Chapter 5

Human Action Recognition with RGB-D Sensors

Enea Cippitelli, Ennio Gambi and Susanna Spinsante

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/68121

Abstract

Human action recognition, also known as HAR, is at the foundation of many different applications related to behavioral analysis, surveillance, and safety, thus it has been a very active research area in the last years. The release of inexpensive RGB-D sensors fostered researchers working in this field because depth data simplify the processing of visual data that could be otherwise difficult using classic RGB devices. Furthermore, the availability of depth data allows to implement solutions that are unobtrusive and privacy preserving with respect to classic video-based analysis. In this scenario, the aim of this chapter is to review the most salient techniques for HAR based on depth signal processing, providing some details on a specific method based on temporal pyramid of key poses, evaluated on the well-known MSR Action3D dataset.

Keywords: kinect, human action recognition, bag of key poses, RGB-D sensors

1. Introduction

The topic known as human action recognition (HAR) has become of interest in the last years mainly because different applications can be developed from the understanding of human behaviors. The technologies used to recognize activities can be varied and based on different approaches [1]. The use of environmental and acoustic sensors allows to infer the activity from the interaction of the user with the environment and the objects located in it, but vision-based solutions [2] and wearable devices [3] are usually the most used technologies to detect human body movements. RGB-D sensors, i.e., Red-Green-Blue and depth sensors, can be considered as enhanced vision-based devices since they can additionally provide depth data that can facilitate the detection of human movements. In fact, depth information may help to improve the performance of HAR algorithms because it is easier to implement a crucial process such as

the extraction of human silhouette, reducing its dependence from shadows, light reflections, and color similarity [4]. Skeleton joints, which can be even exploited to calculate features for action recognition, are extracted from depth data [5].

The aim of this chapter is to discuss HAR algorithms exploiting RGB-D sensors, providing a review of the most salient methods proposed in literature and an overview of nonvision-based devices. A method for HAR exploiting skeleton joints and known as temporal pyramid of key poses is described and experimental results on a well-known RGB-D dataset are provided.

Section 2 of this chapter aims to review methods for human action recognition based on different technologies, with a particular focus on RGB-D data. An algorithm based on histograms of key poses exploiting skeleton joints extracted by Kinect is presented in Section 3. Finally, the last section of the chapter highlights the main conclusion on the proposed topic.

2. Methods and technologies for HAR

HAR methods can be implemented on data gathered from different technologies, which can infer the action from the movements made by the person, or from the interaction with objects or the environment. A review of sensors and technologies for detection of different human activities in smart homes can be found in Ref. [6], where the aim is to face the phenomenon of aging population. Following the same unobtrusive approach, researchers are working also on radio-based techniques [7], where they take advantage of signal attenuation due to the body, and channel fading of wireless radio. Other works have been also published considering wearable devices, such as smartphones, that can be used to collect data and to classify actions [8]. A more general architecture implemented with wearable devices requires the usage of small sensors with sensing and communication capabilities that can acquire data (usually related to acceleration) and send them to a central unit [9].

2.1. Related works on not vision-based devices

HAR based on data generated by environmental devices in home environment may exploit unobtrusive sensors equipping objects with which people usually interact, or other sensors that are installed in the rooms. State-changes sensors, which activate and deactivate if they detect a change, can provide powerful clues about movements in the apartment if placed on windows or doors. If attached to ovens and fridges, or toilet and washing machines, they can reveal kitchen-related activities or activities associated to toileting and doing laundry [10]. Passive infrared sensors (PIRs) detect the presence of a person in a room and a set of activities can be inferred if they are jointly used with other sensors, such as state-changes sensors and flush sensors, to detect the use of the toilet [11]. Multiple binary sensors such as motion detectors, contact switches, break-beam sensors, and pressure mats have been used in Ref. [12]. Using an approach based on particle filter and an ID sensor (RFID) to detect people's identity, the system can reveal information about the occupied rooms and the number of occupants, and recognize if they are moving or not and track their movement. An integrated

http://dx.doi.org/10.5772/68121

platform including PIRs, magnetic sensors, force sensors, gas and smoke detection sensors, water and gas flux meters, power meters connected to some objects has been implemented in a laboratory environment [13]. Some simple activities, such as cooking, sitting, watching TV, can be easily inferred by processing the output data of sensors. Environmental sensors can be installed also in nursing homes, to support and help assistance of Alzheimer's disease patients [14]. In this scenario, even the detection of simple events such as "presence in bed" or "door opening" may be relevant to ensure comfort and safety of patients. Environmental sensors are completely unobtrusive and privacy preserving but they usually require some time for the installation. Furthermore, the amount of information that can be obtained from the sensors is limited, and does not include the extraction of human movements.

Other unobtrusive sensors revealing the interaction with the environment can be audio sensors. In fact, some activities generate sounds that can be captured using one or multiple microphones. Characteristic sounds are generated for example by chatting or reading newspapers activities, as well as drink and food intake events, that can be classified considering their features [15]. Tremblay et al. [16] proposed an algorithm to recognize a limited set of activities from six microphones installed at different positions in a test apartment. Two activities of daily living (ADLs), i.e., breakfast and household, constituted by multiple steps have been recognized with a promising accuracy. Multiple audio sensors in the same apartment could constitute a wireless sensor network (WSN), addressing the challenges of limited amount of memory and processing power of the nodes. However, it has been proven that low complexity features extraction algorithms can be adopted with good performance considering the indoor scenario [17]. Vuegen et al. [18] proposed a WSN constituted by seven nodes placed in different rooms: living room/kitchen, bedroom, bathroom and toilet, covering the entire apartment. A set of 10 ADLs has been recorded considering two test users and an artificial dataset to examine the influence of background noise. Acoustic sensors can be adopted in assistive environments to detect dangerous events such as falls [19, 20].

Radio-based techniques do not require any physical sensing module and they may work without the need of wearing any device, but only exploiting the existing WiFi links between the access point and connected devices. With one access point and three devices, a set of nine in-place activities (such as cooking, eating, washing dishes, etc.) and eight walking activities (distinguishing the direction of movement within the apartment) can be recognized [21]. Another radio-based technique is represented by micro-Doppler signatures (MDS). Commercial radar motes can be used to discern among a small set of activities, such as walking, running, and crawling, with high accuracy values [22]. A larger set of MDS captured from humans performing 18 movements has been collected and presented in Ref. [23]. Activities have been grouped in three categories: stationary, forward-moving and multitarget, and characterized both in free-space and through-wall environments, associating the general properties of the signatures to their phenomenological characteristics. Björklund et al. [24] included a set of five activities (crawling, creeping on their hands and knees, walking, jogging, and running) in their study. They evaluated the performance of an activity recognition algorithm based on a support vector machine (SVM) with features in the time-velocity domain and in the cadence-velocity domain, obtaining comparable results of about 90% of accuracy.

