



Motion analysis: Action detection, recognition and evaluation based on motion capture data

Fotini Patrona*, Anargyros Chatzitofis, Dimitrios Zarpalas, Petros Daras

Information Technologies Institute, Centre for Research and Technology-Hellas, 6th Km Charilaou-Thermi, Thessaloniki GR-57001, Greece



ARTICLE INFO

Article history:

Received 31 January 2017

Revised 25 October 2017

Accepted 9 December 2017

Available online 12 December 2017

Keywords:

Online human action detection
Online human action recognition
Motion evaluation
Kinect
Skeleton data
Automatic joint/angle weighting
Kinetic energy

ABSTRACT

A novel framework, for real-time action detection, recognition and evaluation of motion capture data, is presented in this paper. Pose and kinematics information is used for data description. Automatic and dynamic weighting, altering joint data significance based on action involvement, and Kinetic energy-based descriptor sampling are employed for efficient action segmentation and labelling. The automatically segmented and recognized action instances are subsequently fed to the framework action evaluation component, which compares them with the corresponding reference ones, estimating their similarity. Exploiting fuzzy logic, the framework subsequently gives semantic feedback with instructions on performing the actions more accurately. Experimental results on *MSR-Action3D* and *MSRC12* benchmarking datasets and a new, publicly available one, provide evidence that the proposed framework compares favourably to state-of-the-art methods by 0.5–6% in all three datasets, showing that the proposed method can be effectively used for unsupervised gesture/action training.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Automatic human action recognition and detection constitute two widely studied problems, mainly due to their numerous applications in domains like gaming [1], computer vision [2], animation [3], surveillance [4], man-machine interaction [5], robotics [6], etc. Till recently, the majority of the relevant research works, mostly relied on colour 2D/3D video sequences, RGB image related visual features and Radio Frequency Identification (RFID) sensors. With the advent of low cost, non-intrusive depth sensors like Microsoft Kinect, research efforts are now devoted to the utilization of 3D skeleton joint positions [1], as body part movement related features can be much more representative of actions, and thus, more discriminative. However, issues related to intra-/inter-person variability, random pauses, repetitions and nonlinear stretching characterizing human motion, as well as body part (self-)occlusions along with sensor inaccuracies, also constitute major challenges to be faced.

Recently, research efforts have also been devoted to motion capture data analysis, motion detection and recognition, widely known as human motion evaluation. Research in this field, is mainly exploited in applications for interactive gaming [7], reha-

bilitation [8], self-learning platforms for practising and conquering sports [9], dance [10] and martial arts [11].

In this paper, a framework implementing real time analysis of long action sequences, is proposed. Segmentation of input sequences into their constituent action instances as well as recognition of each action instance are initially performed, followed by evaluation of action execution. As depicted in Fig. 1, motion capture data, and more specifically skeleton 3D joint positions, constitute the only input required. Human action detection/recognition component segments the input sequence into time intervals containing a single action instance each, while also recognizing action instance types. Subsequently, each detected instance is compared to a training sample of the same class, a priori selected and considered as reference, for motion evaluation. Based on the differences between the detected and the reference instance, semantic feedback, indicating ways of performing the actions in a manner more similar to the ground truth (reference) is provided. Moreover, indications concerning the movement of the several body parts can be given until user motion becomes identical to the reference motion instance, or till some specific, user determined, similarity degree.

The action detection/recognition component of the devised framework is inspired by the work of Meshry et al. [12], though introducing significant extensions. In brief, not all 20 joints used by Meshry et al. [12] are utilized, while automatic feature weighting at the frame level, is also employed, with the weights being calculated based on the volumes and areas created by the several body segments, throughout the action sequences. In addition

* Corresponding author.

E-mail address: fotinip@iti.gr (F. Patrona).

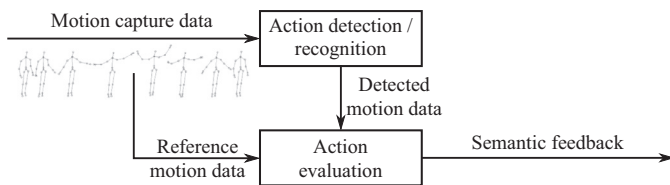


Fig. 1. Proposed human motion analysis method pipeline.

to this, Kinetic energy (KE) is also introduced in the descriptor sampling stage, so that vectors of the most representative action poses can be selected for codebook construction, inspired by Shan and Akella [13]. In this way, descriptive codebooks arise clustering much fewer vectors than the ones needed when randomly selected, resulting in very satisfactory detection and recognition results. Finally, a new dataset is also presented, composed of 656 3D human joint data sequences depicting 15 exercises performed by 15 individuals.

The remainder of the paper is organized as follows. A review of the related work in the field is presented in Section 2, our Action Detection-Recognition-Evaluation framework is described in Section 3, experimental evaluation is detailed in Section 4 and finally, conclusions are drawn in Section 5.

2. Related work

Several research works, addressing both action recognition and detection problems based on motion capture data, have been published in the last decade, either regarding human motion sequences as time series or approaching the two problems from machine learning perspective.

Skeletal pose and kinematics information is employed in [12], aiming for the detection of action instances as the sub-intervals with the maximum score sums in unsegmented motion sequences. Human pose and motion shape analysis on Riemannian shape space are performed in [14], so that motion units, decomposing actions into their constituent basic motions, can be identified. Repeated cycles of such motion units denote repetitions of the same action, thus leading to sequence segmentation. Riemannian geometry is also exploited in [15] framework, in an attempt for handling 3D skeleton data in an execution rate invariant manner. The shapes formed by skeletons and their evolution in time are studied as trajectories, using Kendall's shape framework, while SVM classification is finally performed. Atomic action templates, i.e., key frame tuples, are extracted from 3D human skeleton coordinates, based on KE in [13], and used as spatiotemporal action representations for classification and subsequent segmentation of human motion sequences.

Significantly fewer works focus on action detection, especially the online case, as localizing actions is much more challenging than recognizing them in pre-segmented sequences. One such work is presented in [16], addressing action detection with the aid of sliding window search and SVM classification. Feature selection is performed to cope with the high dimensionality of the features used, and temporal pyramid construction to ensure capturing of multi-scale temporal information. A novel feature, namely Structured Streaming Skeleton (SSS), is proposed in [17], describing skeletons by features denoting the similarity degree of the motion segments ending at some frame with a priori learned movements on the joint level, and thus, segmenting and recognizing motion sequences at the same time. Proposing a new descriptor, capturing both 3D pose and kinematics known as the Moving Pose, Zanfir et al. [18] report low latency action detection and recognition, with the aid of a sliding window of learned size and a modified kNN classifier. Similarly, a fixed sliding window is employed by

Nowozin and Shotton [19], introducing the action point concept for temporally anchoring actions. Skeleton based features are combined with an optimized Hough transform, based on weighting, aspiring to perform both human action segmentation and classification, in [20]. To this end, several skeleton-related features and normalizations are studied. Real time human action segmentation and recognition are achieved in [21], as a result of the combination of two devised frameworks, Kernelized Temporal Cut (KTC) and Dynamic Manifold Warping (DMW), the former performing real-time segmentation handling human motion data as structured time series, and the latter spatiotemporally aligning the time series, thus, calculating similarity between motion data through manifold learning.

