# Effective Descriptors for Human Action Retrieval from 3D Mesh Sequences

Christos Veinidis[*,§], Antonios Danelakis[†,¶], Ioannis Pratikakis[*,‖] and Theoharis Theoharis[†,‡,**]

[*]*Department of Electrical and Computer Engineering*
*Democritus University of Thrace, Xanthi, Greece*

[†]*IDI, Norwegian University of Science and Technology (NTNU)*
*Trondheim, Norway*

[‡]*Department of Informatics and Telecommunications*
*University of Athens, Athens, Greece*
[§]*cveinidi@ee.duth.gr*
[¶]*a.danelakis@gmail.com*
[‖]*ipratika@ee.duth.gr*
[**]*theotheo@ntnu.no; theotheo@di.uoa.gr*

Two novel methods for fully unsupervised human action retrieval using 3D mesh sequences are presented. The first achieves high accuracy but is suitable for sequences consisting of clean meshes, such as artificial sequences or highly post-processed real sequences, while the second one is robust and suitable for noisy meshes, such as those that often result from unprocessed scanning or 3D surface reconstruction errors. The first method uses a spatio-temporal descriptor based on the trajectories of 6 salient points of the human body (i.e. the centroid, the top of the head and the ends of the two upper and two lower limbs) from which a set of kinematic features are extracted. The resulting features are transformed using the wavelet transformation in different scales and a set of statistics are used to obtain the descriptor. An important characteristic of this descriptor is that its length is constant independent of the number of frames in the sequence. The second descriptor consists of two complementary sub-descriptors, one based on the trajectory of the centroid of the human body across frames and the other based on the Hybrid static shape descriptor adapted for mesh sequences. The robustness of the second descriptor derives from the robustness involved in extracting the centroid and the Hybrid sub-descriptors. Performance figures on publicly available real and artificial datasets demonstrate our accuracy and robustness claims and in most cases the results outperform the state-of-the-art.

*Keywords*: 3D mesh sequence; action retrieval; 3D shape representation; 3D shape descriptors; matching.

[§]Corresponding author.

## 1. Introduction

Human action retrieval and recognition is a challenging problem with various applications, such as surveillance, video games, human–computer interaction, etc. This problem has a variety of intrinsic difficulties. Indeed, there are actions which are tagged under a general class and their discrimination requires the extraction of features which characterize the special details of each of them. For example, two of the most common human actions are "walking" and "running". "Running" is essentially a fast version of "walking", so the extraction of the rate of each action is an appropriate feature to distinguish them. Other problems, such as temporal misalignment, the variety of the temporal resolution of an action, or body type variability must also be taken into account when designing a system that accommodates similar actions.

The psychological experiment of Ref. 1, using a number of point lights attached to the human body, has shown that only a small number of critical points on the human body is sufficient for a human to recognize an action. To this end, several technologies have been used to create skeletal data including passive stereo vision, multi-camera systems, time-of-flight cameras, passive optical motion capture systems as well as Kinect-based systems.

In Fig. 1, the operational pipeline of human action retrieval using 3D mesh sequences is shown. The process consists of two stages: the offline stage comprises the descriptor extraction for each sequence and its storage in a database after indexing. The online stage is activated for a *query* mesh sequence; the corresponding descriptor is extracted and the similarities between the query and the indexed

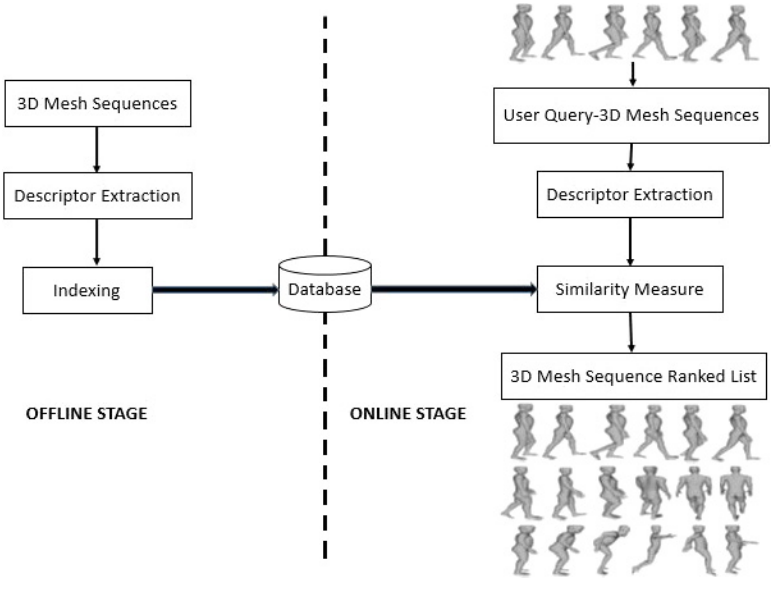


Fig. 1. Typical pipeline of human action retrieval using 3D mesh sequences.

sequences are computed. The output of this pipeline is a ranked list ordered in decreasing similarity.

It is evident from the pipeline that the crucial stages are: (i) the descriptor extraction and (ii) the similarity measure. In this paper, we focus on the representation of the human body as a 3D mesh, since it is more informative, while 3D scanning techniques across time are constantly emerging. Human action retrieval using 3D mesh sequences is an open problem which has hardly been addressed.

In this paper, two alternative methods for *unsupervised* human action retrieval using mesh sequences, are presented. In the first method, the centroid and the five extremities of the human body are determined and a set of kinematic features are extracted from their trajectories. These extremities of the human body is the top of the head and the ends of the upper and the lower limbs, shown in Fig. 2. The method of Ref. 2 is extended using a wavelet-based descriptor which has the advantage of constant size. Due to this, the similarity measure between two mesh sequences can be extracted using accurate distance measures which are defined between vectors with the same number of components.

While the above method is notably accurate, in unprocessed mesh sequences it is often difficult to detect the six extremities. In view of this, we propose a second method which is based on the trajectory of a single point, the centroid of the human body, across frames. The centroid can be robustly extracted as it is an average and is not much influenced by scanning noise. The centroid is complemented by another robust sub-descriptor known as Hybrid,[3] which we have adapted for mesh sequences and reflects geometry using a hybrid scheme that combines 3D and 2D information.

The main contributions of this paper are the following:

- An accurate spatio-temporal descriptor for clean 3D mesh sequences of human actions based on the trajectories of six salient points on the human body.

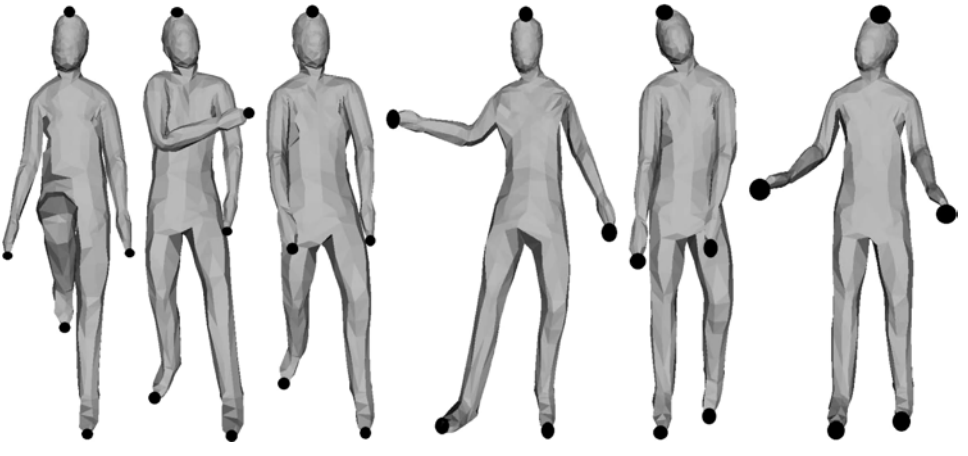

Fig. 2. The five extremities of the human body that are used as salient points.

- A robust descriptor for human action retrieval based on the trajectory of the centroid of 3D human meshes complemented by a Hybrid descriptor. This is particularly applicable on challenging data.

The remainder of this paper is organized as follows: In Sec. 2, related work in 3D mesh sequence retrieval is presented. In Sec. 3, the two alternative methods for human action retrieval are detailed. In Sec. 4, experimental results are presented coupled with an extensive discussion, while in Sec. 5 conclusions are drawn and future work is presented.

## 2. Related Work

In recent years, variations of the traditional human action retrieval problem have been addressed. Still image based human action recognition, i.e. the identification of a person's action or behavior from a single image, constitutes a variation of the same problem.[4–7] Human action prediction[8–10] aims to infer the action from an incomplete video. In Refs. 11–13, the classification, retrieval or prediction of interactions between animated humans and physical objects are addressed.

Many works, such as Refs. 14–21, use skeletons to address the human action retrieval problem. An extensive survey of 3D skeleton-based action classification is Ref. 22. Also, the same problem has been previously addressed using image sequences[23–26] and depth image sequences.[27–29] A recent survey of image-based action recognition is Ref. 30.

Although the research in the case of human action retrieval using 2D sequences and 3D skeletal sequences is extensive, in recent years there are only few works addressing the same problem using 3D meshes for the representation of the human body. In the following, human action retrieval methods, either for motion clips or for whole sequences, using the mesh representation for the human body, are presented.

Concerning motion clip retrieval approaches, in Ref. 31, vector quantization is applied on the mesh of each frame of the sequences producing a set of clusters and a variation of shape distribution of the clusters' centroids[32]; this is utilized as descriptor for each frame of the mesh sequence. The sequence is segmented into motion clips and the final motion clip retrieval process is based on a dynamic programming algorithm. In Ref. 33, the geodesic shape distribution is introduced as a shape descriptor for each frame of a mesh sequence. This descriptor is based on the geodesic distances between clusters' centroids and is combined with the modified shape distribution descriptor, introduced in Ref. 31. The final descriptor for each frame is a weighted mean of these two descriptors. It is worth noting that the experiments in this work are limited to only three mesh sequences. In Ref. 34, a static shape descriptor, which is based on the 10 shortest geodesic paths which connect the protrusions of the human body, namely the Extremal Human Curve (EHC), is introduced. The local extrema of the velocity in time are selected as segmentation points across the actions, in order to segment the sequences into motion clips. These motion clips are later retrieved by applying the Dynamic Time Warping (DTW)

distance[35] on the trajectories of EHC curves resulting in the motion clip matching process. In Ref. 36, the 3D mesh sequences are transformed to 2D sequences by taking 20 projection points and are matched against 2D queries. The descriptors which are extracted here, are based on 2D contours and are called P-Type Fourier Descriptors. For the final matching between the motion clips, a variation of the normalized DTW, assigning exponential costs between the feature vectors of the most distant frames, is applied.

Concerning human action retrieval of whole mesh sequences, we have presented an extensive comparative study of the use of state-of-the-art descriptors at the frame level in our previous work.[37] Each mesh sequence is considered as a curve in the $M$-dimensional space, where $M$ denotes the size of each static descriptor (per mesh frame). The similarity between the actions is evaluated using the DTW algorithm. Additional experiments using the Sakoe band for the DTW computations are performed. In Ref. 2, the trajectories of 6 salient points of the human body (the centroid, the top of the head and the extremities of the upper and the lower limbs) are extracted. A set of kinematic descriptors are used to form the descriptor of the mesh sequences, while a k-means-based algorithm is used to fuse multiple distance matrices.

