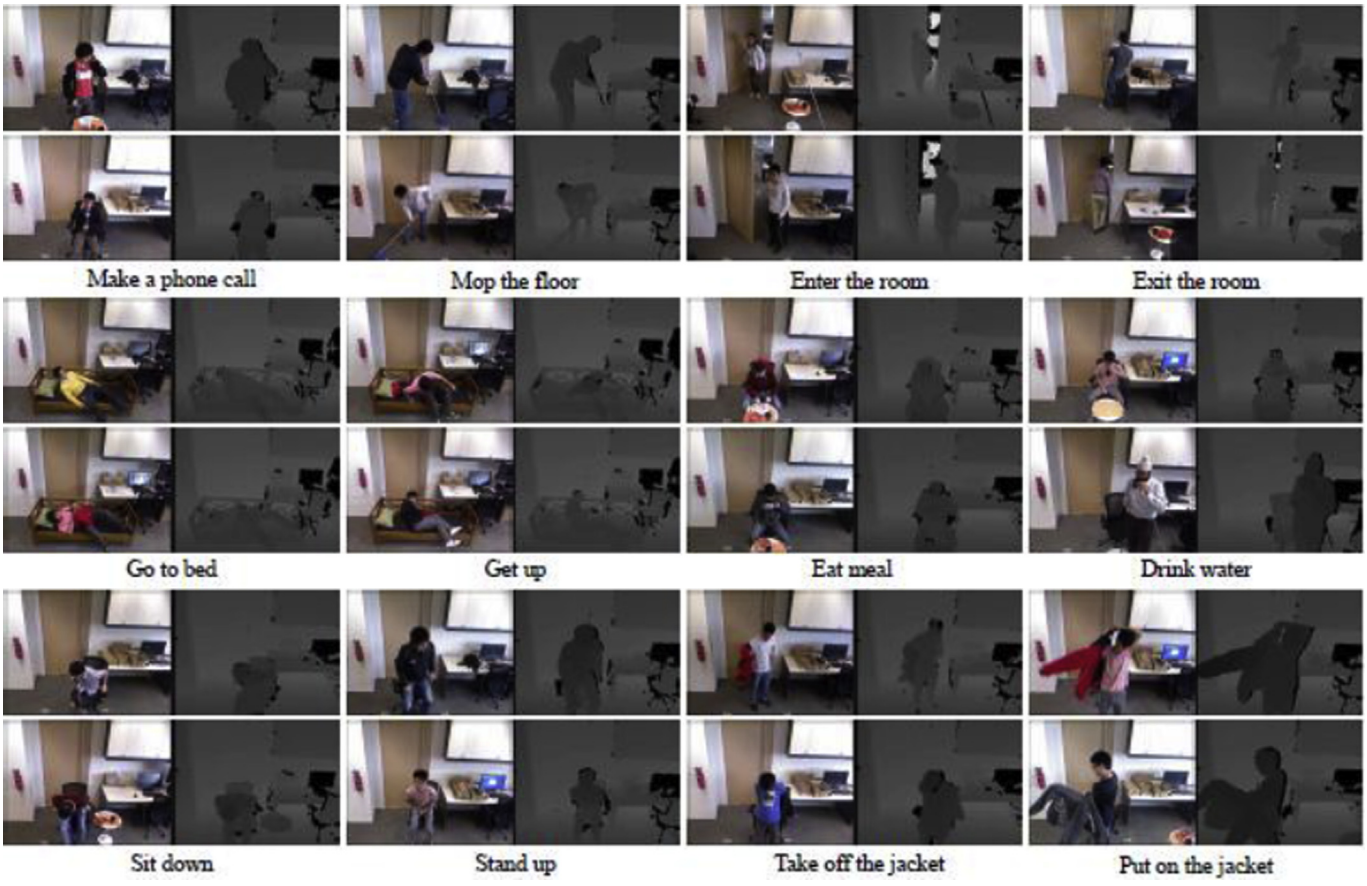

Fig 22. Color and depth frames examples [57].

Table 3
Accuracy comparison on "RGBD-HuDaAct".