Wearable sensors can be used to extract the human movements since they usually provide acceleration data. Considering inertial data, many different features for human action recognition have been proposed, with the aim to reduce the complexity of the features extraction process and to enhance the separation among the classes [25]. Wearable inertial sensors are quite cheap and generate a limited amount of data that can be processed easily with respect to video data, even if they do not provide information about the context. The placement of wearable sensors can be an issue and this step has to be carefully addressed [26]. This choice mainly depends on the movements constituting the set of activities that have to be recognized. The placement on the waist of the subject is close to the center of mass, and can be used to represent activities involving the whole body. With this configuration, sitting, standing, and lying postures can be detected with a high degree of accuracy considering a dataset acquired in a laboratory environment [27]. The placement on the subject's waist, as well as the one on the subject's chest or knee, gives good results with transitional activities also in Ref. [28]. On the other hand, high level activities such as running (in a corridor or on a treadmill) and cycling are revealed mostly by an ear worn sensor, since it measures the change in body posture. The placement of wearable unit on the dominant wrist may help the discrimination of upper body movements constituting for example the activities of brushing teeth, vacuuming, and working at computer [29]. On the other hand, the recognition of gait-related activities, such as normal walking, stair descending, stair ascending, and so on, requires the positioning of the devices on the lower limbs. In particular, even if the shank's sensor could be enough to predict the activities, the usage of other IMUs, placed on tight, foot and waist, can enhance the final accuracy [30]. A multisensor system for activity recognition usually allows to increase the accuracy with respect to a single-sensor system, even if the latter employs a higher sampling rate, more complex features and a more sophisticated classifier [31]. The main drawback is the increasing level of obtrusiveness for the subject being monitored. Furthermore, if it may be acceptable to ask people to wear a device for a limited amount of time, for example to extract some parameters during movement assessment tests [32], it may be unacceptable to request wearing several IMUs to continuously track ADLs.

2.2. Related works on RGB-D sensors

Video-based devices (and especially RGB-D sensors) allow to extract activities from body movements but they are not obtrusive and they do not pose many issues about installation as environmental sensors do. Furthermore, RGB-D sensors do not raise problems related to radiation impact, differently from radar-based techniques, which can limit their acceptability. On the other hand, video-based sensors may be deemed not acceptable for privacy concerns but RGB-D sensors provide not negligible advantages from this point of view. In fact, when the data processing algorithms exploit only depth information, the privacy of the subject is preserved because no plain images are collected, and many details cannot be extracted from depth signal only. Different levels of privacy can be considered according to the user's preferences, thanks to the possibility to extract the human silhouette, or even to represent the human subject only by means of the skeleton [33].

Many different reviews on HAR based on vision sensors have been published in the past, each of which proposing its own taxonomy to classify different approaches [34–36]. Aggarwal and Xia [37], in their review, considered only methods based on 3D data that can be obtained

http://dx.doi.org/10.5772/68121

from three different technologies: marker-based systems, stereo images or range sensors, and organizing the papers in five categories based on the features considered.

The review of action recognition algorithms based on RGB-D sensors is organized considering the data processed by the algorithms, separating methods based on depth data from others exploiting skeleton information. Due to the simple extraction process of the silhouette from depth data, approaches based on this information may exploit features extracted from silhouettes. Li et al. [38] calculate a bag of 3D points from human silhouette, sampling the points on the contours of the planar projections of the 3D depth map. An action graph, where each node is associated to a salient posture, is adopted to explicitly model the dynamics of the actions. Features from 2D silhouettes have been considered in Ref. [39], where an action is modeled as a sequence of key poses, extracted by means of a clustering algorithm, from a training dataset. Dynamic time warping (DTW) is suitable in this case because sequences can be inconsistent in terms of time scale, but they preserve the time order, and DTW can associate an unknown sequence of key poses to the closest sequence in the training set, thus performing the recognition process. Other approaches exploiting depth data considered the extraction of local or holistic descriptors. Local spatio-temporal interest points (STIPs), which have been used with RGB data, can be adapted to depth including additional strategies to reduce the noise typical of depth data, such as the inaccurate identification of objects' borders, or the presence of holes in the frame [40]. A spatio-temporal subdivision of the space in multiple segments has been proposed in Ref. [41], where the occupancy patterns are extracted from a 4D grid. Holistic descriptors, namely histogram of oriented 4D normals (HON4D) and histogram of oriented principal components (HOPC) have been exploited respectively in Refs. [42, 43]. HON4D is based on the orientation of normal surfaces in 4D while HOPC can represent the geometric characteristics of a sequence of 3D points.

Skeleton joints represent a compact and effective description of the human body, for this reason they are assumed and exploited as input data by many action recognition algorithms. Kinect sensor provides 3D coordinates of 20 skeleton joints, thus motion trajectories in a 60-dimensional space can be associated to human motion [44]. A trajectory is the evolution of the positions of joint coordinates along a sequence of frames related to an action. A kNN classifier learns the trajectories of different actions and performs classification. Gaglio et al. [45] proposed an algorithm constituted by three steps: features detection, where the skeleton coordinates are elaborated to extract features; posture analysis, that consists in the detection of salient postures through a clustering algorithm and their classification with a support vector machine (SVM); and activity recognition, where a sequence of postures is modeled by an hidden Markov model (HMM). In Ref. [46], the coordinates of human skeleton models generate body poses and an action can be seen as a sequence of body poses over time. According to this approach, a feature vector is obtained representing each pose in a multidimensional feature space. A movement can be now represented as a trajectory in the feature space, which may constitute a signature of the associated action, if the transformation and features are carefully chosen. An effective representation based on skeleton joints is called APJ3D [47], which is built from 3D joint locations and angles. The key postures are extracted by a k-means clustering algorithm and, following a representation through an improved Fourier temporal pyramid, the recognition task is carried out with random forests. Xia et al. [48] proposed a method to compactly represent human postures with histograms of 3D joints (HOJ3D). The positions of the

joints are translated into a spherical coordinate system and, after a reprojection of the HOJ3D vectors using linear discriminant analysis (LDA), a number of key postures are extracted from training sequences. The temporal evolution of postures is modeled through HMM.

Research on HAR using RGB-D sensors has been fostered by the release of many datasets. An extensive review of the datasets collected for different purposes, going for example from camera tracking and scene reconstruction to pose estimation or semantic reasoning, can be found in Ref. [49]. Another review, which is focused on RGB-D datasets for HAR, has been published in Ref. [50]. In the latter work, the datasets have been organized considering the methods applied for data collection, which can include a single view setup, with one capturing device, a multiview setup with more devices, or a multiperson setup where some interactions among different people are included in the set of classes.

A list of the most used datasets for HAR is provided in **Table 1**, where different features of each dataset are highlighted. Many datasets provide the most important data streams available with a RGB-D device, i.e., the color and depth frames along with skeleton coordinates. They are usually featured by a number of actions between 10 and 20, performed by different subjects (around 10), and repeated 2 or 3 times. Considering the set of actions included in the datasets, they can be used for two main applications that are the detection of daily activities (DA) and the human