Similar in spirit to the action retrieval problem, in Ref. 38, a supervised method for human action recognition from multi-view camera systems, is presented. The proposed descriptors,[39] are based on the optical flow of each of the different views of the 3D human. The optical flow is extracted on each pixel of the 2D sequences and a correspondence between the pixels and the vertices of the meshes is made, using the camera calibration parameters. The final motion vector is extracted as a weighted summation of the motion vector of each view, taking into account the significance and the reliability of each view. The first of the two proposed descriptors, called 3D Motion Context, corresponds to a spherical histogram which is based on the orientation of the velocity vectors. The second proposed descriptor, called Harmonic Motion Context, is a modified version of the 3D Motion Context descriptor using spherical harmonics, so that invariance with respect to the vertical axis is achieved. The similarity between the actions relies upon the normalized correlation coefficients. 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 Ref. 40, the mesh sequences are transformed into voxel sequences. First, an algorithm for human body orientation estimation, based on an estimation of feet direction, is applied. Then, a normalization step to make the method invariant to translation and scaling follows. The $k$-means algorithm is applied to cluster the similar postures of the sequences and the centers of these clusters are called dynemes. Then, for each posture, a vector which is related to the distance from each of the dynemes, is extracted. Each action is represented by the average of the corresponding normalized vectors of the action's postures. Linear Discriminative Analysis (LDA) is applied to reduce the dimensionality of the action representations and the final classification is based on a Support Vector Machine (SVM) classifier.

Human actions representation with 3D mesh sequences has been used to address various other problems, too. For example, in Refs. 41 and 42, a performance evaluation of shape similarity metrics for 3D video sequences of people with unknown temporal correspondence is presented, while in Refs. 43 and 44, the problem of surface matching along mesh sequences is addressed.

## 3. Methodology

### 3.1. *Method based on a fixed length wavelet-based spatio-temporal descriptor for human actions*

#### 3.1.1. *Descriptor extraction*

In this section, the first proposed method of this paper, referred as "Proposed M1", is presented.

Similar to Ref. 2, the trajectories of the six salient points (centroid and extremities of the human body) are extracted.

The descriptor extracted from the trajectory of salient point $i = 1, 2, \ldots, 6$ of mesh sequence $S$ with length $L$ is defined as follows:

$$\mathbf{D}_i^S = ([\mathbf{D}_i^S(m)]_{m=1}^{7}), \tag{1}$$

where $m = 1, 2, \ldots, 7$ denotes the $m$th sub-descriptor, defined as follows:

$$\begin{aligned}
\mathbf{D}_i^S(1) &= [\|\mathbf{v}_{i,h}^t\|, \|\mathbf{v}_{i,v}^t\|]_{t=1}^{L-1},\\
\mathbf{D}_i^S(2) &= [\|\mathbf{d}_{i,h}^t\|, \|\mathbf{d}_{i,v}^t\|]_{t=1}^{L-1},\\
\mathbf{D}_i^S(3) &= [\mathbf{v}_{i,h}^t]_{t=1}^{L-1}\\
\mathbf{D}_i^S(4) &= [\mathbf{v}_{i,v}^t]_{t=1}^{L-1},\\
\mathbf{D}_i^S(5) &= [\mathbf{d}_{i,h}^t]_{t=1}^{L-1},\\
\mathbf{D}_i^S(6) &= [\mathbf{d}_{i,v}^t]_{t=1}^{L-1},\\
\mathbf{D}_i^S(7) &= [\mathbf{k}_i^t]_{t=1}^{L},
\end{aligned}$$

where $\mathbf{v}$ denotes the velocity vector and $\mathbf{d}$ denotes the overall dynamics, defined as

$$\begin{aligned}
\mathbf{v}_i^t &= \mathbf{p}_i^{t+1} - \mathbf{p}_i^t = (\mathbf{v}_{i,h}^t \quad \mathbf{v}_{i,v}^t)\\
&= ([x_{p,i}^{t+1} - x_{p,i}^t \quad z_{p,i}^{t+1} - z_{p,i}^t] \quad y_{p,i}^{t+1} - y_{p,i}^t),
\end{aligned} \tag{2}$$

$$\begin{aligned}
\mathbf{d}_i^t &= \mathbf{p}_i^{t+1} - \mathbf{p}_1^t = (\mathbf{d}_{i,h}^t \quad \mathbf{d}_{i,v}^t)\\
&= ([x_{p,i}^{t+1} - x_{p,i}^1 \quad z_{p,i}^{t+1} - z_{p,i}^1] \quad y_{p,i}^{t+1} - y_{p,i}^1),
\end{aligned} \tag{3}$$

where $t = 1, 2, \ldots, L-1$. $\mathbf{d}_{i,h}^t, \mathbf{d}_{i,v}^t$ denote the horizontal and vertical components of overall dynamics and $\mathbf{v}_{i,h}^t, \mathbf{v}_{i,v}^t$ denote the horizontal and vertical components of velocity vector corresponding to the $i$th salient point, $i = 1, 2, \ldots, 6$. Furthermore, $\mathbf{p}_i^t = [x_{p,i}^t \;\; y_{p,i}^t \;\; z_{p,i}^t]$ denotes the $t$th point of the trajectory corresponding to the $i$th salient point.

In addition, the curvature of each of these trajectories is extracted

$$\mathbf{k}_i^t = \frac{\|\mathbf{p}_i'(t) \times \mathbf{p}_i''(t)\|}{\|\mathbf{p}_i'(t)\|^3}, \tag{4}$$

where $||\mathbf{x}||$ is the magnitude of a vector $\mathbf{x}$ for each salient point $i = 1, 2, \ldots, 6$. Also, $t = 1, 2, \ldots, L$, where $L$ is the length (i.e. the number of frames) of the mesh sequence. It is worth noted that the points $\mathbf{p}_i(t) = [x_{p,i}(t) \;\; y_{p,i}(t) \;\; z_{p,i}(t)]^T$, for each $i = 1, 2, \ldots, 6$ denote points on the normalized trajectories and $\mathbf{p}_i'(t), \mathbf{p}_i''(t)$ are their first and second derivative, respectively.

In order to eliminate the temporal misalignment between pairs of sequences, the pre-alignment step, presented in Ref. 2, is performed.

As the final descriptor given in Eq. (1) consists of seven sub-descriptors for each of six salient points, the total number of sub-descriptors is 42. The wavelet transformation of each of these 42 sub-descriptors of the pre-aligned sequences is computed. The reasoning behind the selection of the wavelet transformation is the fact that it is better at incorporating frequency and time information than other transformations, such as Fourier or Cosine (see Refs. 45 and 46). Gaussian wavelets at 64 different scales are used. At each scale, the statistics shown in Table 1 with the corresponding formulas are computed from the wavelet coefficients. The combination of statistics has been experimentally selected.

The above statistics are used to form a vector at each wavelet scale with respect to each sub-descriptor and the final wavelet-based descriptor for a mesh sequence consists of the concatenation of all these 4D vectors, so the descriptor's size is $64 \times 4 = 256$.

### 3.1.2. *Distance measure*

Let $\mathbf{q}_s = [q_{s1} \;\; q_{s2} \;\; q_{s3} \;\; q_{s4}]^T$ and $\mathbf{t}_s = [t_{s1} \;\; t_{s2} \;\; t_{s3} \;\; t_{s4}]^T$ be the 4D vectors corresponding to sequences $Q$ and $T$, at scale $s$, where $s = 1, 2, \ldots, 64$. The distance between $Q$ and $T$ is defined using the *chi-square* formula:

$$\text{dist}(\mathbf{q}_s, \mathbf{t}_s) = \frac{1}{2} \cdot \sum_{j=1}^{4} \frac{(q_{sj} - t_{sj})^2}{q_{sj} + t_{sj}}. \tag{5}$$

Table 1. The statistics computed from the components of the wavelet coefficients with the corresponding formulas for a vector $\mathbf{x} = [x_1, x_2, \ldots, x_N]^T$.

| Statistic | Formula |
|---|---|
| (1) Mean ($\mu$) | $\frac{1}{N} \cdot \sum_{i=1}^{N} x_i$ |
| (2) Variance ($\sigma^2$) | $\frac{1}{N-1} \cdot \sum_{i=1}^{N} (x_i - \mu)^2$ |
| (3) Root Mean Square | $\sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} x_i^2}$ |
| (4) Kurtosis | $\frac{1}{\sigma^4} \cdot \sum_{i=1}^{N} (x_i - \mu)^4$ |

The distance between the sequences $Q$ and $T$ for each of the 42 sub-descriptors of Eq. (1) is defined as the mean distance of the 4D vectors across all 64 wavelet scales

$$\text{dist}(Q,T) = \frac{1}{64} \cdot \sum_{s=1}^{64} \text{dist}(\mathbf{q}_s, \mathbf{t}_s). \tag{6}$$

Finally, the $k$-means-based algorithm for multiple distance matrix fusion, presented in Ref. 2, is applied in order to compute the distance between the sequences.

### 3.2. *Method based on a robust spatio-temporal descriptor for human actions*

In contrast to the extraction of the five salient points of the previous descriptor, the extraction of the centroid is reliable even in real, unprocessed, datasets, where the representation of the models contains issues such as disconnected parts, glued parts and distortions. In this section, we present the second proposed method of this paper, referred as "Proposed M2".

#### 3.2.1. *Extraction of the first component of "Proposed M2" descriptor: "Proposed M2-C1"*

The first component of this descriptor consists of sub-descriptors $\mathbf{D}^S(1) - \mathbf{D}^S(6)$ of Eq. (1), extracted for the centroid. As the sub-descriptors $\|\mathbf{d}_h^S\|$, $\|\mathbf{d}_v^S\|$, $\|\mathbf{v}_h^S\|$, $\|\mathbf{v}_v^S\|$, $\mathbf{d}_v^S$,$\mathbf{v}_v^S$ in Eq. (1) are 1D trajectories and the sub-descriptors $\mathbf{d}_h^S$, $\mathbf{v}_h^S$ in Eq. (1) are 2D trajectories, all with $L-1$ points, the total size of the descriptor is $10 \times (L-1)$.

#### 3.2.2. *Extraction of the second component of "Proposed M2" descriptor: "Proposed M2-C2"*

Despite its robustness, there are cases where the first component of the "Proposed M2" descriptor is not sufficiently discriminative. For example, if there are two actions where the motion of the centroid is negligible, the discriminative power of the trajectory-based descriptor is limited. To increase its discriminative power, a second part is employed which is the Hybrid descriptor[3] applied on each mesh of the action. Hybrid has previously shown promise in describing 3D mesh sequences and is composed of 2D features based on depth buffers and 3D features based on spherical harmonics. To compensate for rotation, two pose normalization methods, namely CPCA and NPCA,[47] are applied before the extraction of the descriptor. The Hybrid descriptor of a model in frame $t$, is defined as the concatenation of the two pose normalized versions of the 2D and 3D features (4 combined sub-descriptors), as per Eq. (7).

$$\mathbf{h}^t = (2Df_M^{t,\text{CPCA}}, 2Df_M^{t,\text{NPCA}}, 3Df_M^{t,\text{CPCA}}, 3Df_M^{t,\text{NPCA}}), \tag{7}$$

where $*Df_M^{t,*\text{PCA}}$ represents the $*D$ features of the model $M$ at frame $t$ normalized by $*$PCA, where $*D \in \{2D, 3D\}$ and $*\text{PCA} \in \{\text{CPCA}, \text{NPCA}\}$.