RGB-Depth Method	Accuracy
"RGB(IP, HOGHOF), Depth (LDP)" [56]	89.1%
"RGB (IP, HOGHOF), Depth (HOGHOF)" [56]	83.3%
"[Depth (IP), RGB(HOGHOF)], Depth (HOGHOF)" [56]	81.8%
"DLMC(STIPs)" [57]	81.5%

The "DLMC (STIPs)" is 81.5% recognition accuracy. It can be seen that accuracy of "Depth (LDP)" is higher than "Depth-(HOGHOF)". It also shows that the accuracy of "RGB(IP)" is higher than "Depth (IP)", thus "RGB(IP), RGB(HOGHOF), RGB(IP), Depth (LDP)" has the higher accuracy achieved an accuracy of 89.1%.

5. Conclusion

Recently, depth data has been received attention in the human activity recognition field. The main benefits of developing applications using depth map based compared to vision-based are; it is more robust to lighting changes, especially in indoor situations and it has resolved the scale-distance of 2d sensors, making it simpler to create ongoing real-time systems.

This paper studied the different approaches in depth-map activity recognition. Also, it focused on detailed literature about various image representation and feature extraction techniques as a part of activity recognition. The results have been discussed the feature descriptions for human activity recognition using public datasets. For action recognition, the depth motion maps are most effective feature representation technique. Inspired by the great achievement of deep classification model for action recognition using depth map sequences, By rotation and temporal scaling, the volume of training data can be artificially enlarged, from which the convolutional neural networks benefit and obtain better results than training on primitive. The way of pre-trained and fine-tuning is adopted to train ConvNets on small datasets, which achieved best results in most cases. While in gesture recognition, an extra "space-time auto-correlation of gradients" features are also extracted from depth image sequence as corresponding features to cope the loss of the temporal information in the generating of DMMs which achieved best results. The "Polynormal" fisher vector also achieved effective results. Otherwise, PCA and LDA of depth silhouette is a practical approach in daily activity recognition.

References

[1] Shah M, Javed O, Shafique K. Automated visual surveillance in realistic scenarios. IEEE Multimedia 2007;14(1):30–9.
[2] Aggarwal J, Ryoo M. Human activity analysis. ACM Comput Surv Jan. 2011;43(3):1–43.
[3] Chen L, Khalil I. Activity recognition: approaches, practices and trends. In: Activity recognition in pervasive intelligent environments Atlantis ambient and pervasive intelligence, vol. 4; 2011. p. 1–31.
[4] Ivanov Y, Bobick A. Recognition of visual activities and interactions by stochastic parsing. IEEE Trans Pattern Anal Mach Intell 2000;22(8): 852–72.
[5] Stauffer C, Grimson WEL. Learning patterns of activity using real-time tracking. IEEE Trans Pattern Anal Mach Intell Aug 2000;22(8):747–57.
[6] Bodor R, Jackson B, Papanikolopoulos N. Vision-Based Human Tracking and Activity Recognition. Proc. of 11th Mediterranean Conf. on Control and Automation; June 2003.
[7] Ye M, Zhang Q, Wang L, Zhu J, Yang R, Gall J. A Survey on Human Motion Analysis from Depth Data. Lecture Notes in Computer Science Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications; 2013. p. 149–187.