A new descriptor, consisting of pairwise joint angle affinities is introduced in [22], combined with HOG features calculated on depth images, while joint positions constitute the features represented through class specific dictionary learning, exploiting geometry constraints, group sparsity and temporal pyramid matching in [23]. In a project engaging with robotic support for elderly people [24], a novel framework incorporating both human activity recognition and prediction is introduced. Utilizing 3D skeleton data captured by Kinect sensor, Factored Four Way Conditional Restricted Boltzmann Machines (FFW-CRBM) can automatically evaluate their classification performance and retrain themselves if necessary with the aid of sequential Markov chain contrastive divergence (SMCD) training algorithm, which is also introduced. A two-level, hierarchical framework is introduced in [25]. Part-based clustering, is initially performed, splitting action instances based on their dominant body parts, followed by 3D skeleton-based feature extraction only from the aforementioned body parts and subsequent classification through action graphs. Salih et al. [26] propose hierarchical Modified Spherical Harmonics (MSHs) for spatiotemporally modeling skeleton data static poses and selected joint displacements, DTW for the alignment of the MSHs levels and Extreme Learning Machine (ELM) for classification. Covariance matrices of the 3D skeletal joint positions over time are constructed in [27], for the representation of motion sequences sustaining temporal information through the incorporation of matrix computations over temporal sub-intervals and fed to a SVM classifier. Only 12 out of the 20 3D skeletal points estimated from Kinect depth images suffice for the representation of human postures as histograms of 3D joint locations (HOJ3D) and the efficient recognition of actions, encoded as temporal sequences of poses, by HMM classifiers in [28]. Finally, a novel way of human action modeling, encapsulating a new feature, namely Local Occupancy Pattern (LOP) as well as a new representation, retaining pattern temporal order, are presented in [29] by Wang et al.

Research attempts addressing the problem of human motion evaluation are even fewer, with the vast majority made during the last 2 decades. One such approach, with a network of body sensors providing the inertial data required for baseball swings evaluation, employing clustering techniques is presented in [9]. Movement transcripts are generated and inter-segment coordination is measured, based on which user feedback is produced. A marker-based optical motion capture system constitutes the capturing component of the dance training system proposed in [30]. The acquired data are matched to pre-recorded reference ones and compared based on joint positions, velocities and angles, while visual as well as numerical feedback are subsequently provided. In the interactive dancing game presented in [7], user motion data are captured, recognized and analysed so that the virtual partner displayed can be animated appropriately in real time. Joint angle differences are used for frame matching, continuous block matching for temporally aligning the movements and block matching cost for the identification of the deviation of the user movement from the template one. Contrary to all these approaches, marker-less motion

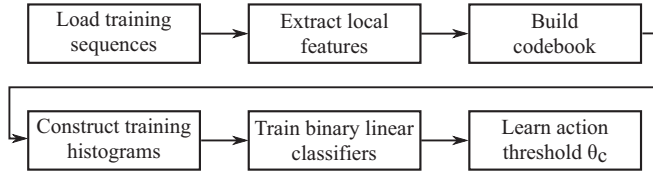


Fig. 2. ELS method training phase overview.

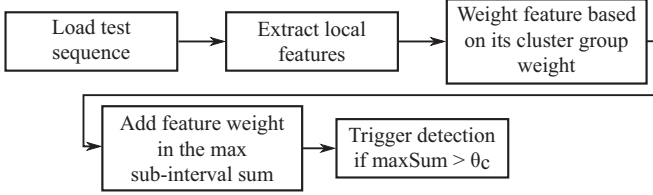


Fig. 3. ELS method testing phase overview.

capture is performed in the automatic dance performance evaluation framework proposed in [10], in which human motion data are represented as sequences of pure quaternions, thus resulting in a framework employing quaternionic vector-signal processing methodologies.

3. Proposed method

3.1. Human action detection/recognition component

As already mentioned, online action detection and recognition constitute the core of the proposed framework, and are based on Efficient Linear Search (ELS) approach [12], diagrammatically presented in Figs. 2 and 3 and briefly detailed in the following paragraphs. To begin with, local features capturing both skeleton and kinematics information at the frame level are calculated on 3D skeleton joint positions. The aforementioned features, called *gesturelets*, form descriptors, emerging by combining the Moving Pose [18] and the angles descriptor [19] with appropriate weighting. All or a subset of them are subsequently clustered into a codebook, so that compact, descriptor invariant action sequence representations can be obtained, exploiting Bag-of-Gesturelets (BoG) model power. The resulting histogram representations are then fed to binary linear classifiers, each one trained to identify a specific action and, based on their weights, thresholds for the detection of each action are estimated (Fig. 2) [12]. In the test phase, presented in Fig. 3, after estimating local features for the sequence under consideration, 1D arrays of length equal to that of the sequence are constructed, and filled with the weighted sums of the sequence gesturelet features for each candidate action. Thus, action interval detection can be achieved by finding the maximum subarray sum [31], while action detection is triggered at some frame with sum exceeding action thresholds, θ_c , estimated during training. Action interval end is signalled by the appearance of negative point scores after action triggering, while slight modifications on this criterion, allow using this method both online and offline [12].

As is, ELS method and especially the Moving Pose descriptor employed, neglect several facts, inextricably linked both to action recording with the aid of depth sensors, such as Kinect, and to human action abundance. To elaborate, it is a matter of fact that depth sensors are quite accurate, especially when there are no (self-)occlusions and the actions are not performed very quickly or abruptly. This may be the case for most of the captured joints, however things tend to be slightly different for limb joints, and especially for the hands and feet, which tend to be inaccurately detected and/or tracked, thus leading to noisy 3D joint position data. To this end, omitting these joints, both from the descriptor and

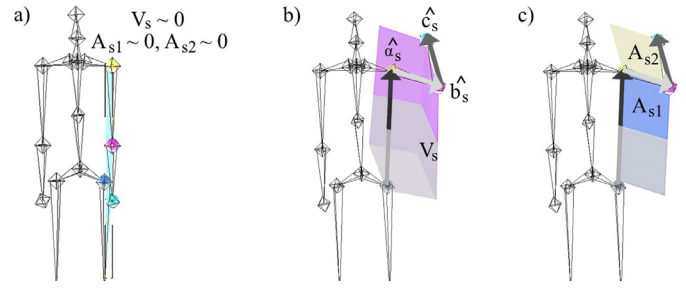


Fig. 4. (a) N-pose or neutral pose; volume V_s , areas A_{s1} and A_{s2} are approximately equal to 0 for all body segments s , (b) vectors \hat{a}_s, \hat{b}_s and \hat{c}_s form volume V_s in a random pose and (c) areas A_{s1} and A_{s2} .

