



HAL
open science

3D Human Motion Analysis Framework for Shape Similarity and Retrieval

Rim Slama, Hazem Wannous, Mohamed Daoudi

► **To cite this version:**

Rim Slama, Hazem Wannous, Mohamed Daoudi. 3D Human Motion Analysis Framework for Shape Similarity and Retrieval. *Image and Vision Computing Journal*, 2014, 32, pp.131-154. hal-00923615

HAL Id: hal-00923615

<https://hal.science/hal-00923615>

Submitted on 3 Jan 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

3D Human Motion Analysis Framework for Shape Similarity and Retrieval

Rim Slama^{a,b}, Hazem Wannous^{a,b}, Mohamed Daoudi^{a,c}

^a*LIFL laboratory (UMR CNRS 8022), Villeneuve d'Ascq, France*

^b*University of Lille 1, Villeneuve d'Ascq, France*

^c*Institut Mines-Télécom / Télécom Lille, Villeneuve d'Ascq, France*

Abstract

3D Shape similarity from video is a challenging problem lying at the heart of many primary research areas in computer graphics and computer vision applications. In this paper, we address within a new framework the problem of 3D shape representation and shape similarity in human video sequences. Our shape representation is formulated using Extremal Human Curve (EHC) descriptor extracted from the body surface. It allows taking benefits from Riemannian geometry in the open curve shape space and therefore computing statistics on it. It also allows subject pose comparison regardless of geometrical transformations and elastic surface change. Shape similarity is performed by an efficient method which takes advantage of a compact EHC representation in open curve shape space and an elastic distance measure. Thanks to these main assets, several important exploitations of the human action analysis are performed: shape similarity computation, video sequence comparison, video segmentation, video clustering, summarization and motion retrieval.

Email addresses: rim.slama@telecom-lille.fr (Rim Slama),
hazem.wannous@telecom-lille.fr (Hazem Wannous),
mohamed.daoudi@telecom-lille.fr (Mohamed Daoudi)

Experiments on both synthetic and real 3D human video sequences show that our approach provides an accurate static and temporal shape similarity for pose retrieval in video, compared with the state-of-the-art approaches. Moreover, local 3D video retrieval is performed using motion segmentation and dynamic time warping (DTW) algorithm in the feature vector space. The obtained results are promising and show the potential of this approach. *Keywords:* Motion analysis, shape similarity, 3D video retrieval, 3D human action.

1 Introduction

While human analysis in 2D image and video has received a great interest during the last two decades, 3D human body is still a little explored field. Relatively few authors have so far reported works on static analysis of 3D human body, but even fewer on 3D human video analysis.

Parallel to this, 3D video sequences of human motion are more and more available. In fact, their acquisition by multiple view reconstruction systems or animation and synthesis approaches [1, 2] received a considerable interest over the past decade, following the pioneering work of Kanade [3].

Most of the recent research topics on 3D video focus mainly on performance, quality improvements and compression methods [4, 2, 5]. Consequently, 3D videos are yet mainly used for display. However, the acquisition of long sequences produces massive amounts of data which necessitates efficient schemes for navigating, browsing, searching, and viewing video data. Hence, we need to develop an efficient and effective descriptor to represent body shape and pose for shape retrieval and video clustering. We also need

17 a motion retrieval system to look for relevant information quickly.

18 3D Human body shape similarity is an important area, recently attracted
19 more attention in the field of human-computer interface (HCI) and computer
20 graphics, with many related research studies. Among these, research started
21 with 3D features have been applied for body pose estimation and 3D video
22 analysis.

23 In this paper, a unified framework providing several processing modules is
24 presented. All viewed within a duality pose/motion approach as summarized
25 in Figure 1 below.

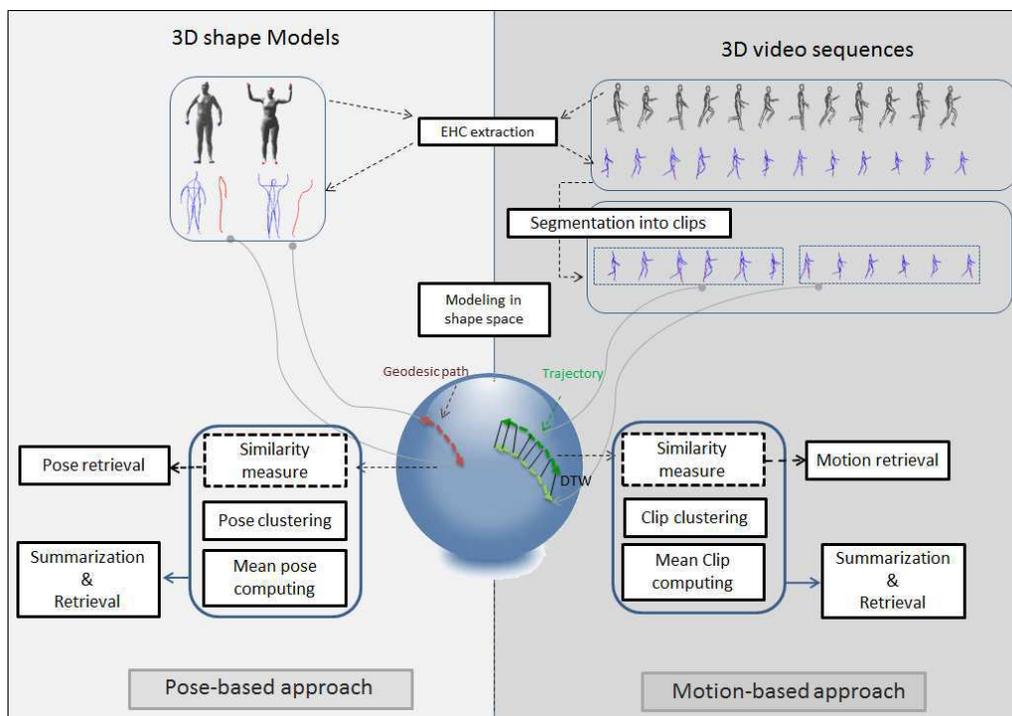


Figure 1: Overview of 3D human motion framework.

26 We first focus on the analysis of human pose and we propose a novel 3D

27 human curve-based shape descriptor called Extremal Human Curves (EHC).
28 This descriptor, extracted on body surface, is based on extremal features
29 and geodesics between them. Every 3D mesh is represented by a collection
30 of these open curves. The mesh to mesh comparison is then performed in
31 a Riemannian shape space using an elastic metric between each two corre-
32 spondent human curves.

33 At this level, our ultimate goal is to be able to perform reliable reduced
34 representation based on geodesic curves for shape and pose similarity metric.
35 Invariant to pose changes, our EHC descriptor allows pose (and motion)
36 comparison of subjects regardless of translation, rotation and scaling. Such
37 descriptor can be employed not only in pose retrieval for video annotation and
38 concatenation but also in motion retrieval, clustering and activity analysis.

39 Second, we are interested in the task of video segmentation and compar-
40 ison between motion segments for video retrieval. As a 3D video of human
41 motion consists of a stream of 3D models, we assume that EHC features
42 are extracted from all 3D shape frames of the sequence, which is further
43 segmented. For direct comparison of video sequences, the motion segmen-
44 tation can play an important role in the dynamic matching by segmenting
45 automatically the continuous 3D video data into small units describing basic
46 movements, called clips.

47 For the segmentation of these units, an analysis of minima on motion
48 vector is performed using the metric employed to compare EHC representa-
49 tions. Finally, the motion retrieval is achieved thanks to the dynamic time
50 warping (DTW) algorithm in the feature vector space.

51 The contributions of this paper are:

- 52 • The proposed surface-based shape descriptor called EHC provides a
53 compact representation of the shape. Thereby, reducing both the re-
54 quired space for storage and the time for comparison. As our descriptor
55 is composed of a collection of local human curves, the EHC can find a
56 number of useful applications lying on body part analysis.
- 57 • The use of video segmentation allows a semantic analysis of the human
58 motion, within a hierarchical structure of three levels "video-clip-pose".
- 59 • The modeling of curves in the shape space manifold allows calculating
60 statistics on shape models and motion clips. Thanks to this latter,
61 templates for the pose/clip are computed as average of a collection
62 of poses/clips. The matching with such templates which represents a
63 class, reduces retrieval complexity algorithm from n to $\log(n)$.
- 64 • The development of a unified framework, viewed as a duality pose/motion,
65 for several processing modules on video retrieval and understanding,
66 where all use the same features and similarity metric.

67 The outline of this paper is as follows: Section 2 discusses related works
68 in the area of static and temporal shape similarity and video retrieval. The
69 extremal curve extraction is presented in section 3. Section 4 describes the
70 pose modeling in shape space and the elastic metric used for curve compari-
71 son. In section 5, our approach used for motion segmentation and retrieval is
72 presented. Section 6 describes video clustering and summarization for motion
73 understanding. In section 7, evaluation of our framework and experimental
74 results for shape similarity, video segmentation and retrieval are performed.
75 Finally, we conclude by a discussion of the limitations of the approach in

76 section 8 and a summarizing of our results issues for future works in the
77 conclusion section.

78 **2. Related works**

79 3D shape representation and similarity have been under investigation for
80 a long time in various research fields (computer vision, computer graphics,
81 robotics) and for various applications (3D object recognition, classification,
82 retrieval). We address below, the most relevant works related to our ap-
83 proach, which only utilize the full-reconstructed 3D data for shape similarity
84 in 3D human video.

85 Most works which address this problem evaluate a similarity metric on
86 static shape descriptors based on the surface or on the volume. Others pro-
87 pose to extend the static approaches to temporal shape descriptors.

88 *2.1. Static descriptors*

89 Some of widely used 3D object representation approaches include: spin
90 images, spherical harmonics, shape context and shape distribution. Johnson
91 et al. [6] propose spin image descriptor, encoding the density of mesh vertices
92 into 2D histogram. Osada et al. [7] use a Shape Distribution, by computing
93 the distance between random points on the surface. Ankerst et al. [8] rep-
94 resent the shape as a volume sampling spherical histogram by partitioning
95 the space containing an object into disjoint cells corresponding to the bins
96 of the histogram. This later is extended with color information by Huang
97 et al. [9]. A similar representation to the Shape Histogram is presented by
98 Kortgen et al. [10] as 3D extended shape context. Kazhdan et al. [11] apply
99 spherical harmonics to describe an object by a set of spherical basis functions

100 representing the shape histogram in a rotation-invariant manner. These ap-
101 proaches use global features to characterize the overall shape and provide a
102 coarse description, that is insufficient to distinguish similarity in 3D video
103 sequence of an object having the same global properties in the time. A com-
104 parison of these shape descriptors combined with self-similarities is made by
105 Huang et al. [12].

106 Other works on the 3D shape similarity can be found in the literature,
107 where surface-based descriptors are often used with a step of features detec-
108 tion. The advantage of these features is that their detection is invariant to
109 pose change. The extremities can be considered as the one among the most
110 important features for the 3D objects. They can be used for extracting a
111 topology description of the object like Reeb-graph descriptor [13] or closed
112 surface-based curves [14, 15, 16]. The extraction and the matching of these
113 features have been widely investigated using different scalar functions from
114 geodesic distances to heat-kernel [17, 18, 19]. Tabia et al. [14] propose to
115 extract arbitrarily closed curves amounting from feature points and use a
116 geodesic distance between curves for 3D object classification. Elkhoury et al.
117 [15] extract the same closed curves but they use heat-kernel distance in the
118 3D object retrieval process.

119 *2.2. Temporal descriptors*

120 Since significant progress in multiple view reconstruction techniques has
121 been made, 3D video sequences of human motion are more and more avail-
122 able. However, the need for handling and processing such data led to several
123 approaches using temporal shape representation and matching.

124 Huang et al. [12] extend the use of static descriptors to temporal ones for

125 frame retrieval, in a 3D human video, using time filtering and shape flows
126 obtained via invariant-rotation shape histograms. Such approaches give a
127 good shape descriptor but usually do not capture any geometrical informa-
128 tion about the 3D human body pose and joint positions/orientations. This
129 prevents using them in certain applications that require accurate estimation
130 of the pose (and the joints in some cases) of the body parts. The temporal
131 similarity in 3D video is addressed also in the case of skeletal motion and is
132 evaluated from difference in joint angle or position together with velocity and
133 acceleration [20]. Huang et al. [21] demonstrate that skeleton-based Reeb-
134 Graph descriptor has a good performance in the task of finding similar poses
135 of the same person in 3D video. Shape similarity is also used for solving the
136 problem of video retrieval by matching frames and comparing correspondent
137 ones using a specified metric. In Yamasaki et al., [22] the modified shape dis-
138 tribution histogram is employed as feature representation of 3D models. The
139 similar motion retrieval is realized by Dynamic Programming matching us-
140 ing the feature vectors and Euclidean distance. The Dynamic Time Warping
141 algorithm (DTW), based on Dynamic Programming and some restrictions,
142 was also widely used to resolve the problem of temporal alignment. Given
143 two time series with different size, DTW finds an optimal match measuring
144 the similarity between these sequences which may vary in time or speed.
145 Thereby, by a frame descriptor and the temporal alignment using DTW,
146 many authors succeed to perform action recognition or sequence matching
147 for indexing [23, 24, 25].

