Noname manuscript No.
(will be inserted by the editor)

Unsupervised Human Action Retrieval using Salient Points in 3D Mesh Sequences

**Christos Veinidis · Ioannis Pratikakis ·
Theoharis Theoharis**

Received: date / Accepted: date

Abstract The problem of human action retrieval based on the representation of the human body as a 3D mesh is addressed. The proposed 3D mesh sequence descriptor is based on a set of trajectories of salient points of the human body: its centroid and its five protrusion ends. The extracted descriptor of the corresponding trajectories incorporates a set of significant features of human motion, such as velocity, total displacement from the initial position and direction. As distance measure, a variation of the Dynamic Time Warping (DTW) algorithm, combined with a $k - means$ based method for multiple distance matrix fusion, is applied. The proposed method is fully unsupervised. Experimental evaluation has been performed on two artificial datasets, one of which is being made publicly available by the authors. The experimentation on these datasets shows that the proposed scheme achieves retrieval performance beyond the state of the art.

Keywords Human action retrieval, 3D mesh sequences, sequence descriptor, Dynamic Time Warping

Christos Veinidis
Department of Electrical and Computer Engineering, Democritus University of Thrace, Xanthi, Greece
E-mail: cveinidi@ee.duth.gr

Ioannis Pratikakis
Department of Electrical and Computer Engineering, Democritus University of Thrace, Xanthi, Greece
E-mail: ipratika@ee.duth.gr

Theoharis Theoharis
Computer Graphics Laboratory, Department of Informatics and Telecommunications, University of Athens, Athens, Greece
and
IDI, Norwegian University of Science and Technology (NTNU), Norway
E-mail: theotheo@idi.ntnu.no

1 Introduction

The problem of retrieving 3D static objects has been widely examined by both unsupervised methods, such as [17], [18] and [21] and by supervised ones, such as [2] and [3]. Recently, the retrieval problem has been extended to the domain of 3D mesh sequences, which are becoming more and more common. A challenging application in this category is human action retrieval and recognition, with applications in surveillance, video games, human-computer interaction etc.

Action retrieval and recognition has recently been addressed in the case of 2D video [33], [14], [35]. In recent years, human action retrieval and recognition using the skeletal representation of the human body has attracted the research interest. The use of a set of critical points to recognize a human action is first introduced by Johansson in [8]. In this work, a psychological experiment where a set of bright spots located on suitable points on the human body, demonstrates that the human optical system is able to discriminate different human actions by following these points across time. The representation of the human body as a skeleton, i.e. as a set of points which represent the joints and the connections between them, has become common. Benchmark datasets have been constructed that contain various actions of human skeletons, such as MSRA-3D [12] and UCF Kinect [13]. Many works have addressed human action retrieval and recognition using the skeletal representation, such as [34], [15], [1], [22], [20] and [28]. A comprehensive survey is presented in [19].

Following the spirit of Johansson's experiment, in this work we tackle the problem of human action retrieval in 3D mesh sequences using the trajectories of a suitably selected set of salient points on the human body. One of these points is the centroid of the human body which is representative of the general trajectory of an action and robust to data defects. However, this provides a very 'abstract' representation of human actions. Thus a set of additional 5 points as shown in Fig. 1 is extracted and their trajectories form the action descriptor. These 5 points correspond to the protrusions of the human body, i.e. the top of the head, the upper limb ends and the lower limb ends. As the human body can be segmented into 6 basic components (head, core body, arms and legs), each of these salient points is a representative of a body segment.

A set of features, which incorporates kinematic information for each segment of the human body, is extracted for each of the 6 aforementioned trajectories. The similarity computation between the sub-descriptors of the trajectories of identical salient points is based on the Dynamic Time Warping (DTW) algorithm. Finally, the resulting similarities are combined using the $k - means$ algorithm to extract the final similarity between two actions. Experimentation is performed using standard retrieval measures.

The main contributions of this paper are the following:

- A new method to extract the trajectories of the salient points of the human body in a consistent way.
- A new trajectory-based descriptor of human body mesh sequences combining the trajectories of the salient points.
- A new distance measure based on DTW and the fusion of multiple distance matrices.

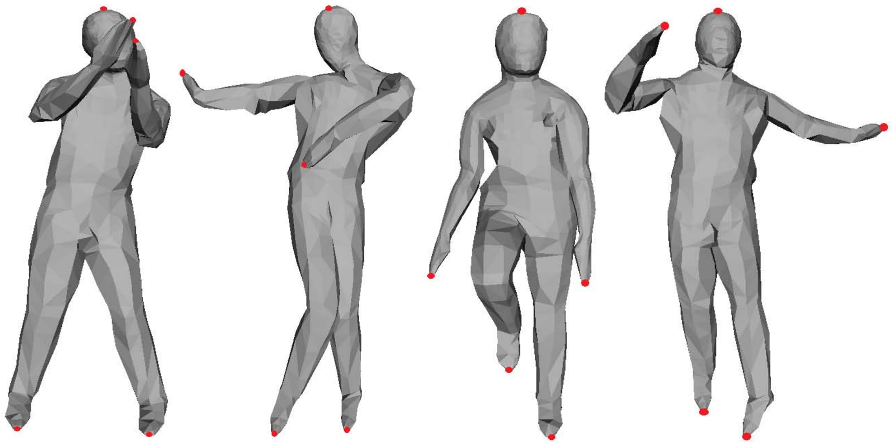


Fig. 1 The 5 protrusions of the human body that are used as salient points.

The remainder of this paper is organized as follows: In section 2, the related work in 3D mesh sequence retrieval is presented. In section 3, the proposed methodology is detailed. Section 4 is dedicated to the presentation and discussion of the experimental results. Finally, in section 5, conclusions are drawn and future work is discussed.

2 Related Work

In this section, related work on human action retrieval, using the 3D mesh representation for human models, is summarized. Both motion clips and full mesh sequences retrieval are considered.

Concerning motion clips retrieval, in [30], the vertices of the mesh in each frame are quantized spatially, forming 1024 spatial clusters. The center of mass of these clusters are used to generate the corresponding shape distribution histogram [16]. The full action sequences are segmented into motion clips based on the distance between the produced shape distribution histograms of successive frames. The motion clips retrieval is based on a Dynamic Programming algorithm, which is a variation of DTW. An enhanced version of this work is presented in [31] where an additional sub-descriptor is considered. This sub-descriptor is produced by quantizing the mesh in each frame into 128 spatial clusters and by constructing a histogram of the geodesic distances computed between the clusters' centroids. The whole mesh sequences are segmented into motion clips using the same methodology as in [30]. Furthermore, the dissimilarity measure between the motion clips is addressed via the use of DTW, where the dissimilarity between the descriptors of the frames is a weighted mean of the dissimilarities between each of the sub-descriptors. In [24], 5 salient points of the human body (the top of the head and the ends of the upper and lower limbs) are extracted in each frame of the sequences and the shortest geodesic paths these points are produced. The set of these geodesic paths is the static descriptor for the meshes in each frame and it is called Extremal Human Curve (EHC). The full sequences are segmented into motion clips using the local extrema of the velocity as the criterion of segmentation. The DTW distance between the EHC descriptors of the motion clips is used for the matching process. Additional experimentation related to data clustering and video summarization has been performed. In [10], the 3D mesh sequences are transformed to 2D sequences of silhouettes. Using the new representation, a descriptor, called P-Type Fourier descriptor, is extracted.The similarity between the motion clips is evaluated using a variation of DTW algorithm.

Concerning approaches with full mesh sequences, a supervised human action recognition method is presented in [6]. In this work, multi-view camera systems are used. The descriptors used in this paper, called 3D Motion Context and Harmonic Motion Context, are presented in [5]. These descriptors are based on the motion vector of the models. The motion vector is extracted using correspondences between the pixels of the cameras and the vertices of the meshes and, finally, combining the motion vectors of each view. The normalized correlation coefficient between the same kind of the two descriptors is used as similarity measure. The classifier is trained by generating a representative set of descriptors for each action class and a reference descriptor is estimated as the average of all descriptors for each action class. In [11], the mesh sequences are transformed to voxel sequences. Using the voxel-based representation, the human body orientation is estimated based on an estimation of feet direction. The models are normalized to be invariant to translation and scaling. Then, the similar postures of the sequences in the training set are clustered using the $k-means$ algorithm. The centers of the resulting clusters are called dynemes. Each dyneme is transformed into a vector containing the distances between the specified dyneme and the posture vectors. These distance vectors are normalized and each action is represented by averaging the distance vectors of all postures of this action. The dimensionality reduction is achieved by using the Linear Discrimination Analysis (LDA). For the classification an SVM classifier is used.

Directly comparable to the proposed approach, is the work in [27]. In this work, a set of state-of-the-art static shape descriptors are compared in human action retrieval framework. Particularly, each mesh in the frames of the sequences is represented by each of these descriptors, leading to sequences of static shape descriptors. The sequences are normalized in each dimension and a temporal filtering is applied. The similarity between the resulting sequences is evaluated using the DTW algorithm. Additional experimentation, using the Sakoe band for DTW computations, are performed.

3 Methodology

Let V be the set of vertices and F be the set of faces of a 3D mesh M, while a 3D mesh sequence S with N frames is denoted as $S =< M_t >$, where $t = 1, 2, ..., N$. The core components of the mesh sequence retrieval pipeline are the descriptors extraction and the similarity measure. To this end, the proposed methodology relies upon a descriptor that incorporates significant features of human motion upon trajectories of human body's salient points. The similarity measure is computed by using a variation of DTW coupled with a $k-means$-based method for multiple distance matrix fusion. In the sequel, a detailed description of the proposed methodology is given.

3.1 Trajectories extraction

The trajectories considered in the proposed methodology comprise those that are produced of five protrusion-oriented salient points and the trajectory of the mesh centroid. In the sequel, the detailed description of the salient points extraction, labeling and use across frames is given.

3.1.1 Protrusion-oriented salient point extraction

In this section, we describe the extraction of salient points $(SP_1, SP_2, ..., SP_5)$ which correspond to the protrusions of the human body, namely the head and the limbs. Starting from a random vertex, the geodesic distances between this vertex to all other vertices of the mesh are computed. The first salient point is the vertex with the maximum distance from the random vertex and the second salient point is determined as the vertex that is most geodesically distant to the first salient point. The third salient point is then the vertex with the maximum cumulative geodesic distance from the first two salient points. This process is continued until five salient points $SP_1, SP_2, ..., SP_5$ are extracted.

3.1.2 Protrusion-oriented salient point labeling

The salient points $SP_1, SP_2, ..., SP_5$ are next labeled according to the point in the human anatomy that they represent. A crucial observation for this labeling is that the feet are the longest protrusions of the human anatomy while the head is the shortest; the discrimination algorithm is thus based on the relative ordering of the cumulative geodesic distances from each of the 5 salient points.