[8] Moeslund TB, Hilton A, Krüger V. A survey of advances in vision-based human motion capture and analysis. Comput Vis Image Understand 2006;104(2–3):90–126.
[9] Poppe R. A survey on vision-based human action recognition. Image Vis Comput 2010;28(6):976–90.
[10] Cameron J, Lasenby J. Estimating human skeleton parameters and configuration in real-time from markered optical motion capture. In: Perales F, Fisher R, editors. Articulated motion and deformable objects, volume 5098 of lecture notes in computer Science; 2008. p. 92–101.
[11] Xia L and Aggarwal J. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. Proc. IEEE Conf. Comput. Vis. Pattern Recog.; 2013. p. 2834–2841.
[12] Li W, Zhang Z, Liu Z. Action recognition based on a bag of 3D points. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops. (San Francisco, CA); 2010. p. 9–14.
[13] Li W, Zhang Z, Liu Z. Expandable data-driven graphical modeling of human actions based on salient postures. IEEE Trans Circ Syst Video Technol 2008;18(11):1499–510.
[14] Vieira AW, Nascimento ER, Oliveira GL, Liu Z, Campos MFM. STOP: space-Time Occupancy Patterns for 3D action recognition from depth map sequences. In: Progress in pattern recognition, image analysis, computer vision, and applications. CIARP 2012, vol. 7441. Berlin, Heidelberg: Springer; 2012. Lecture Notes in Computer Science.
[15] Oliveira GL, Nascimento ER, Vieira AW, Campos MFM. Sparse Spatial Coding: a novel approach for efficient and accurate object recognition. 2012 IEEE Int Conf Robot Autom 2012.
[16] Abdi HCA, Williams LJ. Principal component analysis. Wiley Interdiscipl Rev Comput Stat 2010;2(4):433–59.
[17] Laptev I. On space-time interest points. Int J Comput Vis 2005;64(2–3): 107–23.
[18] Dollar P, Rabaud V, Cottrell G, Belongie S. Behavior Recognition via Sparse Spatio-Temporal Features. 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance; 2005.
[19] Laptev I, Marszalek M, Schmid C, Rozenfeld B. Learning realistic human actions from movies. 2008 IEEE Conference on Computer Vision and Pattern Recognition; 2008.
[20] Klaeser A, Marszalek M, Schmid C. A spatio-temporal descriptor based on 3d-gradients. Procedings of the British Machine Vision Conference 2008; 2008.
[21] Willems G, Tuytelaars T, Gool LV. An efficient dense and scale-invariant spatio-temporal interest point detector. Lecture Notes in Computer Science Computer Vision – ECCV 2008; 2008. p. 650–663.
[22] Cheng Z, Qin L, Ye Y, Huang Q, Tian Q. Human daily action analysis with multi-view and color-depth data. Computer Vision – ECCV 2012. Workshops and Demonstrations Lecture Notes in Computer Science; 2012. p. 52–61.
[23] Wang J, Liu Z, Chorowski J, Chen Z, Wu Y. Robust 3D action recognition with random occupancy patterns. Computer Vision – ECCV 2012 Lecture Notes in Computer Science; 2012. p. 872–885.
[24] Lee H, Battle A, Raina R, Ng AY. Efficient sparse coding algorithms. Proc. 19th Ann. Conf. Neural Information Processing Systems; 2007. p. 801–808.
[25] Jalal A, Uddin MZ, Kim JT, Kim T-S. Recognition of human home activities via depth silhouettes and ℜ transformation for smart homes. Indoor Built Environ 2012;21(1):184–90.
[26] Wang Y, Huang K, Tan T. Human activity recognition based on R transform. 2007 IEEE Conference on Computer Vision and Pattern Recognition; 2007.