Name	Data	Application	Actions	Actors	Times	Samples	Citations	Year
MSR DailyActivity3D [51]	C, D, S	DA	16	10	2	320	614	2012
MSR Action3D [38]	D, S	HCI	20	10	2 or 3	567	603	2010
UTKinect Action [48]	C, D, S	HCI/DA	10	10	2	200	444	2012
MSR ActionPairs [42]	D	DA	6	10	3	180	338	2013
CAD-60 [52]	C, D, S	DA	12	2 + 2	–	60	281	2012
CAD-120 [53]	C, D, S	DA	10	2 + 2	–	120	219	2013
RGBD-HuDaAct [54]	C, D	DA	12	30	2 or 4	1189	211	2011
MSRC-12 KinectGesture [55]	S	HCI	12	30	–	594	197	2012
MSR Gesture3D [56]	D	HCI	12	10	2 or 3	336	159	2012
Berkeley MHAD [57]	C, D, M, Au, Ac	HCI	11	7 + 5	5	~660	110	2013
G3D [58]	C, D, S	HCI	20	10	3	–	61	2012
Florence 3D Action [59]	C, S	DA	9	10	2 or 3	215	54	2012
ACT4 Dataset [60]	C, D	DA	14	24	>1	6844	53	2012
LIRIS Human Activities [61]	C, D	DA	10	21	–	–	49	2012
3D Online Action [62]	C, D, S	DA	7	24	–	–	41	2014
UPCV Action [46]	S	DA	10	20	–	–	39	2014
WorkoutSu-10 Gesture [63]	D, S	DA	10	15	10	1500	32	2013

Name	Data	Application	Actions	Actors	Times	Samples	Citations	Year
KARD [45]	C, D, S	HCI/DA	18	10	3	540	23	2014
UTD-MHAD [64]	C, D, S	HCI	27	8	4	861	22	2015
IAS-Lab Action [65]	C, D, S	DA	15	12	3	540	21	2013
NTU RGB+D [66]	C, D, S, IR	HCI/DA	60	40	–	56880	14	2016

Note: In the column related to data, each label represents the availability of a different type of data: RGB (C), Depth (D), Skeleton (S), Acceleration (Ac), Audio (Au), Mocap (M). The datasets can be oriented to two main applications: Daily Activities (DA) and Human Computer Interaction (HCI).

Table 1. List of the most important RGB-D datasets for Human Action Recognition, listed considering the number of citations according to Google Scholar on January 3rd 2017.

computer interaction (HCI). Datasets belonging to the first group usually include actions like *walking, eating, drinking*, and sometimes they are recorded in a real scenario, which introduces partial occlusions and a complex background [51, 52]. Datasets focused on HCI applications may contain actions like *draw x, draw circle, side kick*, and they are usually captured with a simpler background, even if they can be challenging, due to the similarity of many gestures and to the differences in speeds and way to perform the movement, considering different actors.

The oldest and the newest datasets included in the list are deeply discussed because of their characteristics. MSR Action3D [38] was the first relevant dataset for HAR, it has been released in 2010 and it includes 20 actions that are suitable for HCI. The following activities are included in the dataset: *high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick-up, and throw*. As described in Ref. [38], the dataset has been often evaluated considering three subsets of 8 actions each, namely AS1, AS2, and AS3. As can be noticed from **Table 2**, AS1 and AS2 are built by grouping actions with similar movements, and AS3 includes actions that require more complex movements. From **Figure 1** it is possible to observe sequences of frames constituting two similar actions in AS1: *hammer* and *forward punch*. Sequences of frames from

AS1	AS2	AS3
(a02) *Horizontal arm wave*	(a01) *High arm wave*	(a06) *High throw*
(a03) *Hammer*	(a04) *Hand catch*	(a14) *Forward kick*
(a05) *Forward punch*	(a07) *Draw x*	(a15) *Side kick*
(a06) *High throw*	(a08) *Draw tick*	(a16) *Jogging*
(a10) *Hand clap*	(a09) *Draw circle*	(a17) *Tennis swing*
(a13) *Bend*	(a11) *Two-hand wave*	(a18) *Tennis serve*
(a18) *Tennis serve*	(a12) *Side boxing*	(a19) *Golf swing*
(a20) *Pick-up and throw*	(a14) *Forward kick*	(a20) *Pick-up and throw*

Table 2. Actions constituting the three subsets of MSR Action3D: AS1, AS2, AS3.

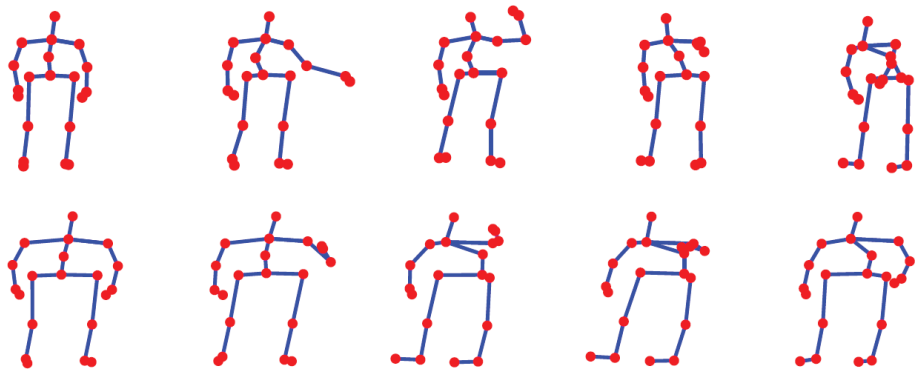

Figure 1. Sequences of frames constituting similar actions in AS1 subset of MSR Action3D: *hammer* (top) and *forward punch* (bottom).

Draw x and *Draw tick*, two similar actions in AS2, are shown in **Figure 2**. The dataset has been collected using a structured light depth sensor and the provided data are represented by depth frames, at a resolution of 320 × 240, and skeleton coordinates. The entire dataset includes 567 sequences but, considering that 10 of them are affected by wrong or missing skeletons, only 557 sequences of skeleton joint coordinates are available. The evaluation method usually adopted on this dataset is called cross-subject test [38] and takes into account samples from actors 1-3-5-7-9 for training, and the remaining data for testing. NTU RGB+D [66] is one of the most recent datasets for HAR and, to the authors' best knowledge, the largest. In fact, it includes 60 different actions that can be grouped in 40 daily actions (*reading, writing, wear jacket, take off jacket*), 9 health-related actions (*falling down, touch head, touch neck*), and 11 interactions (*walking toward each other, walking apart from each other, handshaking*). A number of 40 actors have been recruited to perform the actions multiple times, involving also 17 different setups of the Kinect v2 sensors adopted for data collection. Each

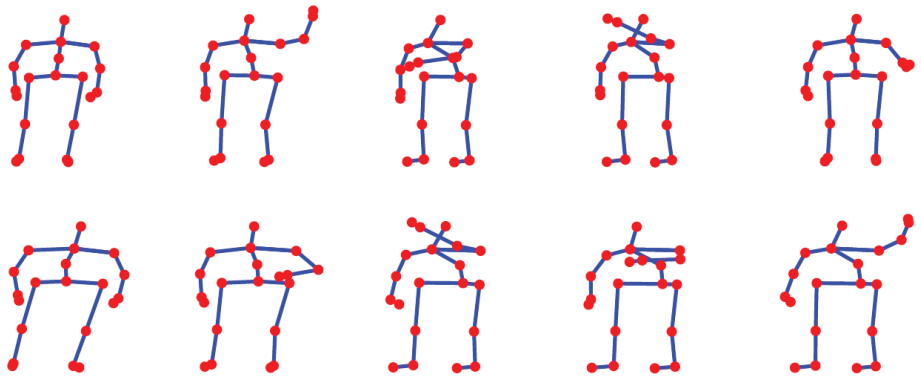

Figure 2. Sequences of frames constituting similar actions in AS2 subset of MSR Action3D: *draw x* (top) and *draw tick* (bottom).

action has been captured from three sensors simultaneously, having three different views of the same scene (0°, +45°, −45° directions). All the data provided by Kinect v2 (RGB, depth, infrared frames and skeleton coordinates) are collected and included in the released dataset. Two evaluation methods have been proposed in Ref. [66], aiming to test the goodness of HAR methods with unseen subjects and new views. In the cross-subject test, a specific list of subjects is used for training and the remaining represent the test data, while in the cross-view test the sequences from devices 2 and 3 are used for training and the ones from camera 1 are adopted for testing.