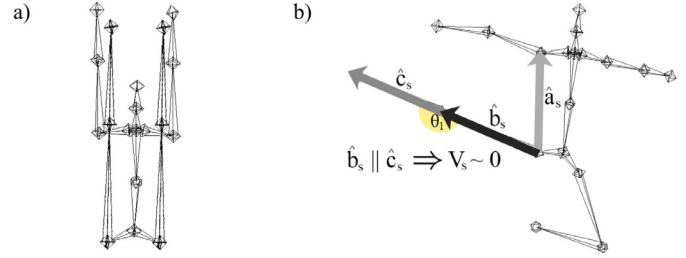


Fig. 5. (a) Body pose with maximum limb activity (all four limbs fully extended). (b) Joint angle conditions allow for correct weighting when the volume formed by vectors \hat{a}_s, \hat{b}_s and \hat{c}_s is eliminated.

from all subsequent calculations, is considered wise also presuming that none of the adopted datasets contains actions in which crucial involvement of these joints is encountered.

On top of that, it should not be disregarded that not all body parts participate in every human action, as well as that even the ones participating are not equally important. Thus, applying joint weighting was attempted, automatically calculating joint weights based on the volumes and the areas formed between the limbs and the main body at the frame level (i.e., dynamically). To cast light upon the automatic weighting extraction procedure, a brief description of the underlying theory is subsequently provided.

Let us consider the standing body pose (Fig. 4(a)) as 'neutral' (N-pose), and its joint positions as reference ones. Pose differences relative to N-pose are used to weight the joints, taking the articulated structure of the human body into account. Thus, the volume and the areas formed between the bone vectors are ideal for giving information regarding the role of a joint during gestures and, therefore, during actions, since their values are approximately equal to 0 in N-pose, as depicted in Fig. 4(a).

Weighting extraction segments the body in 5 parts, one for each of the four limbs and one for the trunk (pelvis, thorax, head). For each part, 3 vectors are formed: \hat{a}_s parallel to the spine, \hat{b}_s for the upper and \hat{c}_s for the lower limb bone, as shown in Fig. 4b. The triple scalar of the unit vectors \hat{a}_s, \hat{b}_s and \hat{c}_s gives the volume $V_s = \hat{a}_s * (\hat{b}_s \times \hat{c}_s)$ for segment s , while $|\hat{a}_s \times \hat{b}_s|$ and $|\hat{b}_s \times \hat{c}_s|$ give areas A_{s1} and A_{s2} , respectively, depicted in Fig. 4(c). Finally, assuming that the maximum weight is assigned to a segment when the corresponding limb is totally stretched and, therefore, the triple scalar tends to zero (Fig. 5(a)), V_s is calculated taking joint angle criteria into account, as shown in Eq. (1).

$$V_s(\theta_1, \theta_2) = \begin{cases} \hat{a}_s * (\hat{b}_s \times \hat{c}_s) & 0^\circ \leq \theta_1 \leq 90^\circ, \theta_2 > 90^\circ \\ 2 - \hat{a}_s * (\hat{b}_s \times \hat{c}_s) & 0^\circ \leq \theta_1 \leq 90^\circ, 0^\circ \leq \theta_2 \leq 90^\circ \\ 3 - \hat{a}_s * (\hat{b}_s \times \hat{c}_s) & \theta_1 > 90^\circ, 0^\circ \leq \theta_2 \leq 90^\circ \\ 4 - \hat{a}_s * (\hat{b}_s \times \hat{c}_s) & \theta_1 > 90^\circ, \theta_2 > 90^\circ \end{cases} \quad (1)$$

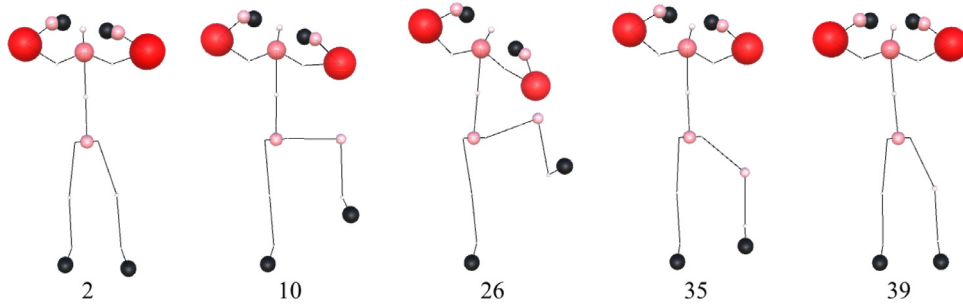


Fig. 6. Key frames of CVD dataset exercise *Standing obliques* with weights presented as spheres on the joints, with their colour and size indicating joint weight value.

Accordingly, $A_{1s}(\theta_1)$ and $A_{2s}(\theta_2)$ are given by Eqs. (2) and (3).

$$A_{1s}(\theta_1) = \begin{cases} |\hat{a}_s \times \hat{b}_s| & 0^\circ \leq \theta_1 \leq 90^\circ \\ 2 - |\hat{a}_s \times \hat{b}_s| & \theta_1 > 90^\circ \end{cases} \quad (2)$$

$$A_{2s}(\theta_2) = \begin{cases} |\hat{b}_s \times \hat{c}_s| & \theta_2 > 90^\circ \\ 2 - |\hat{b}_s \times \hat{c}_s| & 0^\circ \leq \theta_2 \leq 90^\circ \end{cases} \quad (3)$$

Utilizing both the volume V_s and the areas A_{s1} , A_{s2} for joint weighting is necessary, since on the one hand, V_s is the main weighting component, incorporating all 3 vectors and constituting a holistic weighting metric for segment s , and on the other hand, volume elimination due to bone parallel positioning (e.g., Fig. 5(b)), can be balanced using the areas A_{s1} and A_{s2} .

Given the calculated volume V_s and the areas A_{s1} , A_{s2} , Eqs. (4) and (5) give segment s weight w_s at frame f , while the weight $w_{s,j}$ of joint j of segment s is calculated by Eq. (6). Eqs. (7)–(9) give the weights w_{sa} , w_{sb} and w_{sc} of the end joints of the vectors \hat{a} , \hat{b} and \hat{c} , respectively.

$$\hat{a}_s \perp \hat{b}_s \perp \hat{c}_s \rightarrow V_{sMax} = A_{s1Max} = A_{s2Max} = 1 \quad (4)$$

$$w_s = V_s + (1 - V_s) * (A_{1s} + A_{2s})/2 \quad (5)$$

$$w_{s,j} = w_s/N_{s,j} \quad (6)$$

$$w_{sa} = A_{1s}/(A_{1s} + A_{2s}) * w_{s,j} \quad (7)$$

$$w_{sb} = 2 * w_{s,j} \quad (8)$$

$$w_{sc} = A_{2s}/(A_{1s} + A_{2s}) * w_{s,j} \quad (9)$$

The 2D array containing body weights at frame f is given by Eq. (10), with the sum of its elements being equal to 1 (Eq. (11)). $N_s = 5$ denotes the number of the segments, while $N_{s,j}$ the number of the segment joints. Finally, W_f is multiplied by the normalized positions of the segment joints.