We first compute the geodesic distances from each salient point to the others. Let $G_{i_1 i_2}$ for $i_1 = 1, 2, ..., 5$ and $i_2 = 1, 2, ..., 5$ represent the geodesic distance between salient points SP_{i_1} and SP_{i_2}. The cumulative geodesic distances $GD(SP_{i_1})$ using SP_{i_1} as starting point, for $i_1 = 1, 2, ..., 5$, are then evaluated as per Eq. 1:

$$GD(SP_{i_1}) = \sum_{\substack{i_2=1 \\ i_2 \neq i_1}}^{5} G_{i_1 i_2} \tag{1}$$

Since the human body is symmetric, there will be two pairs of cumulative distances that are (approximately) equal and correspond to the ends of the hands and feet. Since the feet are the longest and the head is the shortest protrusion in the human body, the following inequality will be satisfied:

$$GD(SP_{i_1} \equiv TH) < GD(SP_{i_1} \equiv UL) < GD(SP_{i_1} \equiv LL) \tag{2}$$

where TH, UL and LL refer to the top of the head, the ends of the upper limbs (either left or right) and the ends of the upper limbs (either left or right), respectively.

Inequality 2, allows us to label by $SP_1 ... SP_5$ the top of the head, the ends of the upper limbs and the ends of the lower limbs (the ends of limbs are not discriminated between left and right).

3.1.3 Salient points across frames

The algorithm of subsection 3.1.2 is not sufficient to identify corresponding salient points across frames of the sequence, as the cumulative geodesic distances for the left and the right corresponding ends of the limbs are (approximately) equal, due to the symmetry of the human body. Note that the head is uniquely identified as it corresponds to the minimum of the cumulative geodesic distances given in Eq. 1. We need to also detect the remaining salient points across frames and to this end, we assume that there is displacement coherence for the salient points across frames. Without loss of generality, let us suppose that salient points SP_2, SP_3 correspond to the ends of the upper limbs. If $\mathbf{p_{i_1}^{t-1}} = (x_{i_1}^{t-1}, y_{i_1}^{t-1}, z_{i_1}^{t-1})$ and $\mathbf{p_{i_2}^{t}} = (x_{i_2}^{t}, y_{i_2}^{t}, z_{i_2}^{t})$, for $i_1, i_2 = 2, 3$, are the positions of the salient points SP_2 and SP_3 in frames $t-1$ and t, respectively, we need to find which corresponds to which as we move from $t-1$ to t. To this end, four Euclidean distances $(d_{22}, d_{23}, d_{32}, d_{33})$ are computed:

$$d_{i_1 i_2} = \sqrt{(x_{i_1}^{t-1} - x_{i_2}^{t})^2 + (y_{i_1}^{t-1} - y_{i_2}^{t})^2 + (z_{i_1}^{t-1} - z_{i_2}^{t})^2} \tag{3}$$

for each $i_1, i_2 = 2, 3$. These four distances are shown in Fig. 2.

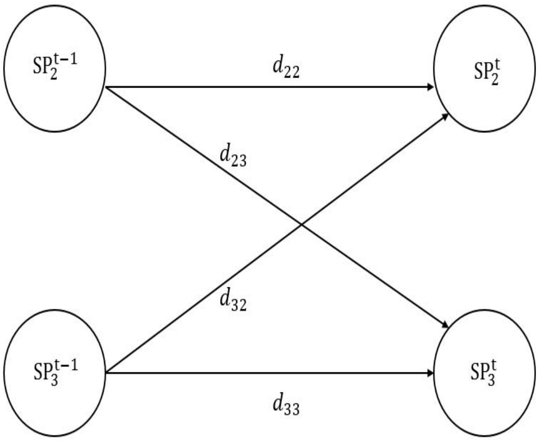


Fig. 2 The four distances $d_{22}, d_{23}, d_{32}, d_{33}$ computed in Eq. 3.

In the case of the ends of the upper limbs, point $\mathbf{p_{i_2}^{t}}, i_2 = 2, 3$ corresponds to $\mathbf{p_{i_1}^{t-1}}, i_1 = 2, 3$ if the following conditions are satisfied:

$$\frac{max\{d_{22}, d_{23}\}}{min\{d_{22}, d_{23}\}} > ratio \tag{4}$$

$$\frac{max\{d_{32}, d_{33}\}}{min\{d_{32}, d_{33}\}} > ratio \tag{5}$$

where, in our experiments, $ratio$ was set to 2. Obviously, two salient points of one frame should not be matched to the *same* salient point of the other frame.

In most cases, the above criterion is sufficient to track each of the left and right salient points across frames. However, in situations where two salient points come close together, the inter-frame distances are no longer discriminative. In such cases, the DTW algorithm [29] between geodesic paths is used. Let $P(SP_1^{t-1}, SP_2^{t-1})$, $P(SP_1^{t-1}, SP_3^{t-1})$ be the geodesic paths from the top of the head to the end of the two upper limbs in frame $t-1$ and $P(SP_1^t, SP_2^t)$, $P(SP_1^t, SP_3^t)$ be the corresponding geodesic paths in frame t. Four DTW distances $\{D_{22}, D_{23}, D_{32}, D_{33}\}$ are computed:

$$D_{i_1 i_2} = DTW(P(SP_1^{t-1}, SP_{i_1}^{t-1}), P(SP_1^t, SP_{i_2}^t)) \tag{6}$$

for $i_1, i_2 = 2, 3$. The pairs of salient points (SP_2^{t-1}, SP_2^t) and (SP_3^{t-1}, SP_3^t) are corresponding if the minimum of the set $\{D_{22}, D_{23}, D_{32}, D_{33}\}$ is D_{22} or D_{33} while the pairs of salient points (SP_2^{t-1}, SP_3^t) and (SP_3^{t-1}, SP_2^t) correspond if the minimum of the set is D_{23} or D_{32}. The same method is used for the ends of the lower limbs.

Note that the DTW-based method can solve the correspondence problem across frames without using the Euclidean distance method. However, DTW distance computation is far more expensive than the corresponding Euclidean distance computation, so the DTW-based method is only resorted to if the conditions expressed by Eq. 4 and 5 are not satisfied.

In Fig. 3, three successive frames of an action are shown. In this case, the distance between the ends of the upper limbs of the model is low and the conditions in Eq. 4 and 5 are not satisfied. However, the corresponding geodesic paths in successive frames do not change significantly and they provide a reliable way to provide a temporal correspondence between the salient points.

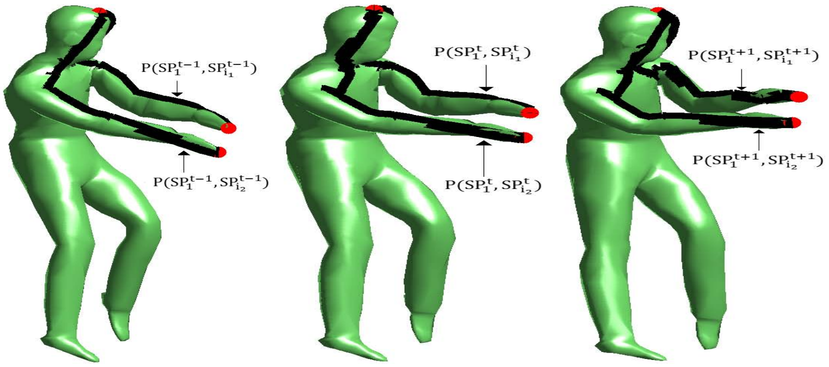

$P(SP_1^{t-1}, SP_{i_1}^{t-1})$
$P(SP_1^{t-1}, SP_{i_2}^{t-1})$
$P(SP_1^{t}, SP_{i_1}^{t})$
$P(SP_1^{t}, SP_{i_2}^{t})$
$P(SP_1^{t+1}, SP_{i_1}^{t+1})$
$P(SP_1^{t+1}, SP_{i_2}^{t+1})$

Fig. 3 Three successive frames with the corresponding geodesic paths between head and the ends of the upper limbs.

In addition to the 5 aforementioned trajectories, the trajectory of the centroid of the human body is also extracted, which is a representative point for the whole mesh. Given a mesh with $\mid V \mid$ vertices x_n, y_n, z_n for $n = 1, 2, ..., \mid V \mid$, the centroid of the mesh in frame t is defined as:

$$\mathbf{p}_6^t = \frac{1}{\mid V \mid} \cdot (\sum_{n=1}^{|V|} x_n, \sum_{n=1}^{|V|} y_n, \sum_{n=1}^{|V|} z_n)^T \tag{7}$$

The trajectory of the centroid of a mesh sequence is defined as the successive positions of the centroid along the sequence.

3.2 Feature extraction

Feature extraction is preceded by a normalization step with respect to translation, scale and direction of motion; a noise filter is also applied in this step on the trajectories of the salient points $\mathbf{p}_i^t, i = 1, 2, ..., 6$.

3.2.1 Normalization step

To make the trajectories invariant to translation, we subtract the mean value of the trajectory coordinates. If $\mathbf{Tr_x}, \mathbf{Tr_y}, \mathbf{Tr_z}$ are the sequences of x, y, z coordinates of the trajectory, respectively, and $mean(\mathbf{Tr_x})$, $mean(\mathbf{Tr_y})$, $mean(\mathbf{Tr_z})$ are their mean values, respectively, the trajectory becomes:

$$\boldsymbol{Tr1} = (\mathbf{Tr_x} - mean(\mathbf{Tr_x}), \quad \mathbf{Tr_y} - mean(\mathbf{Tr_y}), \quad \mathbf{Tr_z} - mean(\mathbf{Tr_z})) \tag{8}$$

where the mean value of each component of ***Tr1*** is zero. For scale invariance, we divide them by the length of the vector that defines the most distant point from the origin. Thus the maximum magnitude value of the trajectories corresponds to one unit. The scale invariant trajectory ***Tr2*** is given by:

$$\boldsymbol{Tr2} = \frac{\boldsymbol{Tr1}}{max\{||\mathbf{p}_i^t||\}_{t=1}^L} \tag{9}$$

where L is the number of frames of the sequence.

For denoising, we next apply a median filter with a window of size 3 frames to each of $\mathbf{Tr2_x}$, $\mathbf{Tr2_y}$ and $\mathbf{Tr2_z}$, thus producing the sequence components $\mathbf{Tr3_x}$, $\mathbf{Tr3_y}$ and $\mathbf{Tr3_z}$ of the filtered trajectory ***Tr3***.

Finally, it is necessary to perform a restricted form of rotation normalization where we normalize with respect to the direction of motion, as actions should not be differentiated if they are performed, for example, eastwards or westwards. Consider the 'horizontal' plane of motion as the xz plane; the maximum motion difference is then assigned to the x axis, while the minimum motion difference to the z axis. The 'vertical' component is assigned to the y axis. To this effect, an exhaustive search of the rotated trajectory about the y axis is performed (by 180^o with 1 degree increments) with the aim of maximizing the difference between the maximum and minimum value on the x axis. This normalization step results in the trajectory ***Tr4***.