3. Human action recognition based on temporal pyramid of key poses

A HAR method that allows to achieve state-of-the-art results has been proposed in Ref. [67] and can be defined as temporal pyramid of key poses. It exploits the bag of key poses model [68] and it adopts a temporal pyramid to model the temporal structure of the key poses constituting an action sequence.

3.1. Algorithm overview

The algorithm based on temporal pyramid of key poses can be represented by the scheme shown in **Figure 3**. It performs four main steps that include the extraction of posture features, the adoption of the bag of key poses model, and the representation of the action sequence through a temporal pyramid of key poses; finally, the classification by a multiclass SVM takes place.

The algorithm takes as an input the coordinates of skeleton joints, that can be seen as a 3-dimensional vector $\mathbf{J}_i$ for the i-th joint of a body with P joints. The aim of the first step is to obtain view- and position-invariant features from the raw coordinates. The feature computation scheme derives from the one proposed in Ref. [69], but here a virtual joint called center-of-mass is introduced. Considering all the skeleton joints stored in the vector $\mathbf{P}_n$ related to the n-th frame of a sequence, the center-of-mass $\mathbf{J}_{cm}$ is calculated by averaging the coordinates of all the P joints. In order to normalize coordinates with respect to the size of the body, the normalization factor s is computed by averaging the L-2 norm between the skeleton joints and $\mathbf{J}_{cm}$, as follows:

$$s = \frac{1}{P}\sum_{i=0}^{P-1}\|\mathbf{J}_i - \mathbf{J}_{cm}\|_2 \tag{1}$$

The normalization with respect to the position of the skeleton is implemented considering the displacement between each joint position and the center-of-mass, normalized by the factor s. Each joint is thus represented by a 3 dimensional vector $\mathbf{d}_i$:

$$\mathbf{d}_i = \frac{\mathbf{J}_i - \mathbf{J}_{cm}}{s} \tag{2}$$

Finally, as can be noticed in the first part of **Figure 3**, each vector $\mathbf{p}_n$ corresponding to the coordinates of the skeleton in the n-th frame, is translated into a vector $\mathbf{f}_n$ which includes the features related to that skeleton.

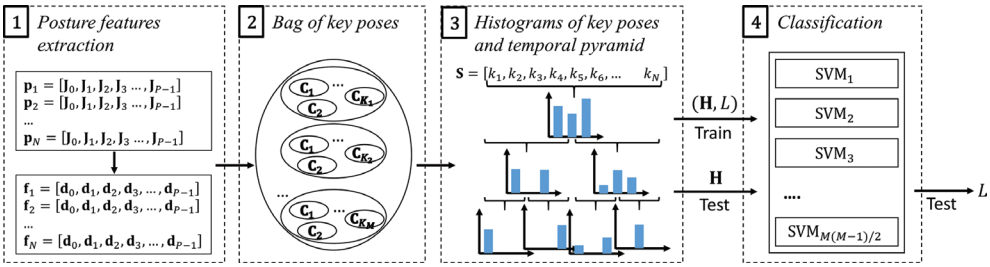

Figure 3. Global scheme of the algorithm based on temporal pyramid of key poses. Step 1 extracts the feature vectors related to posture while step 2 is represented by the bag of key poses model. The third phase exploits the temporal pyramid to model the temporal structure of the sequences and the last step is the classification phase.

Once the features related to the skeleton have been obtained, the bag of key poses method is adopted to extract the most significant postures and the action is then represented as a sequence of key poses. In more detail, the clustering algorithm *k*-means is applied considering separately the training sequences of each class, setting a different number of key poses for each action of the dataset, i.e., K_1 for class 1, K_2 for class 2, up to K_M if the dataset is constituted by M classes. Following the clustering process performed separately for each class, the key poses, which are the centres of the clusters, have to be merged to obtain a unique codebook. Finally, each posture feature vector is associated to the closest key pose in terms of Euclidean distance, and a sequence of key poses $\mathbf{S} = [k_1, k_2, k_3, \ldots, k_N]$ represents an action of N frames.

The temporal structure of an action can be represented with the adoption of a temporal pyramid. The idea is to provide different representations of the action: the most general one is provided at the first level of the pyramid, whereas the most detailed one is given at the last level. For each level, the computation of the histograms of key poses is implemented, having at the end of the process a histogram for each segment at each level. Starting from the consideration of the entire sequence at the first level of the pyramid, two segments are considered in the second level and they are split again in two at the third level, giving a number of seven histograms when three levels are considered. These histograms **H** represent the input data to the final step, which is the classifier.

The classification step aims to associate the data extracted from an unknown data sequence to the correct action label, knowing the training set. In particular, the classifier has to be trained with a set of histograms **H** for which the action labels L are known. Then, in the testing phase, an unknown **H** has to be associated to the corresponding L. A multiclass SVM has been chosen for classification purpose. The approach considered for the implementation of the multiclass scheme is defined as "one-versus-one," where a set of $M(M-1)/2$ binary SVMs are required for a dataset of M classes, each of which has to distinguish two classes. The output class is elected with a voting strategy considering the result of each binary SVM.

3.2. Experimental results and discussion

This method has been evaluated on one of the most used RGB-D dataset for HAR: MSR Action3D [38]. The test scheme adopted is the cross-subject test, described in the previous section.

http://dx.doi.org/10.5772/68121

The algorithm requires to set different parameters in order to be executed, which are the number of key poses per class (clusters), the set of skeleton joints (features) and the set of training sequences (instances). These parameters can be chosen randomly or using some optimization strategies in order to maximize the performance. In this chapter, results are shown using both the options, adopting the optimization process, based on evolutionary [70] and coevolutionary [71] algorithms. These optimization strategies are applied as wrapper methods, associating the fitness of each individual in the population to the accuracy of the action recognition algorithm.

Since the idea is to optimize three parameters, the structure of each individual is constituted by three parts [72]. The first one is related to features, and it is a binary vector of length P, which is the number of joints in a skeleton. A bin is featured by a 1 value if the associated joint has to be considered by the action recognition algorithm; otherwise it is featured by a 0 value. The same approach is used for the part related to training instances, which is therefore represented by a binary vector of length I. Regarding the optimization of the number of key poses, it is necessary to adopt a vector of integer values with a length of M, where each bin is associated to a class of the dataset, and contains the number of its clusters. Crossover and mutation operators have to be used to evolve the population's individuals, and a standard 1-point crossover operator is applied for the subindividuals related to instances and clusters. A specific crossover operator which takes into account the structure of the skeleton joints is applied to the features part. Finally, three different mutation probabilities are considered, for the three parts of the individual.