$$W_f = \begin{pmatrix} w_{1,1} & \dots & w_{1,N_{s,j}} \\ w_{2,1} & \dots & w_{2,N_{s,j}} \\ \vdots & \ddots & \vdots \\ w_{N_s,1} & \dots & w_{N_s,N_{s,j}} \end{pmatrix} \quad (10)$$

$$\sum_{s=1}^{N_s} \sum_{j=1}^{N_{s,j}} w_{s,j} = 1 \quad (11)$$

The underlying weighting theory, can be perceived more easily taking a closer look at Fig. 6, schematically presenting the key

frames of a CVD dataset exercise, namely *Standing Obliques*. As can be seen, starting from a standing position with both arms bent at head height, the right knee is lifted to the side, the trunk is simultaneously bent towards the same side and finally the body returns to its initial pose. The spheres positioned over the joints represent the corresponding weights, with their colour becoming more intense and their size increasing relative to the weight, while the four black ones denote the joints not taken into consideration (i.e., the hands and the feet) since they provide no useful information for any of the studied actions.

Also looking at Fig. 7, the joint with the most significant change is the right knee, since the right leg starts being close to its neutral pose, bends and reverts to its original pose. Accordingly, the weights of the other two right leg joints, namely the hip and the ankle, increase as the leg is lifted and afterwards decrease. Moreover, due to the fact that during the action the angles formed by the two arms also slightly change, elbow weights are also modified, with the left one increasing and the right decreasing.

Apart from the above, it has also been noticed that in the original ELS implementation, if the training set descriptors are fewer than 500,00, they all take part in the BoG step, in which codebook construction is performed, while otherwise they are randomly subsampled. Considering the case of an action being performed once quickly and once slowly, it becomes obvious that redundant descriptors, which cannot substantially contribute to the creation of the most descriptive codebook possible, emerge. However, they do participate in the procedure and, in fact, they are more likely to be selected than other descriptors which may be more discriminative, due to their abundance. A more elaborate and action-oriented way of selecting the descriptors to be clustered into the codebook was thus required. To this end, the idea of exploiting the information provided by KE values [13], was adopted.

As already known, the KE of a point object is given by:

$$E_k = \frac{1}{2} m v^2. \quad (12)$$

Considering human skeleton as an ensemble of point objects (i.e., the joints) whose mass is of no interest, KE at frame f can be calculated as the sum of the KEs of its joints, j . Denoting joint positions by p , KE of some joint j at frame f is, thus, given by Eq. (13) while entire skeleton KE by (14).

$$E_f^j = \frac{1}{2} (v_f^j)^2 = \frac{1}{2} \left(\frac{p_f^j - p_{f-1}^j}{\Delta T} \right)^2 \quad (13)$$

$$E_f = \frac{1}{2} \sum_{j=1}^N \left(\frac{p_f^j - p_{f-1}^j}{\Delta T} \right)^2 \quad (14)$$

Taking the above into consideration, it can be easily deduced that in the case of human skeleton, KE is zeroed/minimized when all the joints are stationary, as is the case right before every motion

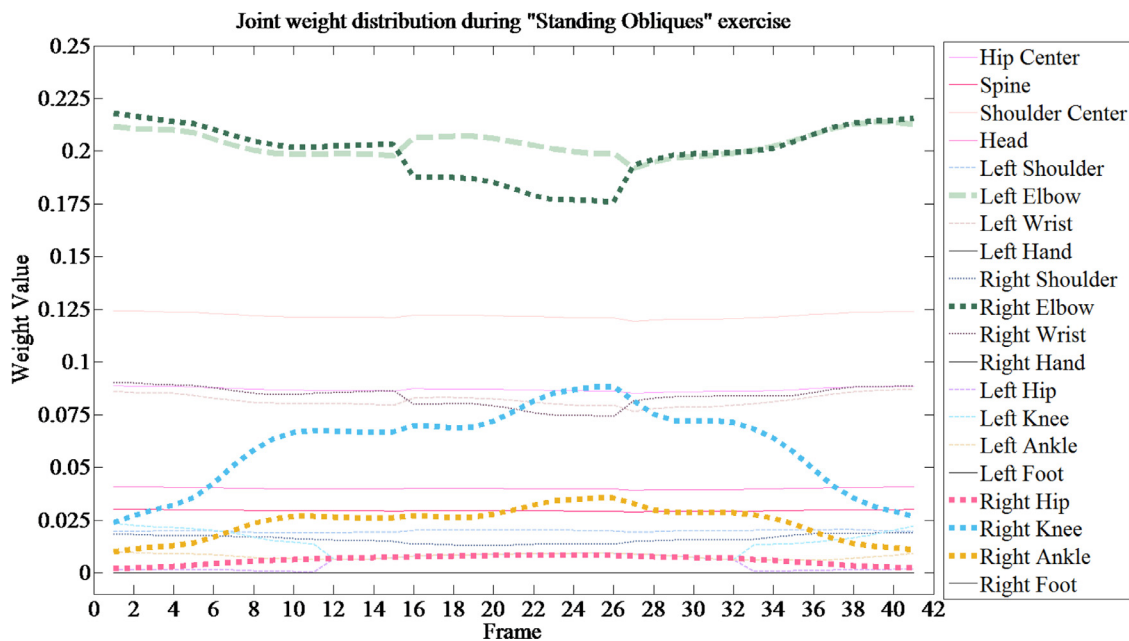


Fig. 7. Weight distribution plot of CVD dataset exercise *Standing obliques* emphasizing on the joints mostly involved in the action.

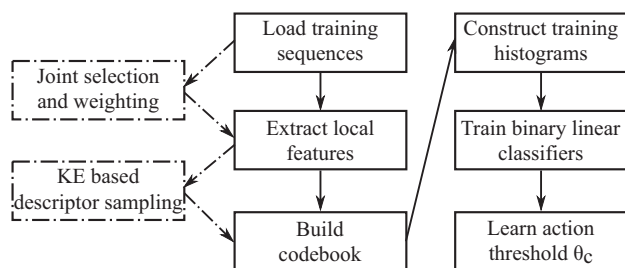


Fig. 8. ELS method contribution overview (addition of the dash-dotted joint selection/weighting and KE-based descriptor sampling).

direction change. As also proven by Shan and Akella [13], KE local minima can help identify extremal positions in actions, which can be regarded as key poses for subsequent action description, resulting in satisfactory action detection and recognition. Skeleton KE local minima and maxima are, thus, also calculated by the proposed framework at the frame level, so that the respective descriptors can be selected for codebook construction.

In detail, making the assumption that all the motion sequences have to be represented by the same number of descriptors, D , during codebook construction, D is estimated by the division of the number of descriptors of the entire dataset and the number of descriptors to be kept (e.g., half the descriptors). Afterwards, KE local minima and maxima are calculated for each motion sequence, and only the ones closer to the sequence global minimum or maximum KE values are retained. In case the corresponding descriptors are less than D , some more are randomly selected, while if they are more, the ones corresponding to frames with KE value closest to the mean sequence KE are discarded. In this way, selection of the most important descriptors as well as reduction in the number of the descriptors to be clustered, are attempted, thus resulting in the clustering only of the most descriptive features, avoiding redundant ones and reducing the time required for the clustering procedure.