A normalization step is used to set the values in each dimension of each of the four components of Eq. (7) in the interval $[0, 1]$. Furthermore, temporal filtering is applied to each aligned version of each sub-descriptor vector of each frame resulting in a final averaged descriptor taking into account the corresponding versions of sub-descriptors of the frames in a neighborhood $N$. So, if $\mathbf{hd}^t$ is one of the 4 components of $\mathbf{h}^t$, then its temporally filtered version $\mathbf{hd}_f^t$ is given by

$$\mathbf{hd}_f^t = \frac{1}{2N+1} \cdot \sum_{k=-N}^{N} \mathbf{hd}^t. \tag{8}$$

Thus, the second component of the "Proposed M2" descriptor of a sequence is given by the equation

$$\mathbf{H}^S = (2D\mathbf{F}^{S,\mathrm{CPCA}}, 2D\mathbf{F}^{S,\mathrm{NPCA}}, 3D\mathbf{F}^{S,\mathrm{CPCA}}, 3D\mathbf{F}^{S,\mathrm{NPCA}}), \tag{9}$$

where $*D\mathbf{F}^{*\mathrm{PCA}}$ denotes the sub-descriptor of the sequence $S$ after temporal filtering and normalization for which $*D \in \{2D, 3D\}$ and $*\mathrm{PCA} \in \{\mathrm{CPCA}, \mathrm{NPCA}\}$.

### 3.2.3. *Distance measure*

Each mesh sequence of a dataset is used in turn as query and the aim is to retrieve all other mesh sequences that belong to the same class. To this end, the distance between each mesh sequence with every other is evaluated, forming a *distance matrix* for a specific dataset. In our method, seven distance measures are used so there are seven distances, which must be combined to give the final distance matrix. Let $L_Q$ and $L_T$ be the length of sequences $Q$ and $T$, respectively. Each of these corresponds to a row and column of a symmetric distance matrix.

For the first part of the "Proposed M2" descriptor, the DTW value between each sub-descriptor, normalized by the minimum of $L_Q - 1$ and $L_T - 1$, is used as distance measure. The distance between the centroid part of two descriptors is based on the following equations:

$$\mathrm{Dist}_1 = \mathrm{DTW}(D^Q(1), D^T(1))/L_{\min}, \tag{10}$$

$$\mathrm{Dist}_2 = \mathrm{DTW}(D^Q(2), D^T(2))/L_{\min}, \tag{11}$$

$$\mathrm{Dist}_3 = ((w_h)^p \cdot \mathrm{DTW}(D^Q(3), D^T(3))/L_{\min}, \tag{12}$$

$$\mathrm{Dist}_4 = ((w_v)^p \cdot \mathrm{DTW}(D^Q(4), D^T(4))/L_{\min}, \tag{13}$$

$$\mathrm{Dist}_5 = ((w_{dh})^p \cdot \mathrm{DTW}(D^Q(5), D^T(5))/L_{\min}, \tag{14}$$

$$\mathrm{Dist}_6 = ((w_{dv})^p \cdot \mathrm{DTW}(D^Q(6), D^T(6))/L_{\min}, \tag{15}$$

where $L_{\min} = \min\{L_Q, L_T\} - 1$ and

$$w_{dh} = \frac{\max\{\mathrm{dif}_{1Q}, \mathrm{dif}_{1T}\}}{\min\{\mathrm{dif}_{1Q}, \mathrm{dif}_{1T}\}}, \quad w_{dv} = \frac{\max\{\mathrm{dif}_{2Q}, \mathrm{dif}_{2T}\}}{\min\{\mathrm{dif}_{2Q}, \mathrm{dif}_{2T}\}} \tag{16}$$

$$w_h = \frac{\max\{\text{dif}_{3Q}, \text{dif}_{3T}\}}{\min\{\text{dif}_{3Q}, \text{dif}_{3T}\}}, \quad w_v = \frac{\max\{\text{dif}_{4Q}, \text{dif}_{4T}\}}{\min\{\text{dif}_{4Q}, \text{dif}_{4T}\}} \tag{17}$$

and

$$\text{dif}_{mS} = (\max_t \|\mathbf{nrm}^t(m)\| - \min_t \|\mathbf{nrm}^t(m)\|)_S \tag{18}$$

for $m = 1, 2, 3, 4$, where

$$\mathbf{nrm}^t = [\|\mathbf{d}_h^t\| \quad \|\mathbf{d}_v^t\| \quad \|\mathbf{v}_h^t\| \quad \|\mathbf{v}_v^t\|]^T. \tag{19}$$

The weights given in Eqs. (16)–(17) are used to further discriminate the actions proportionally to the contribution of their horizontal and vertical components. In our experiments, parameter $p$ of Eqs. (12)–(15) is set to 3.

The above process aims to incorporate various properties of the motion using the trajectories of the centroid of each action. Some of these characteristics are the rate, the direction and the total displacement from the initial position. DTW is a distance measure which minimizes the effects of phase differences.

For the second part of the "Proposed M2" descriptor, the distance between the Hybrid part of the mesh sequences $Q$ and $T$ is evaluated as

$$\text{Dist}_7 = (\text{dist}_{2D} + \text{dist}_{3D}) / \min\{L_Q, L_T\}, \tag{20}$$

where $\text{dist}_{2D}$, $\text{dist}_{3D}$ is the distance between the 2D and 3D sub-descriptors:

$$\text{dist}_{2D} = \min_k \{\text{DTW}(2D\mathbf{F}^{Q,k}, 2D\mathbf{F}^{T,k})\}, \tag{21}$$

$$\text{dist}_{3D} = \min_k \{\text{DTW}(3D\mathbf{F}^{Q,k}, 3D\mathbf{F}^{T,k})\}, \tag{22}$$

where $k \in \{\text{CPCA}, \text{NPCA}\}$.

Before combination, all distance matrices are normalized to the interval $[0, 1]$. Then, for each element *dist* in each distance matrix the following transformation is applied

$$N\_\text{dist} = \frac{1}{\log\left(\frac{1}{\text{dist}}\right)}. \tag{23}$$

This transformation is increasing, so the relative ordering between the distances is maintained. Additionally, the logarithmic function maps from the $[0, 1]$ interval to the $[0, +\infty)$ interval and possesses a more discriminative resolution between. The final distance matrix is given by the following equation:

$$\text{Dist\_final} = w \cdot \sum_{i=1}^{6} N\_\text{Dist}_i + (1 - w) \cdot N\_\text{Dist}_7, \tag{24}$$

where the weight $w$, between centroid and Hybrid features, is experimentally set to 0.82. $N\_\text{Dist}_i$, $i = 1, 2, \ldots, 7$ is the distance matrix which is produced after the application of Eq. (23) on the corresponding distance matrices $\text{Dist}_i$.

### 3.3. *Method selection criteria*

There are some situations where the "Proposed M1" method is not applicable or relevant. The corresponding limitations result from the fact that the extraction of geodesic distances and paths is not feasible or reliable. The existence of these limitations in data can be detected as follows, and in such cases the "Proposed M2" method should be opted for:

- If there are disconnected parts among the input meshes (as shown in Fig. 5), the computation of geodesic distances and paths is not feasible. So, if there are more than one connected components in the input meshes, then the "Proposed M1" method is not applicable.
- If there are protrusions which do not belong to the human body (Fig. 6), the computation of geodesic distances and paths is feasible but not reliable. The appearance of such a protrusion, may lead to the extraction of a salient point in this protrusion. If the protrusion does not exist in all frames of a sequence, then the salient point will likely be extracted in different regions of the human body, in consecutive frames. A suitable criterion to observe this situation is to extract the Euclidean distances of the salient points in consecutive frames. The corresponding values of these distances must not surpass a threshold.
- In the case of "glued" parts in the human body, as shown in Fig. 4, geodesic distances are not reliable for identifying the extremities of the human body. To detect such situations we may employ the heat kernel signature[48] which is commonly used for detecting extremities that correspond to local maxima in the heat kernel map. This signature can be examined in the neighborhood of the initially detected extremities for validating them. In such erroneous cases, fewer than five extremities will be finally validated and then one could switch to the second methodology.

## 4. Experimental Evaluation

Experimental evaluation is performed using standard retrieval performance measures: precision-recall graphs, Nearest Neighbor, First Tier, Second Tier and Discounted Cumulative Gain[49]; the data used for the experimentation are three challenging, publicly available, datasets.

### 4.1. *Datasets*

The first dataset contains real data, while the two other datasets, namely USurrey-artificial and DUTH-artificial, contain artificial data.

#### 4.1.1. *i3DPost-Real dataset*

The first dataset consists of real data.[50,51] Each of eight models have performed 10 actions. These actions are: (1) "walking", (2) "jogging", (3) "jumping",

(4) “bending”, (5) “hand-waving”, (6) “jumping-in-place”, (7) “sitting down-and-standing up”, (8) “running-and-falling”, (9) “walking and sitting”, (10) “running-jumping and walking”. Also, in this dataset there are two interactions between two models: “handshaking” and “pulling”. In our experiments the two interactions have been left out and only the actions (1)–(10), which have been performed by one person only, are used. Actions (1)–(6) are basic human motions. Actions (7)–(10) are combinations of basic actions that are implemented successively. The number of frames per sequence is not always the same (ranging from 55 to 125). Example frames of this dataset are shown in Fig. 3. The mesh sequences of the i3DPost-Real dataset are available in http://kahlan.eps.surrey.ac.uk/i3dpost_action/.

The sequences in this dataset have a number of defects, such as:

- There are parts of the human body that are “glued” to each other and thus different parts of the human body are not distinct. This fact limits the choices of descriptors that can be applied. For example, the geodesic distances between two vertices of a mesh cannot be computed consistently across frames of a sequence, as they depend on whether body parts are glued together. For example, in Fig. 4, the geodesic paths, which connect the lower limbs of the human body, are shown for

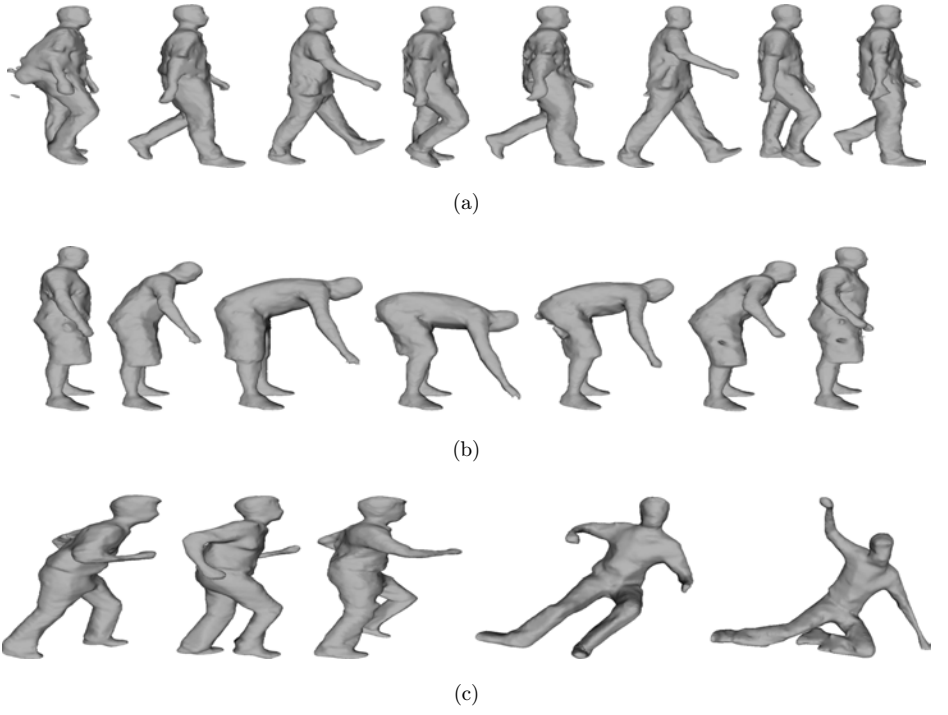

Fig. 3. Example frames from the i3DPost-Real dataset for the actions: (a) “walking”, (b) “bending”, (c) “running-falling”.