148 Recently, Tung et al. [13] propose a topology dictionary for video un-
149 derstanding and summarizing. Using the Multi-resolution Reeb Graph as

150 a relevant descriptor for the shape in video stream for clustering. In this
151 approach, they perform a clustering of the video frames into pose clusters
152 and then they represent the whole sequence with a Markov motion graph in
153 order to model the topology change states.

154

155 From the above review, we can identify certain issues in order to consider in
156 our approach. Most of these works have attempted to use global description
157 of the model ignoring the local details. The similarity metric is usually cal-
158 culated directly on descriptors whereas the notion of motion is incorporated
159 by time convolution of the distance metric itself computed from static poses.
160 The video sequence is considered as a succession of frames in time and not a
161 succession of elementary motions (or gestures).

162 On one hand, the extremities feature points used in many state-of-the-art
163 algorithms can be considered as an important compact semantic represen-
164 tation of human posture. On the other hand, the shape analysis of curves
165 extracted from human body mesh allows representing the shape variations.
166 Choosing some representative curves of the body surface may provide an
167 efficient and a compact representation of human shape.

168 Our approach has several benefits: (1) the EHC descriptor can be con-
169 sidered as a surface skeletal based representation, which allows to describe
170 surface deformations of the human posture. As it is composed by a collec-
171 tion of local extremal open 3D curves, a body part representation can be
172 performed; (2) the motion analysis is incorporated in two ways, firstly by
173 time convolution of the distance metric vectors for pose retrieval in video
174 sequence, and secondly by employing motion segmentation and the notion of

175 clips; (3) the video segmentation allows the localization of transition states in
176 the video, in order to analyze the local dynamic of the motion, representing
177 an atomic action or gesture; and (4) an original idea is proposed to represent
178 a clip as a trajectory composed of a collection of successive frames viewed
179 as points in shape space. Finally, the video segmentation and clustering are
180 exploited in content-based summarization and motion retrieval.

181 **3. Extremal Curves**

182 We aim to represent a body shape as a skeleton based shape representa-
183 tion. This skeleton will be extracted on the surface of the mesh by connecting
184 features located on the extremities of the body. The main idea behind the use
185 of this representation is to analyze pose variation with elastic deformation of
186 the body, using representative curves on the surface.

187 *3.1. Feature point detection*

188 Feature points refer to the points of the surface located at the extremity of
189 its prominent components. They are useful in many applications, including
190 deformation transfer, mesh retrieval, texture mapping and segmentation. In
191 our approach, feature points are used to represent a new pose descriptor
192 based on curves connecting each two extremities. Several approaches have
193 been proposed in the literature to extract feature points; Mortara et al. [26]
194 select as features points the vertices where Gaussian curvature exceeds a given
195 threshold. Unfortunately, this method can miss feature points because of the
196 threshold parameter and cannot resolve extraction on constant curvature
197 areas. Katz et al. [27] develop an algorithm based on multidimensional
198 scaling, in quadratic execution complexity. Another approach more robust,

199 is proposed by Tierny et al. [28] to detect extremal points, based on geodesic
 200 distance evaluation. This approach is used successfully to detect the body
 201 extremities, since it is stable and invariant to geometrical transformations
 202 and model pose. The extraction process can be summarized as the following:

203 Let v_1 and v_2 be the most geodesic distant vertices on a connected tri-
 204 angulated surface S of a human body. These two vertices are the farthest
 205 on S , and can be computed using Tree Diameter algorithm (Lazarus et al.
 206 [29]). Now, let f_1 and f_2 be two scalar functions defined on each vertex v of
 207 the surface S as follows:

$$f_1(v) = g(v, v_1) \setminus f_2(v) = g(v, v_2) \quad (1)$$

208 where $g(x, y)$ is the geodesic distance between points x and y on the
 209 surface. Let E_1 and E_2 be respectively the sets of extrema vertices (minima
 210 and maxima) of f_1 and f_2 on S (calculated in a predefined neighbourhood).
 211 We define the set of feature points of the surface of human body S as the
 212 intersection of E_1 and E_2 . Concretely, we perform a crossed analysis in order
 213 to purge non-isolated extrema, as illustrated in Figure 2 (top). The f_1 local
 214 extrema are displayed in blue color, f_2 local extrema are displayed in red
 215 color and feature points resulting from their intersection are displayed in
 216 mallow color. Figure 2 (bottom) shows different persons from three different
 217 datasets where feature extraction is stable despite change in shape, pose and
 218 clothing for each actor.

219 3.2. Body curve extraction

220 Let M be a body surface and $E = \{e_1, e_2, e_3, e_4, e_5\}$ a set of feature points
 221 on the body representing the output of the extraction process. Let β denotes

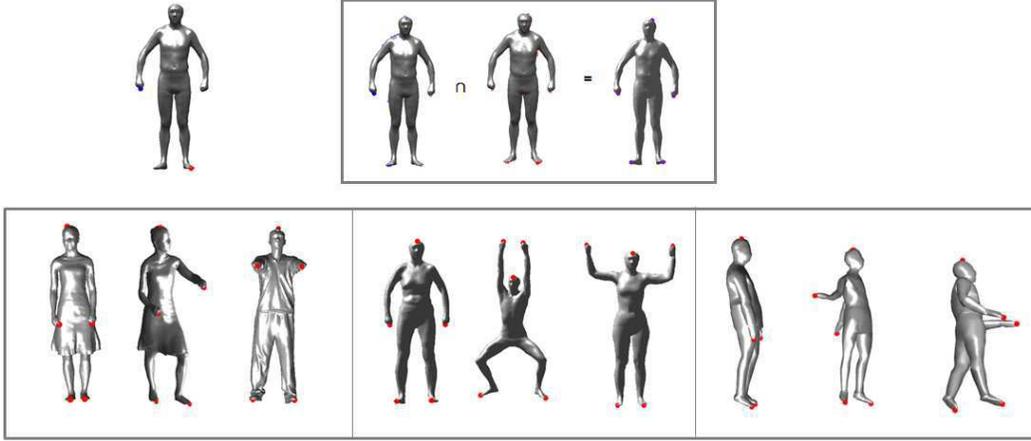


Figure 2: Extremity points of the 3D human body. (top) extracting process, (bottom) different human body subjects in different poses.

222 the open curve on M which joints two feature points of M $\{e_i, e_j\}$. To ob-
 223 tain β , we seek for geodesic path P_{ij} , whose length is shortest while passing
 224 through the surface of the mesh, between e_i and e_j . We repeat this step to
 225 extract extremal curves from the body surface ten times so that we do all
 226 possible paths between elements of E . As illustrated in the top of Figure 3,
 227 the body posture is approximated by using these extremal curves $M \sim \bigcup \beta_{ij}$,
 228 and we can categorize these curves into 5 categories (Figure 3 bottom):

- 229 • Curves connecting hand and foot on the same side: for controlling the
 230 movement of the left/right half of the body.
- 231 • Curves between hands and between feet: for controlling the movement
 232 of the upper/lower body.
- 233 • Curves connecting crossed hand and foot: for controlling the movement

234

of the crossed limbs.

235

- Curves between head and feet: for controlling the movement of right/left foot.

236

237

- Curves between head and hands: for controlling the movement of right/left hands.

238

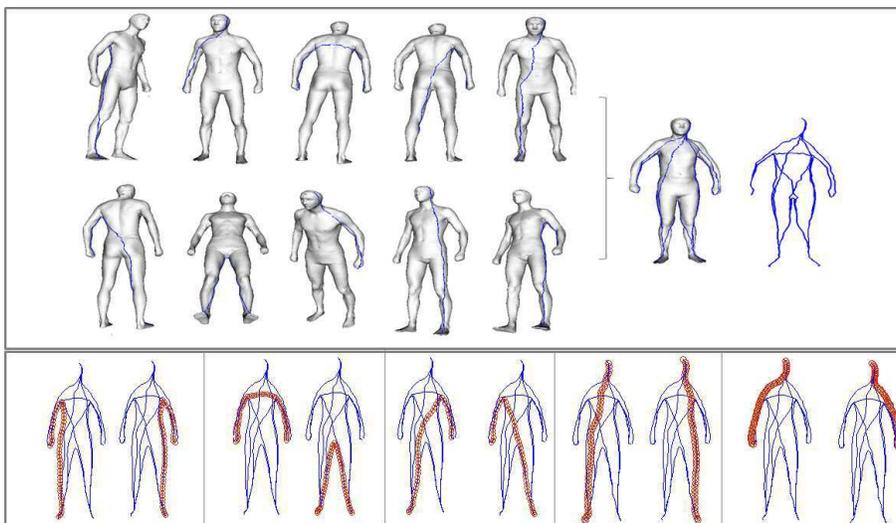


Figure 3: Body representation as a collection of extremal curves.

239

Note that modeling objects with curves is recently carried out for several

240

applications; Abdelkader et al. [30] use closed curves extracted from human

241

silhouettes to characterize human poses in 2D videos for action recognition.

242

Drira et al. [31] use open curves extracted from nose tip and face surface as

243

a surface parametrization for 3D face recognition.

244

In our approach, we have chosen to represent the body pose by a col-

245

lection of curves for two reasons. Firstly, these curves connect limbs and

246 give obviously a good representation of the body shape and pose, using a
247 reduced representation of the mesh surface. Secondly, this representation
248 allows studying the shape variation using Riemannian geometry by project-
249 ing these curves in the shape space of curves and using its elastic metric
250 introduced by Joshi et al. [32].

251 4. Pose modeling in shape space

252 In order to compare the similarity between two human body postures,
253 we must quantify the change of shape between correspondent curves. To do
254 this, the metric used to compare shape of curves can be computed inside an
255 open curve shape space.

256 In the last few years, many approaches have been developed to analyze
257 shapes of 2-D curves. We can cite approaches based on Fourier descriptors,
258 moments or the median axis. More recent works in this area consider a formal
259 definition of shape spaces as a Riemannian manifold of infinite dimension
260 on which they can use the classic tools for statistical analysis. The recent
261 results of Michor et al. [33], Klassen et al. [34] and Yezzi et al. [35] show
262 the efficiency of this approach for 2-D curves. Joshi et al. [32] have recently
263 proposed a generalization of this work to the case of curves defined in \mathbb{R}^n .
264 We adopt this work to our problem since our 3-D curves are defined in \mathbb{R}^3 .

265 4.1. Elastic distance

266 While human body is an elastic shape, its surface can be simply affected
267 by a stretch (raising hand) or a bind (squatting). In order to analyze human
268 curves independently to this elasticity, an elastic metric is needed within a
269 shape space framework.

270 Let $\beta : I \rightarrow \mathbb{R}^3$, for $I = [0, 1]$, represents an extremal curve obtained as de-
 271 scribed above. To analyze its shape, we shall represent it mathematically us-
 272 ing a *square-root velocity function* (SRVF), denoted by $q(t) \doteq \dot{\beta}(t)/\sqrt{\|\dot{\beta}(t)\|}$.
 273 $q(t)$ is a special function introduced by Joshi et al.[32] that captures the shape
 274 of β and is particularly convenient for shape analysis.

275 The set of all unit-length curves in \mathbb{R}^3 is given by $\mathcal{C} = \{q : I \rightarrow \mathbb{R}^3 \mid \|q\| =$
 276 $1\} \subset \mathbb{L}^2(I, \mathbb{R}^3)$, where using \mathbb{L}^2 -metric on its tangent spaces, \mathcal{C} becomes a
 277 Riemannian manifold.

278 **Proposition 4.1.** *Having two open curves represented by their SRVF, q_1*
 279 *and q_2 , the shortest geodesic between them in the shape space of open curves*
 280 *is given by: $\alpha(\tau) = \frac{1}{\sin(\theta)} (\sin((1 - \tau)\theta)q_1 + \sin(\theta\tau)q_2^*)$,*
 281 *and the geodesic distance is given by: $d_s(q_1, q_2) \doteq \cos^{-1}(\langle q_1, q_2^* \rangle)$.*

282 where q_2^* is the optimal element associated with the optimal rotation O^*
 283 and re-parametrization γ^* of the second curve.

284 This defined distance allows comparing shape curves regardless of iso-
 285 metric and elastic deformation. In Figure 4, a geodesic path between each
 286 corresponding two extremal curves, taken from two human bodies doing dif-
 287 ferent poses, is computed in shape space.

288 For the left model, the person’s arm is down and for the right model it
 289 is raised. The geodesic path between each two curves is shown in the shape
 290 space. This evolution looks very natural under the elastic matching.