[27] Aradhya VNM, Kumar GH, Noushath S. Fisher linear discriminant analysis and connectionist model for efficient image recognition. Studies in Computational Intelligence Speech, Audio, Image and Biomedical Signal Processing using Neural Networks. 2008. p. 269−82.
[28] Xia L, Chen CC, Aggarwal JK. View invariant human action recognition using histograms of 3D joints. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops; 2012.
[29] Khan Z, Sohn W. Abnormal human activity recognition system based on R-transform and kernel discriminant technique for elderly home care. IEEE Trans Consum Electron 2011;57(4):1843−50.
[30] Lu J, Plataniotis KN, Venetsanopoulos AN, Wang J. An efficient kernel discriminant analysis method. Pattern Recogn 2005;38(10):1788−90.
[31] Kurakin A, Zhang Z, Liu Z. A real time system for dynamic hand gesture recognition with a depth sensor. EUSIPCO; 2012.
[32] Oreifej O and Liu Z. HON4D: histogram of oriented 4D normals for activity recognition from depth sequences. 2013 IEEE Conference on Computer Vision and Pattern Recognition; 2013.
[33] Zhang H and Parker LE. 4-dimensional local spatio-temporal features for human activity recognition. 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems; 2011.
[34] Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans Pattern Anal Mach Intell July 2002;24(7):881−92.
[35] Lakshminarayanan B and Raich R. Inference in supervised latent dirichlet allocation. 2011 IEEE International Workshop on Machine Learning for Signal Processing; 2011.
[36] Yang X, Tian Y. Super normal vector for activity recognition using depth sequences. 2014 IEEE Conf Comput Vis Pattern Recogn 2014.
[37] Perronnin F, Dance C. Fisher kernels on visual vocabularies for image categorization. 2007 IEEE Conf Comput Vis Pattern Recogn 2007.
[38] Yang X and Tian Y. Polynormal fisher vector for activity recognition from depth sequences. SIGGRAPH Asia 2014 Autonomous Virtual Humans and Social Robot for Telepresence on - SIGGRAPH ASIA 14; 2014.
[39] Yang X, Zhang C, Tian Y. Recognizing actions using depth motion maps-based histograms of oriented gradients. Proceedings of the 20th ACM international conference on Multimedia - MM 12; 2012.
[40] Dalal N, Triggs B. Histograms of oriented gradients for human detection. IEEE Comput Soc Conf Comput Vis Pattern Recogn (CVPR05) 2005.
[41] Megavannan V, Agarwal B, Babu RV. Human action recognition using depth maps. 2012 International Conference on Signal Processing and Communications (SPCOM); 2012.
[42] Chen C, Liu K, Kehtarnavaz N. Real-time human action recognition based on depth motion maps. J Real Time Image Process Nov. 2013; 12(1):155−63.
[43] Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 2002;24(7):971−87.
[44] Chen C, Jafari R, Kehtarnavaz N. Action recognition from depth sequences using depth motion maps-based local binary patterns. 2015 IEEE Winter Conference on Applications of Computer Vision; 2015.
[45] Kim D, Yun W, Yoon H, Kim J. Action Recognition with Depth Maps Using HOG Descriptors of Multi-view Motion Appearance and History. The Eighth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, UBICOMM; 2014.
[46] Chen C, Hou Z, Zhang B, Jiang J, Yang Y. Gradient local auto-correlations and extreme learning machine for depth-based activity recognition. Advances in Visual Computing Lecture Notes in Computer Science; 2015. p. 613−623.
[47] Kobayashi T, Otsu N. Image feature extraction using gradient local auto-correlations. Lecture Notes in Computer Science Computer Vision − ECCV 2008; 2008. p. 346−358.
[48] Huang G-B, Zhu Q-Y, Siew C-K. Extreme learning machine: theory and applications. Neurocomputing 2006;70(1−3):489−501.
[49] Chen C, Zhou L, Guo J, Li W, Su H, Guo F. Gabor-filtering-based completed local binary Patterns for Land-Use Scene Classification. 2015 IEEE International Conference on Multimedia Big Data; 2015.
[50] Li W, Chen C, Su H, Du Q. Local binary patterns and extreme learning machine for hyperspectral imagery classification. IEEE Trans Geosci Rem Sens 2015;53(7):3681−93.
[51] Chen C, Zhang B, Hou Z, Jiang J, Liu M, Yang Y. Action recognition from depth sequences using weighted fusion of 2D and 3D auto-correlation of gradients features. Multimed Tool Appl 2016;76(3): 4651−69.
[52] Kobayashi T, Otsu N. Motion recognition using local auto-correlation of space−time gradients. Pattern Recogn Lett 2012;33(9):1188−95.
[53] Wang P, Li W, Gao Z, Zhang J, Tang C, Ogunbona P. Deep convolutional neural networks for action recognition using depth map sequences. arXiv preprint arXiv:1501.04686; 2015.
[54] Liu H, He Q, Liu M. Human action recognition using adaptive hierarchical depth motion maps and gabor filter. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.
[55] Lei J, Ren X, Fox D. Fine-grained kitchen activity recognition using RGB-D. Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp 12; 2012.
[56] Yang Z, Liu Z, Yang L, Cheng H. Combing rgb and depth map features for human activity recognition. Signal Information Processing Association Annual Summit and Conference (APSIPA ASC); 2012. p. 1−4.
[57] Ni B, Wang G, Moulin P. RGBD-HuDaAct: a color-depth video database for human daily activity recognition. Consumer Depth Cameras for Computer Vision; 2013. p. 193−208.
[58] Liu L, Shao L. Learning discriminative representations from RGB-D video data, Proc. IJCAI; 2013.
[59] Xia L, Gori I, Aggarwal J, Ryoo M. Robot-centric activity recognition from first-person RGB-d videos. 2015 IEEE Winter Conference on Applications of Computer Vision; 2015.
[60] Shahroudy A, Liu J, Ng TT, Wang G. NTU RGB+D: a large scale dataset for 3D human activity analysis. 2016 IEEE conference on computer vision and pattern recognition (CVPR). (Las Vegas, NV); 2016. p. 1010−1019.
[61] Wang P, Li W, Gao Zhimin, Zhang Y, Tang C, Ogunbona P. Scene flow to action map: a new representation for RGB-D based action recognition with convolutional neural networks. arXiv preprint arXiv:1702.08652; 2017.