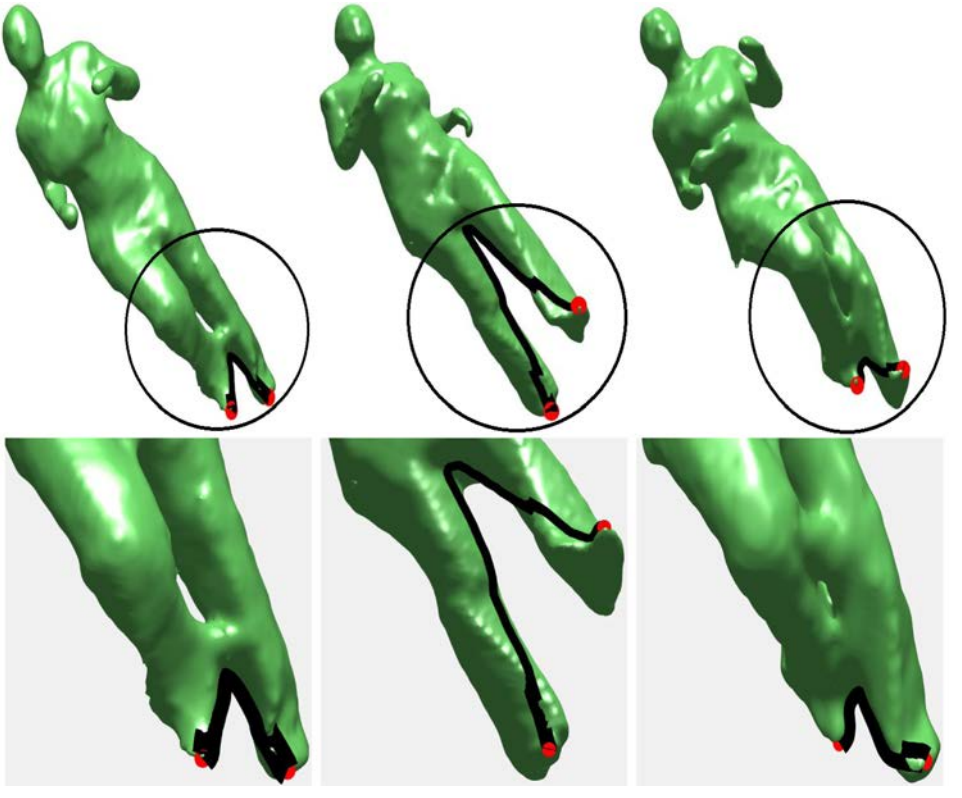

Fig. 4. Example frames from the i3DPost-Real dataset for the action "jogging" of Woman1 with the corresponding geodesic paths between lower limbs and their zoomed versions.

three different frames of the same sequence ("jogging" action of the model Woman1).

- There are parts of the human body which are disconnected from the rest of the body. This gives rise to problems, e.g. geodesic distance computation between disconnected components is not feasible. In Fig. 5, some examples of the action "sitting down-standing up" of the model Man6 with disconnected parts are shown.

Fig. 5. Example frames from the i3DPost-Real dataset for the action "sitting down-standing up" of Man6 with disconnected parts.

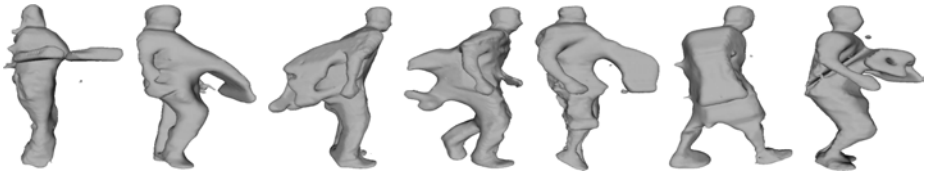

Fig. 6. Example frames from the i3DPost-Real dataset with high level of distortion.

- The quality of several 3D reconstructed models is poor, so that the shape of the relevant models does not relate to human figures. In Fig. 6, some examples of such models are shown.

The study of these problems led to the centroid-based part of the "Proposed M2" descriptor, as the centroid can be extracted more robustly than the rest of the salient points.

#### 4.1.2. *USurrey-artificial dataset*

The second dataset contains artificial data.[52,42] The number of models is 14 and each model has performed 28 different actions, so the total number of actions (sequences) which the dataset contains is 392. 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 different types of walking, seven are different types of running and four are other actions. Specifically, 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". The above enumeration of the actions is used for all relevant confusion matrices. In Fig. 7, example frames of this dataset are shown.

#### 4.1.3. *DUTH-artificial dataset*

The third dataset also contains artificial data. The number of models is 6 and each model has performed 10 different actions, so the total number of actions (sequences) which the dataset contains is 60. Each of the 6 models has different body type and the actions in this dataset are: (1) "hop on left foot", (2) "jumping", (3) "jumping forward", (4) "jumping-Turn", (5) "running", (6) "walking-90° turn left", (7) "walking-90° turn right", (8) "walking", (9) "walking with arms out — balancing", (10) "washing window". The majority of these actions belong to two general classes, namely jumping and walking. The general class jumping comprises the actions "jumping", "jumping forward" and "jumping-Turn", while the general class walking comprises the actions "hop on left foot", "running", "walking-90° turn left", "walking-90° turn

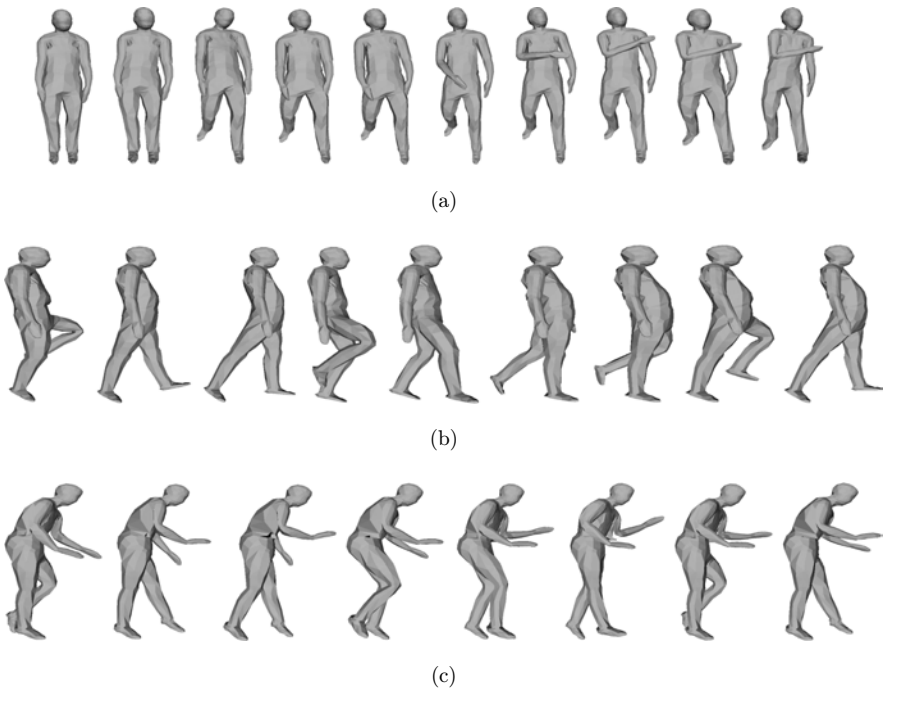

Fig. 7. Example frames from the USurrey-artificial dataset for the actions: (a) "shotarm", (b) "walkcowboy", (c) "sneak".

right", "walking" and "walking with arms out — balancing". This dataset is publicly available and the corresponding data can be found at: https://vc.ee.duth.gr/cmu-duth-mesh/. Example actions of this dataset are shown in Fig. 8.

The number of frames per sequence in the DUTH-artificial dataset is different (it ranges from 21 to 250). The actions contained in this dataset include common human actions and their variations, making the corresponding retrieval problem more challenging. It is denoted that this dataset was originally introduced in Ref. 2, where a description of its construction process is given.

### 4.2. *Results*

Let us first define some key metrics. *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. Precision-recall diagrams are often used in the evaluation of retrieval methods and show how these values relate; ideally a method should be at the $[1, 1]$ point. In addition a number of standard scalar measures are used[49]:

- *Nearest Neighbor (NN)*: The percentage of queries where the closest match belongs to the query class.

(a)

(b)

(c)

Fig. 8. Example frames from the DUTH-artificial dataset for the actions: (a) "hop on left foot", (b) "jumping", (c) "walking".

- *First Tier (FT)*: The recall for the $(C-1)$ closest matches where $C$ is the cardinality of the query's class.
- *Second Tier (ST)*: The recall for the $2 \cdot (C-1)$ closest matches where $C$ 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 scalar metrics are in the interval $[0, 1]$.

In Table 2, the retrieval results for the "Proposed M2" method are compared to the state-of-the-art[37] using the four scalar metrics on the i3DPost-Real dataset.

In Ref. 37, six static shape descriptors are extracted for each mesh of the human action sequences and DTW is used as similarity measure between the sequences of descriptors. These six static shape descriptors are the following:

- **The Hybrid descriptor**, which is composed of by two sub-descriptors. Each of these sub-descriptors, namely 2D and 3D, are also used separately. The 2D sub-descriptor is based on the Fourier coefficients of the projections of the 3D mesh

Table 2. Experimental retrieval results on the i3DPost-Real dataset.

| Method | *N* | NN | FT | ST | DCG |
|---|---|---|---|---|---|
| 2D[3,37] | 1 | 0.663 | 0.563 | 0.759 | 0.783 |
| 3D[3,37] | 14 | 0.925 | 0.727 | 0.861 | 0.885 |
| Hybrid[3,37] | 6 | 0.850 | 0.745 | 0.889 | 0.890 |
| PANORAMA[37,53] | 15 | 0.725 | 0.550 | 0.730 | 0.789 |
| Shape Dist.[32,37] | 1 | 0.775 | 0.516 | 0.655 | 0.759 |
| Spin Images[37,54] | 6 | 0.663 | 0.432 | 0.595 | 0.696 |
| Salient Points+DTW[2] | — | — | — | — | — |
| Proposed M1 | — | — | — | — | — |
| Proposed M2 | **7** | **0.975** | **0.829** | **0.966** | **0.949** |

*Notes*: 2D, 3D indicate the 2D and the 3D part of the Hybrid descriptor respectively. The column labeled N indicates the optimal value of this parameter used in the Hybrid descriptor.

onto each of the six faces of a cube. The 3D sub-descriptor is based on the spherical harmonic coefficients of a set of spherical functions. These functions are defined by determining the intersections of the models surface with a set of concentric spheres with increasing radii.

- **The shape distribution.** The Euclidean distance between a set of randomly selected points on the mesh surface is computed and a histogram of these distances is created.
- **The PANORAMA descriptor.** The mesh model is projected on the lateral surface of a cylinder, which includes the model. A set of depth images is produced, using two types of projections, one of which is based on the position and on the orientation of points on the mesh and their normals respectively. The final PANORAMA descriptor is based on the 2D Fourier and 2D Wavelet coefficients of these depth images.
- **The Spin Images descriptor.** A local coordinate system, based on the position and the normal of the vertices of the mesh model, is constructed in order to transform the 3D space of vertices to multiple 2D spaces.

The corresponding precision-recall diagrams are shown in Fig. 9. As can be seen, this method outperforms the state-of-the-art in terms of retrieval accuracy, demonstrating the robustness of the centroid-based descriptor on challenging datasets. The "Proposed M1" method as well as the method presented in Ref. 2 [referred as "Salient Points+DTW"] are not applicable on this dataset due to the geodesic-based extraction of the protrusions of the human body that they use.

In Ref. 2, the trajectories of six salient points of the human body and a set of kinematic features are extracted from these trajectories and the DTW algorithm is used to evaluate the distance between such descriptors. The main limitation of this method is that the extraction method of the six salient points is geodesic-based and cannot be applied on low quality mesh sequences.