291 4.2. Static shape similarity

292 The elastic metric applied on extremal curve-based descriptors can be
 293 used to define a similarity measure. Given two 3D meshes x, y and their

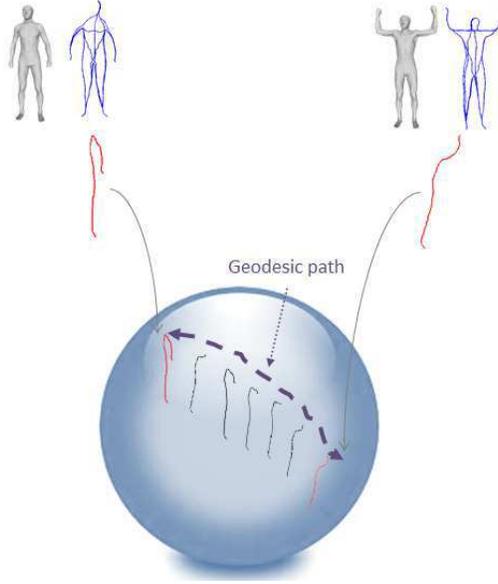


Figure 4: Geodesic path between two extremal human curves of neutral pose with raised hands.

294 descriptors $x' = \{q_1^x, q_2^x, q_3^x, \dots, q_N^x\}$ and $y' = \{q_1^y, q_2^y, q_3^y, \dots, q_N^y\}$, the mesh-to-
 295 mesh similarity can be represented by the curve pairwise distances and can
 296 be defined as follows:

$$297 \quad s(x, y) = d(x', y') , \quad (2)$$

$$d(x', y') = \frac{\sum_{i=1}^N d(\beta_i^x, \beta_i^y)}{N} = \frac{\sum_{i=1}^N d_s(q_i^x, q_i^y)}{N}. \quad (3)$$

298 where N is the number of curves used to describe the mesh. The mean of
 299 curve distances between two descriptors captures the similarity between their
 300 mesh poses. In case of shape change in even one curve, the global distance
 301 is affected and it increases indicating that the poses are different. In order
 302 to have a global distance, an arithmetic distance can be computed in order
 303 to compare human poses.

304 *4.3. Average poses*

305 The use of EHC descriptor to represent the human pose by a collection
 306 of 3D open curves allows analyzing the human shape using the geometrical
 307 framework. It also allows computing some related statistics like "average"
 308 of several extremal human curves. Such an average, called Karcher Mean, is
 309 introduced by Srivastava et al. [36]. It can be computed between different
 310 poses to represent the intermediate pose, or between similar poses done by
 311 several actors to represent a template for similar poses.

312 We are interested in defining a notion of "mean" for a given set of human
 313 postures in the same cluster of poses for the goal of fast pose retrieval.

314 To compute the average of EHC representation, we need only to know
 315 how to compute an average for one extremal human curve. The Riemannian
 316 structure defined on the shape space \mathcal{S} enables us to perform such statistical
 317 analysis for computing average and variance for each 3D open curve on body
 318 surface. The intrinsic average or the Karcher Mean utilizes the intrinsic
 319 geometry of the manifold to define and compute a mean on that manifold.
 320 In order to calculate the Karcher Mean of extremal human curves $\{q_1^\alpha, \dots, q_n^\alpha\}$
 321 in \mathcal{S} , we define the variance function as:

$$\mathcal{V} : \mathcal{S} \rightarrow \mathbb{R}, \mathcal{V}(\mu) = \sum_{i=1}^n d_{\mathcal{S}}(q_i^\alpha, \mu)^2 \quad (4)$$

322 The Karcher Mean is then defined by:

$$\bar{q}^\alpha = \arg \min_{\mu \in \mathcal{S}} \mathcal{V}(\mu) \quad (5)$$

323 The gradient of \mathcal{V} is used in the tangent space $T_\mu(\mathcal{S})$ to iteratively update
 324 the current mean μ . \bar{q}^α is an element of \mathcal{S} that has the smallest geodesic

325 path length from all given extremal human curves for the index α .

326 An example of using the Karcher Mean to compute average curve for 6
327 extremal human curves connecting hand and foot from the same side is shown
328 in the top of Figure 5, and several examples of using the Karcher mean to
329 compute average EHC representation are shown in the bottom of this figure.

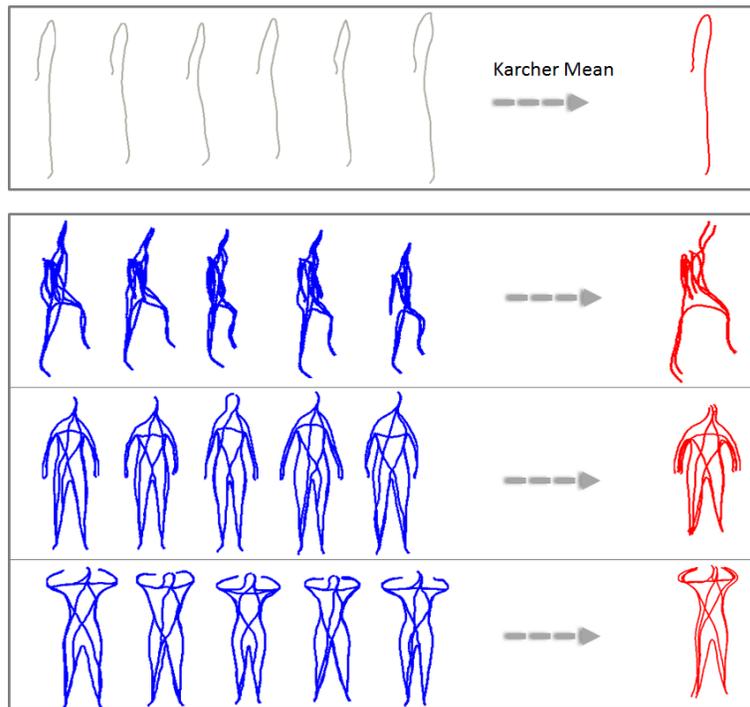


Figure 5: Example of Karcher Mean computation. (top) Mean curve for six extremal human curves: curve connecting hand and foot from the same side. (bottom) Example of average poses computed using Karcher mean.

330 5. Motion segmentation and matching

331 Based on our EHC representation of the shape model, it is possible to
332 compare two video sequences by matching all pairwise correspondent ex-
333 tremal curves inside their frames, using the geodesic distance in the shape
334 space. However, a sequence of human action can be composed of several
335 distinct actions, and each one can be repeated several times. Therefore, the
336 motion segmentation can play an important role in the dynamic matching
337 by dividing the whole 3D video data into small, meaningful and manageable
338 elementary actions called clips. EHC descriptor will be employed to segment
339 continuous sequences into clips.

340 5.1. Motion segmentation

341 Video segmentation has been studied for various applications, such as
342 gesture recognition, motion synthesis and indexing, browsing and retrieval.
343 A vast amount of works in video segmentation has been performed for 2D
344 video [37], where usually the object segmentation is firstly performed before
345 the movement analysis. In Rui et al. [38], an optical flow of moving objects
346 is used and motion discontinuities in trajectories of basis coefficient over time
347 are detected. However, in Wang et al. [39], break points were considered as
348 local minima in motion and local maxima in direction change.

349 Motion segmentation is strongly applied in several algorithms using 3D
350 motion capture feature points trackable within the whole sequence, to seg-
351 ment the video. Detected local minima in motion (Shiratori et al. [40]) or
352 extrema (Kahol et al. [41]) are used in motion segmentation for kinematic
353 parameters.

354 Most of works on the 3D video segmentation use the motion capture
355 data, and very few of them were applied to dynamic 3D mesh. One of them
356 is presented by Xu et al. [42], where a histogram of distance among vertexes
357 on 3D mesh is generated to perform the segmentation through thresholding
358 step defined empirically. In Yamasaki et al. [43], the motion segmentation is
359 automatically conducted by analyzing the degree of motion using modified
360 shape distribution for mainly japanese dances. These sequences of motion
361 are paused for a moment and then they are consider as segmentation points.
362 Huang et al. [44] propose an automatic key-frame extraction method for
363 3D video summarization. To do so, they compute the self similarity matrix
364 using volume-sampling spherical Shape Histogram descriptor. Then, they
365 construct a graph based on this self similarity matrix and define a set of key
366 frames as the shortest path of this graph.

367 In our work, we propose an approach fully automatic to segment a 3D
368 video efficiently without making neither thresholding step nor assumption on
369 the motion's nature. In motion segmentation, the purpose is to split automat-
370 ically the continuous sequence into segments which exhibit basic movements,
371 called clips. As we need to extract meaningful clips, the segmentation is
372 overly fine and can be considered as finding the alphabet of motion. For a
373 meaningful segmentation, motion speed is an important factor. In fact, when
374 human changes motion type or direction, the motion speed becomes small
375 and this results in dips in velocity. We exploit this latter by finding the local
376 minima for the change in type of motion and local maxima for the change
377 in direction. The extrema detected on velocity curve should be selected as
378 segment points (see Figure 6). We show frames detected as maxima (the ac-

379 tor changes the foot's direction) on the top of the plot, and frames detected
 380 as minima (the actor raise the other foot) on the bottom. In this work, we
 381 consider only the change in type of motion as a meaningful clip. Thus, clips
 382 with slight variations and a small number of frames are avoided.

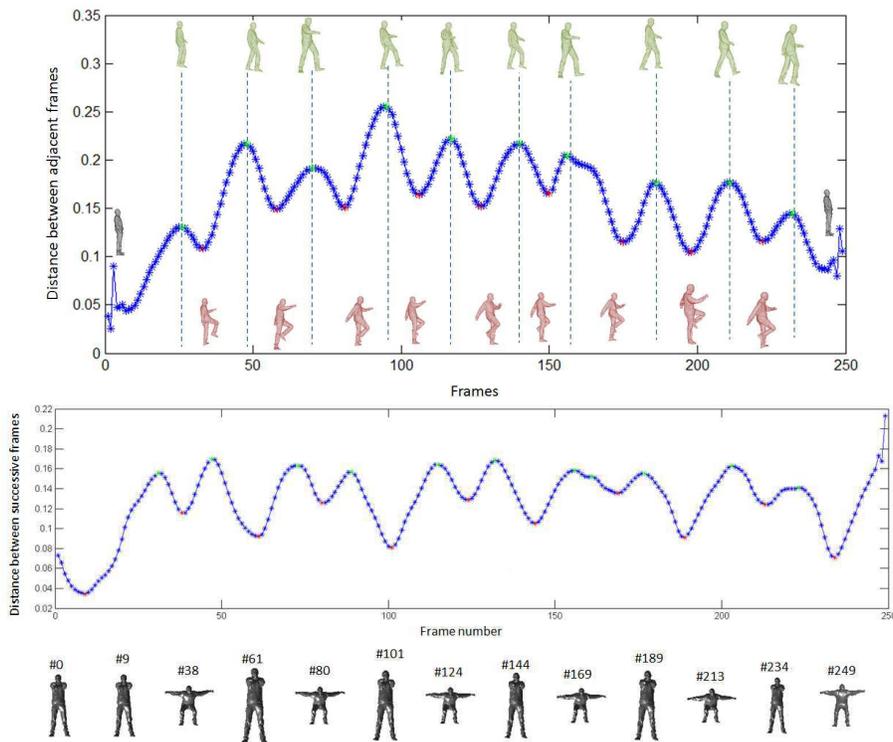


Figure 6: Segmentation of a 3D sequence into motion clips. Feature vector and detected frames as local extrema are presented at the top of the figure and detected frames as minima are at the bottom.

383 Note that optimum local minimum, that detect precise break points where
 384 the motion changes, is selected in a predefined neighbourhood. For this
 385 raison, we fix a size of window to test the efficiency of the local minimum in

386 this condition. To calculate the speed variation, distance between each two
387 successive EHC in the sequence is computed. The variations of the sequence
388 are represented in vector of speed and a further smoothing filter is applied
389 to obtain the final degree of motion vector.

390 5.2. *Clip matching*

391 To seek for similar clips, we need to encode gestures in a specific repre-
392 sentation that we can compare regardless to certain variations. In fact, two
393 motions are considered similar even if there are changes in the shape of the
394 actor and the speed of the action execution. This problem is similar to time-
395 series retrieval where a distance metric is used to look for, in a database, the
396 sequences whose distance to the query is below a threshold value. Each clip
397 is represented as a temporal sequence of human poses, characterized by EHC
398 representation associated to shape model. Then, extremal curves are tracked
399 in each sequence to characterize a trajectory of each curve in the shape space
400 as illustrated in Figure 7 (top). Finally, the trajectories of each curve are
401 matched and a similarity score is obtained. However, due to the variations
402 in execution rates of the same clip, two trajectories do not necessarily have
403 the same length. Therefore, a temporal alignment of these trajectories is
404 crucial before computing the global similarity measure, as shown in Figure 7
405 (bottom).