These two contributions are presented with dashed-dotted arrows and rectangles in Fig. 8, depicting the proposed method training phase.

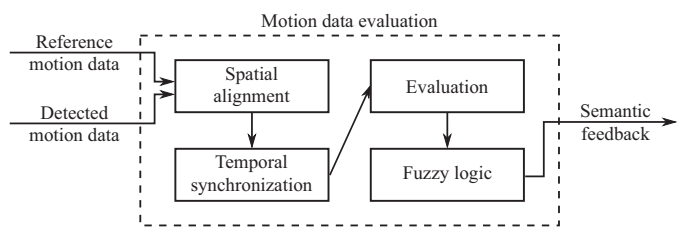


Fig. 9. Proposed human motion evaluation component pipeline.


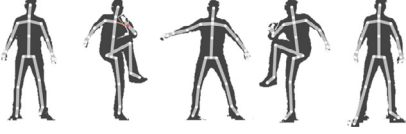

3.2. Human action evaluation component

Motion capture data are segmented into action instances and labelled by the action recognition/detection component, and thus action evaluation can subsequently be performed. More specifically, the captured motion data are spatiotemporally compared to reference actions, originating from the training set, so that semantic feedback can be provided to the user along with instructions for improving action execution form.

For this purpose, the motion data processing pipeline depicted in Fig. 9, is proposed. Initially, spatiotemporal alignment of the detected and reference actions is performed. Before data analysis, bone normalization is applied, as a preprocessing step aiming to prevent erroneous analysis due to body structure differences. Afterwards, spatial action alignment is achieved using the rotational offset between the bodies, extracted by the shoulders and torso 3D positions [32]. It is worth noting that the sequences are of the same action, thus the comparison is not affected by spatial alignment and is performed exclusively on the 8 limb joints, i.e., the elbows, wrists, knees and ankles.

Applying multivariate Dynamic Time Warping (m-DTW) on the 3D joint positions and KEs, motion data are temporally aligned and synchronization feedback can be obtained. The 3D joint position and linear velocity errors are subsequently calculated and normalized, enabling statistical joint error analysis. Error statistics are fed to a fuzzy logic engine developed to produce semantic feedback, including information regarding the action similarity to the reference as well as the most erroneous limbs and joints, while

Table 1
Sample CVD dataset exercises.

Exercise	Key poses
Standing gluteus medius	
Side step knee-elbow	
Lateral lunge	

also providing instructions for improving action performance. It is worth noting that the proposed method can function even with slightly different actions, since the analysis is based on normalized error values, while a minimum similarity threshold can be manually set, so that the actions can be considered similar.

4. Experiments

Experimental evaluation of the proposed method on two public datasets and a new one, introduced in this paper, is presented in the following sections.

4.1. CVD exercise dataset

In the context of *PATHway H2020* project, funded by the European Union and aiming to aid cardiac rehabilitation (CR) teaching patients how to manage their cardiovascular disease (CVD), a new dataset was collected, composed of exercises specifically selected to help patients sustain a minimum of physical activity and get healthier through it. A subset of this dataset, consisting of 15 exercises performed by 15 subjects (4 female and 11 male) 2–3 times each, resulting in a total of 656 sequences and 28,979 frames, was used for the evaluation of the proposed framework.¹ Data capturing was performed at 30 frames per second (fps) by one Kinect sensor facing the user. The exercise sequences were pre-segmented and the subjects were advised to perform unilateral exercises using their right limbs for some repetitions and their left limbs for the rest. Key frames of some of the exercises are presented in [Table 1](#), along with the corresponding tracked skeletons.

4.2. MSRC-12 Dataset

The Microsoft Research Cambridge-12 dataset [33] is a large, public gesture dataset for action detection, comprising of 594 unsegmented sequences depicting 30 subjects performing 12 gestures. It was captured by Kinect sensor, at 30 fps and contains 6244 gesture instances and 719,359 frames in total. The 3D positions of 20 body joints are provided along with annotation files, denoting the action points [19] at which action detection should be triggered. More specifically, the included gestures are: *beat both*, *bow*, *change weapon*, *duck*, *goggles*, *had enough*, *kick*, *lift outstretched arms*, *push right*, *shoot*, *throw*, *wind it up* and the data were acquired giving different kinds of instructions to the participants.

¹ CVD dataset is publicly available at <http://vcliti.gr/maadre>.

Table 2
Comparison results of action recognition experiments on the CVD dataset.

Method	Classification Accuracy	
	pre-segmented	auto-segmented
Meshry et al. [12]	97.76%	97.19%
Proposed	98.54%	98.04%

Table 3

Comparison results of action recognition experiments on half the descriptors of the pre-segmented CVD dataset, applying KE sampling and joint weighting, separately.

Method	Classification Accuracy
Meshry et al. [12]	93.28%
KE sampling (only)	96.30%
Weighting (only)	98.04%

4.3. MSR-Action3D dataset

The MSR-Action3D dataset [34] is a public, benchmark, depth map and skeleton sequence dataset for action recognition, consisting of 567 pre-segmented sequences of 20 actions, performed by 10 subjects 2–3 times. It was captured with the aid of a depth camera using infra-red light at 15 fps and contains 23,797 samples frames with 20 3D joint positions for each of them. The constituting actions are: *bend*, *draw circle*, *draw tick*, *draw x*, *forward kick*, *forward punch*, *golf swing*, *hammer*, *hand catch*, *hand clap*, *hand wave*, *high arm wave*, *high throw*, *horizontal arm wave*, *jogging*, *pickup & throw*, *side kick*, *tennis serve*, *tennis swing*, *two side-boxing*, and subjects were asked to perform unilateral exercises with their right arm or leg. Alike most works reporting results on this dataset, 10 sequences were omitted, due to missing or highly erroneous data, thus resulting in 557 sequences in total [29].

4.4. Human action detection/recognition experimental results

4.4.1. CVD Dataset

The cross-subject experimental setup used in [12] was adopted for the evaluation of the proposed method on the CVD dataset. For the action recognition experiments, half of the available subjects, namely {1, 3, 5, 7, 9, 11, 13} constituted the training set, and the remaining the test set. The same experiment was performed 5 times, with a different codebook each, and mean classification accuracy is reported. [Table 2](#) summarizes our method performance both on the pre-segmented dataset and on the segments that emerged after automatic detection on the randomly concatenated test set samples.

Taking a closer look at [Table 2](#), it can be noticed that in both cases the proposed approach outperforms ELS, using half the descriptors employed by the latter, chosen based on their KE, and applying joint weighting. Taking this into account, the importance of exploiting motion information as well as the discriminative power of weighting are highlighted, as also confirmed by [Table 3](#), in which experiments applying KE based sampling and joint weighting, independently, are presented.