3.2.2 Action sequence descriptor

The mesh sequence features are based on the velocity vector and the overall dynamics vector of a sequence which are complemented by the sequence of relative positions of the salient points in each frame. In particular, the proposed features are detailed as follows:

- The vector of overall dynamics of each salient point $\mathbf{p_i}$, for each $i = 1, 2, ..., 6$, with respect to the initial frame is defined as in [32]:
$$\mathbf{d}_i^t = \mathbf{p}_i^{t+1} - \mathbf{p}_i^1 = (x_{p_i}^t - x_{p_i}^1, y_{p_i}^t - y_{p_i}^1, z_{p_i}^t - z_{p_i}^1) \tag{10}$$
for each frame $t = 1, 2, ..., L-1$ and $i = 1, 2, ..., 6$. This vector represents the displacement of each salient point $\mathbf{p_i}$, for each $i = 1, 2, ..., 6$, with respect to its initial position.
The horizontal and vertical components of the overall dynamics vector are $\mathbf{d}_{i,h}^t = (x_{p_i}^t - x_{p_i}^1, z_{p_i}^t - z_{p_i}^1)$ and $\mathbf{d}_{i,v}^t = (y_{p_i}^t - y_{p_i}^1)$ for each $t = 1, 2, ..., L-1$ and $i = 1, 2, ..., 6$.
- The velocity vector is defined as:
$$\mathbf{v}_i^t = \mathbf{p}_i^{t+1} - \mathbf{p}_i^t = (x_{p_i}^{t+1} - x_{p_i}^t, y_{p_i}^{t+1} - y_{p_i}^t, z_{p_i}^{t+1} - z_{p_i}^t) \tag{11}$$
and its horizontal and vertical components are $\mathbf{v}_{i,h}^t = (x_{p_i}^{t+1} - x_{p_i}^t, z_{p_i}^{t+1} - z_{p_i}^t)$ and $\mathbf{v}_{i,v}^t = (y_{p_i}^{t+1} - y_{p_i}^t)$ respectively for each $t = 1, 2, ..., L-1$ and $i = 1, 2, ..., 6$. Thus, $\mathbf{d}_{i,h}$ and $\mathbf{v}_{i,h}$ are 2D time trajectories, while $\mathbf{d}_{i,v}$ and $\mathbf{v}_{i,v}$ are 1D time trajectories, for each $i = 1, 2, ..., 6$.
- The pairwise differences between the salient points in each frame. Specifically, this feature vector is defined in frame t as follows:
$$\mathbf{pd}^t = \mathbf{p}_{i_1}^t - \mathbf{p}_{i_2}^t \tag{12}$$
where $i_1, i_2 = 1, 2, ..., 6$. This feature has $6^2 - 6 = 30$ components for each frame (all potential differences between the salient points except the differences each salient point's from itself).

The final descriptor of the sequence S is:

$$\mathbf{D}_i^S = (\mathbf{D}_{i,1}^S, \mathbf{D}_2^S) = ([\mathbf{D}_{i,1}^S(1), \mathbf{D}_{i,1}^S(2), \mathbf{D}_{i,1}^S(3), \mathbf{D}_{i,1}^S(4), \mathbf{D}_{i,1}^S(5), \mathbf{D}_{i,1}^S(6)], \mathbf{D}_2^S) \tag{13}$$

for each salient point $i = 1, 2, ..., 6$, where

$$\mathbf{D}_{i,1}^S = [([||\mathbf{v}_{i,h}^t||,\ ||\mathbf{v}_{i,v}^t||], [||\mathbf{d}_{i,h}^t||,\ ||\mathbf{d}_{i,v}^t||], \mathbf{v}_{i,h}^t, \mathbf{v}_{i,v}^t, \mathbf{d}_{i,h}^t, \mathbf{d}_{i,v}^t)]_{t=1}^{L-1} \tag{14}$$

where $||\mathbf{x}||$ is the magnitude of a vector $\mathbf{x}$ and for each salient point $i = 1, 2, ..., 6$. Also:

$$\mathbf{D}_2^S = [\mathbf{pd}^t]_{t=1}^L \tag{15}$$

For the sake of clarity, the points $\mathbf{p_i}$, for each $i = 1, 2, ..., 6$, denote the points of the normalized trajectories ***Tr4***, as described in subsection 3.2.1.

3.3 Similarity measure

The DTW algorithm is used as a distance measure between two mesh sequences. In this algorithm, a minimum cost path is determined on an $M \times N$ grid, where M, N are the number of frames of two sequences respectively [29]. This optimal path is called warping path. However, the DTW algorithm enforces that the starting point of the warping path is $(1, 1)$ and the ending point is (M, N). This is a serious limitation of DTW in cases where two sequences which belong to the same action class are *temporally misaligned*, i.e. one of the two sequences is temporally delayed against the other.

To overcome this restriction, the DTW algorithm is applied on pre-aligned sequences using a sample cross-correlation criterion. The sample cross-correlation function is evaluated between the amplitude of the velocity vector $||\mathbf{v}_6||$ of the centroid of the query-sequence Q and the target-sequence T.

Let l^* be the point where the sample cross-correlation function is maximized. We can distinguish three cases:

- If $l^* = 0$, the two sequences are well aligned and the DTW algorithm is applied between the descriptors of Q and T without changes.
- If $l^* > 0$, the best alignment between the descriptors of Q and T occurs when we skip the first $l^* - 1$ items from the second sequence.
- If $l^* < 0$, the best alignment between the descriptors of Q and T occurs when we skip the first $|l^*| - 1$ items from the first sequence.

In other words, if $l^* > 0$, the starting point on the grid for DTW is $(1, l^*)$ while if $l^* < 0$, the starting point is $(|l^*|, 1)$; only if $l^* = 0$ is the starting point $(1, 1)$ as in the original DTW. In Fig. 4 the sample cross-correlation function values between the velocity magnitudes of the centroid trajectory of two actions is shown.

In this case, the sample cross-correlation function is maximized when $lag = 8$, so the first 7 frames of the second sequence are skipped and then the DTW distance between the two sequences is computed.

Let L'_Q and L'_T be the length of the sequences Q and T, respectively, after the pre-alignment based on sample cross-correlation. The distances between the sub-descriptors $\mathbf{D}^Q_{i,1}(r)$ of Q and $\mathbf{D}^T_{i,1}(r)$ of T are first computed as shown below.

$$Dist_{i,r}(Q,T) = DTW(\mathbf{D}^Q_{i,1}(r), \mathbf{D}^T_{i,1}(r)) \; / \; min\{L'_Q - 1, L'_T - 1\} \tag{16}$$

for $r = 1, 2, ..., 6$ and for each salient point $i = 1, 2, ..., 6$.

Distances $Dist_3 - Dist_6$ are weighted by the weights $w_3 - w_6$, where

$$w_{i,r} = \frac{max\{dif^Q_{i,r}, dif^T_{i,r}\}}{min\{dif^Q_{i,r}, dif^T_{i,r}\}} \tag{17}$$

and

$$dif^S_{i,r} = \max_t ||\mathbf{D}^S_{i,1}(r)|| - \min_t ||\mathbf{D}^S_{i,1}(r)|| \tag{18}$$

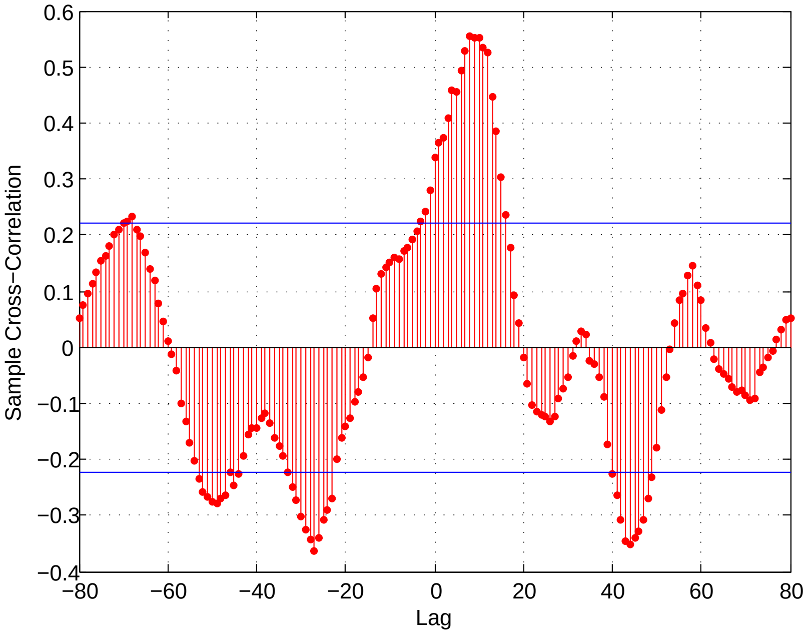


Fig. 4 The sample cross-correlation values between the velocity magnitudes of the centroid trajectory of two actions.

for $i = 1, 2, 3, 4, 5, 6$, $r = 3, 4, 5, 6$ and $S \in \{Q, T\}$. The weights of Eq. 17 aim to further discriminate the actions where the horizontal component of the corresponding sub-descriptors dominate the vertical component and vice versa.

The distance between sub-descriptor $\mathbf{D}_2$, given in Eq. 15, is computed as follows:

$$Dist_7(Q,T) = DTW(\mathbf{D}_2^Q\ ,\mathbf{D}_2^T)\ /\ min\{L'_Q, L'_T\} \tag{19}$$

Setting each sequence of the datasets as query, a set of distance matrices is created using the above process. Since $\mathbf{D}_1$ (Eq. 14) consists of 6 different sub-descriptors and there are 6 salient points, 36 distance matrices related to $\mathbf{D}_1$ are created. Also, since there are 30 pairwise distances between the 6 salient points (without taking into account the differences each salient point's from itself), 30 distance matrices related to $\mathbf{D}_2$ are created. Thus, 66 distance matrices are produced in total.

To generate a single final distance matrix from these 66 distance matrices, the following process is followed: The $k - means$ algorithm with 10 clusters is applied for the values contained in each of the 66 distance matrices produced for the given query sequence. The n^{th} row of each of the 66 distance matrices relates to the distances of the sequence with index n to all other sequences, which includes the zero value i.e. the distance of the given sequence from itself. The distance values which are contained in the same cluster as the zero value determine the sequences whose distance from the query is smallest, i.e. the ones that are most similar.