In addition to the evolutionary algorithm, a cooperative coevolutionary optimization method can be also implemented. The main difference between evolutionary and coevolutionary approaches is in the organization of the population of individuals. In particular, in the latter case, each subindividual is part of a different population, thus generating a set of three populations. The selection of one element from each population is necessary to execute the action recognition algorithm and to extract the fitness value, which is associated to each subindividual. Crossover and mutation operators can be applied according to the same considerations made for the evolutionary computation. In order to improve the performance of the optimization process, different priorities are given to the individuals of the populations. In particular, in the populations related to features and instances, the individuals with a lower number of ones are preferred, while in the populations related to clusters, the individuals featuring a lower number of key poses are favored.

The three parameter selection methods can be described as follows:

- Random selection: the number of clusters required by the bag of key poses method is selected randomly within the interval [4, 26] for the subsets AS1 and AS2 and the interval [44, 76] for AS3. All the skeleton joints and training instances are included in the processing.
- Evolutionary optimization: the evolutionary algorithm selects the best combination of skeleton joints and *clusters*, considering all the training sequences. The same intervals adopted in the random selection are used for the optimization of the number of key poses.
- Coevolutionary optimization: the optimization method selects all the parameters required by the HAR algorithm: *features, clusters*, and *instances*. In this case, the intervals for *clusters* optimization are [4, 16] for AS1 and AS2, and [4, 64] for AS3.

The results are summarized in **Table 3**, where it can be noticed that, for all the parameters selection methods, the best results are obtained for AS3, AS1, and finally AS2. In fact, as already stated, subsets AS1 and AS2 group have similar gestures (**Figures 1** and **2**). More in detail, from **Figure 2** it is quite evident that *Draw x* and *Draw tick* involve the same poses, and the main cue to differentiate them is their order.

	AS1	AS2	AS3	Avg
Random selection	95.24	86.61	95.5	92.45
Evolutionary optimization	95.24	90.18	100	95.14
Coevolutionary optimization	95.24	91.96	98.2	95.13

Table 3. Results in terms of accuracy (%) obtained on MSR Action3D by the method based on temporal pyramid of key poses.

An average accuracy of 92.45% can be achieved considering the random selection of number of key poses. The subset AS2 is the most critical one, with an accuracy of 86.61% due to the aforementioned reasons. Considering evolutionary optimization, where the evaluated parameters are the number of key poses and the set of skeleton joints, there is a noticeable improvement in AS2 and AS3, and the HAR algorithm shows an average accuracy of 95.14%. Similar average results are obtained with the adoption of the coevolutionary optimization method, including also the set of training instances in the optimization process. In particular, there is a further improvement in AS2, which shows an accuracy of 91.96%, while a suboptimal result (98.2%) is achieved in AS3.

Table 4 aims to compare the results obtained by different HAR methods on MSR Action3D considering the cross-subject evaluation protocol and averaging the results on AS1, AS2, and

	AS1	AS2	AS3	Avg
Li et al. [38]	72.9	71.9	79.2	74.67
Chaaraoui et al. [68]	92.38	86.61	96.4	91.8
Lo Presti et al. [73]	90.29	95.15	93.29	92.91
Tao and Vidal [74]	89.81	93.57	97.03	93.5
Du et al. [75]	93.3	94.64	95.5	94.49
Temporal pyramid of key poses	95.24	90.18	100	95.4
Lillo et al. [76]	94.3	92.9	99.1	95.4
Xu et al. [77]	99.1	92.9	96.4	96.1
Liang et al. [78]	98.1	92.9	99.1	96.7
Shahroudy et al. [79]	–	–	–	98.2

Table 4. Results in terms of accuracy (%) obtained by main HAR algorithms evaluated on cross-subject tests.

AS3 [38]. Only the works in which the use of cross-subject test with actors 1-3-5-7-9 for training and the rest for testing is clearly stated are included in the table.

Some recently published works outperform the performance achieved by the method based on temporal pyramid of key poses. Lillo et al. [76] proposed an activity recognition method based on three levels of abstraction. The first level is dedicated to learning the most representative primitives related to body motion. The poses are combined to compose atomic actions at the mid-level, and more atomic actions are combined to create more complex activities at the top-level. As input data, the aforementioned proposal exploits angles and planes from segments extracted from joint coordinates, adding also histograms of optical flow calculated from RGB patches centered at the joint locations. Xu et al. [77] proposed the adoption of depth motion map (DMM), which is computed from the differences among consecutive maps, to describe the dynamic feature of an action. In addition to this method, the depth static model (DSM) can describe the static feature of an action. The so-called TPDM-SPHOG descriptor encodes DMMs and DSM represented by a temporal pyramid and histogram of oriented gradient (HOG) extracted using a spatial pyramid. DMM and multiscale HOG descriptors are also exploited by Liang et al. [78], and they are combined with local space-time auto-correlation of gradients (STACOG), which compensates the loss of temporal information. l_2-regularized collaborative representation classification (CRC) is adopted to take as inputs for the proposed descriptors and classify the actions. In Ref. [79], a joint sparse regression learning method, which models each action as a combination of multimodal features from body parts, is proposed. In fact, each skeleton is separated into a number of parts and different features, related to the movement and local depth information, are extracted from each part. A small number of active parts for each action class are selected through group sparsity regularization. A hierarchical mixed norm, which includes three levels of regularization over learning weights, is integrated into the learning and selection framework.

The comparison of the algorithm based on temporal pyramid of key poses to other approaches achieving higher accuracies on MSR Action3D allows to conclude that all the considered works exploit not only skeleton data but also RGB or depth information. One approach is based on the extraction of the most important postures considering skeleton joints and RGB data [76], DMM and HOG descriptors calculated from depth data are exploited by more papers [77, 78], and a heterogeneous set of depth and skeleton-based features has been considered in Ref. [79].

4. Conclusion

Human action recognition performed exploiting data collected by RGB-D devices has been an active research field and many researchers are developing algorithms exploiting the properties and characteristics of depth sensors. The main advantages in using this technology include unobtrusiveness and privacy preservation, differently from video-based solutions; additionally, it does not extract movements from interaction with objects, as environmental sensors do, and it does not require the subject to wear any device, differently from systems based on wearable technologies.

Among the HAR algorithms based on RGB-D data, the chapter provided a detailed discussion of a method exploiting a temporal pyramid of key poses that has been able to achieve state-of-the-art results on the well-known MSR Action3D dataset.

Author details

Enea Cippitelli*, Ennio Gambi and Susanna Spinsante

*Address all correspondence to: e.cippitelli@univpm.it

Department of Information Engineering, Polytechnic University of Marche, Ancona, Italy

References

[1] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, Z. Yu. Sensor-based activity recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 2012;**42**(6):790-808. DOI: 10.1109/TSMCC.2012.2198883

[2] R. Poppe. A survey on vision-based human action recognition. Image and Vision Computing. 2010;**28**(6):976-990. DOI: 10.1016/j.imavis.2009.11.014

[3] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, P. Havinga. Activity Recognition Using Inertial Sensing for Healthcare, Wellbeing and Sports Applications: A Survey. In: 23th International Conference on Architecture of Computing Systems. Hannover, Germany. 2010. pp. 1-10.

[4] T. D'Orazio, R. Marani, V. Renò, G. Cicirelli. Recent trends in gesture recognition: How depth data has improved classical approaches. Image and Vision Computing. 2016;**52**: 56-72. DOI: 10.1016/j.imavis.2016.05.007

[5] J. Shotton, A. Fitzgibbon, A. Blake, A. Kipman, M. Finocchio, R. Moore, T. Sharp. Real-time Human Pose Recognition in Parts from a Single Depth Image. In: CVPR; Colorado Springs, CO. June; 2011.