While ELS accuracy reported in [Table 2](#) was achieved using all the dataset descriptors, the corresponding accuracy in [Table 3](#) achieved employing only half the descriptors, as is also the case with the proposed method, is substantially lower. Exploiting KE for descriptor sampling, instead of performing it at random, though, seems to aid in retaining the most important descriptors and, thus, the most discriminative information. Similarly, the contribution of joint weighting on efficient action recognition, is highlighted by the fact that the proposed method outperforms ELS not

Table 4
Detection mean F-score results on CVD dataset.

	Meshry et al. [12]	Proposed
$\Delta = 333$ ms	0.85	0.89
Overlap 0.2	0.91	0.96

Table 5

Comparison mean F-score and standard deviation results of action detection experiments on the different modalities of MSRC-12 dataset at 0.2 overlap ratios.

	Sharaf et al. [16]	Meshry et al. [12]	Proposed
Video - Text	0.684 ± 0.074	0.921 ± 0.126	0.983 ± 0.008
Image - Text	0.687 ± 0.099	0.894 ± 0.085	0.905 ± 0.007
Text	0.558 ± 0.092	0.788 ± 0.139	0.851 ± 0.012
Video	0.669 ± 0.082	0.895 ± 0.068	0.927 ± 0.009
Image	0.598 ± 0.082	0.858 ± 0.086	0.894 ± 0.010
Overall	0.639	0.871	0.912

only in case they both use half the dataset descriptors (Table 3), but even when the latter is applied on the entire descriptor set, in which case its accuracy is 97.79% (see Table 2).

Splitting of dataset subjects in two halves was also employed for action detection. The subjects used for training in this case were {1, 2, 3, 4, 5, 6, 7} and the experiment was run 100 times, randomly alternating sample concatenation order. The mean F-scores reported in the first row of Table 4 were obtained regarding as accurate those detections triggered within 10 frames (i.e., 333 ms) from the ground truth segment end. In the following row, the overlap between the ground truth action interval and the detected one constitutes the factor determining whether the detection can be considered as positive or not. Again, the proposed approach outperforms ELS.

4.4.2. MSRC-12 Dataset

Adhering to the experimental setup introduced by Fothergill et al. [33], for each of the instruction modalities of MSRC-12 dataset, the ‘leave-persons-out’ protocol was adopted, keeping the minimum subject subset containing all the gestures as test set and employing all the others for training. Ten such experiments were performed and the resulting mean F-score for the test sets is reported. The dataset annotation provided by Hussein et al. [27], considering action instances to begin at the annotation start and end at the frame annotated as action point, is also employed, so that comparable results can be obtained.

Table 5 summarizes detection results for relative temporal overlap percentage to the ground truth annotation at least 0.2. It can be easily observed that in all modalities the proposed approach outperforms both [12,16], achieving state-of-the-art results. Furthermore, it should be noted, that the mean F-score standard deviations of our method are the lowest reported, which is indicative of the robustness of the method, highlighting the fact that the results are not affected by the test data to a great extent.

4.4.3. MSR-Action3D dataset

Following the cross-subject experimental setup introduced in [34], action recognition results obtained on the pre-segmented MSR-Action3D dataset using half of the subjects (i.e., {1, 3, 5, 7, 9}) for training and the remaining half for testing, are reported in Table 6. Mean average results over the 3 Action Sets on 5 runs with different codebooks are presented and alike [12], parameter fine tuning is performed on the training set and used universally.

It should be noted that the results achieved by the proposed method, applying joint weighting and using only half of the available descriptors are comparable to the state-of-the-art reported by Luo et al. [23], slightly outperforming them, while also outper-

Table 6

Comparison results of action recognition experiments on the pre-segmented MSR-Action3D dataset.

Method	Classification Accuracy
Li et al. [34]	74.70%
Wang et al. [29]	88.20%
Hussein et al. [27]	90.53%
Zanfir et al. [18]	91.70%
Meshry et al. [12]	96.05%
Luo et al. [23]	96.70%
Proposed	96.77%

Table 7

Comparison results of action recognition experiments on MSR-Action3D dataset after automatic segmentation.

Method	Classification Accuracy
Shan et al. [13]	84.0%
Sharaf et al. [16]	91.1%
Proposed	96.2%

Table 8

Overlap detection mean average precision and F-score on MSR-Action3D dataset at 0.2 overlap ratio.

Method	MAP	F-score
Meshry et al. [12]	0.902	0.930
Proposed	0.915	0.919

Table 9

Mean action recognition results on MSR-Action3D dataset Action Sets.

Experiment	AS1	AS2	AS3
pre-segmented	94.7%	96.2%	96.3%
automatically segmented	94.9%	95.8%	96.1%

forming ELS method [12], which does not perform descriptor sampling for codebook construction in this dataset.

Similar recognition accuracy was obtained after randomly concatenating sequences and performing action recognition on the automatically detected segments, as reported in Table 7. As was also the case with the CVD dataset (see Table 2), the results on the pre-segmented data are slightly higher than those on the automatic segments. This can be attributed to the fact that the segments may contain irrelevant frames, while other important frames may be missing, making them ‘unclear’ and thus burdening recognition.

As far as action detection for 0.2 overlap ratio is concerned, Table 8 shows that the proposed method outperforms ELS in detection mean average precision, while the opposite happens with F-score, which remains competitive, though. The reason for this, is that the devised technique for joint weighting is more effective on actions with a greater range of motion than the ones contained in the MSR-Action3D dataset. This was also verified by the fact that Action Set 3, which contains the most complex and ‘active’ gestures was the one on which the proposed approach achieved the highest recognition results in all the performed experiments, as shown in Table 9.

To sum up, the proposed human action detection and recognition approach has been proven to be very effective on all three datasets, outperforming other methods and achieving state-of-the-art results. Thus, automatic and dynamic weighting can be considered to enhance the descriptor discriminative power and KE-based descriptor sampling also seems to result in more descriptive codebooks.

Table 10

Semantic feedback and corresponding frames. The detected and reference skeletons are depicted with red and white colours, respectively. Bold text indicates the fuzzy inferences.