Let Q^j_{row} be the row of the j^{th} distance matrix, $j = 1, 2, ..., 66$, containing the distances between the sequence Q and the target sequences T. We proceed in producing an ordering of the distances in this row. Our aim is to divide the values in this ordering so that close values are kept together. For this purpose, a $k - means$ algorithm is applied wherein the cluster which contains the items that are grouped together with the item having zero distance (the distance to itself) determines the weighting factor as shown in Eq. 20. In particular, this factor is computed by taking into account the cardinality of the chosen cluster $C^{Q,j}$.

$$S(Q,T) = \sum_{j=1}^{66} m_j \cdot \frac{1}{card(C^{Q,j})} \tag{20}$$

where $card(C^{Q,j})$ is the cardinality of $C^{Q,j}$ and:

$$m_j = \begin{cases} 0, & T \notin C^{Q,j} \\ \\ 1, & T \in C^{Q,j} \end{cases} \tag{21}$$

for each $j = 1, 2, ..., 66$. In other words, the weights are only applied if the target sequence is an element of $C^{Q,j}, j = 1, 2, ..., 66$. As a consequence of the weighting, the smaller the $card(C^{Q,j}), j = 1, 2, ..., 66$, the larger its expected discriminative power and thus its weight.

4 Experimental results

4.1 Datasets

Two artificial datasets have been used for evaluation, named the USurrey dataset and the DUTH dataset. The DUTH dataset is presented for first time in this paper and is being made publicly available for the research community.

4.1.1 DUTH dataset

The DUTH dataset is a new dataset containing 60 mesh sequences in total. Each of 6 models, both men and women, has performed 10 actions. The corresponding mesh sequences consist of a number of frames which ranges from 21 to 250. The 10 action classes are the following: (1) “hop on left foot”, (2) “jumping”, (3) “jumping forward”, (4) “jumping-Turn”, (5) “running”, (6) “walking-90 degrees turn left”, (7) “walking-90 degrees turn right”, (8) “walking”, (9) “walking with arms out - balancing”, (10) “washing window”. The corresponding data are made publicly available through [37] and example frames are shown in Fig. 5.

For the process of dataset creation, we transformed publicly available motion files [38] from BVH format to mesh sequences. In order to produce human models with different characteristics, such as height, weight, age etc, the open source software MakeHuman [39] was used. In order to animate the 3D characters produced by MakeHuman, the free and open source Blender 3D software suite [40] was used.

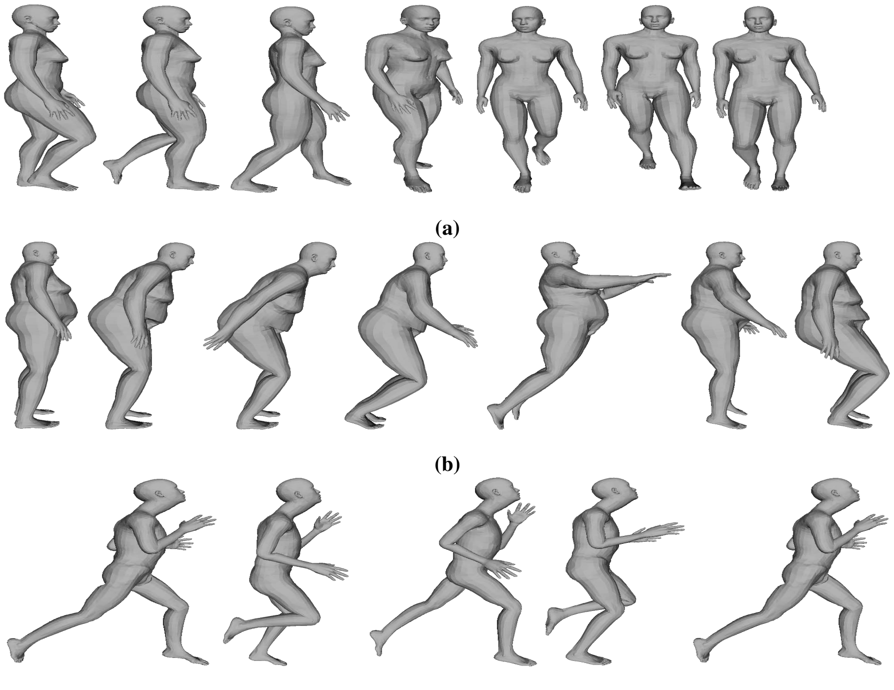
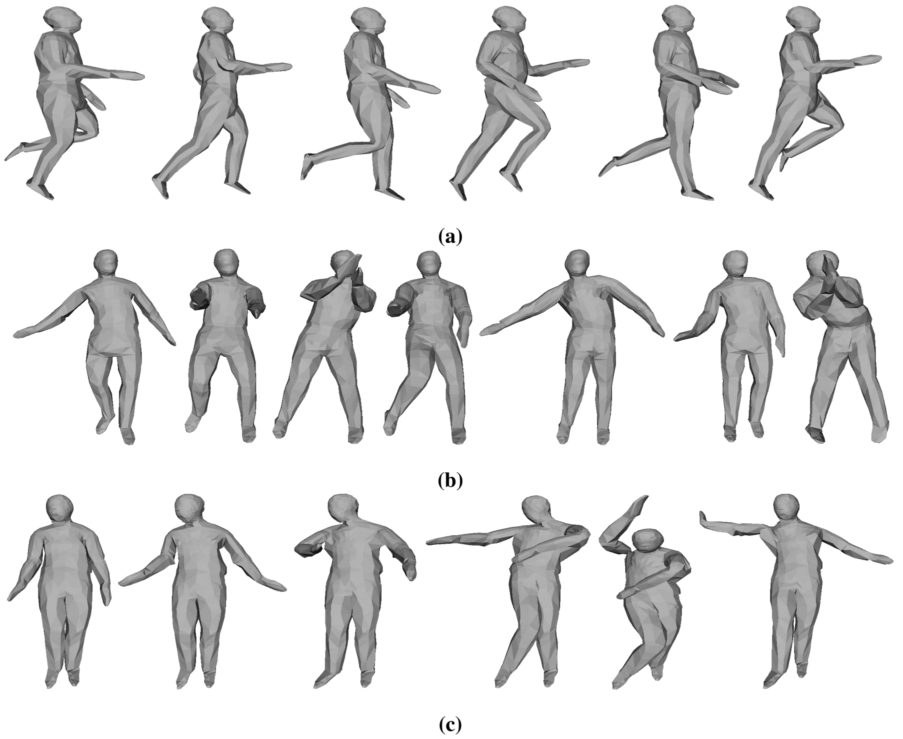


Fig. 5 Example frames from the DUTH-Artificial dataset for the actions: **(a)** “walking-90 degrees turn right”, **(b)** “jumping forward”, **(c)** “running”.

4.1.2 USurrey dataset

The USurrey dataset [25], [7] contains 392 mesh sequences in total. Specifically, each of 14 models, both men and women, has performed 28 actions. All mesh sequences consist of 100 frames and each frame consists of the same numbers of faces and vertices. Among the 28 actions, 17 are variations of the general action “walking”, 7 are variations of the general action “running” and 4 are other actions. The actions in this dataset are the following: (1) “faint”, (2) “fastrun”, (3) “fastwalk”, (4) “rocknroll”, (5) “runcircleleft”, (6) “runcircleright”, (7) “runturnleft”, (8) “runturnright”, (9) “shotarm”, (10) “slorun”, (11) “slowalk”, (12) “sneak”, (13) “sprint”, (14) “vogue”, (15) “walkcircleleft”, (16) “walkcircleright”, (17) “walkcool”, (18) “walkcowboy”, (19) “walkdainty”, (20) “walkelderly”, (21) “walkmacho”, (22) “walkmarch”, (23) “walkmickey”, (24) “walksexy”, (25) “walktired”, (26) “walktoddler”, (27) “walkturnleft”, (28) “walkturnright”. In Fig. 6, example frames of the artificial dataset are shown.



(a)

(b)

(c)

Fig. 6 Example frames from the USurrey-Artificial dataset for the actions: **(a)** “fastrun”, **(b)** “rocknroll”, **(c)** “vogue”.

4.2 Experimental evaluation

Performance evaluation in terms of retrieval was based on the following standard scalar measures:

- Nearest Neighbor (NN) : The percentage of queries where the closest match belongs to the query class.
- First Tier (FT) : The recall value for the $(C-1)$ closest matches were C is the cardinality of the query’s class.
- Second Tier (ST) : The recall value for the $2 \cdot (C-1)$ closest matches were C is the cardinality of the query’s class.
- Discounted Cumulative Gain (DCG) : A statistical measure which places more weight on correct results near the front of the retrieval list, under the assumption that a user is less likely to consider elements near the end of the list.

The values of the above metrics are in the interval $[0,1]$. Quantitative results can also be shown using precision-recall diagrams. Precision is the fraction of the retrieved sequences which belong to the same class as the query over the total number of retrieved sequences. Recall is the fraction of the retrieved sequences which belong to the same class as the query over the total number of sequences which belong to the same class as the query. For the experimental evaluation using these measures, the software offered by the University of Princeton was used [23].

In Table 1 the retrieval results using the four scalar metrics are given for the DUTH dataset. The corresponding precision-recall diagrams are shown in Fig. 7. The retrieval results of other methods are related to the optimal values presented in [27].

Table 1 Experimental retrieval results on the DUTH dataset (scalar metrics).

Method	NN	FT	ST	DCG
2D sub-descriptor of Hybrid [27], [17]	0.617	0.390	0.533	0.643
3D sub-descriptor of Hybrid [27], [17]	0.750	0.527	0.717	0.763
Hybrid [27], [17]	0.733	0.547	0.703	0.761
PANORAMA [27], [18]	0.717	0.553	0.650	0.748
Shape Distribution [27], [16]	0.633	0.367	0.563	0.638
Spin Images [27], [9]	0.517	0.337	0.537	0.602
Proposed Method	**0.967**	**0.767**	**0.863**	**0.907**

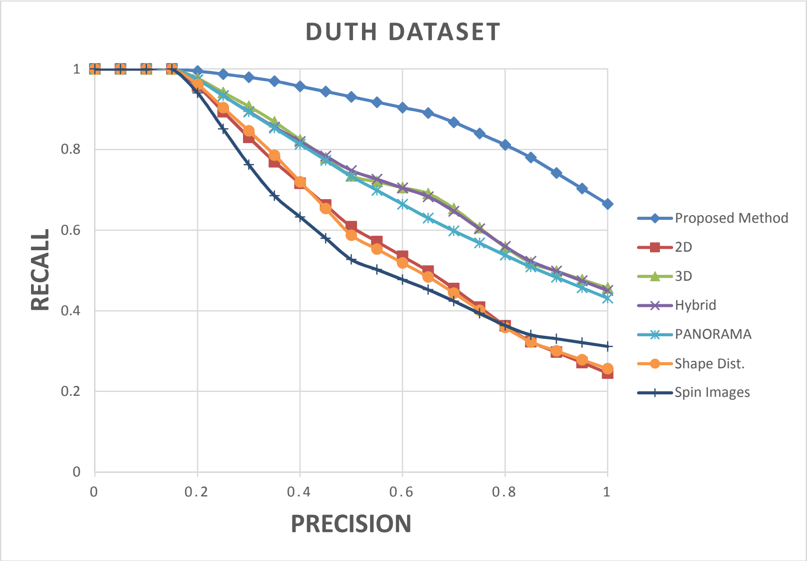


Fig. 7 Precision-recall diagrams for the DUTH dataset.