406 In order to solve the temporal variation problem, we use DTW algorithm
407 (Giorgino et al. [45]). This algorithm is used to find optimal non-linear
408 warping function to match a given time-series with another one, while ad-
409 hering to certain restrictions such as the monotonicity of the warping in the
410 time domain. The optimization process is usually performed using dynamic

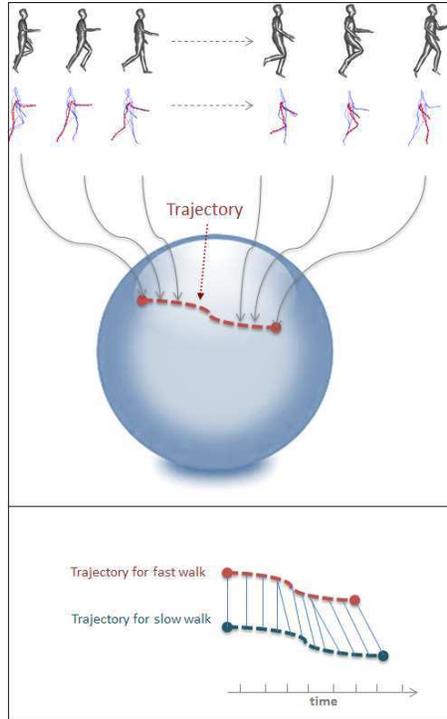


Figure 7: Graphical illustration of a sequence of shapes obtained during a walking action (top). Alignment process between trajectories of same curve index using DTW (bottom).

411 programming approaches given a measure of similarity between the features
 412 of the two sequences at different time instants. The global accumulated costs
 413 along the path define a global distance between the query clip and the motion
 414 segments found in the database. Since DTW can operate with any measure
 415 of similarity between different temporal features, we adapt it to features
 416 that reside on Riemannian manifolds. Hence, we use the geodesic distance
 417 between different shape points $d_s(q_i, q_j)$ as a distance function between the
 418 shape features at different time instants.

419 In practice, the first step is to follow independently curve variation in
420 time resulting on N trajectories in the shape space. In fact, each frame
421 in the 3D video sequence can be represented by a predetermined number
422 (N) of extremal curves, splitting the sequence into N parts, where each one
423 represents the trajectory of an open curve in the shape space. Then, DTW
424 will be applied in the feature space for each tracked curve index. The distance
425 between two clips is then the average distance given by each comparison
426 between corresponding trajectories.

427 5.3. Average clip

428 Based on the two algorithms, Karcher Mean and DTW, we can extend
429 the notion of “mean” of a set of human poses to the “mean” of trajectories
430 of poses in order to compute an ”average” of several clips.

431 Let N be the number of clips represented by N trajectories $T_1, T_2 \dots T_N$.
432 For a specific human curve index, we look for the mean trajectory that has
433 the minimum distance to the all N trajectories. As shown in Algorithm 1,
434 the mean trajectory is given by computing the non-linear warping functions
435 and setting iteratively the template as the Karcher Mean of the N warped
436 trajectories represented in the Riemannian shape space.

437 6. Video summarization and retrieval

438 In order to represent compactly a video sequence, we need to know how
439 to exploit the redundancy of information over time. However, when this in-
440 formation should be extracted from motion and not from frames separately,
441 the challenge is then about complex matching processes required to find ge-
442 ometric relations between consecutive data stream elements. We therefore

Algorithm 1 Computing trajectory template

Require: N trajectories from N clips $T_1, T_2 \dots T_N$

Initialization: chose randomly one of the N input trajectories as an initial guess of the mean trajectory T_{mean}

repeat

for $i=1 : N$ **do**

 find optimal path p^* using DTW to warp T_i to T_{mean}

end for

 Update T_{mean} as the Karcher Mean of all N warped trajectories

until Convergence

443 propose to use EHC to represent a pose and a trajectory as key descriptors
444 characterizing geometric data stream. Based on EHC representation, we de-
445 velop several processing modules as clustering, summarization and retrieval.

446 *6.1. Data clustering*

447 Let V denotes a video stream of human sequence containing elements
448 $\{e_i\}_{i=1\dots k}$, where e can be a frame or a clip. To cluster V , the data set is
449 recursively split into subsets C_t and R_t as described in the following recursive
450 algorithm:

Algorithm 2 Data clustering

Require: $V\{e_i\}_{i=1\dots k}$;

Ensure: $C_0 = \emptyset$; $R_0 = \{e_1, \dots, e_k\}$;

if $(R_t \neq \emptyset) \&\& (t \leq k)$ **then**

$C_t = \{f \in R_{t-1} : dist(e_t, f) < Th\}$;

$R_t = R_{t-1} \setminus C_t$;

end if

451 The result of clustering is contained in $C_{t=1\dots k}$ where C_t is a subset of V

452 representing a cluster containing similar elements to e_t . For each iteration of
 453 clustering steps, from $t = 1 \dots K$, the closest matches to e_t are retrieved and
 454 indexed with the same cluster reference as e_t . Any visited element e_t already
 455 assigned to a cluster in C during iteration step is considered as already clas-
 456 sified and is not processed subsequently. We regroup not empty sub sets C_t
 457 in l clusters $\{c_1, \dots, c_l\}$ (with $l \leq k$). Similarities between elements of V are
 458 evaluated using a similarity distance $dist$ allowing to compare the elements
 459 of V . The threshold Th is defined experimentally .

460 If we consider the video V as a long stream of 3D meshes, the clusters that
 461 should be obtained must gather models with similar poses. In this case, the
 462 EHC feature vector is used as an abstraction for every mesh and the similarity
 463 distance is the elastic metric computed between each pair of human poses.
 464 Motion can be incorporated in this similarity by applying a simple time
 465 filter on static similarity measure with a window size chosen experimentally
 466 [46]. The use of temporal filter integrates consecutive frames in a fixed time
 467 window, thus allowing the detection of individual poses while taking into
 468 account smooth transitions. Note that the pose invariance property of the
 469 EHC allows us to compare poses (and motions) of subjects regardless of
 470 translation, rotation and scaling .

471 The video V can also be considered as a stream of clips resulting from
 472 the video segmentation approach and clusters here gathers clips with similar
 473 repeated atomic actions. In this case: (1) the feature vector used as abstrac-
 474 tion for each clip is a trajectory on shape space of extremal human curves;
 475 and (2) the similarity distance, used to compare clips, is based on the DTW
 476 algorithm.

477 *6.2. Content-based summarization*

478 Our approach for video summarization is based on three steps: First, the
479 whole video is segmented and clustered into several clusters of clips. Second,
480 only the most significant clip (the nearest one to all cluster elements) of each
481 cluster is kept. Third, we construct a subsequence, from the starting video,
482 where this representative clips of each cluster are concatenated. Finally,
483 This new subsequence, is clustered into clusters of poses, and only most
484 representative poses are kept to describe the dataset.

485 This summarization allows a reduction of dimension for the original dataset
486 where we can display only main clips if we stop on third step, or to display
487 key frames if we continue summarization process until pose clustering.

488 *6.3. Pose and motion Retrieval*

489 As in a classical retrieval procedure, in response to a given query, an
490 ordered list of responses that the algorithm found nearest to the query is
491 given. Then to evaluate the algorithm, this ranked list is analyzed. Whatever
492 the given query, pose or clip, the crucial point in the retrieval system is the
493 notion of "similarity" employed to compare different objects.

494 For content-based pose retrieval, thanks to the static shape similarity, we
495 are able to compare human poses using their extremal human curve descrip-
496 tors and decide if two poses are similar or not. In this scenario, the query
497 consists of a 3D human shape model in a given pose and the response is 3D
498 human bodies more similar in pose to the query. We advocate the usage of
499 the EHC to represent the 3D human shape model in a given pose and then
500 comparison between each pair of models using the elastic metric defined in

501 the Proposition 4.1. This system can find a number of utilities like pose-
502 based searching and facilitate retrieval of efficient information as subjects in
503 same poses in the database of 3D models scanned in different poses [47, 48].

504 Note also that identifying frames with similar shape and pose can be used
505 potentially for concatenative human motion synthesis. Concatenate existing
506 3D video sequences allows the construction of a novel character animation. A
507 good descriptor that much correctly correspondent frames allows the synthe-
508 sis of videos with smooth transitions and finding best frames to summarize
509 the video. However, extension of static shape descriptor to include tem-
510 poral motion information is required to remove the ambiguities inherent in
511 static shape descriptor for comparing 3D video sequences of similar shape.
512 Therefore, the static shape descriptor can be extended to the time domain
513 by applying a simple time filter with a window size like $2N_t + 1$. This time
514 filter is a way of incorporating motion in the similarity measure, as so-called
515 temporal similarity, also used by Huang et al. [12]. The temporal similarity
516 is presented in the following Equation:

$$s_{ij}^t = \frac{1}{2N_t + 1} \sum_{k=-N_t}^{N_t} s(i + k, j + j) \quad (6)$$

517 where s is the frame-to-frame similarity matrix and N_t is a time filter with
518 window size $2N_t + 1$.

519 For content-based motion retrieval, we advocate the usage of the EHC
520 representation, where a query consists of a trajectories representing a clip on
521 the shape space. As response to this specific query, our approach looks in the
522 sequence for most similar trajectories and returns an ordered list of similar
523 ones using the process of motion clip explained in section 5.2.

524 7. Experimental results

525 To show the practical relevance of our method, we perform an experimen-
526 tal evaluation on several databases (summarized in Table 1) and compare it
527 to the most efficient descriptors of the state-of-the-art methods. We first eval-
528 uate our descriptor for shape similarity application over public static shape
529 database [48] and evaluate the results against Spherical Harmonic descriptor
530 [11]. Secondly, we measure the efficiency of our descriptor to capture the
531 shape similarity in 3D video sequences of different actors and motions from
532 other public 3D synthetic [12] and real [49, 50] video databases. We evaluate
533 this later against Temporal Shape Histogram [12], Multi-resolution Reeb-
534 graph [21] and other classic shape descriptors, using provided Ground Truth.
535 Motion segmentation into clips and clip matching performance are tested on
536 several video sequences of different people doing different motions. Finally,
537 we evaluate our clustering and summarization approach for pose/clip-based
538 video retrieval.

539 7.1. Feature matching

540 The extraction and comparison of our curves requires the identification of
541 feature end-points as head, right/left hand and right/left foot, which is not
542 affordable in practice. This requirement is important to perform the curve
543 matching separately between models. In order to overcome this problem,
544 our method is based on two benefits from the morphology of the human
545 body. First, we deduce that geodesic path connecting each one of the hand
546 end-points and the head end-point is shortest among all possible geodesics
547 between the five end-points. Second, the geodesic path connecting right hand

Dataset	Motions/Poses	Number of frames
Dataset (1) [48]: 144 subjects (59 men/55 women)	18 static poses (1 neutral done by all subjects and 17 other different poses)	\emptyset
Dataset (2) [12]: 14 people (10 men and 4 women)	28 motions: sneak, walk (slow, fast, turn left/right, circle left/right, cool, cowboy, elderly, tired, macho, march, mickey, sexy,dainty), run (slow, fast,turn right/left, circle left/right), sprint, vogue, faint, rockn’roll, shoot.	392 seq, 39200f (100f per seq.)
Dataset (3) [49]: 3 people (2 men and 1 woman)	6 motions: 2×cran, 2×marche, 2×squat, 1×handstand, 1×samba, 1×swing.	1582f (on average 226 ± 48 per seq.)
Dataset (4) [50]: Roxanne	Game character motion: walk	32 f

Table 1: Summarization of data used for all experimental tests.

548 to left foot end-points or left hand to right foot end-points is the longest. The
549 first observation allows to identify precisely the end-point corresponding to
550 the head, the two end-points connected to this later corresponding to the
551 hands without distinguishing between right and left. The second one allows
552 the identification of the couple of hand/foot as corresponding to same side of
553 the body without distinguishing between right and left. A prior knowledge
554 on the direction of the posture of the human body for static pose and in the
555 starting frame for video sequence has allowed to distinguish between left and

556 right. Once the end-points are correctly detected from the starting frame
557 in the video sequence, a simple algorithm of end-point tracking over time is
558 performed.

559 *7.2. Static shape similarity*

560 The protocol and the dataset used to validate the experiments are firstly
561 presented and then, the results following this protocol are analyzed and com-
562 pared to those obtained by other approaches.

563 *7.2.1. Evaluation methodology*

564 To assess the performance of the EHC for static shape similarity, sev-
565 eral experiments were performed on a statistical shape database [48]. This
566 database, summarized in Table 1 (1st row), is challenging for human body
567 shape and pose retrieval as it is realistic shape database captured with a
568 3D laser scanner. It contains more than hundred subjects doing more than
569 thirty different poses. We perform our descriptor on a subset of 338 shape
570 models obtained from 144 subjects 59 male and 55 female aged between 17
571 and 61 years. There are 18 consistent poses (p0, p1, p2, p3, p4, p5, p6, p7,
572 p8, p9, p10, p11, p12, p13, p16, p28, p29, p32). Some poses are illustrated
573 in Figure 8. Each pose represents a class where at least 4 different subjects
574 do the same pose.

575 For evaluation, we use Recall/Precision plot in addition to the three
576 statistics which indicate the percentage of the top K matches that belong to
577 the same pose class as the query pose:

- 578 • The nearest neighbor statistic (NN): it provides an indication to how
579 well a nearest neighbor classifier would perform (here $K = 1$).