Semantic Feedback	Detected Frames
<p>Draw X - High Similarity!</p>	<p>The highest <i>POSITION</i> error is detected at Left Wrist, at the Latest temporal phase (frame 26) of the movement. Please, position your Left Wrist Left and Down, at this time instance.</p> <p>The highest <i>VELOCITY</i> error is detected at Left Wrist, at the Latest temporal phase (frame 26) of the movement. Please, move your Left Wrist Down and Backwards at this time instance.</p>
<p>Standing gluteus medius - High Similarity!</p>	<p>The highest <i>POSITION</i> error is detected at Left Wrist, at the Latest temporal phase (frame 38) of the movement. Please, position your Left Wrist Left and Up, at this time instance.</p> <p>The highest <i>VELOCITY</i> error is detected at Left Ankle, at the Early temporal phase (frame 10) of the movement. Please, move your Left Ankle Right and Forward at this time instance.</p>
<p>Bend - Low Similarity!</p>	<p>The highest <i>POSITION</i> error is detected at Right Wrist, at the Latest temporal phase (frame 13) of the movement. Please, position your Right Wrist Right and Down and Forward at this time instance.</p> <p>The highest <i>VELOCITY</i> error is detected at Left Wrist, at the Latest temporal phase (frame 13) of the movement. Please, move your Left Wrist Down and Backwards at this time instance.</p>

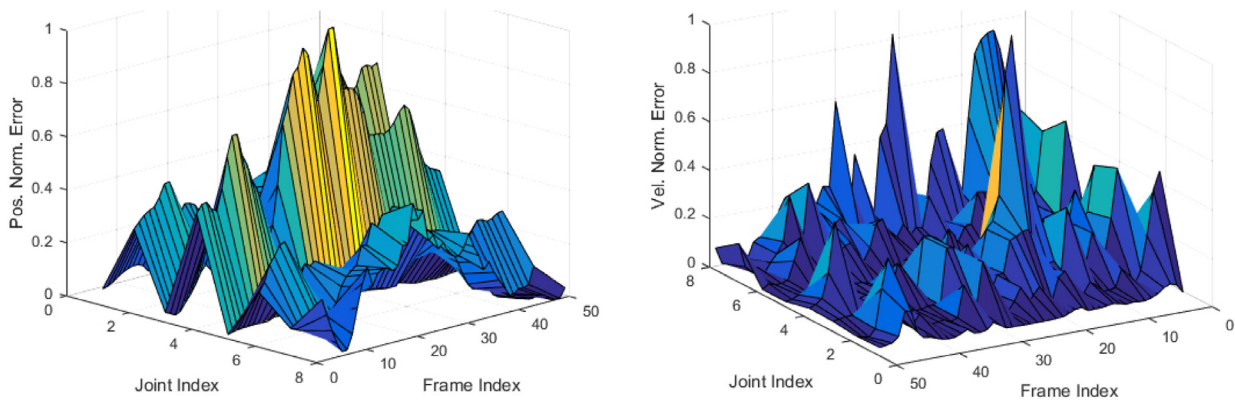


Fig. 10. Surfaces of similar actions. The mean and the standard deviation of all the normalized joint errors are high.

4.5. Human action evaluation experimental results

Feeding detected and reference actions to the evaluation component, the motion data are analyzed and, using the proposed fuzzy engine, semantic feedback is retrieved, as described in Section 3.2. For demonstration purposes, 3 different actions from the 3 presented datasets are evaluated, with the semantic feedback and the corresponding frames presented in Table 10. In order to facilitate reader understanding of joint velocities, the flow of the joint movements is also rendered.

The first evaluated action is MSR-Action3D Draw X. As shown in Table 10, both the position and velocity highest errors are detected at frame 26, and thus the semantic feedback provided instructs the user on how to perform the action for achieving higher similarity with the reference motion at this time instance. Even though the two actions seem to differ significantly at the frame with the highest errors, watching the corresponding supplementary video,² it can be easily perceived that the spatiotemporally aligned actions are of high similarity (as evaluated by the system).

CVD dataset *Standing gluteus medius* exercise is subsequently evaluated. It is worth mentioning that after spatiotemporal alignment, the errors between most of the joints are numerically close to the maximum mean joint error. In this case, the mean normalized position and velocity errors are high, while semantic feedback is retrieved for further performance improvement. Feeding the motion data calculated statistics (total as well as per joint mean and standard deviation of the joint motion features) to the fuzzy engine, allows similarity calculation. In particular, one of the fuzzy rules, based on the assumption that the performed actions have mean normalized error close to the maximum mean joint error and high standard deviation, as depicted in Fig. 10, stipulates that the similarity is high. *Bend* action from the MSRC-12 dataset is finally evaluated. In this case, erroneous motion capture has occurred due to self-occlusion when bending, resulting in high detected joint errors and, therefore, low similarity.

5. Conclusions

In this paper, a novel framework for motion analysis is proposed, performing action detection/recognition and evaluation based on motion capture data, employing pose and kinematics information for data description. Motion specific characteristics are exploited for efficient data weighting and KE features are introduced in the construction of the BoG-based data representation. Evaluation of the automatically detected and recognized action instances is subsequently performed, and semantic feedback is pro-

vided in the form of instructions for conquering the performed actions. Experimental results on three public datasets denote the effectiveness of the proposed framework, since it outperforms recently proposed state-of-the-art methods by 0.5–6%. Experiments performed applying weighting and KE sampling, separately, highlight the significance of the framework's contribution. The performed evaluation has also been proven accurate, providing insightful feedback, which can be easily perceived and adopted in order to master the actions. Thus, the proposed framework can be used for unsupervised physical exercise training in several applications, since it can perform a complete motion analysis, based on motion data captured with any of-the-shelf device.

Acknowledgement

This work was supported by the European Union's Horizon 2020 Framework Programme for Research and Innovation Action under Grant Agreement no. 643491.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [10.1016/j.patcog.2017.12.007](https://doi.org/10.1016/j.patcog.2017.12.007).

References

- [1] W. Shen, K. Deng, X. Bai, T. Leyvand, B. Guo, Z. Tu, Exemplar-based human action pose correction, *IEEE Trans. Cybern.* 44 (7) (2014) 1053–1066.
- [2] M. Vrigkas, C. Nikou, I.A. Kakadiaris, A review of human activity recognition methods, *Front. Rob. AI* 2 (2015) 28.
- [3] J. Lee, J. Chai, P.S. Reitsma, J.K. Hodgins, N.S. Pollard, Interactive control of avatars animated with human motion data, in: *ACM Transactions on Graphics (TOG)*, 21, ACM, 2002, pp. 491–500.
- [4] W. Lin, M.-T. Sun, R. Poovendran, Z. Zhang, Group event detection with a varying number of group members for video surveillance, *IEEE Trans. Circuits Syst. Video Technol.* 20 (8) (2010) 1057–1067.
- [5] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, E. Mavroudi, A. Katsamanis, A. Tsiami, P. Maragos, Multimodal human action recognition in assistive human-robot interaction, in: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2702–2706.
- [6] L. Piyathilaka, S. Kodagoda, Human activity recognition for domestic robots, in: *Field and Service Robotics*, Springer, 2015, pp. 395–408.
- [7] J.K. Tang, J.C. Chan, H. Leung, Interactive dancing game with real-time recognition of continuous dance moves from 3D human motion capture, in: *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication*, ACM, 2011, p. 50.
- [8] H. Nicolau, T. Guerreiro, R. Pereira, D. Gonçalves, J. Jorge, Computer-assisted rehabilitation: towards effective evaluation, *Int. J. Cognit. Perform. Support* 1 (1) (2013) 11–26.
- [9] H. Ghasemzadeh, R. Jafari, Coordination analysis of human movements with body sensor networks: a signal processing model to evaluate baseball swings, *IEEE Sens. J.* 11 (3) (2011) 603–610.
- [10] D.S. Alexiadis, P. Daras, Quaternionic signal processing techniques for automatic evaluation of dance performances from MoCap data, *IEEE Trans Multimedia* 16 (5) (2014) 1391–1406.