In Table 3, the retrieval results for both of the proposed methods using the four basic scalar metrics on the USurrey-artificial dataset are given. The corresponding

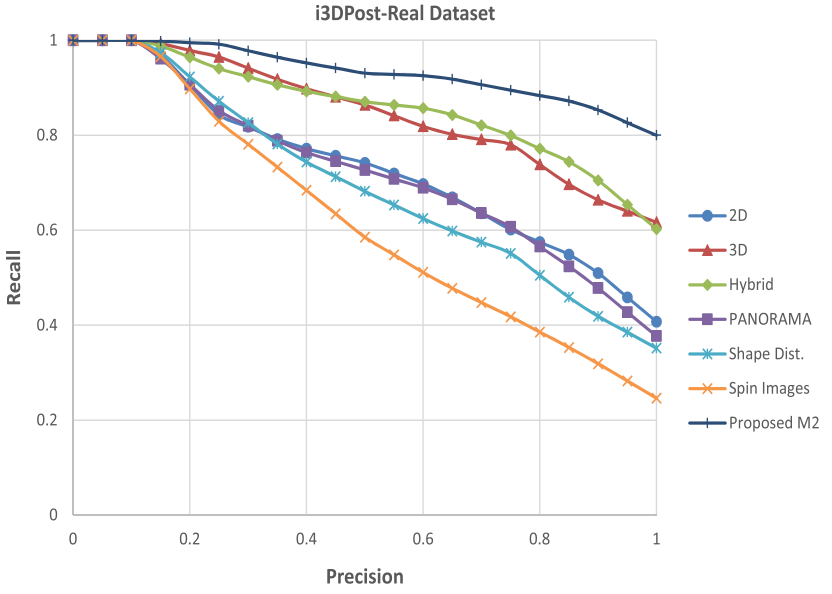


Fig. 9. Precision-recall diagrams for the i3DPost-Real dataset.

Table 3. Experimental retrieval results on the USurrey-artificial dataset with full sequences.

| Method | $N$ | NN | FT | ST | DCG |
|---|---|---|---|---|---|
| 2D[3,37] | 9 | 0.995 | 0.979 | **1.000** | 0.997 |
| 3D[3,37] | 8 | **1.000** | 0.983 | 0.999 | 0.999 |
| Hybrid[3,37] | 9 | 0.980 | 0.968 | 0.999 | 0.994 |
| PANORAMA[37,53] | 0 | 0.985 | 0.973 | **1.000** | 0.992 |
| Shape Dist.[32,37] | 9 | 0.921 | 0.889 | 0.972 | 0.956 |
| Spin Images[37,54] | 8 | **1.000** | 0.871 | 0.941 | 0.972 |
| Salient Points+DTW[2] | — | **1.000** | 0.998 | **1.000** | **1.000** |
| Proposed M1 | — | **1.000** | **1.000** | **1.000** | **1.000** |
| Proposed M2 | 9 | **1.000** | **1.000** | **1.000** | **1.000** |

*Notes*: 2D, 3D indicate the 2D and the 3D part of the Hybrid descriptor respectively. The column labeled N indicates the optimal value of this parameter used in the Hybrid descriptor.

precision-recall diagrams are shown in Fig. 10. The retrieval performance is ideal, as all scalar metrics are equal to 1.00 and the recall value is 1.00 for all precision values, for the two proposed methods and for the method presented in Ref. 2.

In order to examine the robustness of the method, an additional experiment using the USurrey-artificial dataset has been performed. Specifically, the initial part of the sequences in this dataset has been truncated by a random number of frames, between 0 and 50. By reducing the number of frames in the two sequences being compared, we reduce any explicit correspondence between their frames. In this experiment the

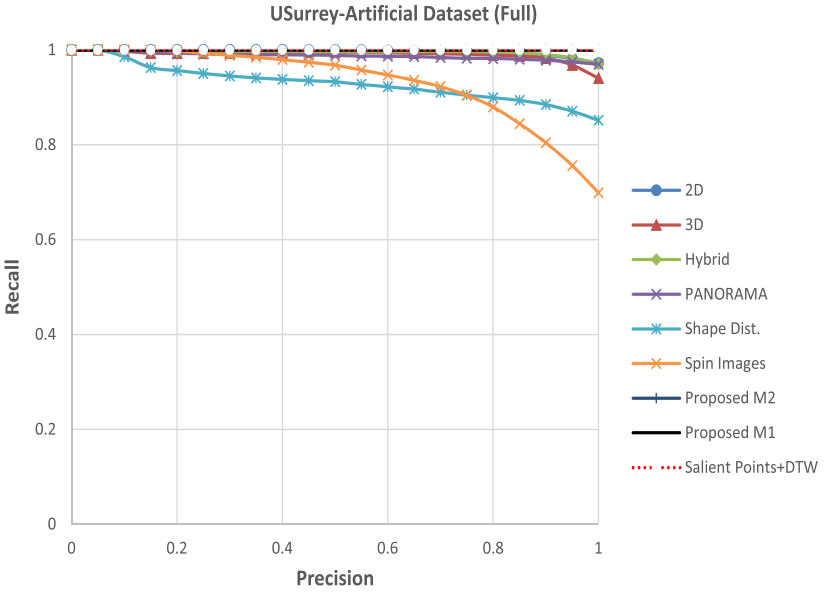


Fig. 10. Precision-recall diagrams for the USurrey-artificial dataset with full sequences.

sequences do not only end up with different lengths but also a part of the corresponding actions is lost through truncation. In Table 4, the retrieval results for both of the proposed methods on these truncated sequences are given, while the corresponding precision-recall diagrams are shown in Fig. 11. The same experiment has been repeated eight times, with different random lengths for the sequence truncations. The retrieval results correspond to the average performance of the eight experiments. The results indicate that the "Proposed M1" method is more robust against truncation compared to all other methods. The crucial step behind the success of this algorithm is that all descriptors have the same length, so the proposed

Table 4. Experimental retrieval results on the USurrey-artificial dataset with truncated sequences.

| Method | $N$ | NN | FT | ST | DCG |
|---|---|---|---|---|---|
| 2D[3,37] | 0 | 0.997 | 0.924 | 0.983 | 0.986 |
| 3D[3,37] | 0 | 0.946 | 0.894 | 0.991 | 0.973 |
| Hybrid[3,37] | 0 | 0.982 | 0.883 | 0.973 | 0.973 |
| PANORAMA[37,53] | 3 | 0.946 | 0.902 | 0.994 | 0.973 |
| Shape Dist.[32,37] | 0 | 0.890 | 0.797 | 0.903 | 0.926 |
| Spin Images[37,54] | 1 | 0.993 | 0.771 | 0.870 | 0.937 |
| Salient Points+DTW[2] | — | **1.000** | 0.962 | 0.987 | 0.994 |
| Proposed M1 | — | **1.000** | **0.987** | **0.998** | **0.999** |
| Proposed M2 | 0 | **1.000** | 0.957 | 0.989 | 0.994 |

*Notes*: 2D, 3D indicate the 2D and the 3D part of the Hybrid descriptor respectively. The column labeled $N$ indicates the optimal value of this parameter used in the Hybrid descriptor.

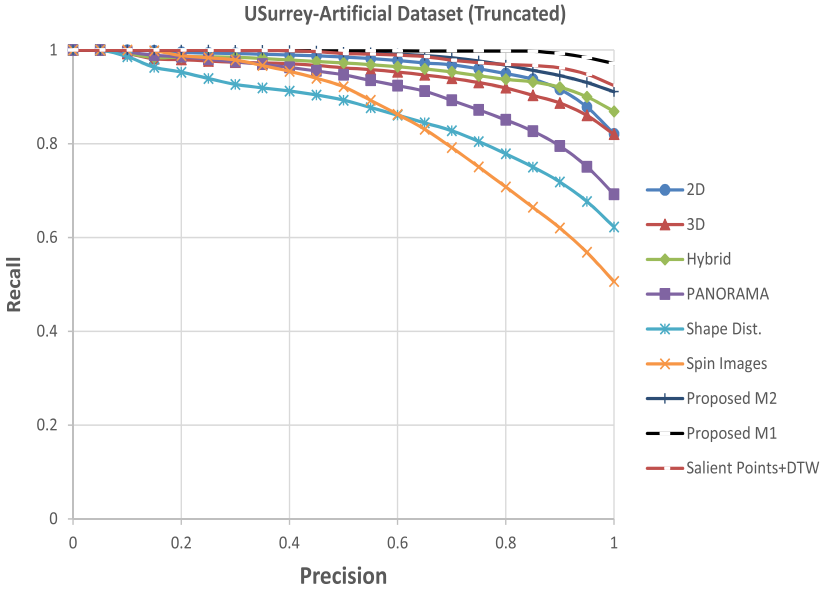


Fig. 11. Precision-recall diagrams for the USurrey-artificial dataset with truncated sequences.

pre-processing step combined with the usage of the chi-square distance (which is feasible because of the same length of the descriptors) results in the maintenance of the 1–1 frame correspondence between actions which belong to the same class.

In Table 5, the retrieval results for both of the proposed methods using the four basic scalar metrics on the DUTH-artificial dataset are given. The corresponding precision-recall diagrams are shown in Fig. 12. In this dataset, the retrieval performance, although high in absolute numbers, is the lowest among the three datasets for all descriptors, due to the intra-class nature of this dataset, as reported in Sec. 4.1.3.

Table 5. Experimental retrieval results on the DUTH-artificial dataset.

| Method | $N$ | NN | FT | ST | DCG |
|---|---|---|---|---|---|
| 2D[3,37] | 11 | 0.617 | 0.390 | 0.533 | 0.643 |
| 3D[3,37] | 3 | 0.750 | 0.527 | 0.717 | 0.763 |
| Hybrid[3,37] | 4 | 0.733 | 0.547 | 0.703 | 0.761 |
| PANORAMA[37,53] | 10 | 0.717 | 0.553 | 0.650 | 0.748 |
| Shape Dist.[32,37] | 6 | 0.633 | 0.367 | 0.563 | 0.638 |
| Spin Images[37,54] | 4 | 0.517 | 0.337 | 0.537 | 0.602 |
| Salient Points+DTW[2] | — | **0.967** | **0.767** | **0.863** | **0.907** |
| Proposed M1 | — | 0.917 | 0.673 | 0.790 | 0.851 |
| Proposed M2 | 15 | 0.950 | 0.720 | 0.837 | 0.877 |

*Notes*: 2D, 3D indicate the 2D and the 3D part of the Hybrid descriptor respectively. The column labeled $N$ indicates the optimal value of this parameter used in the Hybrid descriptor.

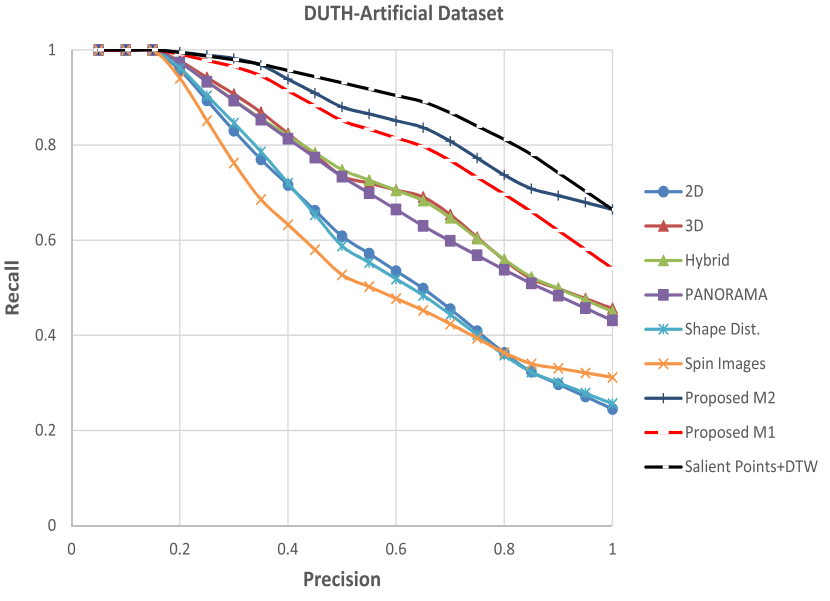


Fig. 12. Precision-recall diagrams for the DUTH-artificial dataset.