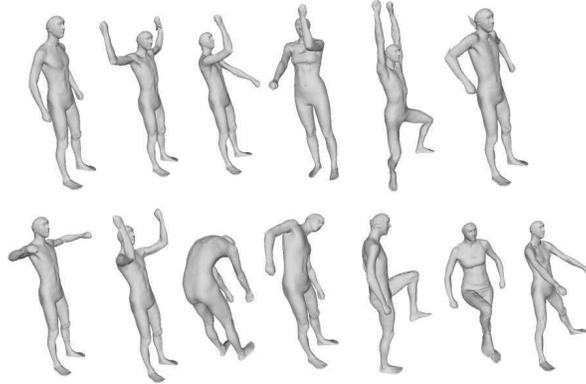


Figure 8: Example of body poses in the static human dataset [48].

- 580 • The first tier statistic (FT): it indicates the recall for the smallest K
581 that could possibly include 100% of the models in the query class.
- 582 • The second tier statistic (ST): it provides the same type of result, but
583 it is a little less stringent (i.e., K is twice as big).
- 584 • E-Measures: it is a composite measure of precision and recall for a fixed
585 number of retrieved results.

586 We note here that these statistics will be used for static and video retrieval
587 evaluations.

588 7.2.2. Curve selection

589 From five feature endpoints, we have extracted ten extremal curves rep-
590 resenting the human body shape model. According to the human poses,
591 extremal curves exhibit different performance and some curves are more effi-
592 cient to capture the shape similarity between two poses. Our shape descriptor

593 can be seen as a concatenation of ten curve representations and the similarity
 594 between two shape models doing two different poses, is represented by a vec-
 595 tor of ten elastic distance values. Before all tests, we analyze the performance
 596 of all possible combinations of curves on the shape similarity measurements.
 597 A Sequential Forward Selection method, applied on elastic distance values
 598 and coupled with ST statistic, has been used to select the best combination
 599 of curves among all possible ones (1013 combinations according to Eq. 7):

$$\sum_{k=2}^n C_n^k = \sum_{k=2}^n \frac{n!}{k!(n-k)!} \quad (7)$$

600 where n is equal to 10 and it represent the number of curves.

601 Experiment of pose-based retrieval on the dataset [48] shows that the best
 602 combination is obtained by the five curves: right hand to right foot, left hand
 603 to left foot, left hand to right hand, left foot to right foot, and head to the
 604 right foot (Figure 9).

605 The selected five curves seem to be the most stable ones and they are
 606 sufficient to represent at best the body like a skeleton on the surface. There-
 607 fore, the elimination of five curves allows to eliminate the ambiguity due to
 608 the redundancy of some curves on the body parts.

609 7.2.3. Result analysis

610 The self similarity matrix obtained from the mean elastic distance of the
 611 five selected curves is shown in the Figure 10.

612 This matrix demonstrates that similar poses have a small distance (cold
 613 color) and that this distance increases with the degree of the change be-
 614 tween poses (hot color). This allows pose classification or pose retrieval by

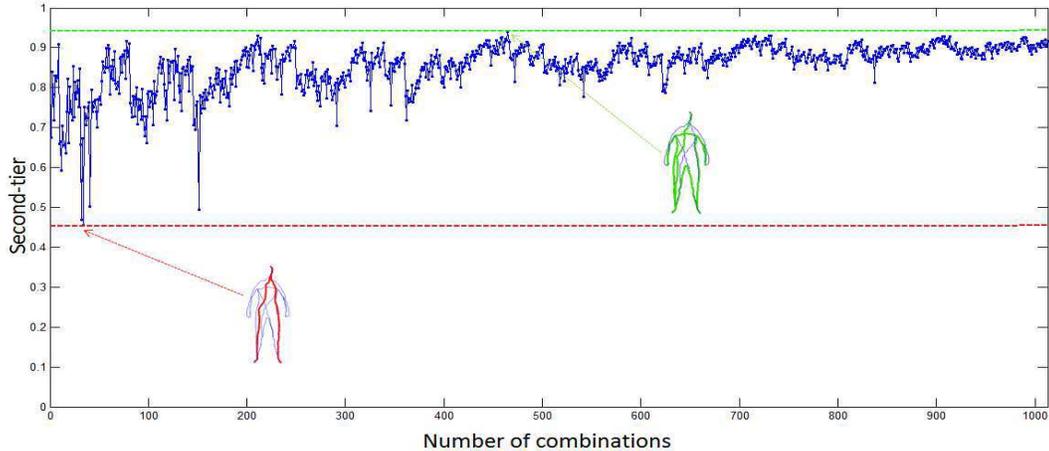


Figure 9: Second-Tier statistic for all combinations of curves. The best combination is obtained by 5 curves (green) and the worst combination is obtained by 2 curves (red).

615 comparing models using their extremal curve representation and the elastic
 616 metric.

617 From a quantitative point of view, we present the Recall/Precision plot
 618 obtained by EHC compared to the popular Spherical Harmonic (SH) descrip-
 619 tor with optimal parameter setting ($N_s = 32$ and $N_b = 16$) [8]. This plot and
 620 accuracy rates (NN, FT and ST) reported in Table 2 show that our approach
 621 provides better retrieval precision. EHC using only the five selected curves
 622 outperforms SH and EHC using the 10 curves to retrieve models with the
 623 same pose.

624 Note finally that the accuracies of retrieval ranks for some poses are rela-
 625 tively low. Such ambiguities can be noticed in the case of comparison between
 626 neutral pose and a pose where subjects just twist their body to the left, or

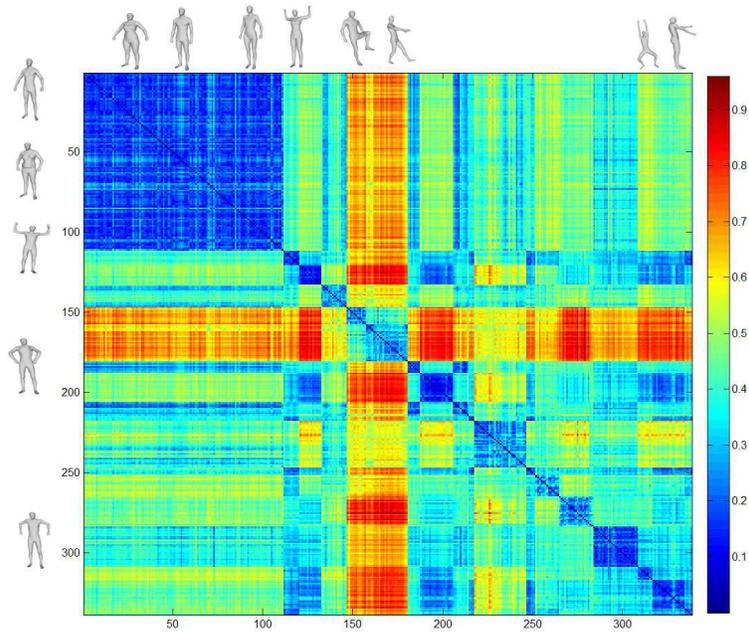


Figure 10: Confusion similarity matrix. The matrix contains pose dissimilarity computation between models of a 3D humans in different poses. More the color is cold more the two poses are similar.

Approach	NN(%)	FT(%)	ST(%)	E-Measure(%)
SH	71.0	57.9	75.5	41.3
EHC 10 curves	80.3	75.5	85.2	42.5
EHC 5 curves	84.8	77.2	89.1	43.0

Table 2: Retrieval statistics for pose based retrieval experiment

627 twist their torso to look around.

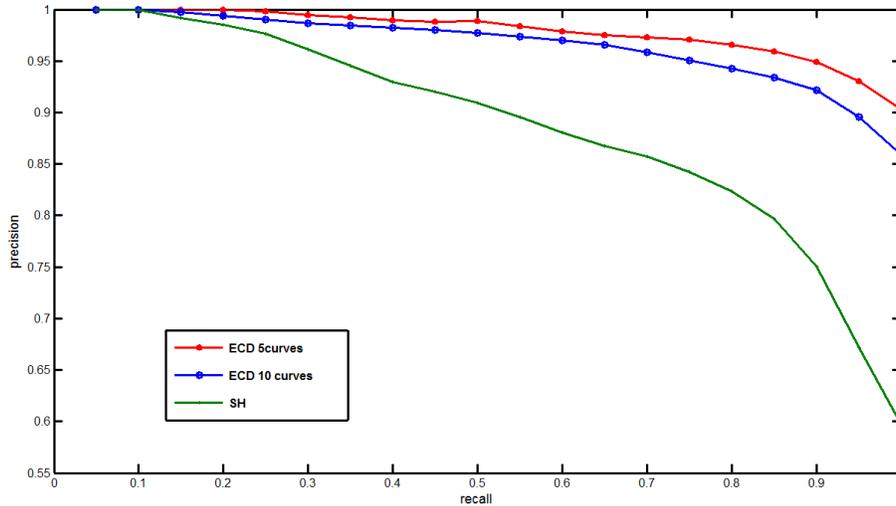


Figure 11: Precision-recall plot for pose-based retrieval.

628 *7.3. Temporal shape similarity for 3D video sequences*

629 We firstly present the protocol and the dataset used in these experiments
 630 and then, the results following this protocol are analyzed and compared to
 631 the most relevant state-of-the-art approaches.

632 *7.3.1. Evaluation methodology*

633 The recognition performance of the temporal shape descriptor is evaluated
 634 using a ground-truth dataset from a synthetic 3D video sequences proposed
 635 by Huang et al. [12] and a real captured 3D video sequences of people [49].
 636 As described in Table 1 (2nd row), the synthetic data is obtained by 14 people
 637 (10 men and 4 women) performing 28 motions. Each sequence is composed
 638 of 100 frames and the whole dataset contains a total of 39200 frames.

639 Given the known correspondences, a temporal ground-truth similarity is
 640 computed between each two surfaces. The known correspondence is only

641 used to compute this ground truth similarity. Having two Mesh X and Y
 642 with N vertices $x_i \in X$ and $y_i \in Y$, a temporal-ground truth C_T is computed
 643 by combining a shape similarity C_p and a temporal similarity C_v as follows:

$$\begin{aligned}
 C_T(X, Y) &= (1 - \alpha)C_p(x_i, y_j) + \alpha C_v(x_i, y_j) \\
 C_p(X, Y) &= \frac{1}{N} \sum_{k=1}^N d(x_i, y_j) \\
 C_v(X, Y) &= \frac{1}{N} \sum_{k=1}^N d(\dot{x}_i, \dot{y}_j)
 \end{aligned} \tag{8}$$

644 where d is an Euclidean distance, \dot{x}_i and \dot{y}_j are the derivation of x and
 645 y between next and current frame. the parameter α is used to balance the
 646 equation and it is set to 0.5 . In order to identify frames as similar or
 647 dissimilar, the temporal ground truth similarity matrix is binarized using a
 648 threshold set to 0.3 similarly to Huang et al. [12].

649 Finally, recognition performance is evaluated using the Receiver-Operator-
 650 Characteristic (ROC) curves, created by plotting the fraction of true-positive
 651 rate (TPR) against the fraction of false-positive rate (FPR), at various
 652 threshold settings. The true and false dissimilarity compare the predicted
 653 similarity between two frames, against the ground-truth similarity.

654 An example of self-similarity matrix computed using temporal ground-
 655 truth descriptor, static and temporal descriptors are shown in Figure 12.
 656 This figure illustrates also the effect of time filtering with increasing temporal
 657 window size for EHC descriptors on a periodic walking motion.

658 7.3.2. Result analysis

659 A comparison is made between our Temporal Extremal Human Curve
 660 (TEHC) and several descriptors from the state-of-the-art: Shape Distribu-
 661 tion (SD) , Spin Image (SI) , Spherical Harmonics Representation (SHR),

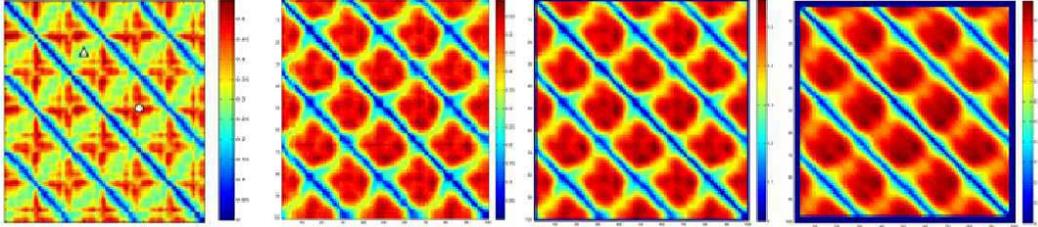


Figure 12: Similarity measure for "Fast Walk" motion in a straight line compared with itself. Coldest colors indicate most similar frames. 1st matrix: temporal Ground-Truth (TGT). 2nd, 3rd and 4th matrix: self-similarity matrix computed with Temporal EHC with window size 3, 5 and 7 respectively.