In Table 2 the retrieval results using the four scalar metrics are given for the USurrey dataset. The corresponding precision-recall diagrams are shown in Fig. 8. The retrieval results of the other methods are related to the optimal values presented in [27]. As can be seen, the proposed method achieves ideal, or in some metrics almost ideal, results, outperforming competing methods.

Table 2 Experimental retrieval results on the USurrey dataset with full sequences (scalar metrics).

Method	NN	FT	ST	DCG
2D sub-descriptor of Hybrid [27], [17]	0.995	0.979	**1.000**	0.997
3D sub-descriptor of Hybrid [27], [17]	**1.000**	0.983	0.999	0.999
Hybrid [27], [17]	0.980	0.968	0.999	0.994
PANORAMA [27], [18]	0.985	0.973	**1.000**	0.992
Shape Distribution [27], [16]	0.921	0.889	0.972	0.956
Spin Images [27], [9]	**1.000**	0.871	0.941	0.972
Proposed Method	**1.000**	**0.998**	**1.000**	**1.000**

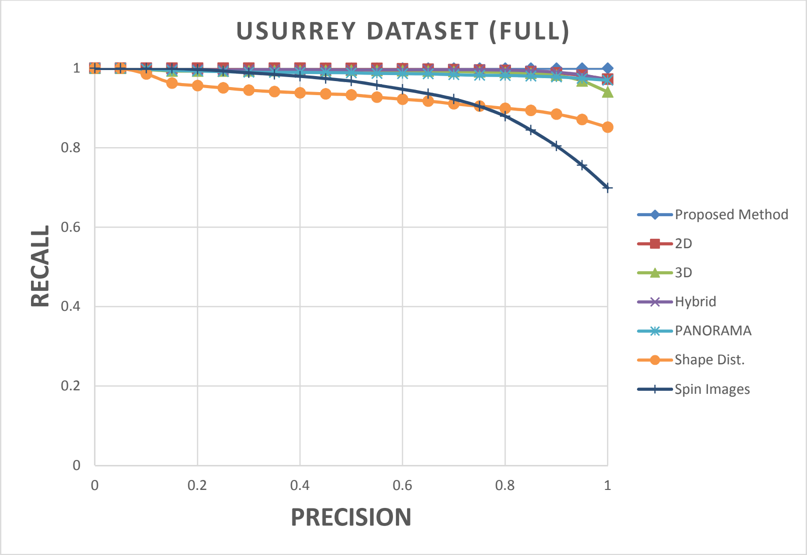


Fig. 8 Precision-recall diagrams for the USurrey dataset with full sequences.

In order to cancel the 1-1 correspondence between the frames of the actions of the USurrey dataset, an additional experiment has been performed. Specifically, corresponding sequences have been truncated by a random number of frames (up to 50) from the starting frame. Thus, part of each action is missing, so the sequences are temporally misaligned even though they may belong to same class. In Table 3, the retrieval results using the four scalar metrics are given and the corresponding precision-recall diagrams are shown in Fig. 9. The retrieval results of the other methods are related to the optimal values presented in [27]. Again, the proposed method almost always performs best and the reduction in retrieval performance is low compared to the corresponding results using the full sequences of the same dataset.

4.3 Ablation study

The proposed descriptor for a 3D mesh sequence, given in Eq. 13, is composed of a set of sub-descriptors for each salient point. In this section, an ablation study that reflects the contribution of each of the proposed sub-descriptors is presented.

Table 3 Experimental retrieval results on the USurrey dataset with truncated sequences (scalar metrics).

Method	NN	FT	ST	DCG
2D sub-descriptor of Hybrid [27], [17]	0.997	0.924	0.983	0.986
3D sub-descriptor of Hybrid [27], [17]	0.946	0.894	0.991	0.973
Hybrid [27], [17]	0.982	0.883	0.973	0.973
PANORAMA [27], [18]	0.946	0.902	**0.994**	0.973
Shape Distribution [27], [16]	0.890	0.797	0.903	0.926
Spin Images [27], [9]	0.993	0.771	0.870	0.937
Proposed Method	**1.000**	**0.962**	0.987	**0.994**

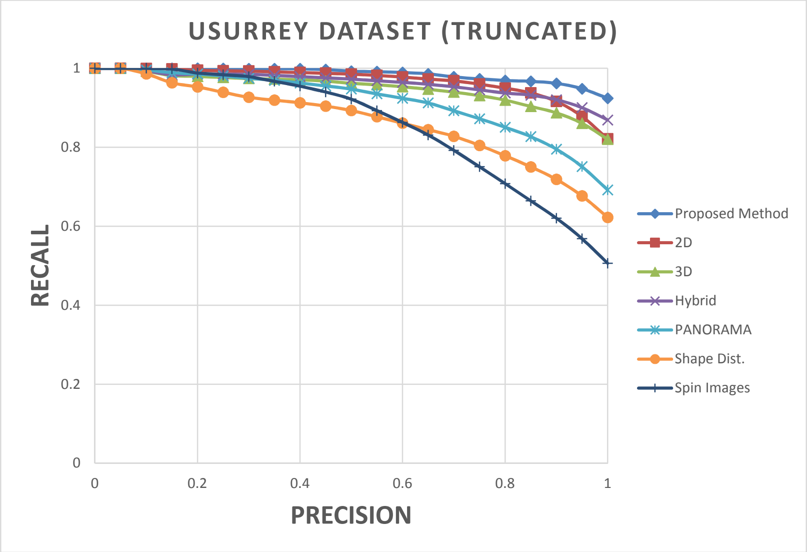


Fig. 9 Precision-recall diagrams for the USurrey dataset with truncated sequences.

The ablation study was realized by following a leave-one-out method on the full set of sub-descriptors for several rounds. At the end of each round we determine the least important sub-descriptor which is not further considered at the next round. The least important sub-descriptor is the one which, if left out, has the minimal effect on retrieval performance degradation. Initially, each time one sub-descriptor is omitted and the retrieval performance using the remaining sub-descriptors given in Eq. 13, is evaluated. As the number of sub-descriptors is 7, the number of initial experiments is also 7, where each experiment comprises the 6 remaining sub-descriptors. By this way, a set of 7 evaluations of retrieval performance is extracted. Finding the part of the descriptor of Eq. 13 with the maximum retrieval performance, we determine the single sub-descriptor which is the most useless in this step of the ablation study. This sub-descriptor is omitted and in the second step of the ablation algorithm the number of experiments is 6 as each experiment consists of 5 (out of 6) sub-descriptors. This process is repeated until only the two most useful sub-descriptors finally remain.

In order to measure the retrieval performance, precision-recall diagrams are used. In Fig. 10 - 12 the precision-recall diagrams resulting from the ablation study are shown. In successive precision - recall diagrams the sub-descriptor having the worst influence on the total retrieval results, is ignored. The legend on the right of each figure shows the sub-descriptor which is ignored each time. The enumeration of sub-descriptors is compatible with the enumeration in Eq. 14 and 15, i.e. $Subr$ is the r^{th} sub-descriptor of Eq. 14, $r = 1, 2, ..., 6$, while $Sub7$ represents $\mathbf{D}_2^S$, defined in Eq. 15. It is obvious that in all cases the sub-descriptor which provides the major gain is $Sub2$, i.e. the sub-descriptor which is related to the magnitude of the overall dynamics vector. As shown in Fig. 11, the retrieval performance is almost ideal in all steps of the ablation study. In the other two cases, the least significant sub-descriptor is $Sub1$, i.e. the sub-descriptor related to the magnitude of the velocity vector, as it is the first sub-descriptor which is ignored.

Concerning the selection of the centroid of magnitude of velocity vector as the critical point for pre-alignment, the intuition behind this selection is that the centroid is a point that describes reliably the whole movement of human body. In this case, all the 6 cases, one for each salient point, have been tested as the critical points for the desired pre-alignment. The above experiments have been performed for both the datasets used (also, for the two different versions of the USurrey dataset). The precision-recall diagrams using the other salient points in this criterion, are shown in Fig. 13 - 15. In the cases of the DUTH dataset and the USurrey dataset with full sequences, there are not significant variations in retrieval performance, while in the case of the USurrey dataset with truncated sequences the selection of the centroid is important.

4.4 Discussion

The proposed method achieves almost ideal retrieval results on the artificial dataset of USurrey and compares favorably to the state-of-the-art on the new artificial DUTH dataset that is being made publicly available as part of this paper for the benefit of the research community.

The majority of the actions included in the DUTH dataset belong to two main categories: variations of the general class “jumping” and variations of the general class “walking”. In Fig. 16 the relevant confusion matrix for the actions of the DUTH dataset is shown.

Most retrieval misses occur among actions which belong to the same general class (i.e. among the classes “jumping”-“jumping Forward”-“jumping-Turn” and the classes “hop on left foot”-“walking-90 degrees turn left”-“walking-90 degrees turn right”-“walking”). In particular, the majority of retrieval misses occurs for actions which are related to the general class “walking”. This is rather expected as intra-class variations are likely to be smaller than inter-class variations and fortunately these misses are few as per the performance figures.