The method presented in Ref. 2 tops the performance list in retrieval accuracy terms, closely followed by the two proposed methods.

### 4.3. *Discussion*

In a field that is driven by datasets, we have presented two methods that achieve optimal or near optimal performance on real and artificial datasets. In this section, the corresponding retrieval results are individually discussed for the two proposed methods.

#### 4.3.1. *Retrieval results related to the "Proposed M1" method*

In the case of the USurrey-artificial dataset, there exists a 1–1 correspondence between the sequences of the same class, that facilitated ideal or near-ideal performance of the "Proposed M1" method (Tables 3 and 4).

In Fig. 13, the confusion matrix related to the DUTH-artificial dataset is shown. In the case of USurrey-artificial dataset, the corresponding confusion matrices are not shown, as the retrieval results are ideal using the full sequences and almost ideal using the truncated sequences.

In the case of the DUTH-artificial dataset, there were more challenges, including different actions from the ones in USurrey-artificial dataset as well as variability in the shape of the human bodies reflecting variations in height, age and weight.

In particular, in the case of the action "running" the retrieval performance is beyond 90%. This means that the "Proposed M1" method has strong discriminative

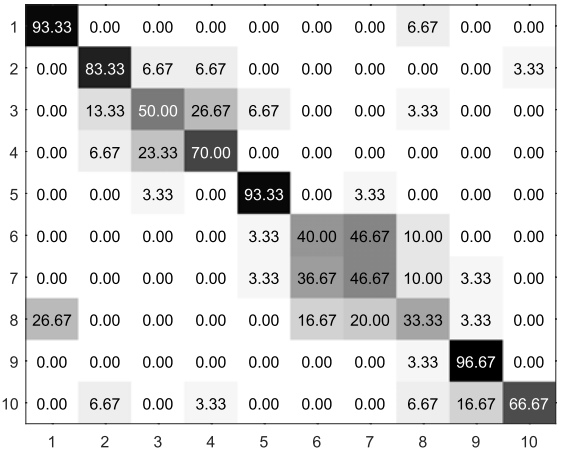


Fig. 13. The confusion matrix related to the DUTH-artificial dataset using "Proposed M1" method. The enumeration of the actions is compatible with the enumeration given in Sec. 4.1.3.

power in the case of actions which have significantly different rate from other actions. The highest retrieval performance is related to the action "walking with arms out — balancing". The main feature of this action is that the arms of the human are extended and the body swings while it simultaneously moves. The proposed approximation, which uses the trajectories of the upper limbs and the head, is suitable to discriminate this action from the others. To this effect, the retrieval performance for the action class "hop on left foot" is beyond 90%. In this case, the incorporation of the trajectories of the lower limbs of the human body in the proposed descriptor, is significant for successful retrieval of this action.

The most retrieval misses occur between the pairs of actions "walking-90° turn left" and "walking-90° turn right" as well as "jumping forward" and "jumping-Turn". These actions differ only in the direction while the other kinematic features, such as the rate and the total displacement of the initial position are identical, so their discrimination is difficult.

In Table 6, the retrieval results, using different distance measures for the DUTH-artificial dataset, are shown. As can be seen, the selection of chi-square distance as distance measure between the mesh sequences is experimentally justified and it is applicable because the descriptors have been extracted so as to have the same length.

Table 6. Experimental retrieval results on the DUTH-artificial dataset using different distance types.

| Distance type | NN | FT | ST | DCG |
|---|---|---|---|---|
| City Block | 0.750 | 0.477 | 0.580 | 0.703 |
| Euclidean | 0.767 | 0.447 | 0.567 | 0.694 |
| Square of Euclidean | 0.750 | 0.567 | 0.737 | 0.784 |
| Chi Square | **0.917** | **0.673** | **0.790** | **0.851** |

Table 7. Relative differences of the retrieval results using the pre-alignment step of the method.

| Dataset | NN | FT | ST | DCG |
|---|---|---|---|---|
| USurrey-artificial (truncated) | +0.010 | +0.278 | +0.181 | +0.143 |

*Note*: Positive differences denote improvement of the retrieval results.

The differences on the USurrey-artificial dataset are negligible and are thus not presented.

The significance of the pre-alignment step used in the "Proposed M1" method is experimentally highlighted in Table 7. In this Table, the differences between the same method with or without the pre-alignment step using the truncated version of USurrey-artificial dataset, are shown (positive differences indicate the improvement of the retrieval results using the pre-alignment step in the method). In this case, the improvement of the retrieval results is significant, as the pre-alignment step restores the 1–1 correspondence between the frames. In the cases of the full version of mesh sequences for the USurrey-artificial dataset and the DUTH-artificial dataset, the differences are not significant and are not presented.

Finally, the run times (in seconds) using the "Proposed M1" descriptor are presented in Table 8. The row "Extraction of descriptors" is referred to the mean run time for the extraction of the descriptor of one mesh sequence, while the row "Computation of distances" is referred to the mean run time for the computation of a distance between a pair of extracted descriptors. All experiments took place using a hybrid scheme, with MATLAB (version 2015b) and C code, on a machine with 16 GB memory and a CPU at 3.5 GHz.

#### 4.3.2. *Retrieval results related to the "Proposed M2" method*

**i3DPost-Real dataset:** In Fig. 14(a), the confusion matrix, using the i3DPost-Real dataset and the "Proposed M2" method, is shown. Generally, the retrieval performance is adequate for the majority of the classes of this dataset as the proposed method incorporates features that seem to discriminate these actions well. As can be seen, there are specific actions where the majority of retrieval misses occur:

- In the action "jogging" the FT value is 66.07% and the only action corresponding to retrieval misses is "walking". Similarly, in the action "walking" the FT value is

Table 8. Mean run times (in seconds) using the "Proposed M1" descriptor, for each dataset.

| Dataset | USurrey-artificial (full) | DUTH-artificial |
|---|---|---|
| Extraction of descriptors | 0.293 | 0.100 |
| Computation of distances[a] | 1.080 | 0.943 |

*Note*: [a]In one scale of the Wavelet Transform and for the trajectory of a specific salient point, for all sub-descriptors.

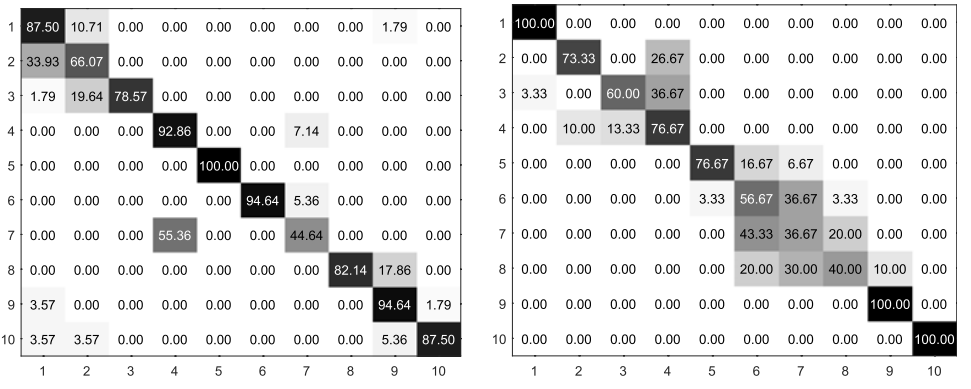


Fig. 14. The confusion matrix related to (a) the i3DPost-Real dataset and (b) the DUTH-artificial dataset for the "Proposed M2" method. The enumeration of the actions is compatible with the enumeration in Secs. 4.1.1 and 4.1.3, respectively.

87.50% and the majority of the retrieval misses correspond to the action "jogging". These two actions belong to the general class walking, where the action "jogging" incorporates a faster movement of the model than the action "walking". Additionally, the magnitude of the vertical component of the trajectory of the action "jogging" is a bit larger than the vertical component in the action "walking". Note that "jogging" differs from "running" in that "running" incorporates a faster movement of the model than "jogging". The "Proposed M2" method incorporates the feature of the rate of an action, so the discrimination across these actions is adequate.

- In the action "sitting down and standing up" the FT value is 44.64% and the only action corresponding to retrieval misses is "bending". In these two actions the vertical component dominates and the horizontal component is not sufficient to discriminate them adequately. In Fig. 15, some examples of the trajectories of the centroid for the actions "sitting down and standing up" and "bending" are shown.
- In the action "jumping" the FT value is 78.57% and the majority of the retrieval misses correspond to the action "jogging". These two actions have the common feature that the horizontal component of both is linear and their vertical component is sinusoidal. Normally, the vertical component of the corresponding trajectories of the centroid is more important in the action "jumping" than in the action "jogging". However, there are models who perform the action "jumping" without jumping high and the vertical components of the two actions are comparable, while the horizontal components are similar, too. In Fig. 16, some examples of the trajectories of the centroid of these two actions are shown.

In the other cases, the retrieval misses are low and thus retrieval accuracy is high. This implies that the set of features which are taken into account in the extraction of

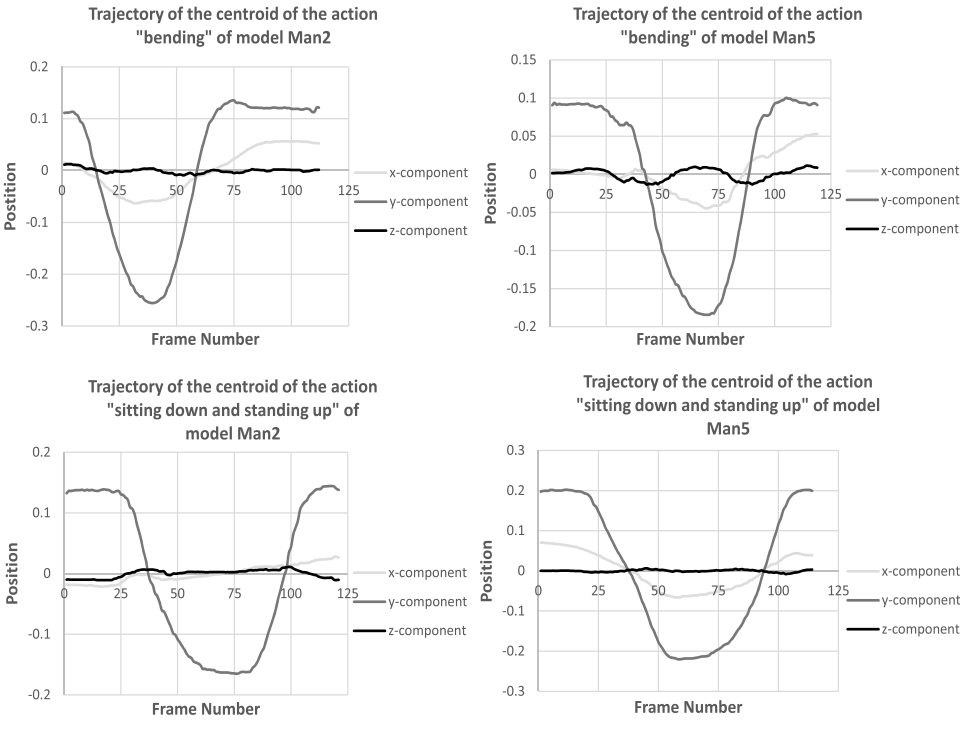


Fig. 15. Examples of the trajectories of the centroid for the actions "sitting down and standing up" and "bending" from the i3DPost-Real dataset.

the proposed descriptor of M2 method is sufficient to discriminate many common actions. Another important property of the extracted descriptor is that its two parts, namely the centroid-based and the Hybrid-based sub-descriptors, are complementary, as can be attested by the fact that the final retrieval performance is significantly higher than the retrieval performance of each of the two sub-descriptors individually. In Table 9, the retrieval performance of the two sub-descriptors and the final descriptor is summarized.