662 two Shape-flow descriptors, the global / local frame alignment Shape His-
 663 tograms (SHvrG / SHvrS) (Huang et al. [12]) and Reeb-Graph as skeleton
 664 based shape descriptors (aMRG) (Tung et al. [51]). Note that a spectral
 665 representation was also evaluated by Huang et al. [21] which is the Multi-
 666 Dimensional Scaling (MDS). Huang et al. [12] evaluated the performances
 667 of all these descriptors for the purpose of shape similarity.

668 The effectiveness of our descriptor have been evaluated by varying temporal
 669 window and comparing it to the most relevant state-of-the-art descriptors
 670 [12] as shown in the plot of ROC curves in Figure 13.

671 Several observations can be made on the obtained results: (i) Our descrip-
 672 tor outperforms classic shape descriptors (SI, SHR, SD) and shows compet-
 673 itive results with SHvrS and aMRG. We also notice that recognition perfor-
 674 mance of EHC increases with the increase of the window size of time-filter
 675 like any other descriptor. In fact, time-filter reduces the minima in the anti-
 676 diagonal direction, resulting from motion in the static descriptor (Figure 13).

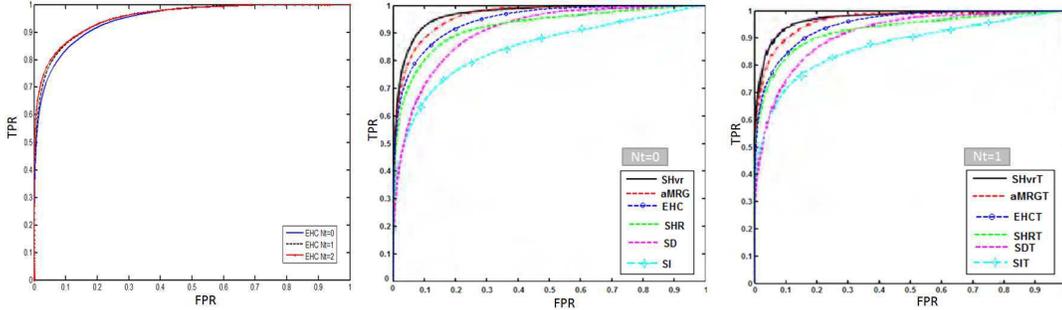


Figure 13: Evaluation of ROC curve for static and time-filtered descriptors on self-similarity across 14 people doing 28 motions. From right to left: ROC curves obtained by our TEHC descriptor with three different values of windows size N_t , ROC curve obtained by our EHC descriptor compared to different algorithms and ROC curves obtained with $N_t= 1$.

677 Multiframe shape-flow matching required in SHvrS allows the descriptor to
 678 be more robust but the computational cost will increase by the size of se-
 679 lected time window.

680 (ii) EHC descriptor by its simple representation, demonstrates a comparable
 681 recognition performance to aMRG. It is efficient as the curve extraction is
 682 instantaneous and robust as the curve representation is invariant to elastic
 683 and geometric changes thanks to the use of the elastic metric.

684 (iii) The result analysis for each motion shows that EHC gives a smooth
 685 rates that are stable and not affected by the complexity of the motion. Such
 686 complex motions are rockn’roll, vogue dance, faint, shot arm (Figure 14).
 687 However, this is not the case for SHvrS where performance recognition falls
 688 suddenly with complex motions as illustrated in Figure 15.

689

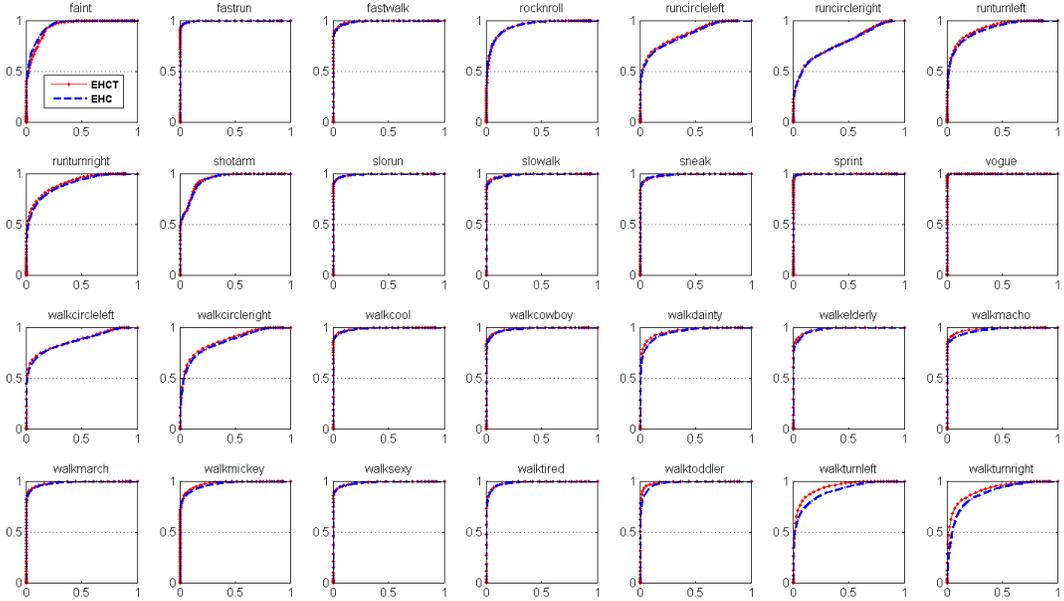


Figure 14: Evaluation of EHC descriptor against Temporal Ground Truth (TGT) for $N_t=0$ and $N_t=1$. ROC performance for 28 motions across 14 people.

690 We also applied the time filtering EHC descriptor on two real captured 3D
 691 video sequences of people. The first sequence is extracted from the dataset
 692 [49] described in Table 1 (3^{rd} row). The second one is extracted from real
 693 data reconstructed by multiple camera video [50] and described in Table 1
 694 (4^{th} row).

695 Inter-person similarity across two people in a walking motion with an
 696 example similarity curve are shown in Figure 16 (a). Our temporal similarity
 697 measure identifies correctly similar frames across different people. These
 698 similar frames are located in the minima of the similarity curve. In addition,

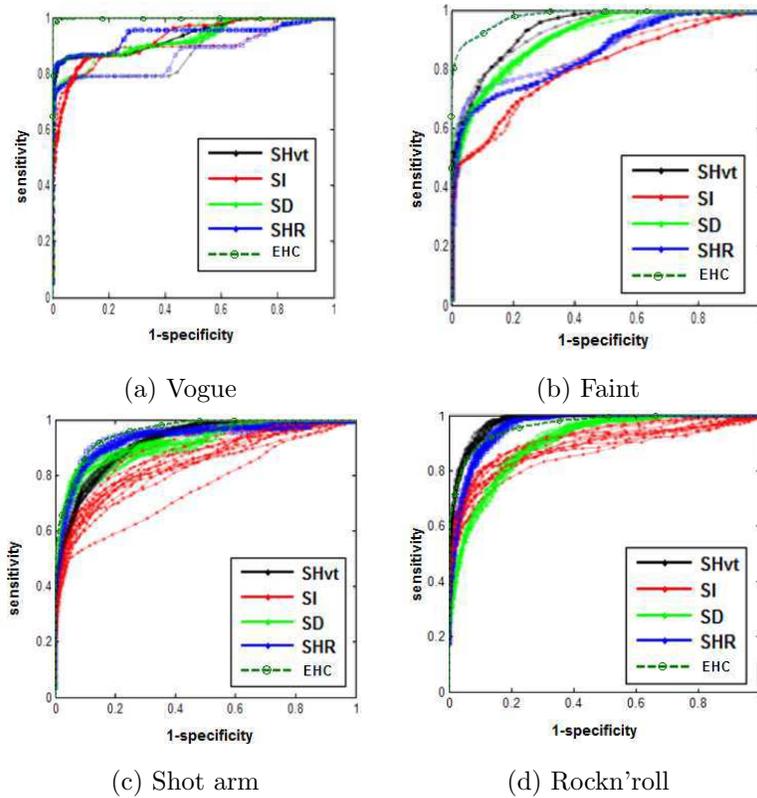


Figure 15: Evaluation of ROC curves for complex motions with $N_t=3$.

699 despite the topology change and the reconstruction noise as shown in Figure
 700 16 (b), our algorithm succeed to identify correctly the frame in the sequence
 701 similar to the query.

702 7.4. Motion segmentation and retrieval

703 In this section, we evaluate temporal shape similarity descriptor. Details
 704 about the computation of the ground truth descriptor are given in addition
 705 to the description of the different datasets used for evaluation. The results
 706 obtained by our approach, compared to those of different state-of-the-art

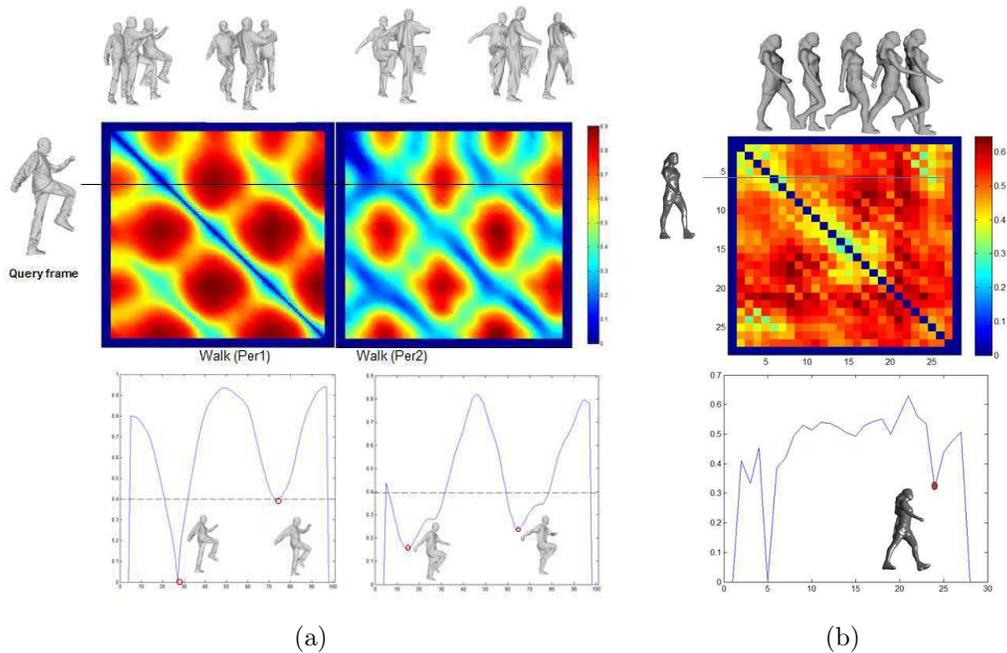


Figure 16: Inter-person similarity measure for real sequences. Similarity matrix, curve and example frames for (a) Walk motion across two actors [49] (b) walk motion for Roxanne [50] Game Character Walk .

707 descriptors, are then discussed.

708 7.4.1. Evaluation methodology

709 The two datasets (2) and (3) presented in Table 1 are used in these experi-
 710 ments. From the synthetic dataset [12], we have chosen 14 different motions:
 711 walk (slow, fast, circle left/right, cowboy, march, mickey), run (slow, fast,
 712 circle left/right), sprint, and rockn’roll. These motions are performed by two
 713 actors (a woman and a man) making a total of 28 motions (2800 frames).
 714 They are chosen for their interesting challenges as: (i) change in execution

715 rate (slow/fast motions) (ii) change in direction while moving (walking in
716 straight line, moving in circle and turning left and right) (iii) change in
717 shape (a woman and a man). We used these motion sequences for both
718 segmentation and retrieval experiment.

719 To validate the segmentation step, we segment all these 3D video se-
720 quences with the proposed approach and then compare results to manual
721 segmentation ground-truth. In the retrieval process, each query clip is com-
722 pared to all other clips obtained by the segmentation of sequences. Finally,
723 the statistics (NN, FT, ST and E-measure) are used for the evaluation.

724 *7.4.2. Analysis of motion segmentation result*

725 Plotting the distance between EHC representation of successive frames
726 gives a very noisy curve. The break points from this curve do not define
727 semantic clips and the extracting of minima leads to an over-segmentation of
728 the sequence (see Figure 17 (top)). To obtain more significant local minima,
729 we convolve the curve with a time-filter allowing to take into account the
730 motion variation, not only between two successive frames but also in a time
731 window. The motion degree after convolution is shown in Figure 17 (bottom).
732 Break points are more precise and delimits significant clips corresponding to
733 step change in the video sequence. In order to evaluate its efficiency, we
734 apply our segmentation method on the whole dataset (3) described in Table
735 1 (3rd row) and then compare the results to a manual segmentation of the
736 base done carefully .