Many retrieval false negatives occur between the actions “walking-90 degrees turn left” and “walking-90 degrees turn right”. These actions differ only in the direction of motion of the model. Part of the proposed descriptor is based on the velocity and dynamics vectors and the vector of pairwise differences; the phase of these vectors incorporates information about the motion direction, coupling the parts of the

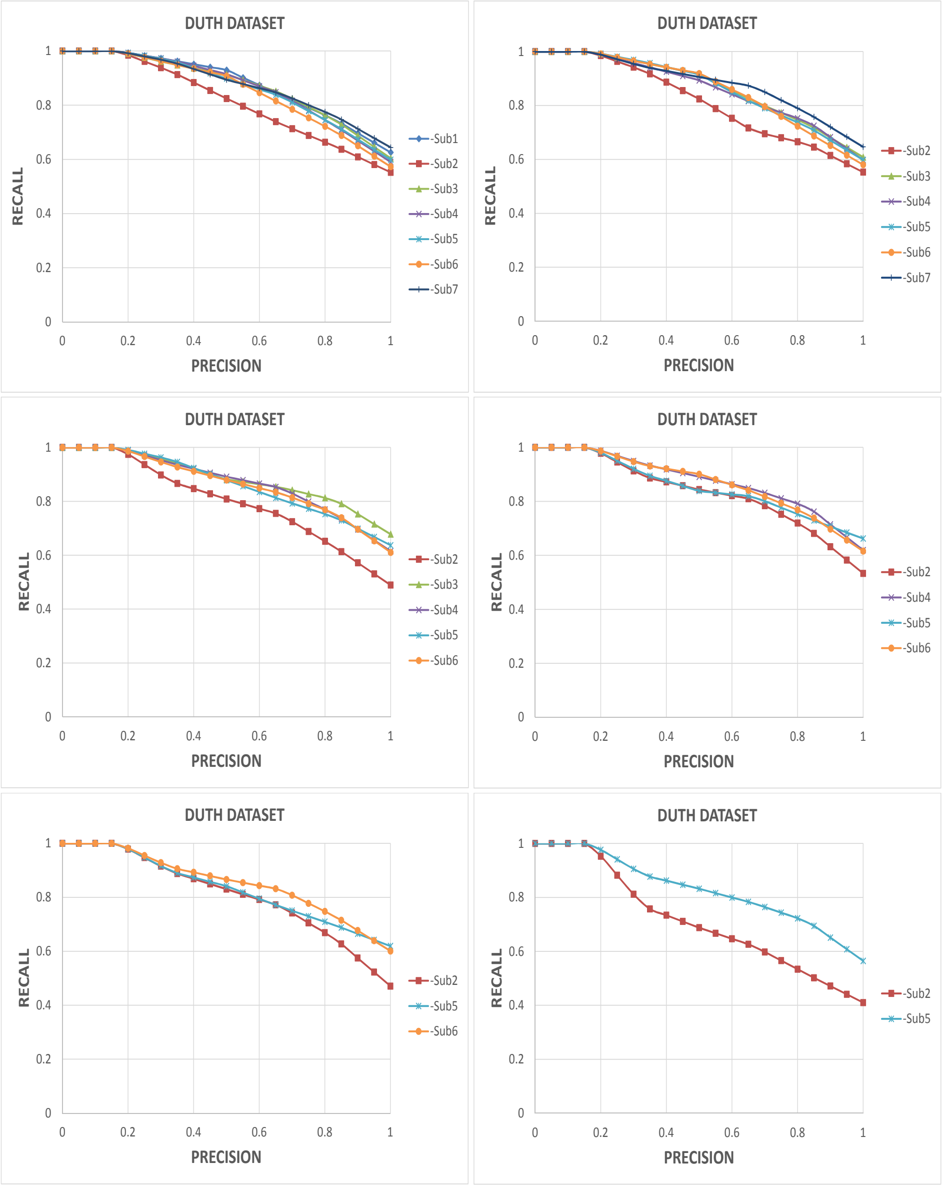


Fig. 10 Ablation study related to the sub-descriptors for the DUTH dataset.

proposed descriptor which are based on the amplitudes of velocity and dynamics vectors. Additionally, the amplitude of the vectors which describe the relative motion between the salient points is identical in these action classes.

Some other misses are related to the action “hop on left foot” which are wrongly paired with the action “walking”. In this case, the main differences between the two action classes are observable on only some of the salient points.

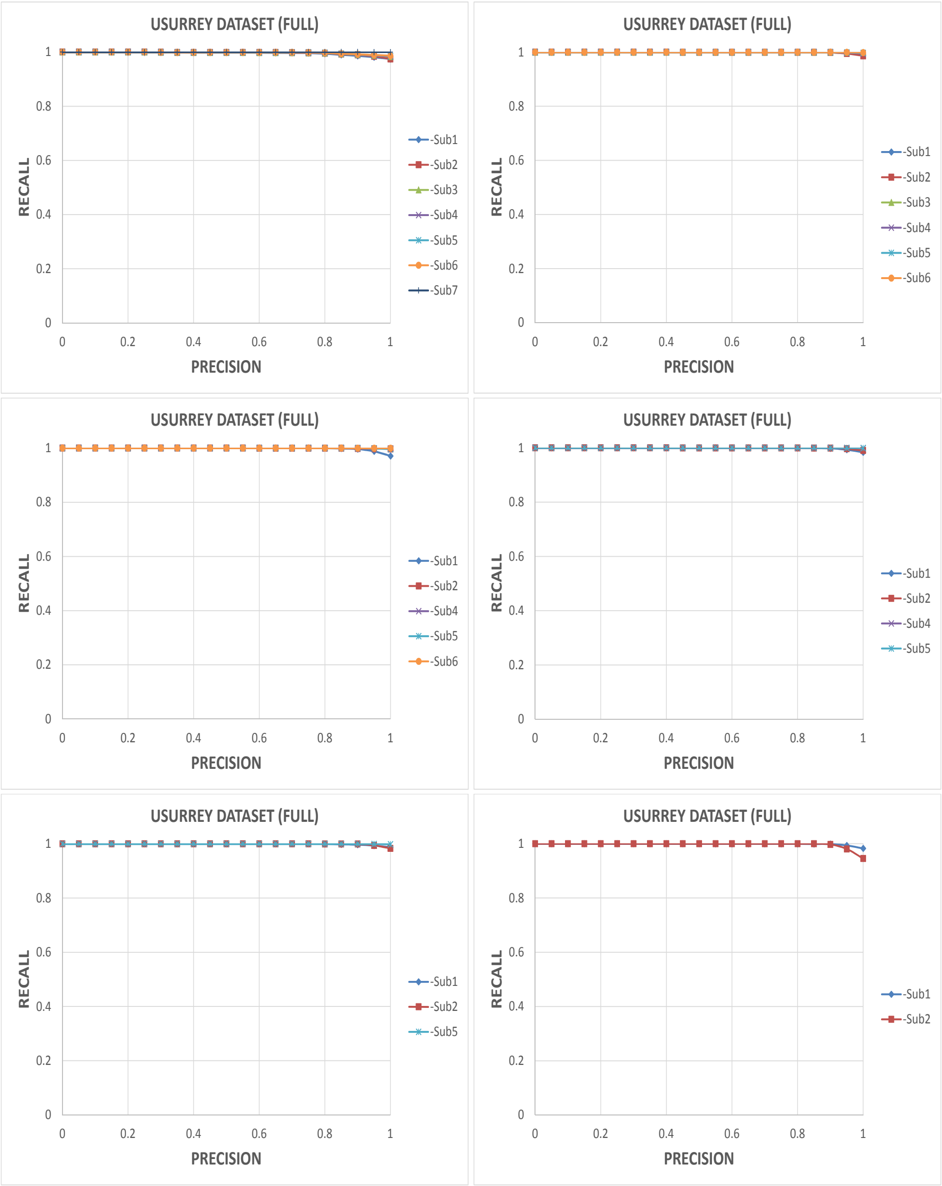


Fig. 11 Ablation study related to the sub-descriptors for the USurrey dataset with full sequences.

It is notable that all examples of the action “running” are retrieved successfully. This shows that the combination of the proposed descriptor, which is based on the total displacement from the model’s initial position and its velocity, in addition to the DTW-based distance measure, is discriminative with respect to actions which have different rates. The retrieval results for the action “walking with arms out-balancing” are ideal, too. The main feature of this action is that the hands of the moving human

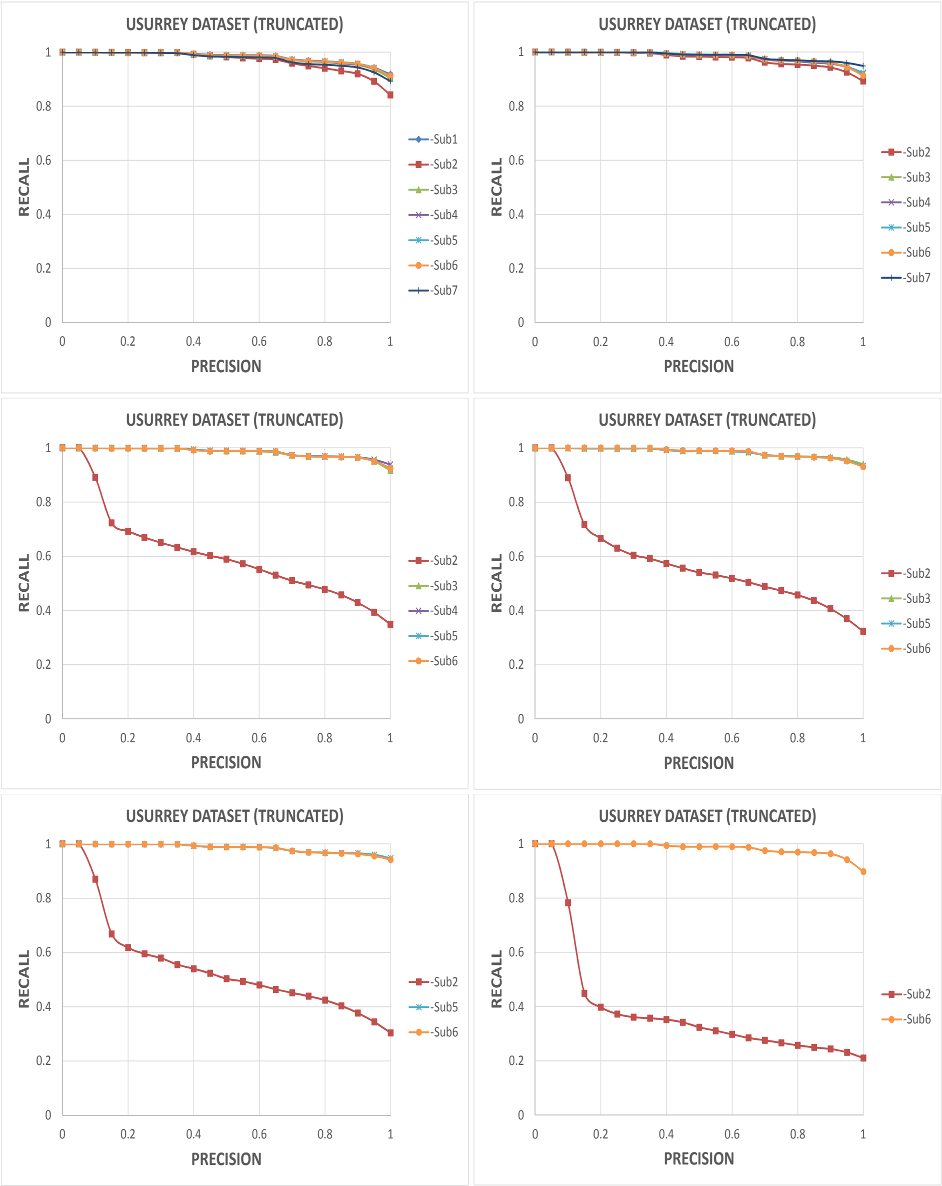


Fig. 12 Ablation study related to the sub-descriptors for the USurrey dataset with truncated sequences.

are outstretched, so the usage of the trajectories of the ends of the corresponding limbs increases the discrimination power concerning this action.