[6] Q. Ni, A. B. García Hernando, I. P. de la Cruz. The elderly's independent living in smart homes: A characterization of activities and sensing infrastructure survey to facilitate services development. Sensors. 2015;**15**(5):11312-11362. DOI: 10.3390/s150511312

[7] S. Wang, G. Zhou. A review on radio based activity recognition. Digital Communications and Networks. 2015;**1**(1):20-29. DOI: 10.1016/j.dcan.2015.02.006

[8] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, P. J. Havinga. A survey of online activity recognition using mobile phones. Sensors. 2015;**15**(1):2059-2085. DOI: 10.3390/s150102059.

[9] D. Lara, M. A. Labrador. A survey on human activity recognition using wearable sensors. IEEE Communications Surveys Tutorials. 2013;**15**(3):1192-1209. DOI: 10.1109/SURV.2012.110112.00192

[10] E. Munguia Tapia, S. S. Intille, K. Larson. Activity Recognition in the Home Using Simple and Ubiquitous Sensors. In: A. Ferscha, F. Mattern, editors. Pervasive Computing, Second International Conference, PERVASIVE 2004, Linz/Vienna, Austria, April 21-23, 2004. Proceedings. Lecture Notes in Computer Science ed. Springer Berlin Heidelberg; 2004. pp. 158-175. DOI: 10.1007/978-3-540-24646-6_10

[11] F. J. Ordóñez, P. de Toledo, A. Sanchis. Activity recognition using hybrid generative/discriminative models on home environments using binary sensors. Sensors. 2013;**13**(5): 5460-5477. DOI: 10.3390/s130505460

[12] D. H. Wilson, C. Atkeson. Simultaneous Tracking and Activity Recognition (Star) Using Many Anonymous, Binary Sensors. In: H. W. Gellersen, R. Want, A. Schmidt, editors. Pervasive Computing: Third International Conference, Pervasive 2005, Munich, Germany, May 8-13, 2005. Proceedings. Lecture Notes in Computer Science ed. Springer Berlin Heidelberg; 2005. pp. 62-79. DOI: 10.1007/11428572_5

[13] S. Spinsante, E. Gambi, A. De Santis, L. Montanini, G. Pelliccioni, L. Raffaeli, G. Rascioni. Design and Implementation of a Smart Home Technological Platform for the Delivery of AAL Services: From Requirements to Field Experience. In: F. Florez-Revuelta, A. A. Chaaraoui, editors. Active and Assisted Living: Technologies and Applications. London (UK): The Institution of Engineering and Technology; 2016. pp. 433-456.

[14] L. Montanini, L. Raffaeli, A. De Santis, A. Del Campo, C. Chiatti, G. Rascioni, E. Gambi, S. Spinsante. Overnight Supervision of Alzheimer's Disease Patients in Nursing Homes - System Development and Field Trial. In: Proceedings of the International Conference on Information and Communication Technologies for Ageing Well and e-Health; 21-22 April 2016. Rome (IT); 2016. pp. 15-25. DOI: 10.5220/0005790000150025

[15] J. M. Sim, Y. Lee, O. Kwon. Acoustic sensor based recognition of human activity in everyday life for smart home services. International Journal of Distributed Sensor Networks. 2015;**11**(9) DOI: 10.1155/2015/679123

[16] S. Tremblay, D. Fortin-Simard, E. Blackburn-Verrault, S. Gaboury, B. Bouchard. Exploiting Environment Sounds for Activity Recognition in Smart Homes. In: 2nd AAAI Workshop on Artificial Intelligence Applied to Assistive Technologies and Smart Environments (ATSE '15); January 25-26; Austin (TX). 2015.

[17] E. L. Salomons, P. J. M. Havinga. A survey on the feasibility of sound classification on wireless sensor nodes. Sensors. 2015;**15**(4):7462-7498. DOI: 10.3390/s150407462

[18] L. Vuegen, B. Van Den Broeck, P. Karsmakers, H. Van hamme, B. Vanrumste. Energy Efficient Monitoring of Activities of Daily Living Using Wireless Acoustic Sensor Networks in Clean and Noisy Conditions. In: Signal Processing Conference (EUSIPCO), 2015 23rd European, Nice, France; 31 Aug.–4 Sept.; 2015. pp. 449-453. DOI: 10.1109/EUSIPCO.2015.7362423

[19] M. Salman Khan, M. Yu, P. Feng, L. Wang, J. Chambers. An unsupervised acoustic fall detection system using source separation for sound interference suppression. Signal Processing. 2015;**110**:199-210. DOI: 10.1016/j.sigpro.2014.08.021

[20] Y. Li, Z. Zeng, M. Popescu, K. C. Ho. Acoustic Fall Detection Using a Circular Microphone Array. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology; 31 Aug.–4 Sept; Buenos Aires, Argentina; 2010. pp. 2242-2245. DOI: 10.1109/IEMBS.2010.5627368

[21] Y. Wang, J. Liu, Y. Chen, M. Gruteser, J. Yang, H. Liu. E-eyes: Device Free Location-Oriented Activity Identification Using Fine-Grained Wifi Signatures. In: Proceedings of the 20th Annual International Conference on Mobile Computing and Networking; Maui, Hawaii, USA. 2014. pp. 617-628. DOI: 10.1145/2639108.2639143

[22] B. Çaglıyan, S. Z. Gürbüz. Micro-doppler-based human activity classification using the mote-scale bumblebee radar. IEEE Geoscience and Remote Sensing Letters. 2015;**12**(10):2135-2139. DOI: 10.1109/LGRS.2015.2452946

[23] R. M. Narayanan, M. Zenaldin. Radar micro-Doppler signatures of various human activities. IET Radar, Sonar & Navigation. 2015;**9**(9):1205-1215. DOI: 10.1049/iet-rsn.2015.0173

[24] S. Björklund, H. Petersson, G. Hendeby. Features for micro-Doppler based activity classification. IET Radar, Sonar & Navigation. 2015;**9**(9):1181-1187. DOI: 10.1049/iet-rsn.2015.0084

[25] R. Damaševičius, M. Vasiljevas, J. Šalkevičius, M. Woźniak. Human activity recognition in AAL environments using random projections. Computational and Mathematical Methods in Medicine. 2016;**2016**:Article ID 4073584, 17 pages. DOI: 10.1155/2016/4073584

[26] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, Y. Amirat. Physical human activity recognition using wearable sensors. Sensors. 2015;**15**(12):31314-31338. DOI: 10.3390/s151229858

[27] D. M. Karantonis, M. R. Narayanan, M. Mathie, N. H. Lovell, B. G. Celler. Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. IEEE Transactions on Information Technology in Biomedicine. 2006;**10**(1):156-167. DOI: 10.1109/TITB.2005.856864

[28] L. Atallah, B. Lo, R. King, G. Z. Yang. Sensor Placement for Activity Detection Using Wearable Accelerometers. In: International Conference on Body Sensor Networks, Singapore; 2010. pp. 24-29. DOI: 10.1109/BSN.2010.23

[29] J.-Y. Yang, J.-S.Wang, Y.-P. Chen. Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers. Pattern Recognition Letters. 2008;**29**(16):2213-2220. DOI: 10.1016/j.patrec.2008.08.002