737 We performed the clip segmentation for all window size values from 1
738 to 11 over a representative set of clips extracted from the dataset (3) [49].
739 Compared to manual ground truth, the best segmentation is obtained using

740 a window size of 5. This value is then fixed for the rest of the tests. The
 741 segmentation of the dataset(3) gives 83 segmented clips (78 correct clips and
 742 5 incorrect clips). This can be explained by the fact that the 5 failing clips
 743 are short. They contain about 6 frames at most and do not describe atomic
 744 significant actions. Otherwise, the a total of 144 clips have been obtained by
 745 the segmentation of the 14 motions taken from the dataset (2) described in
 746 Table 1 (2nd raw) performed by two actors.

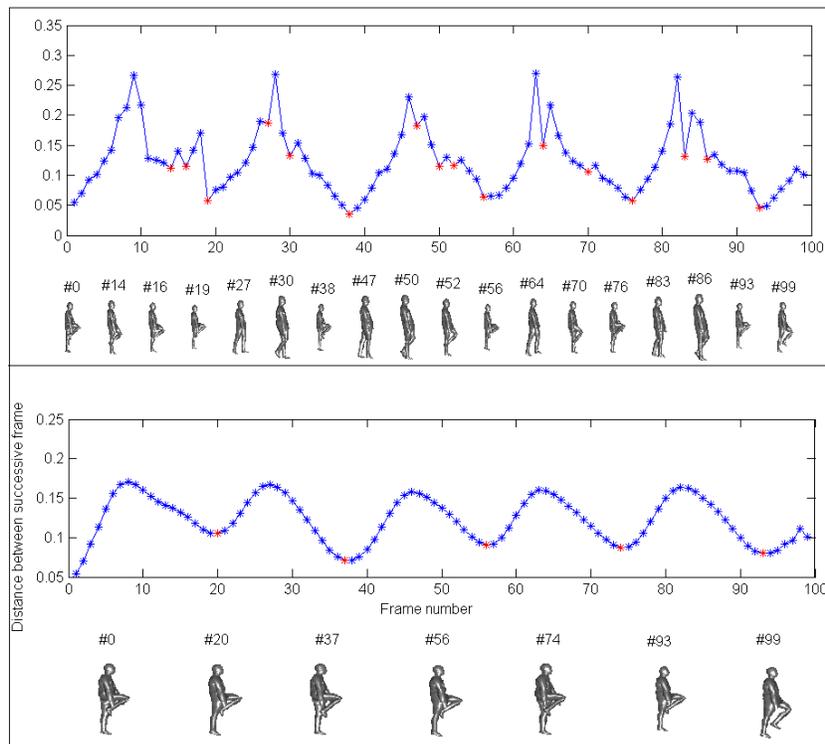


Figure 17: Speed curve smoothing.

747 Figure 18 shows some results of motion segmentation on a "slow walk"
 748 and a "fast walk" motions. Although the walk speed increases, the motion

749 segmentation remains significant and does not change and corresponds to the
 750 step change of the actor. The Rockn'roll dance motion segmentation is also
 751 illustrated in Figure 18 (bottom). Thanks to the selection of local minima
 752 in a precise neighborhood, only significant break points are detected.

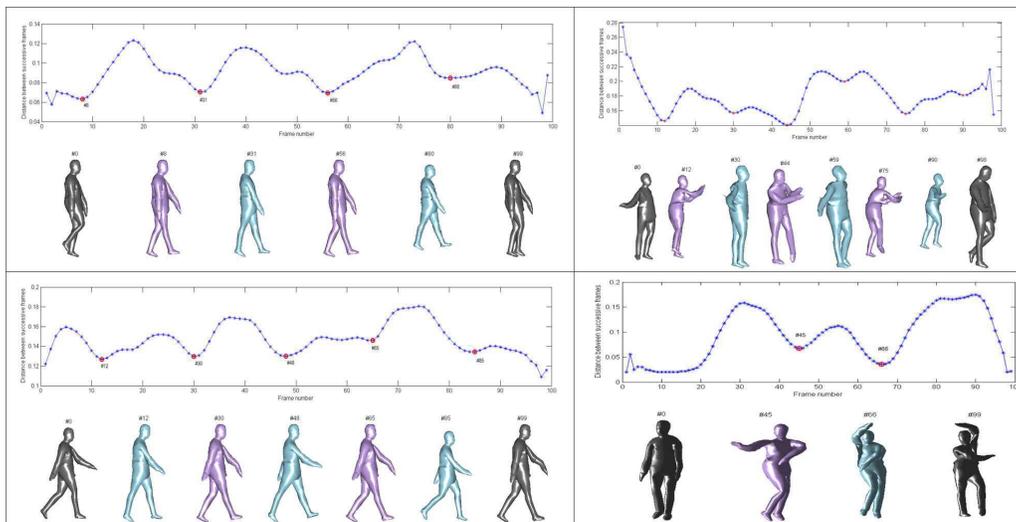


Figure 18: Various motion segmentation. From right top to left bottom, motions are: slow walk, Rockn'roll dance, fast walk, vogue dance.

753 7.4.3. Analysis of motion retrieval result

754 The motion segmentation method, applied on 14 motion sequences from
 755 the dataset (2) and performed by a man and a woman, gives a total of
 756 144 clips. These clips, with an average number of frames per clip equal to
 757 15, are categorized into 14 classes. The motion sequences consist mainly
 758 of different styles of walking, running and some dancing sequences. Classes
 759 grouped together represent different styles of walking, running and dancing
 760 steps. For example, a step change in a walk may represent a class and groups

761 similar clips done with different speed and in different trajectories. We notice
762 that right to left change step is grouped in a different class than left to light
763 change step.

764 The similarity metric represented by elastic measure values between each
765 pair of clips allows us to generate a confusion matrix for all classes of clips,
766 in order to evaluate the recognition performance by computing dynamic re-
767 trieval measures thanks to a manually annotated ground truth. An example
768 of the matrix representing the similarity evaluation score among clips in
769 sequences performed by a female actress against the clips of sequences of
770 motions performed by a male actor is shown in Figure 19. More the color is
771 cold more the clips are similar.

772 Thanks to the use of DTW, it is noticed that similarity score between
773 same clips done in different speeds is small (see Figure 19). The matching
774 between the clip representing change in step in a slow walk motion composed
775 of 25 frames and a fast walk motion, composed of 18 frames, is small.

776 Besides, our approach succeed to retrieve clips within motions done in
777 different ways. For example, the walk circle clips can be matched with the
778 clips of slow walk motion done in a straight line (see Figure 19). This explains
779 why the use of an elastic metric, to compare and match trajectories, makes
780 the process independent to rotation. Although the actors performing the
781 motions are different, it is observed that similar clips yield smaller similarity
782 score. Like it is shown in "Rockn'roll" dance motion, steps of the dance
783 performed by different actors are correctly retrieved.

784 It is demonstrated that 79.26% of similar motion clips are included in
785 the first tier and 93% of clips are correctly retrieved in the second tier. It

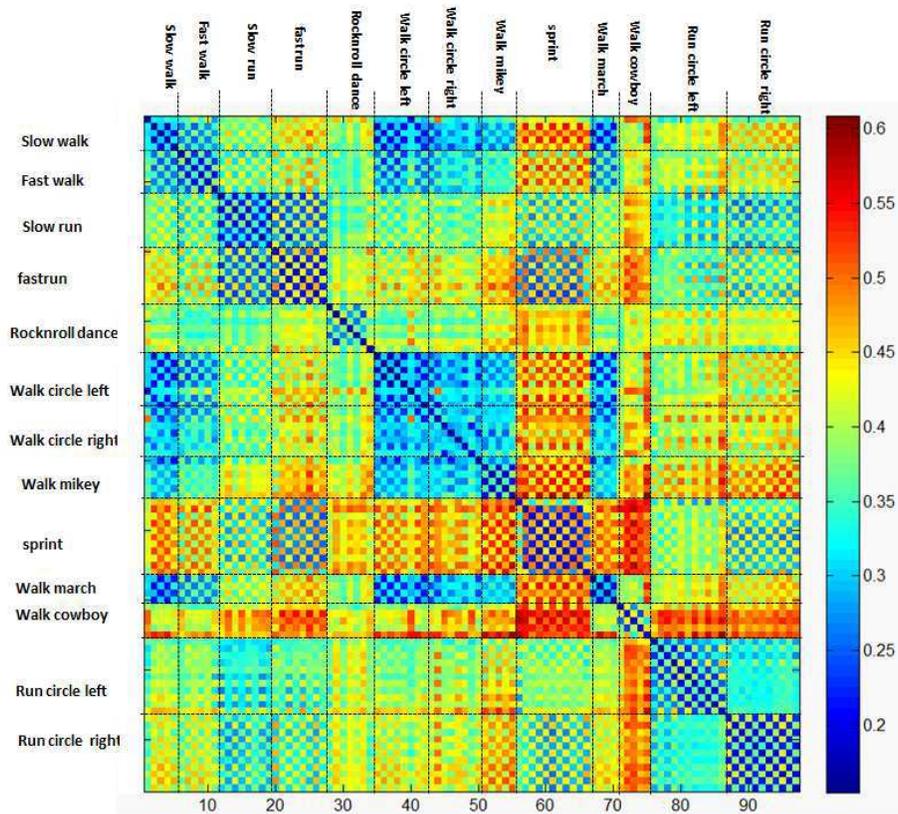


Figure 19: Similarity matrix evaluation between clips. More the color is cold more the two clips are similar.

786 is a rather good performance considering that only such low-level feature as
 787 the EHC is utilized in the matching. This can be explained by the fact that
 788 geodesics are not completely invariant to the topology changes. Thereby, the
 789 extracted sequential curves that represent the trajectory tend to change the
 790 path on the models for certain motions and therefore mislead the matching
 791 performed by DTW.

792 We also apply our retrieval approach to a real captured 3D video sequence

793 from the real dataset (3) described in Table 1 (3rd row). Self similarity
 794 example with an actor in a walking motion (walking in circular way) and its
 795 similarity curve are shown in Figure 20. For the query clip presented at the
 796 right of the figure, retrieved clips are found correctly in the sequence when
 797 the actor is turning.

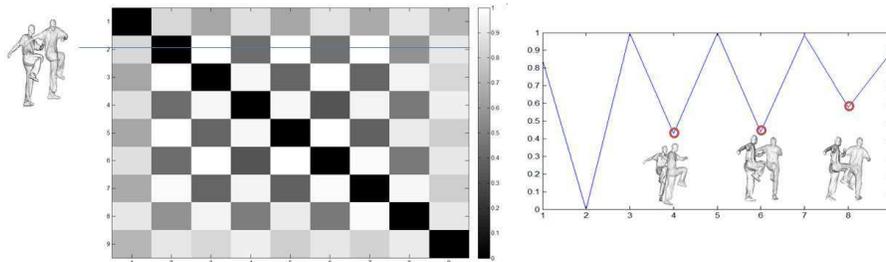


Figure 20: Experimental results for 3D video retrieval using motion of "walk in circle".

798 *7.5. Data summarization and content-based retrieval*

799 In this section, we firstly conducted multiple experimental trials by ana-
 800 lyzing the video clustering method on two aspects: the pose-based clustering
 801 and the clip-based clustering. Secondly, we evaluate the impact of the sum-
 802 marization process on the retrieval system by comparing the results with and
 803 without using clustering.

804 *7.5.1. Content-based summarization*

805 The performance of the content-based summarization approach is eval-
 806 uated for pose and clip data. To validate the pose-based summarization,
 807 we use a composed long sequence of a subject performing walk and squat

808 motions from the dataset (3). For clip-based summarization experiment, the
 809 same 28 motions used for video segmentation and the retrieval have been
 810 used.

811 The effectiveness of clustering process is evaluated by the number of clus-
 812 ters found which should allow the identification of eventual redundant pat-
 813 terns. The threshold Th in the Algorithm 2 is set accordingly to the values of
 814 the similarity function. The distances computed between descriptors (EHC
 815 for pose and trajectory of EHC for clip) are normalized to return values in
 816 the range $[0, 1]$, and Th was then defined experimentally. An optimal set-
 817 ting of Th should return a set of clusters similar to what a "hand-made"
 818 ground-truth classification would perform. The Figure 21 shows the cluster-
 819 ing result obtained from the composed long sequence. The number of clusters
 820 decreases with the increase of the threshold Th . We obtain the best result
 821 for $Th = 0.5$ with 51 clusters partitioned as the bar diagram shown in the
 right of the Figure 21.

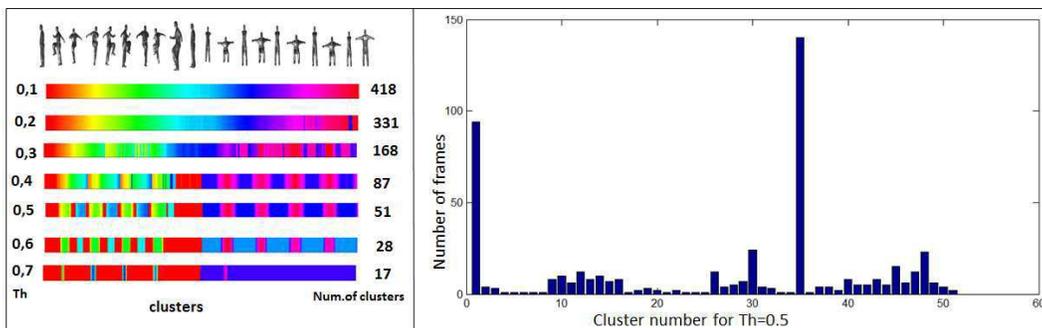


Figure 21: Frame clustering process with respect to threshold Th .