² Supplementary videos can be found at <http://vcl.iti.gr/maadre>.

- [11] D.Y. Kwon, M. Gross, Combining body sensors and visual sensors for motion training, in: Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology, ACM, 2005, pp. 94–101.
- [12] M. Meshry, M.E. Hussein, M. Torki, Linear-time online action detection from 3D skeletal data using bags of gesturelets, in: IEEE Winter Conference on Applications of Computer Vision (WACV), 2016, pp. 1–9.
- [13] J. Shan, S. Akella, 3D human action segmentation and recognition using pose kinetic energy, in: IEEE International Workshop on Advanced Robotics and its Social Impacts, 2014, pp. 69–75.
- [14] M. Devanne, H. Wannous, P. Pala, S. Berretti, M. Daoudi, A. Del Bimbo, Combined shape analysis of human poses and motion units for action segmentation and recognition, in: 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 7, 2015, pp. 1–6.
- [15] B.B. Amor, J. Su, A. Srivastava, Action recognition using rate-invariant analysis of skeletal shape trajectories, *IEEE Trans. Pattern Anal. Mach. Intell.* 38 (1) (2016) 1–13.
- [16] A. Sharaf, M. Torki, M.E. Hussein, M. El-Saban, Real-time multi-scale action detection from 3D skeleton data, in: IEEE Winter Conference on Applications of Computer Vision, 2015, pp. 998–1005.
- [17] X. Zhao, X. Li, C. Pang, Q.Z. Sheng, S. Wang, M. Ye, Structured streaming skeleton—a new feature for online human gesture recognition, *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* 11 (1s) (2014) 22.
- [18] M. Zanfir, M. Leordeanu, C. Sminchisescu, The moving pose: an efficient 3D kinematics descriptor for low-latency action recognition and detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2752–2759.
- [19] S. Nowozin, J. Shotton, Action Points: A Representation for Low-Latency Online Human Action Recognition, Technical Report MSR-TR-2012-68, Microsoft Research Cambridge, 2012.
- [20] A. Chan-Hon-Tong, C. Achard, L. Lucat, Simultaneous segmentation and classification of human actions in video streams using deeply optimized Hough transform, *Pattern Recognit.* 47 (12) (2014) 3807–3818.
- [21] D. Gong, G. Medioni, X. Zhao, Structured time series analysis for human action segmentation and recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* 36 (7) (2014) 1414–1427.
- [22] E. Ohn-Bar, M. Trivedi, Joint angles similarities and HOG² for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 465–470.
- [23] J. Luo, W. Wang, H. Qi, Group sparsity and geometry constrained dictionary learning for action recognition from depth maps, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1809–1816.
- [24] D.C. Mocanu, H.B. Ammar, D. Lowet, K. Driessens, A. Liotta, G. Weiss, K. Tuyls, Factored four way conditional restricted Boltzmann machines for activity recognition, *Pattern Recognit. Lett.* 66 (2015) 100–108.
- [25] H. Chen, G. Wang, J.-H. Xue, L. He, A novel hierarchical framework for human action recognition, *Pattern Recognit.* 55 (2016) 148–159.
- [26] A.A.A. Salih, C. Youssef, Spatiotemporal representation of 3D skeleton joints-based action recognition using modified spherical harmonics, *Pattern Recognit. Lett.* 83 (2016) 32–41.
- [27] M.E. Hussein, M. Torki, M.A. Gowayed, M. El-Saban, Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations, in: International Joint Conference on Artificial Intelligence (IJCAI), 13, 2013, pp. 2466–2472.
- [28] L. Xia, C.-C. Chen, J. Aggarwal, View invariant human action recognition using histograms of 3D joints, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012, pp. 20–27.
- [29] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2012.
- [30] J.C. Chan, H. Leung, J.K. Tang, T. Komura, A virtual reality dance training system using motion capture technology, *IEEE Trans. Learn. Technol.* 4 (2) (2011) 187–195.
- [31] J. Bentley, Programming pearls: algorithm design techniques, *Commun. ACM* 27 (9) (1984) 865–873.
- [32] S. Asteriadis, A. Chatzitofis, D. Zarpalas, D.S. Alexiadis, P. Daras, Estimating human motion from multiple kinect sensors, in: Proceedings of the 6th International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications, ACM, 2013, p. 3.
- [33] S. Fothergill, H. 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, pp. 1737–1746.
- [34] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3D points, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, 2010, pp. 9–14.

Fotini Patrona was born in Siatista, Kozani, Greece, on May 13, 1990. She received the B.Sc. degree in Applied Informatics from University of Macedonia, Thessaloniki, Greece in 2012 and the M.Sc. degree in Digital Media from Aristotle University of Thessaloniki, Greece in 2014.

She was a research assistant at the Artificial Intelligence and Information Analysis laboratory of the Department of Informatics in Aristotle University of Thessaloniki from 2013 to 2015 and at the Centre of Research & Technology - Hellas/Information Technologies Institute (CERTH/ITI) from 2016 to 2017.

She has participated in 4 research projects financed by national and European funds, her research interests include image and video processing, 2D/3D computer vision and pattern recognition and has co-authored 3 articles in refereed journals.

Anargyros Chatzitofis completed his studies at the Electrical & Computer Engineering School, in the National Technical University of Athens (NTUA), where, currently, he is studying for a Ph.D. in Computer Science.

Since 2012, he has been working as a Research Assistant at the Information Technologies Institute (ITI) of the Centre for Research and Technology Hellas (CERTH).

His main research interests include data acquisition and processing from several sensors, multimedia and stereo/multi-view processing, focusing on human motion capturing and analysis, by using depth sensors, Wireless Inertial Measurement Units and RGB-cameras. From 2013 to 2016, he (co)authored 1 paper in refereed journal and 10 international conference papers.

Dimitrios Zarpalas has joined the Information Technologies Institute in 2007, and is currently working as a post-doctoral Research Associate.

His current research interests include real time tele-immersion applications (3D reconstruction of moving humans and their compression), 3D computer vision, 3D medical image processing, shape analysis of anatomical structures, 3D object recognition, motion capturing and evaluation, while in the past has also worked in indexing, search and retrieval and classification of 3D objects and 3D model watermarking.

His involvement with those research areas has led to the co-authoring of 1 book chapter, 9 articles in refereed journals and 30 papers in international conferences. He has been involved in more than 10 research projects funded by EC, and Greek Secretariat of Research and Technology. He is a member of the Technical Chamber of Greece.

Petros Daras is a Principal Researcher at the Information Technologies Institute (ITI) of the Centre for Research and Technology Hellas (CERTH).

His main research interests include multimedia processing, multimedia and multimodal search engines, 3D reconstruction from multiple sensors, dynamic mesh coding, medical image processing and bioinformatics.

He has co-authored more than 40 papers in refereed journals, 29 book chapters and more than 100 papers in international conferences. He has served as a regular reviewer for a number of international journals and conferences.