**USurrey-artificial dataset:** As the results in the USurrey-artificial dataset using the full sequences are ideal, all the entries in the main diagonal of the corresponding confusion matrix are equal to 100.00%, so this confusion matrix is not presented. In the case of the truncated sequences, the decrease in retrieval performance is not significant (thus the confusion matrix is also not presented) and occurs due to the loss of the one-to-one correspondence between the frames of the sequences. The few misses in this case occur between related actions, where the number of frames differs significantly.

Generally, the complementarity of the proposed sub-descriptors is maintained in this dataset. In Tables 10 and 11, the retrieval performance of the two

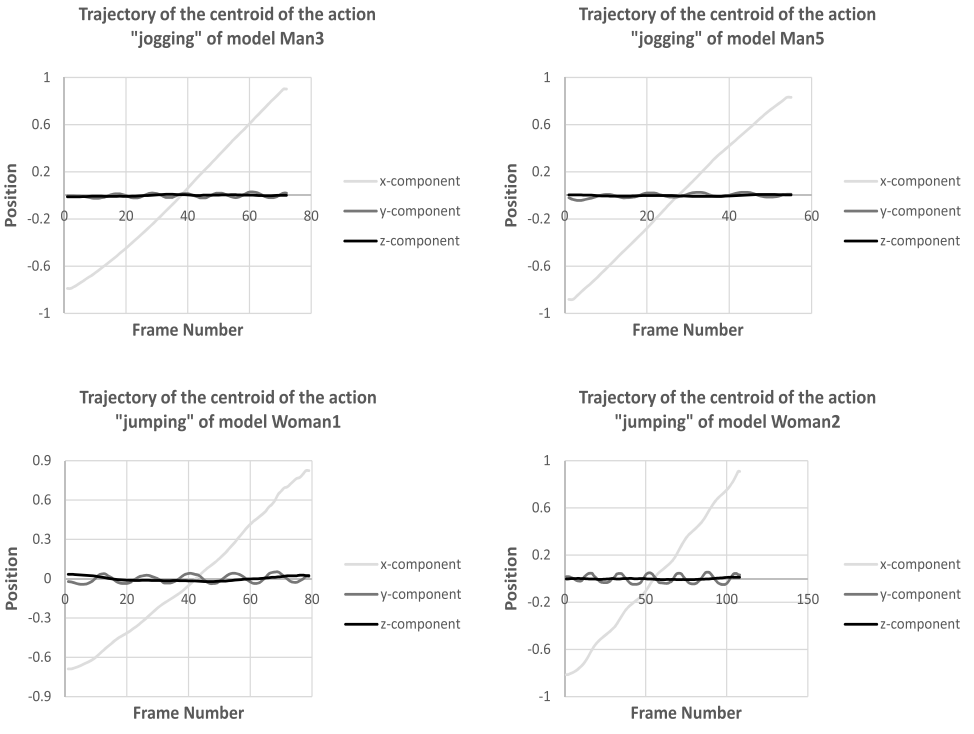


Fig. 16. Examples of the trajectories of the centroid for the actions "jogging" and "jumping" from the i3DPost-Real dataset.

Table 9. Experimental retrieval results using the final "Proposed M2" descriptor and its two components on the i3DPost-Real dataset.

| Method | NN | FT | ST | DCG |
|---|---|---|---|---|
| Proposed M2-C1 | 0.887 | 0.743 | 0.950 | 0.904 |
| Proposed M2-C2 | 0.850 | 0.739 | 0.893 | 0.888 |
| Proposed M2 | **0.975** | **0.829** | **0.966** | **0.949** |

Table 10. Experimental retrieval results using the final "Proposed M2" descriptor and its two components on the USurrey-artificial dataset, using the full sequences.

| Method | NN | FT | ST | DCG |
|---|---|---|---|---|
| Proposed M2-C1 | **1.000** | **1.000** | **1.000** | **1.000** |
| Proposed M2-C2 | 0.995 | 0.979 | 1.000 | 0.997 |
| Proposed M2 | **1.000** | **1.000** | **1.000** | **1.000** |

Table 11. Experimental retrieval results using the final "Proposed M2" descriptor and its two components on the USurrey-artificial dataset, using the truncated sequences.

| Method | NN | FT | ST | DCG |
|---|---|---|---|---|
| Proposed M2-C1 | 0.999 | 0.825 | 0.891 | 0.959 |
| Proposed M2-C2 | 0.967 | 0.910 | **0.993** | 0.979 |
| Proposed M2 | **1.000** | **0.957** | 0.989 | **0.994** |

sub-descriptors and the final descriptor, using the USurrey-artificial dataset with full and truncated sequences, respectively, is summarized.

**DUTH-artificial dataset:** In Fig. 14(b), the confusion matrix for the DUTH-artificial dataset using the "Proposed M2" method is shown. Some remarks:

- As reported in Sec. 4.1.3, the actions "jumping", "jumping Forward" and "jumping-Turn" belong to the general class jumping, so there are many retrieval misses among these classes. Theoretically, in the action "jumping" the models perform only vertical movement. In the action "jumping Forward" the models perform both vertical and horizontal movement. In the action "jumping-Turn" the models perform both vertical and horizontal movement and, simultaneously, turn the body.
  Example trajectories of the centroid of these actions are shown in Fig. 17. As can be seen, the shape of the trajectories is similar in all these cases. Note that in these cases, the horizontal component of the action "jumping" (first row in Fig. 17) is not negligible.
- As mentioned in Sec. 4.1.3, the actions "walking", "running", "walking-90° turn left", "walking-90° turn right", "hop on left foot" and "walking with arms out-balancing" belong to the general class walking. As can be seen in Fig. 14(b), the retrieval performance for the action "hop on left foot" is ideal (as also for the action "washing window") and almost ideal for the action "walking with arms out-balancing". In these actions the body poses of the models across the sequences differ significantly from the poses of a model which walks or runs. The FT value for the action "running" is 76.67% and the retrieval misses correspond to the actions "walking-90° turn left" and "walking-90° turn right". The majority of these misses occur for the actions "walking-90° turn left" of Model1 and Model2. In these instances the models do not turn 90° left, but the corresponding trajectories are more smooth. The horizontal component of these trajectories approximate more closely the horizontal component of a linear trajectory, such as the trajectory of the "running" action.

The complementarity of the two parts which constitute the final 'Proposed M2' descriptor is also investigated for this dataset. As shown in Table 12 retrieval performance is increased with the composition (except for the ST value).

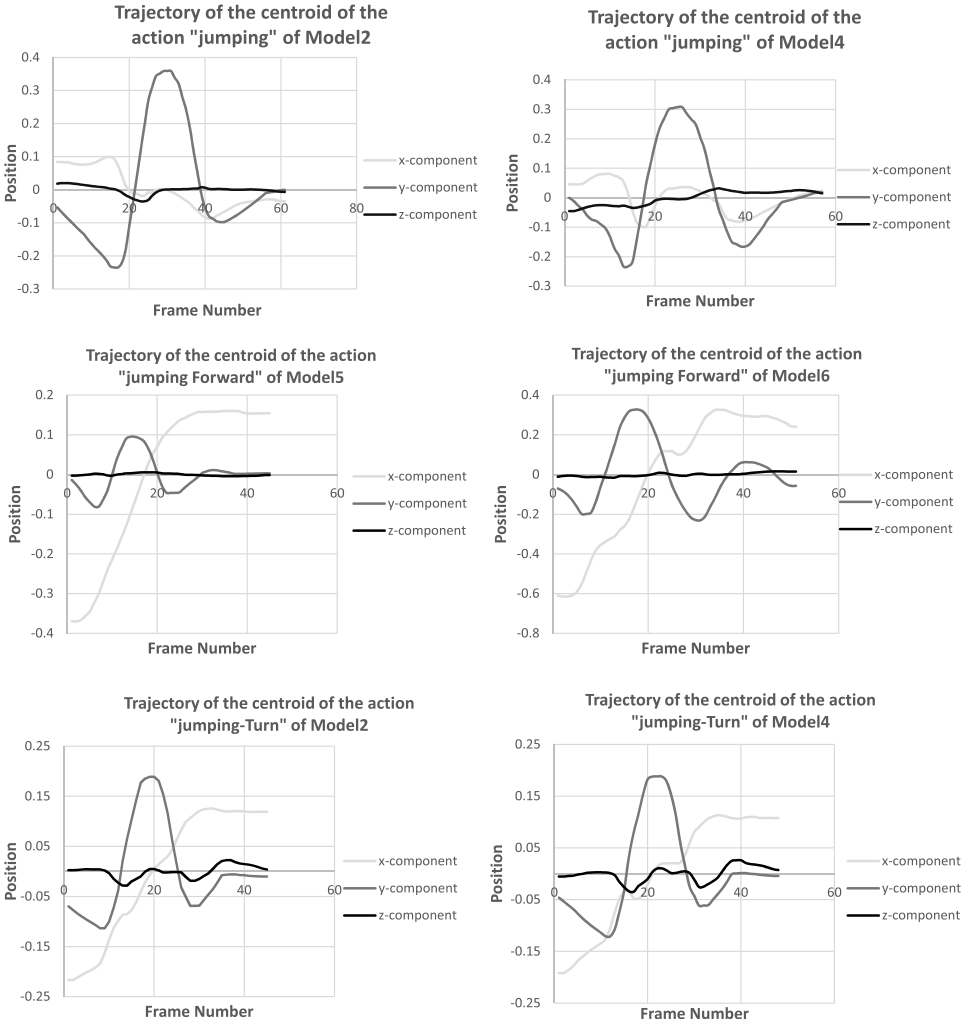


Fig. 17. Example trajectories of the centroid of the actions "jumping", "jumping Forward" and "jumping-Turn" from the DUTH-artificial dataset.

Table 12. Experimental retrieval results using the final "Proposed M2" descriptor and its two components on the DUTH-artificial dataset.

| Method | NN | FT | ST | DCG |
|---|---|---|---|---|
| Proposed M2-C1 | 0.917 | 0.707 | **0.843** | 0.868 |
| Proposed M2-C2 | 0.767 | 0.520 | 0.660 | 0.743 |
| Proposed M2 | **0.950** | **0.720** | 0.837 | **0.877** |

Table 13. Mean run times (in seconds) using the "Proposed M2" descriptor, for each of its components and for each dataset.

| Dataset | i3DPost-Real | USurrey-artificial (full) | DUTH-artificial |
|---|---|---|---|
| Extraction of C1 | 0.094 | 0.093 | 0.080 |
| Distances between C1 | 0.061 | 0.090 | 0.105 |
| Extraction of C2 | 0.203 | 0.053 | 0.097 |
| Distances between C2 | 0.184 | 0.202 | 0.242 |

Finally, the run times (in seconds) using the "Proposed M2" descriptor, for each of its component separately, are presented in Table 13. The rows "Extraction of C1" and "Extraction of C2" are referred to the mean run time for the extraction of the components C1 and C2 of the "Proposed M2" descriptor of one mesh sequence, respectively. The rows "Distances between C1" and "Distances between C2" are referred to the mean run time for the computation of the distance between a pair of extracted descriptors, using the components C1 and C2 of the "Proposed M2" descriptor, respectively. All experiments took place using a hybrid scheme, with MATLAB (version 2015b) and C code, on a machine with 16 GB memory and a CPU at 3.5 GHz.