With regard to the actions which are variations of the general class “jumping”, most misses are observed on the couples “jumping”-“jumping Forward” and “jumping Forward”-“jumping-Turn”. In the action “jumping” the models perform only vertical movement. In the action “jumping Forward” the models perform both vertical

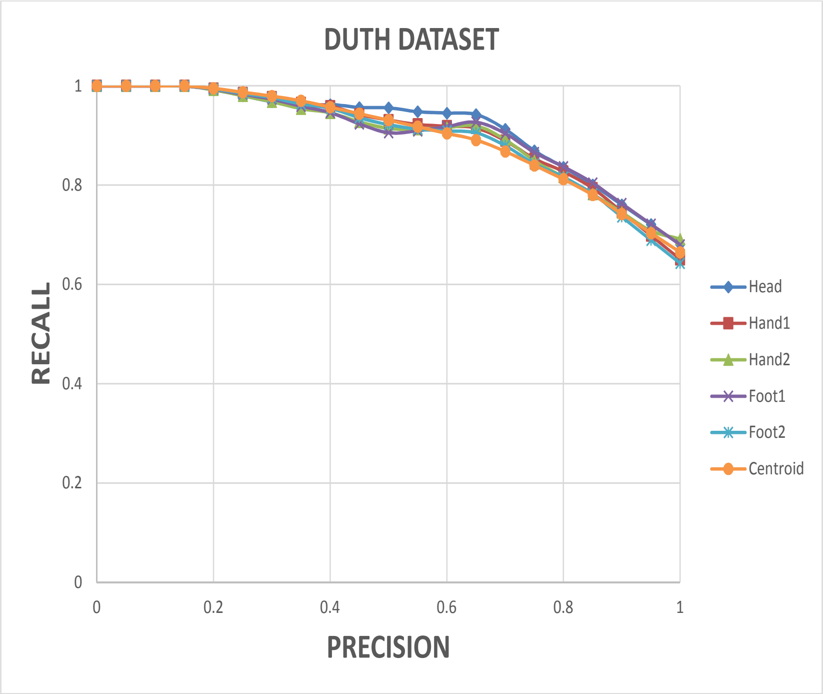


Fig. 13 Precision-recall diagrams for the different selections of the salient point for pre-alignment for DUTH dataset.

and horizontal movement. In the action “jumping-Turn” the models perform both vertical and horizontal movement and, simultaneously, turn the body. The proposed descriptor is able to discriminate satisfactorily the aforementioned classes.

In the case of the USurrey dataset, the retrieval results are almost ideal, both when using the full sequences and when using the truncated sequences, so the corresponding confusion matrices are not presented. Action retrieval using the modified version of the USurrey dataset, i.e. using the truncated sequences, is a more challenging problem, as initial frames are arbitrarily missing thus lacking the 1-1 correspondence between frames. Additionally, all sequences in the initial version of the dataset have the same frame number. The main difficulty in the retrieval problem using the modified (truncated) sequences is the temporal misalignment between the sequences. The proposed pre-alignment step in the similarity evaluation is based on the maximization of sample cross-correlation. This is the crucial step to eliminate the temporal misalignment between the sequences, leading to excellent retrieval results.

Finally, the $k - means$ based method to produce a final distance matrix aims to assign to each descriptor a weight proportional to its discriminative power. This is a novelty in the area of fusing multiple distance matrices that leads to a unique set of distances between the actions.

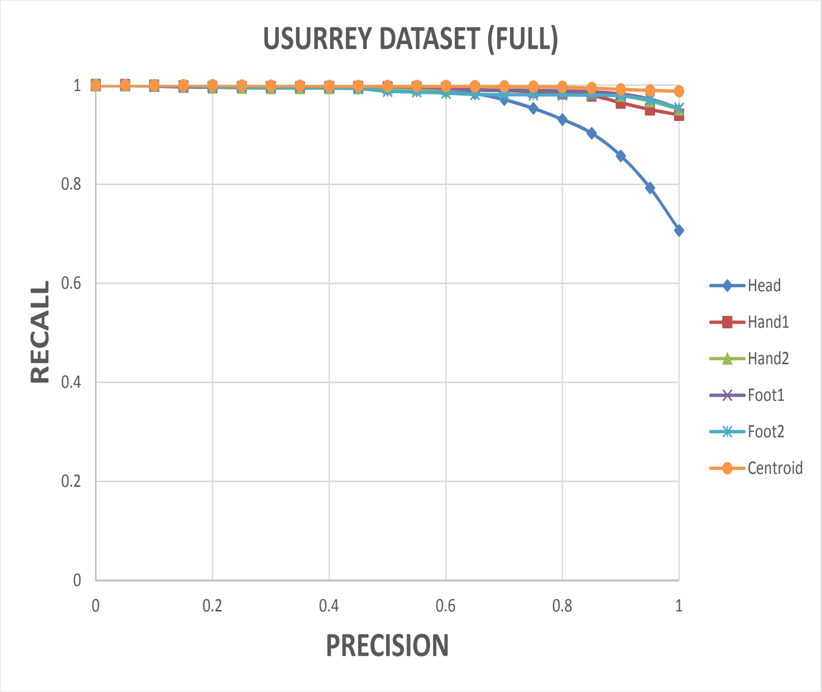


Fig. 14 Precision-recall diagrams for the different selections of the salient point for pre-alignment for USurrey dataset with full sequences.

5 Conclusions and Future Work

A novel unsupervised method for human action retrieval from 3D mesh sequences is presented. It is based on the trajectories of 6 salient points of the human body. These are 5 salient points that correspond to the ends of protrusions of the human body, i.e. head and limbs, and the centroid of the human body. The centroid of the trajectory is utilized to incorporate in the proposed descriptor the general features of the various actions and the trajectories of the 5 salient points are used to incorporate the details of human motion. The evaluation of the proposed methodology was realized in a consistent evaluation framework wherein the results prove that a limited number of representative points of the human body can be used to discriminate a wide variety of human actions, in a fully unsupervised framework.

Naturally, there are some limitations to the proposed method. Since it is based on salient points, it is difficult to apply it to datasets with real data where the salient points may not be reliably identifiable. Such a dataset is provided by the University of Surrey, i3DPost [26], [4]. The corresponding data are available through [36]. This dataset contains some meshes where parts of the human body are ’glued’ together, thus messing up the geodesic paths. Example meshes with this defect contained in i3DPost with the corresponding geodesic paths between ends of the lower limbs are shown in Fig. 17. Additionally, in this dataset there are meshes with disconnected

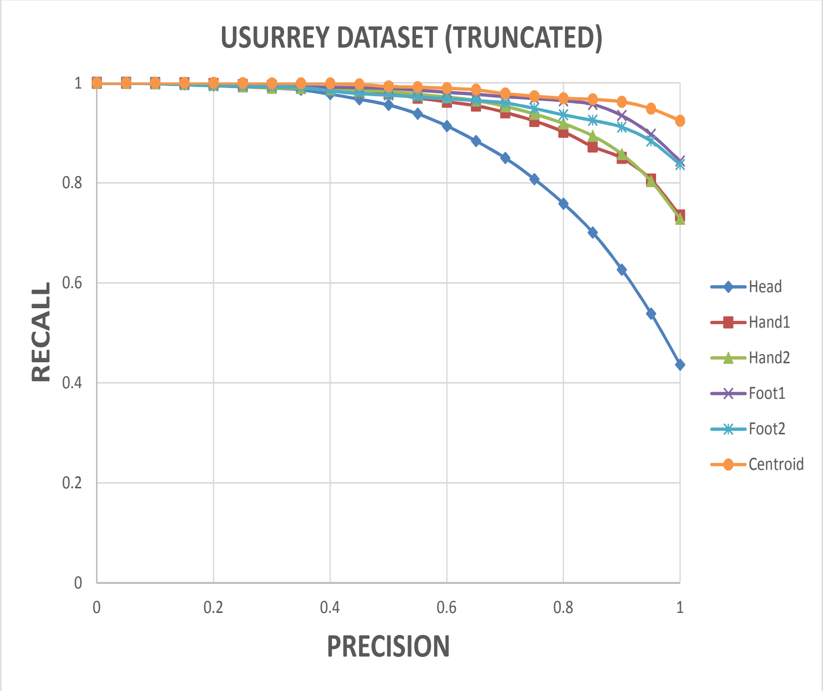
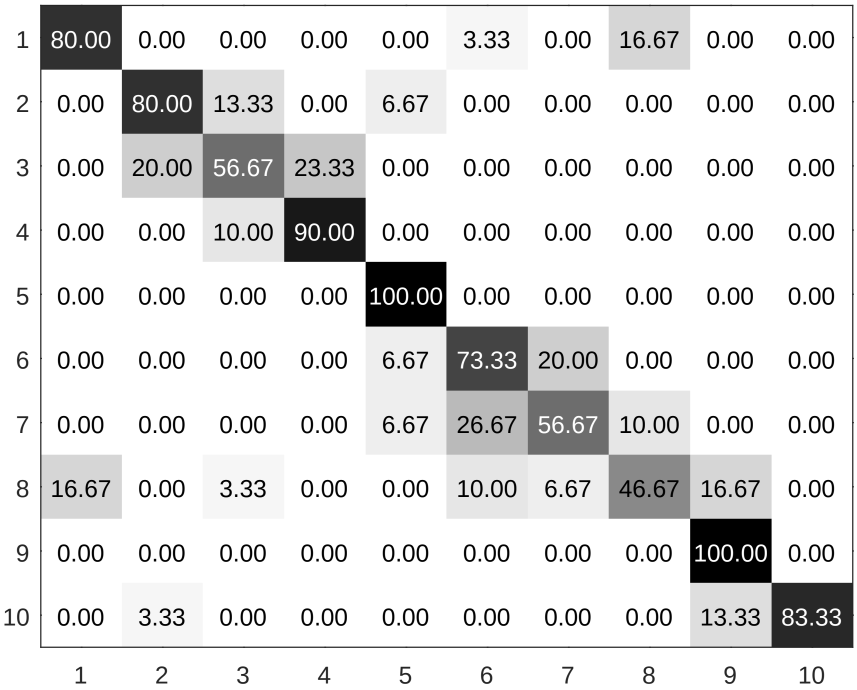


Fig. 15 Precision-recall diagrams for the different selections of the salient point for pre-alignment for USurrey dataset with truncated sequences.

parts, as in Fig. 18, or with non desirable bumps, as in Fig. 19. Geodesic-based methods are unreliable or infeasible in such cases. A potential extension of this work is to create a new method for salient point detection, that is more robust to the defects of real data.