[30] M. M. Hamdi, M. I. Awad, M. M. Abdelhameed, F. A. Tolbah. Lower Limb Gait Activity Recognition Using Inertial Measurement Units for Rehabilitation Robotics. In: International Conference on Advanced Robotics (ICAR); Istanbul. 2015. pp. 316-322. DOI: 10.1109/ICAR.2015.7251474

[31] L. Gao, A. Bourke, J. Nelson. Evaluation of accelerometer based multi-sensor versus single-sensor activity recognition systems. Medical Engineering & Physics. 2014;**36**(6):779-785. DOI: 10.1016/j.medengphy.2014.02.012

[32] E. Cippitelli, S. Gasparrini, E. Gambi, S. Spinsante, J. Wahslen, I. Orhan, T. Lindh. Time Synchronization and Data Fusion for Rgb-Depth Cameras and Inertial Sensors in AAL Applications. In: 2015 IEEE International Conference on Communication Workshop (ICCW); London. 2015. pp. 265-270. DOI: 10.1109/ICCW.2015.7247189

[33] J. R. Padilla-Lopez, A. A. Chaaraoui, F. Gu, F. Florez-Revuelta. Visual privacy by context: Proposal and evaluation of a level-based visualisation scheme. Sensors. 2015;**15**(6):12959-12982. DOI: 10.3390/s150612959

[34] J. K. Aggarwal, Q. Cai. Human motion analysis: A review. Computer Vision and Image Understanding. 1999;**73**(3):428-440. DOI: 10.1006/cviu.1998.0744

[35] P. Turaga, R. Chellappa, V. S. Subrahmanian, O. Udrea. Machine recognition of human activities: A survey. IEEE Transactions on Circuits and Systems for Video Technology. 2008;**18**(11):1473-1488. DOI: 10.1109/TCSVT.2008.2005594

[36] A. A. Chaaraoui, P. Climent-Perez, F. Florez-Revuelta. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. Expert Systems with Applications. 2012;**39**(12):10873-10888. DOI: 10.1016/j.eswa.2012.03.005

[37] J. K. Aggarwal, L. Xia. Human activity recognition from 3D data: A review. Pattern Recognition Letters. 2014;**48**:70-80. DOI: 10.1016/j.patrec.2014.04.011

[38] W. Li, Z. Zhang, Z. Liu. Action Recognition Based on a Bag of 3d Points. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2010. San Francisco, CA; pp. 9-14. DOI: 10.1109/CVPRW.2010.5543273

[39] A. A. Chaaraoui, P. Climent-Perez, F. Florez-Revuelta. Silhouette-based human action recognition using sequences of key poses. Pattern Recognition Letters. 2013;**34**(15):1799-1807. DOI: 10.1016/j.patrec.2013.01.021

[40] L. Xia, J. K. Aggarwal. Spatio-temporal Depth Cuboid Similarity Feature for Activity Recognition Using Depth Camera. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition; Portland. 2013. pp. 2834-2841. DOI: 10.1109/CVPR.2013.365

[41] A. Vieira, E. Nascimento, G. Oliveira, Z. Liu, M. Campos. STOP: Space-Time Occupancy Patterns for 3D Action Recognition from Depth Map Sequences. In: L. Alvarez, M. Mejail, L. Gomez, J. Jacobo, editors. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Lecture Notes in Computer Science ed. Springer Berlin Heidelberg; 2012. pp. 252-259. DOI: 10.1007/978-3-642-33275-3_31

[42] O. Oreifej, Z. Liu. HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition; Portland. 2013. pp. 716-723. DOI: 10.1109/CVPR.2013.98

[43] H. Rahmani, A. Mahmood, D. Q Huynh, A. Mian. HOPC: Histogram of Oriented Principal Components of 3D Pointclouds for Action Recognition. In: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars, editors. Computer Vision - ECCV 2014. Lecture Notes in Computer Science ed. Springer International Publishing, Cham; 2014. pp. 742-757. DOI: 10.1007/978-3-319-10605-2_48

[44] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, A. Del Bimbo. Space-Time Pose Representation for 3D Human Action Recognition. In: A. Petrosino, L. Maddalena, P. Pala, editors. New Trends in Image Analysis and Processing - ICIAP 2013. Lecture Notes in Computer Science ed. Springer, Berlin, Heidelberg; 2013. pp. 456-464. DOI: 10.1007/978-3-642-41190-8_49

[45] S. Gaglio, G. L. Re, M. Morana. Human activity recognition process using 3-d posture data. IEEE Transactions on Human-Machine Systems. 2015;**45**(5):586-597. DOI: 10.1109/THMS.2014.2377111

[46] I. Theodorakopoulos, D. Kastaniotis, G. Economou, S. Fotopoulos. Pose-based human action recognition via sparse representation in dissimilarity space. Journal of Visual Communication and Image Representation. 2014;**25**(1):12-23. DOI: 10.1016/j.jvcir.2013.03.008

[47] L. Gan, F. Chen. Human action recognition using apj3d and random forests. Journal of Software. 2013;**8**(9):2238-2245. DOI: 10.4304/jsw.8.9.2238-2245

[48] L. Xia, C.-C. Chen, J. Aggarwal. View Invariant Human Action Recognition Using Histograms of 3d Joints. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2012. Providence, Rhode Island; pp. 20-27. DOI: 10.1109/CVPRW.2012.6239233

[49] M. Firman. RGBD datasets: Past, present and future. CoRR. 2016;**abs/1604.00999**

[50] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, C. Tang. RGB-D-based action recognition datasets: A survey. Pattern Recognition. 2016;**60**:86-105. DOI: 10.1016/j.patcog.2016.05.019

[51] J. Wang, Z. Liu, Y. Wu, J. Yuan. Mining Actionlet Ensemble for Action Recognition with Depth Cameras. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2012. Providence, Rhode Island; pp. 1290-1297. DOI: 10.1109/CVPR.2012.6247813

[52] J. Sung, C. Ponce, B. Selman, A. Saxena. Unstructured Human Activity Detection from Rgbd Images. In: 2012 IEEE Conference on Robotics and Automation (ICRA); 2012. St. Paul, Minnesota; pp. 842-849. DOI: 10.1109/ICRA.2012.6224591

[53] H. S. Koppula, R. Gupta, A. Saxena. Learning human activities and object affordances from RGB-D videos. International Journal of Robotics Research. 2013;**32**(8):915-970. DOI: 10.1177/0278364913478446

[54] B. Ni, G. Wang, P. Moulin. RGBD-HuDaAct: A Color-Depth Video Database for Human Daily Activity Recognition. In: 2011 IEEE International Conference on Computer Vision Workshops; 2011. Barcelona, Spain; pp. 1147-1153. DOI: 10.1109/ICCVW.2011.6130379

[55] S. Fothergill, H. M. Mentis, P. Kohli, S. Nowozin. Instructing People for Training Gestural Interactive Systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; ACM, 2012. Austin, TX; pp. 1737-1746. DOI: 10.1145/2207676.2208303

[56] A. Kurakin, Z. Zhang, Z. Liu. A Real Time System for Dynamic Hand Gesture Recognition with a Depth Sensor. In: Proceedings of the 20th European Signal Processing Conference (EUSIPCO); 2012. Bucharest, Romania; pp. 1975-1979.