## 5. Conclusions and Future Work

Two distinct methods for human action retrieval using 3D mesh sequences were presented. In the first method, an accurate spatio-temporal descriptor, which, is of constant length, for clean 3D mesh sequences of human actions is presented. In the second method, a robust descriptor for the retrieval of human actions represented as 3D mesh sequences, which is suitable for noisy meshes, such as those that often result from unprocessed scanning or 3D surface reconstruction errors, is proposed. This descriptor consists of two sub-descriptors, which are experimentally proven to be complementary.

The main advantages of the proposed methods are that they are *fully unsupervised*, so there is not dependence on training data, and have high retrieval performance on all publicly available datasets.

As machine learning has shown advantages in many recognition-related tasks over the past decade, the exploitation of the corresponding technologies is a potential extension of this work. Deep learning-based methods for human action recognition[55–57] have led to improved recognition and classification performance. A survey for human action recognition based on deep-learning is presented in Ref. 58. However, the application of such techniques requires the existence of huge training datasets labeled with the ground truth. This pre-assumes the construction of these huge datasets, which are not currently available in the case of 4D human actions (capture time and space usage being some of the reasons).

## References


1. G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception and Psychophysics* **14**(2), 201–211 (1973).
2. C. Veinidis, I. Pratikakis and T. Theoharis, "Unsupervised human action retrieval using salient points in 3D mesh sequences," *Multimedia Tools and Applications* (2018), doi: 10.1007/s11042-018-5855-2.
3. 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.
4. B. Yao and L. Fei-Fei, "Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses," *IEEE Trans. Pattern Anal. Mach. Intell.* **99**, 1691–1703 (2012).
5. F. Sener, C. Bas and N. Ikizler-Cinbis, "On recognizing actions in still images via multiple features," in *European Conference on Computer Vision Workshops and Demonstrations* (Springer, 2012), pp. 263–272.
6. F. S. Khan, M. A. Rao, J. van de Weijer, A. D. Bagdanov, A. Lopez and M. Felsberg, "Coloring action recognititon in still images," *Int. J. Comput. Vis.* 1–17 (2013).
7. G. Guo and A. Lai, "A survey on still image based human action recognition," *Pattern Recognition* **47**(10), 3343–3361 (2014), doi: https://doi.org/10.1016/j.patcog.2014.04.018.
8. M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *ICCV* (2011).
9. Y. Kong, D. Kit and Y. Fu, "A discriminative model with multiple temporal scales for action prediction," in *ECCV* (2014).
10. Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *Computer Vision and Pattern Recognition* **13**(9) (2018).
11. R. Hu, O. van Kaick, B. Wu, H. Huang, A. Shamir and H. Zhang, "Learning how objects function via co-analysis of interactions," *ACM Trans. Graph.* **35**(4), 1–47 (2016), doi: 10.1145/2897824.2925870.
12. S. Pirk, V. Krs, K. Hu, S. D. Rajasekaran, H. Kang, Y. Yoshiyasu, B. Benes and L. J. Guibas, "Understanding and exploiting object interaction landscapes," *ACM Trans. Graph.* **36**(3) 2017, doi: 10.1145/3083725.
13. X. Zhao, M. G. Choi and T. Komura, "Character-object interaction retrieval using the interaction bisector surface," *Computer Graphics Forum* **36**, 119–129 (2017).
14. J. Wang, Z. Liu, Y. Wu and J. Yuan, "Learning actionlet ensemble for 3D human action recognition," *TPAMI* **36**(5), 914–927 (2014).
15. F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal and R. Bajcsy, "Sequence of the most informative joints (SMIJ)," *Journal of Visual Communication and Image Representation* **25**(1), 24–38 (2014), doi: 10.1016/j.jvcir.2013.04.007.
16. G. Evangelidis, G. Singh and R. Horaud, "Skeletal quads: Human action recognition using joint quadruples," in *Proc. IEEE Intl. Conf. Pattern Recog.* (2014), pp. 1–6.
17. A. Shahroudy, G. Wang, T.-T. Ng and Q. Yang, "Multimodal multipart learning for action recognition in depth videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **38**(10), 2123–2129 (2016), doi: 10.1109/TPAMI.2015.2505295.
18. R. Qiao, L. Liu, C. Shen and A. van den Hengel, "Learning discriminative trajectorylet detector sets for accurate skeleton-based action recognition," *Pattern Recognition* **66**, 202–212 (2017), doi: http://dx.doi.org/10.1016/j.patcog.2017.01.015.
19. M. E. Hussein, M. Torki, M. A. Gowayyed and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations," in *IJCAI* 13, 2466–2472 (2013).

20. X. Yang and Y. Tian, "Effective 3D action recognition using eigenjoints," *Journal of Visual Communication and Image Representation* **25**(1), 2–11 (2014).
21. R. Vemulapalli, F. Arrate and 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), pp. 588–595.
22. L. L. Presti and M. L. Cascia, "3D skeleton-based human action classification: A survey," *Pattern Recognition* **53**, 130–147 (2016), doi: http://dx.doi.org/10.1016/j.patcog.2015.11.019.
23. H. Wang, A. Klaser, C. Schmid and C. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision* **103**(1), 60–79 (2013), doi: 10.1007/s11263-012-0594-8.
24. P. Matikainen, M. Hebert and R. Sukthankar, "Representing pairwise spatial and temporal relations for action recognition," in *ECCV* (2010), pp. 508–521.
25. D. Weinland and E. Boyer, "Action recognition using exemplar-based embedding," in *IEEE Conference on Computer Vision and Pattern Recognition — CVPR* (2008), pp. 1–7.
26. H. Jiang and D. R. Martin, "Finding actions using shape flows," in *Computer Vision — ECCV* (2008), pp. 278–292.
27. H. Zhang and L. E. Parker, "4-Dimensional local spatio-temporal features for human activity recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (2011), pp. 2044–2049, doi: 10.1109/IROS.2011.6094489.
28. O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 716–723, doi: 10.1109/CVPR.2013.98.
29. X. Yang, C. Zhang and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proceedings of the 20th ACM International Conference on Multimedia* (2012), pp. 1057–1060, doi: 10.1145/2393347.2396382.
30. G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri and B. P. Buckles, "Advances in human action recognition: A survey," *CoRR* (2015).
31. T. Yamasaki and K. Aizawa, "Motion segmentation and retrieval for 3D video based on modified shape distribution," *EURASIP Journal on Applied Signal Processing* **2007**(1), 211–222 (2017), doi: 10.1155/2007/59535.
32. R. Osada, T. Funkhouser, B. Chazelle and D. Dobkin, "Shape distributions," *ACM Trans Graph (TOG)* **21**(4), 807–832 (2002), doi: 10.1145/571647.571648.
33. T. Yamasaki and K. Aizawa, "A euclidean-geodesic shape distribution for retrieval of time-varying mesh sequences," in *IEEE ICME* (2009), pp. 846–849.
34. R. Slama, H. Wannous and M. Daoudi, "3D human motion analysis framework for shape similarity and retrieval," *Image Vision Computing* **32**(2), 131–154 (2014).
35. M. Vlachos, M. Hadjieleftheriou, D. Gunopulos and E. Keogh, "Indexing multidimensional time-series," *VLDB J.* **15**(1), 1–20 (2006), doi: 10.1007/s00778-004-0144-2.
36. D. Kasai, T. Yamasaki and K. Aizawa, "Retrieval of time-varying mesh and motion capture data using 2D video queries based on silhouette shape descriptors," in *IEEE ICME* (2009), pp. 854–857, doi: 10.1109/ICME.2009.5202629.
37. C. Veinidis, I. Pratikakis and T. Theoharis, "On the retrieval of 3D mesh sequences of human actions," *Multimedia Tools and Applications* **76**(2), 2059–2085 (2017).
38. M. B. Holte, T. B. Moeslund, N. Nikolaidis and I. Pitas, "3D human action recognition for multi-view camera systems," in *Proceedings of the 3DIMPVT* (2011).
39. M. B. Holte, T. B. Moeslund and P. Fihl, "View-invariant gesture recognition using 3D optical flow and harmonic motion context," *Computer Vision and Image Understanding* **114**(12), 1353–1361 (2010), doi: 10.1016/635 j.cviu.2010.07.012.

40. K. Kelgeorgiadis and N. Nikolaidis, "Human action recognition in 3D motion sequences," in *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO)* (IEEE, 2014).
41. P. Huang, T. Tung, S. Nobuhara, A. Hilton and T. Matsuyama, "Comparison of skeleton and non-skeleton shape descriptors for 3D video," in *Proceedings of the 3DPVT International Symposium* (2010).
42. P. Huang, A. Hilton and J. Starck, "Shape similarity for 3D video sequences of people," *International Journal of Computer Vision* **89**(2–3), 362–381 (2010).
43. A. Zaharescu, E. Boyer, K. Varanasi and R. Horaud, "Surface feature detection and description with applications to mesh matching," in *IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp. 373–380, doi: 10.1109/CVPR.2009.5206748.
44. C. Budd and A. Hilton, "Temporal alignment of 3D video sequences using shape and appearance," in *Conference on Visual Media Production* (2010), pp. 114–122, doi: 10.1109/CVMP.2010.22.
45. A. C. Bovik, *Handbook of Image and Video Processing (Communications, Networking and Multimedia)* (Academic Press, Orlando, 2005).
46. R. Q. Quiroga, O. W. Sakowitz, E. Basar and M. Schrmann, "Wavelet transform in the analysis of the frequency composition of evoked potentials," *Brain Res. Protoc.* **8**(1), 16–24 (2001).
47. P. Papadakis, I. Pratikakis, S. Perantonis and T. Theoharis, "Efficient 3D shape matching and retrieval using a concrete radialized spherical projection representation, in *Pattern Recognition* (2007), doi: 10.1016/j.patcog.2006.12.026.
48. M. Alexa, M. Kazhdan, J. Sun, M. Ovsjanikov and L. Guibas, "A concise and provably informative multi-scale signature based on heat diffusion," *Computer Graphics Forum* (2009).
49. P. Shilane, P. Min, M. Kazhdan and T. Funkhouser, "The princeton shape benchmark," in *Shape Modeling International* (2004), pp. 167–178.
50. N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis and I. Pitas, "The i3Dpost multi-view and 3D human action/interaction," in *Proc. CVMP* (2009), pp. 159–168.
51. J. Starck and A. Hilton, "Surface capture for performance based animation," *IEEE Computer Graphics and Applications* **27**(3), 21–31 (2007).
52. J. Starck and A. Hilton, "Model-based multiple view reconstruction of people," in *Proceedings of the Ninth International Conference on Computer Vision* (2003), pp. 915–922.
53. P. Papadakis, I. Pratikakis, T. Theoharis and S. Perantonis, "PANORAMA: A 3D shape descriptor based on panoramic views for unsupervised 3D object retrieval," *International Journal of Computer Vision* **89**, 177–192 (2010).
54. A. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3D scenes," *IEEE Trans on PAMI* **21**(5), 433–449 (1999).
55. Q. Ke, M. Bennamoun, S. An, F. Sohel and F. Boussaid, "A new representation of skeleton sequences for 3D action recognition," *Computer Vision and Pattern Recognition* (2017).
56. L. Wang, Y. Qiao and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," *Computer Vision and Pattern Recognition* (2015).
57. G. Varol, I. Laptev and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**(6), 1510–1517 (2018), doi: 10.1109/TPAMI.2017.2712608.
58. M. Koohzadi and N. M. Charkari, Survey on deep learning methods in human action recognition, *IET Computer Vision* **11**(8), 623–632 (2017), doi: 10.1049/iet-cvi.2016.0355.