References

1. G. Evangelidis, G. Singh, R. Horaud, Skeletal quads: Human action recognition using joint quadruples, in Proc. IEEE Intl. Conf. Pattern Recog. (2014) 1-6
2. Y. Gao, M. Wang, R. Ji, X. Wu, Q. Dai (2012) 3D object retrieval and recognition with hypergraph analysis. IEEE Trans Image Process 21(9):42904303. doi:10.1109/TIP.2012.2199502
3. Y. Gao, M. Wang, R. Ji, X. Wu, Q. Dai (2014) 3D object retrieval with Hausdorff distance learning. IEEE Trans Ind Electron 61(4):20882098. doi:10.1109/TIE.2013.2262760
4. N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, I. Pitas, The i3Dpost multi-view and 3D human action/interaction, Proc. CVMP (2009) 159-168.
5. M. B. Holte, T. B. Moeslund, P. Fihl, View-invariant gesture recognition using 3D optical flow and harmonic motion context, Computer Vision and Image Understanding 114 (12) (December, 2010) 1353-1361. doi:10.1016/635 j.cviu.2010.07.012
6. M. Holte, T.Moeslund, N. Nikolaidis, I. Pitas, 3D human action recognition for multi-view camera systems, in Proceedings of the 3DIMPVT, 2011
7. P. Huang, A. Hilton, J. Starck, Shape similarity for 3D video sequences of people, International Journal of Computer Vision 89 (2-3) (2010) 362-381.



	1	2	3	4	5	6	7	8	9	10
1	80.00	0.00	0.00	0.00	0.00	3.33	0.00	16.67	0.00	0.00
2	0.00	80.00	13.33	0.00	6.67	0.00	0.00	0.00	0.00	0.00
3	0.00	20.00	56.67	23.33	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	10.00	90.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	6.67	73.33	20.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	6.67	26.67	56.67	10.00	0.00	0.00
8	16.67	0.00	3.33	0.00	0.00	10.00	6.67	46.67	16.67	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00
10	0.00	3.33	0.00	0.00	0.00	0.00	0.00	0.00	13.33	83.33

Fig. 16 The confusion matrix related to the DUTH dataset. The enumeration of the actions is compatible with the enumeration given in subsection 4.1.1.

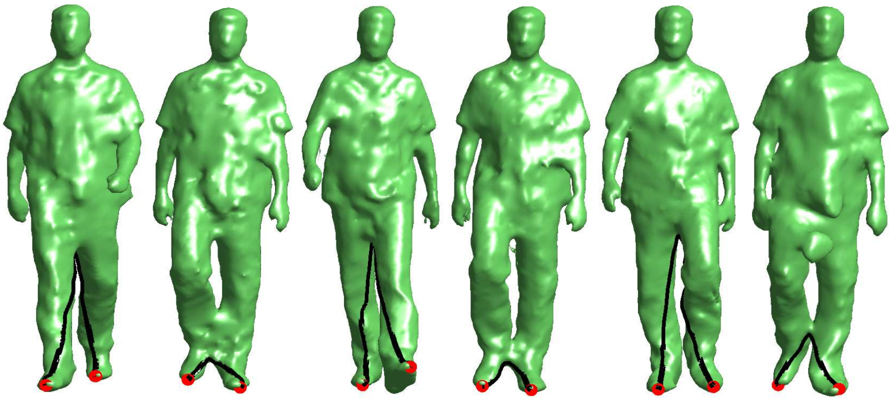


Fig. 17 Example frames from the i3DPost dataset for the action "walking" of Man1 with the corresponding geodesic paths between ends of the lower limbs.

8. G. Johansson, Visual perception of biological motion and a model for its analysis, Perception and Psychophysics 14 (2) (June 1973) 201-211

9. Johnson A, Hebert M (1999) Using spin images for efficient object recognition in cluttered 3D scenes. IEEE Trans on PAMI 21(5):433-449

10. D. Kasai, T. Yamasaki, Kiyoharu Aizawa, Retrieval of time-varying mesh and motion capture data using 2D video queries based on silhouette shape descriptors, IEEE ICME (2009) 854-857,

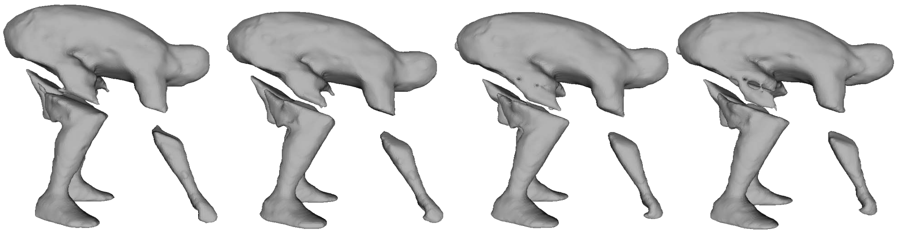


Fig. 18 Example frames from the i3DPost dataset for the action "bending" of Man6 with disconnected parts.

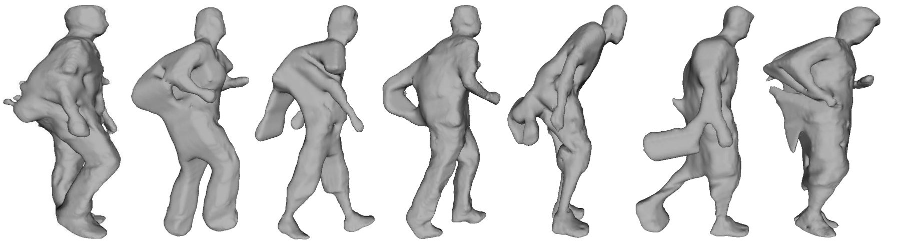


Fig. 19 Example frames from the i3DPost dataset with high level of distortion.

doi:10.1109/ICME.2009.5202629

11. K. Kelgeorgiadis, N. Nikolaidis, Human action recognition in 3D motion sequences, in Proceedings of the 22nd European Signal Processing Conference (EUSIPCO), 2014 [IEEE]
12. W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3D points,in: Proceedings of Computer Vision and Pattern Recognition Workshops (CVPRW),IEEE, San Francisco, CA, USA, 2010, pp. 914, http://dx.doi.org/10.1109/CVPRW.2010.5543273.
13. S. Z. Masood, C. Ellis, M. F. Tappen, J. J. LaViola, R. Sukthankar, Exploring the trade-off between accuracy and observational latency in action recognition, Int. J. Comput. Vis. 101(3) (2013), pp. 420436, http://dx.doi.org/10.1007/s11263-012-0550-7.
14. P. Matikainen, M. Hebert, R. Sukthankar, Representing pairwise spatial and temporal relations for action recognition, ECCV (2010) 508-521
15. F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy, Sequence of the most informative joints (smij), Journal of Visual Communication and Image Representation 25 (1), January 2014, 24-38. doi:10.1016/j.jvcir.2013.04.007
16. R. Osada, T. Funkhouser, B. Chazelle, D. Dobkin (2002) Shape distributions. ACM Trans Graph (TOG) 21:807832. doi:10.1145/571647.571648
17. P. Papadakis, I. Pratikakis, T. Theoharis, G. Passalis and S. Perantonis, 3D object retrieval using an efficient and compact hybrid shape descriptor, In Eurographics 2008 Workshop on 3D Object Retrieval. doi:10.2312/3DOR/3DOR08/009-016
18. P. Papadakis, I. Pratikakis, T. Theoharis, G. Passalis, S. Perantonis, PANORAMA: A 3D shape descriptor based on panoramic views for unsupervised 3D object retrieval, International Journal of Computer Vision 89 (2010) 177-192.
19. L. L. Presti, M. L. Cascia, 3D skeleton-based human action classification: A survey, Pattern Recognition 53 (May 2016) 130-147. doi:http://dx.doi.org/10.1016/j.patcog.2015.11.019
20. R. Qiao, L. Liu, C. Shen, A. van den Hengel, Learning discriminative trajectorylet detector sets for accurate skeleton-based action recognition, Pattern Recognition 66 (2017) 202-212. doi:http://dx.doi.org/10.1016/j.patcog.2017.01.015
21. K. Sfikas, I. Pratikakis, A. Koutsoudis, M. Savelonas, T. Theoharis (2014) Partial matching of 3D cultural heritage objects using panoramic views, multimedia tools and applications, In press. Springer. doi:10.1007/s11042-014-2069-0
22. A. Shahroudy, G. Wang, T.-T. Ng, Q. Yang, Multimodal multipart learning for action recognition in depth videos, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (10), October 2016, 2123-2129.doi:10.1109/TPAMI.2015.2505295

23. P. Shilane, P. Min, M. Kazhdan, T. Funkhouser, The princeton shape benchmark, In: Shape Modeling International (2004) 167-178
24. R. Slama, H. Wannous, M. Daoudi, 3D human motion analysis framework for shape similarity and retrieval, Image Vision Computing 32 (2) (February 2014) 131-154
25. J. Starck, K. Aizawa, Model-based multiple view reconstruction of people, In: Proceedings of the ninth International Conference on Computer Vision (2003) 915-922.
26. J. Starck, A. Hilton, Surface capture for performance based animation, IEEE Computer Graphics and Applications 27 (3) (2007) 21-31
27. C. Veinidis, I. Pratikakis, T. Theoharis, On the retrieval of 3D mesh sequences of human actions, Multimedia Tools and Applications 76 (2), January 2017, 2059-2085
28. R. Vemulapalli, F. Arrate, R. Chellappa, Human action recognition by representing 3D skeletons as points in a lie group, In IEEE Conference on Computer Vision and Pattern Recognition (2014) 588-595
29. M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, E. Keogh, Indexing multidimensional time-series, VLDB J 15 (1) (2006) 1-20. doi:10.1007/s00778-004-0144-2.
30. T. Yamasaki, K. Aizawa, Motion segmentation and retrieval for 3D video based on modified shape distribution, EURASIP Journal on Applied Signal Processing 2007 (1) (2007) 211-222. doi:10.1155/2007/59535
31. T. Yamasaki, K. Aizawa, A euclidean-geodesic shape distribution for retrieval of time-varying mesh sequences, IEEE ICME (2009) 846-849
32. X. Yang, Y. Tian, Effective 3D action recognition using eigenjoints, Journal of Visual Communication and Image Representation 25 (1) (January 2014) 2-11
33. H. Wang, A. Klaser, C. Schmid, C. Liu, Dense trajectories and motion boundary descriptors for action recognition, International Journal of Computer Vision 103 (1) (2013) 60-79. doi:10.1007/s11263-012-0594-8
34. J.Wang, Z. Liu, Y.Wu, J. Yuan, Learning actionlet ensemble for 3D human action recognition, TPAMI 36 (5) (2014) 914-927
35. D. Weinland, E. Boyer, Action recognition using exemplar-based embedding, IEEE Conference on Computer Vision and Pattern Recognition, CVPR (June 2008) 1-7.
36. http://kahlan.eps.surrey.ac.uk/i3dpost_action/
37. https://vc.ee.duth.gr/cmu-duth-mesh/
38. http://mocap.cs.cmu.edu/
39. http://www.makehuman.org/
40. https://www.blender.org/