[57] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy. Berkeley MHAD: A Comprehensive Multimodal Human Action Database. In: 2013 IEEE Workshop on Applications of Computer Vision (WACV); 2013. Clearwater, Florida; pp. 53-60. DOI: 10.1109/WACV.2013.6474999

[58] V. Bloom, D. Makris, V. Argyriou. G3D: A Gaming Action Dataset and Real Time Action Recognition Evaluation Framework. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2012. Providence, Rhode Island; pp. 7-12. DOI: 10.1109/CVPRW.2012.6239175

[59] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, P. Pala. Recognizing Actions from Depth Cameras as Weakly Aligned Multi-part Bag-of-Poses. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2013. Portland, Oregon; pp. 479-485. DOI: 10.1109/CVPRW.2013.77

[60] Z. Cheng, L. Qin, Y. Ye, Q. Huang, Q. Tian. Human Daily Action Analysis with Multi-view and Color-Depth Data. In: A. Fusiello, V. Murino, R. Cucchiara, editors. Computer Vision - ECCV 2012. Workshops and Demonstrations. Lecture Notes in Computer Science ed. Springer, Berlin, Heidelberg; 2012. pp. 52-61. DOI: 10.1007/978-3-642-33868-7_6

[61] C. Wolf, J. Mille, E. Lombardi, O. Celiktutan, M. Jiu, M. Baccouche, E. Dellandréa, C.-E. Bichot, C. Garcia, B. Sankur. The LIRIS Human Activities Dataset and the ICPR 2012 Human Activities Recognition and Localization Competition. In: LIRIS Laboratory, Tech. Rep. RR-LIRIS-2012-004, March 2012.

[62] G. Yu, Z. Liu, J. Yuan. Discriminative Orderlet Mining for Real-Time Recognition of Human-Object Interaction. In: D. Cremers, I. Reid, H. Saito, M.-H. Yang, editors. Computer Vision – ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part V. Image Processing, Computer Vision, Pattern Recognition, and Graphics ed. Springer International Publishing, Cham; 2014. pp. 50-65. DOI: 10.1007/978-3-319-16814-2_4

[63] F. Negin, F. Özdemir, C. B. Akgül, K. A. Yüksel, A. Erçil. A Decision Forest Based Feature Selection Framework for Action Recognition from RGB-Depth Cameras. In: M. Kamel, A. Campilho, editors. Image Analysis and Recognition. Lecture Notes in Computer Science ed. Munich: Springer, Berlin, Heidelberg; 2013. pp. 648-657. DOI: 10.1007/978-3-642-39094-4_74

[64] C. Chen, R. Jafari, N. Kehtarnavaz. UTD-MHAD: A Multimodal Dataset for Human Action Recognition Utilizing a Depth Camera and a Wearable Inertial Sensor. In: 2015 IEEE International Conference on Image Processing (ICIP); 2015. Québec City, Canada; pp. 168-172. DOI: 10.1109/ICIP.2015.7350781

[65] M. Munaro, G. Ballin, S. Michieletto, E. Menegatti. 3D flow estimation for human action recognition from colored point clouds. Biologically Inspired Cognitive Architectures. 2013;**5**:42-51. DOI: 10.1016/j.bica.2013.05.008

[66] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. Las Vegas, NV; pp. 1010-1019. DOI: 10.1109/CVPR.2016.115

[67] E. Cippitelli, E. Gambi, S. Spinsante, F. Florez-Revuelta. Human Action Recognition Based on Temporal Pyramid of Key Poses Using RGB-D Sensors. In: J. Blanc-Talon, C. Distante, W. Philips, D. Popescu, P. Scheunders, editors. Advanced Concepts for Intelligent Vision Systems: 17th International Conference, ACIVS 2016, Lecce, Italy, October 24-27, 2016, Proceedings. Lecture Notes in Computer Science, Springer International Publishing; 2016. pp. 510-521. DOI: 10.1007/978-3-319-48680-2_45

[68] A. A. Chaaraoui, J. R. Padilla-López, F. Flórez-Revuelta. Fusion of Skeletal and Silhouette-Based Features for Human Action Recognition with RGB-D Devices. In: 2013 IEEE International Conference on Computer Vision Workshops; 2013. Sydney, Australia; pp. 91-97. DOI: 10.1109/ICCVW.2013.19

[69] E. Cippitelli, S. Gasparrini, E. Gambi, and S. Spinsante. A human activity recognition system using skeleton data from RGBD sensors. Computational Intelligence and Neuroscience. 2016;**2016**, Article ID 4351435, 14 pages. DOI: 10.1155/2016/4351435

[70] A. A. Chaaraoui, J. R. Padilla-Lopez, P. Climent-Perez, F. Florez-Revuelta. Evolutionary joint selection to improve human action recognition with RGB-D devices. Expert Systems with Applications. 2014;**41**(3):786-794. DOI: 10.1016/j.eswa.2013.08.009

[71] A. A. Chaaraoui, F. Florez-Revuelta. Optimizing human action recognition based on a cooperative coevolutionary algorithm. Engineering Applications of Artificial Intelligence. 2014;**31**:116-125. DOI: 10.1016/j.engappai.2013.10.003

[72] A. A. Chaaraoui, F. Florez-Revuelta. Adaptive human action recognition with an evolving bag of key poses. IEEE Transactions on Autonomous Mental Development. 2014;**6**(2):139-152. DOI: 10.1109/TAMD.2014.2315676

[73] L. Lo Presti, M. L. Cascia, S. Sclaroff, O. Camps. Hankelet-based dynamical systems modeling for 3d action recognition. Image and Vision Computing. 2015;**44**:29-43. DOI: 10.1016/j.imavis.2015.09.007

[74] L. Tao, R. Vidal. Moving Poselets: A Discriminative and Interpretable Skeletal Motion Representation for Action Recognition. In: 2015 IEEE International Conference on Computer Vision Workshop (ICCVW); 2015. Santiago, Chile; pp. 303-311. DOI: 10.1109/ICCVW.2015.48

[75] Y. Du, W. Wang, L. Wang. Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015. Boston, Massachusetts; pp. 1110-1118. DOI: 10.1109/CVPR.2015.7298714

[76] I. Lillo, J. C. Niebles, A. Soto. Sparse composition of body poses and atomic actions for human activity recognition in RGB-D videos. Image and Vision Computing. 2017;**59**:63-75. DOI: 10.1016/j.imavis.2016.11.004

[77] H. Xu, E. Chen, C. Liang, L. Qi, L. Guan. Spatio-Temporal Pyramid Model Based on Depth Maps for Action Recognition. In: 2016 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP); 2015. Montreal, Canada; pp. 1-6. DOI: 10.1109/MMSP.2015.7340806

[78] C. Liang, L. Qi, E. Chen, L. Guan. Depth-Based Action Recognition Using Multiscale Sub-Actions Depth Motion Maps and Local Auto-Correlation of Space-Time Gradients. In: 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS); 2016. Niagara Falls, NY; pp. 1-7. DOI: 10.1109/BTAS.2016.7791167

[79] A. Shahroudy, T. T. Ng, Q. Yang, G. Wang. Multimodal multipart learning for action recognition in depth videos. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2016;**38**(10):2123-2129. DOI: 10.1109/TPAMI.2015.2505295