822

823 Pose-based clustering process can be improved by increasing the window

824 size of the time filter as shown in Figure 22.

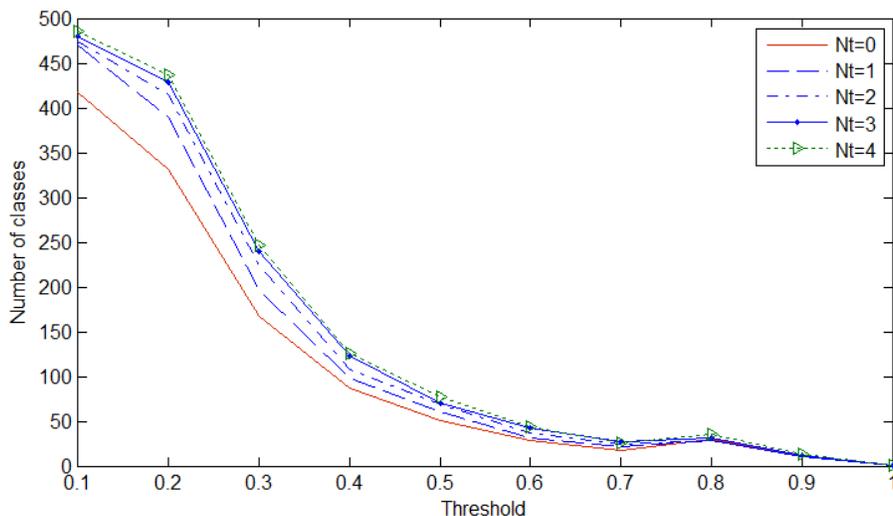


Figure 22: Frame clustering with respect to a threshold and with different window size.

825 We notice from this figure that for a $Th = 0.2$, the number of clusters
826 varies from 330 to 440 and a good compromise is obtained for $Nt = 3$.

827 Furthermore, clustering is applied on 14 motions extracted from the
828 dataset (3) and performed by two actors (a man and a woman) in order
829 to evaluate the efficiency of the clip-based clustering. By decreasing the
830 threshold Th of the clustering algorithm, we obtain more clusters. Experi-
831 mentally, we set Th to 0.43 and obtain 23 clusters from initially 110 clips for
832 the first actor and 26 clusters for the second one (see Figure 23). We notice
833 that clips representing sprint or running steps are clustered together.

834 The video summarization process can be used efficiently in hierarchical
835 structure, starting by video segmentation into clips, followed by clip-based

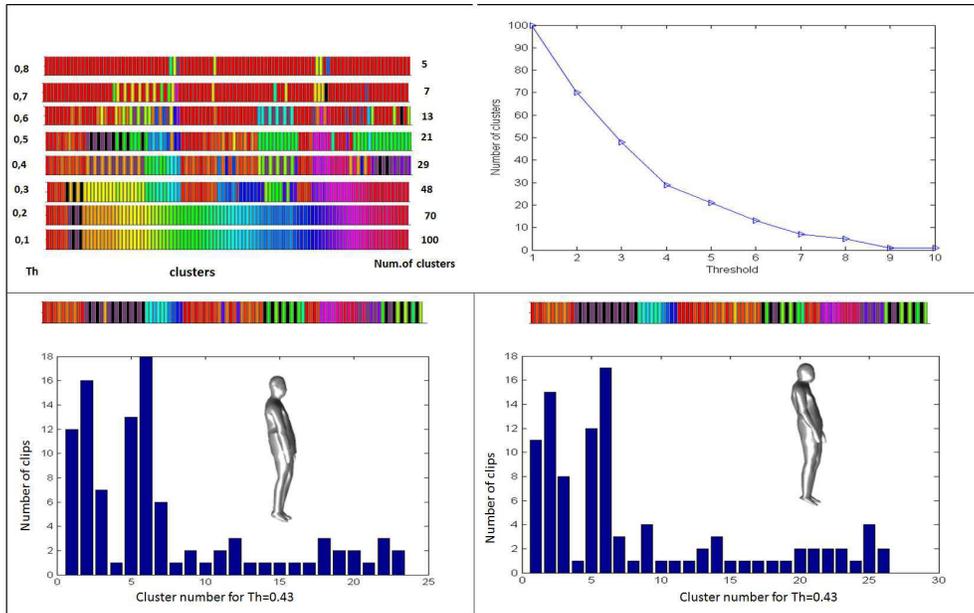


Figure 23: Clustering clips from a sequence of two actors performing 14 motions from the dataset (3) for a total of 1400 frames, with respect to Th . In second row, the variation of clip number in each cluster is presented.

836 clustering and then a pose-based clustering performed on the frames of all
 837 represented clusters of the clips resulting from the last step. The effectiveness
 838 of our summarization process is shown in Figure 24 for the sequence of a real
 839 actor performing walking and squatting motion. From 500 frames segmented
 840 into 18 clips, the clustering process gives 6 clusters. The new subsequence
 841 containing 6 clips (most representative clip in each cluster) and 180 frames
 842 is then clustered into 41 clusters where each one represent a class of pose.

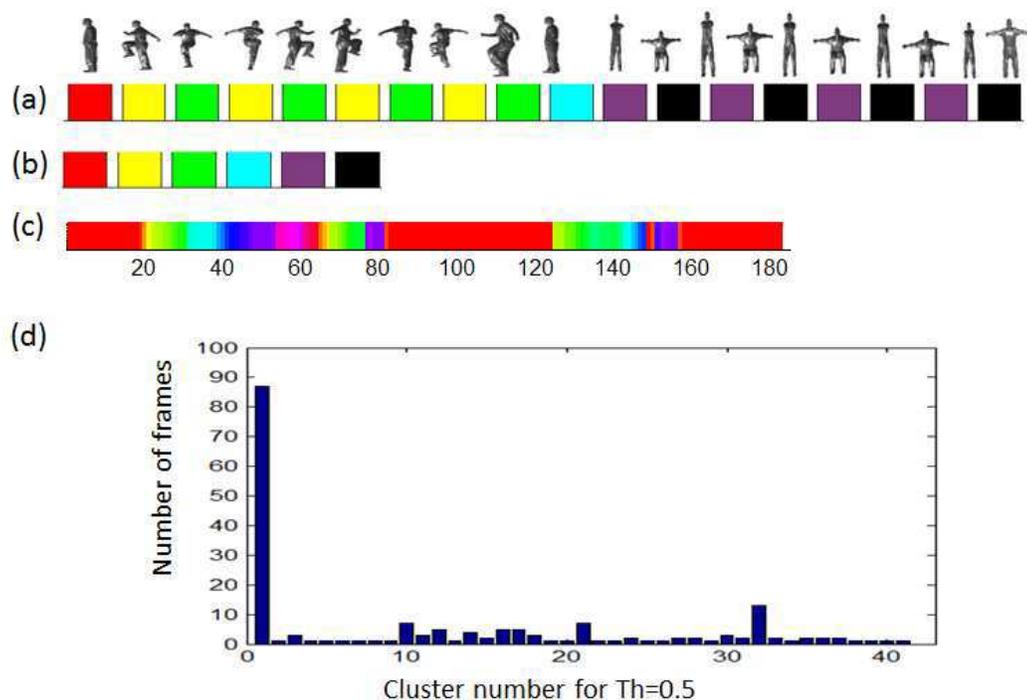


Figure 24: Summarization process: (a) for a sequence of 500 frames segmented into 18 clips, the clustering process returns 6 clusters of clips using $Th = 0.38$ (b) subsequence of clustered clips (180 frames) where each cluster is represented by only one clip chosen as the Karcher Mean clip of the cluster, (c) clustering of subsequence into 41 clusters of frames using $Th = 0.5$, (d) distribution of the number of frames in clusters.

843 7.5.2. Hierarchical data retrieval

844 For a mesh model of 1 MB size, the size of the 3D video sequence grows
 845 linearly of 1 MB per frame. Hence, the video retrieval becomes very difficult
 846 in long sequences. Within our framework, we propose to combine the data
 847 clustering approach with the content-based retrieval in order to perform an

848 hierarchical retrieval.

849 The clustering approach gathers models with similar poses/clips in clus-
850 ters. If we consider the element of cluster as a pose, clusters are firstly
851 performed over the entire sequence in order to gather frames with similar
852 poses and then a template model is obtained for each cluster by computing
853 its Karcher Mean as described in section 4.3. The retrieval system can then
854 be described as an hierarchical structure composed of two levels, the first one
855 containing templates and the second one containing all models of the dataset.
856 In view of this structure, a natural way is to start at the top, compare the
857 query with the template of each cluster and proceed down the branch that
858 leads to the closest shape.

859 We reconsider the same experiments for pose based retrieval in section 7.2
860 by applying the hierarchical approach to the dataset summarized in Table 1
861 (1st row) . Each query model is compared to each one of the template models
862 representing the clusters. The elastic measure values are used to generate a
863 confusion matrix for all classes of pose, in order to evaluate the recognition
864 performance by computing statistic retrieval measures thanks to the provided
865 ground truth. The matrix of comparison in the first level (model-template
866 comparison), is shown in Figure 25.

867 If we compare this matrix to that already obtained for the same dataset
868 without the use of summarization (Figure 10), you can easily notice the
869 effectiveness of the summarization. The main advantage of this approach is
870 the reduction of computation time which complexity pass from n to $\log(n)$
871 while keeping relevant information. Retrieval performances obtained from
872 this matrix for FT, ST and E-Measure are respectively 84.5% , 88.2% and

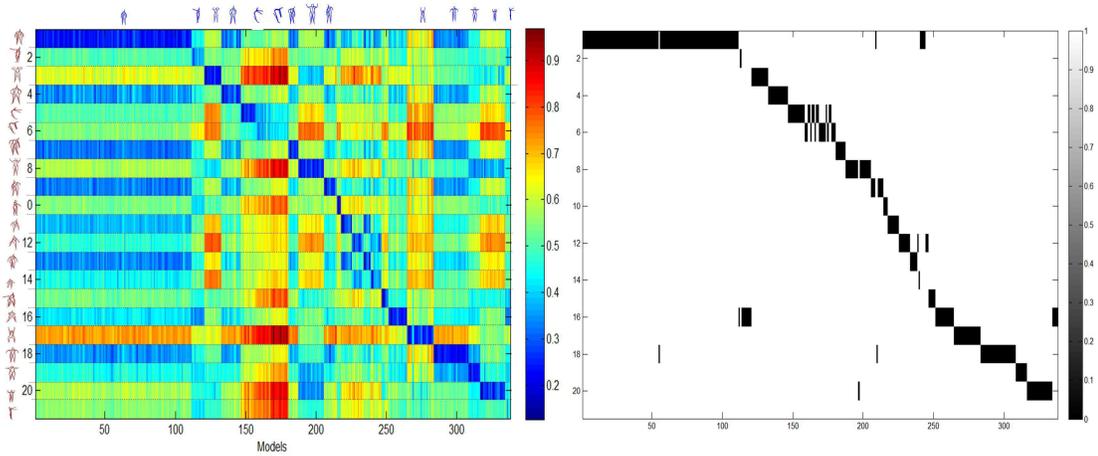


Figure 25: Similarity matrix and its binarization for template pose of each class against all models in the dataset.

873 43.6%. Comparing these results to those in Table 2, a small improvement is
 874 achieved for classic retrieval scenario in term of second tier.

875 In term of pose classification, the obtained accuracy is about 90.24%.
 876 Models of the class #2 are the most ones affected by misclassification and are
 877 assigned to the class #16. Looking at these two classes, we perceive that their
 878 poses are close to each other, both represent people with hands outstretched.
 879 The only difference is that one does with open legs and the other with closed
 880 ones.

881

882 Finally, if we consider the element of cluster as a clip, where a video
 883 segmentation is firstly performed on the whole sequence. In this case, the
 884 template model is a "mean" clip obtained for each cluster of clips by com-
 885 puting its Karcher Mean (see Algorithm 1). The retrieval system can then

886 be viewed as above with hierarchical structure. As experimental test, we
 887 performed a similar experimentation on the 14 motions performed by two
 888 actors as already evaluated in the section 7.3. In this experimentation, each
 889 query is a clip compared to each one of the template models representing
 890 the clusters of clips. The similarity measure values obtained by DTW algo-
 891 rithm between clips are used to generate a confusion matrix for all classes of
 892 clips, in order to evaluate the recognition performance by computing statis-
 893 tic retrieval measures thanks to the provided ground truth. The matrix of
 894 comparison in the first level (model-template comparison) is shown in Figure
 895 26.

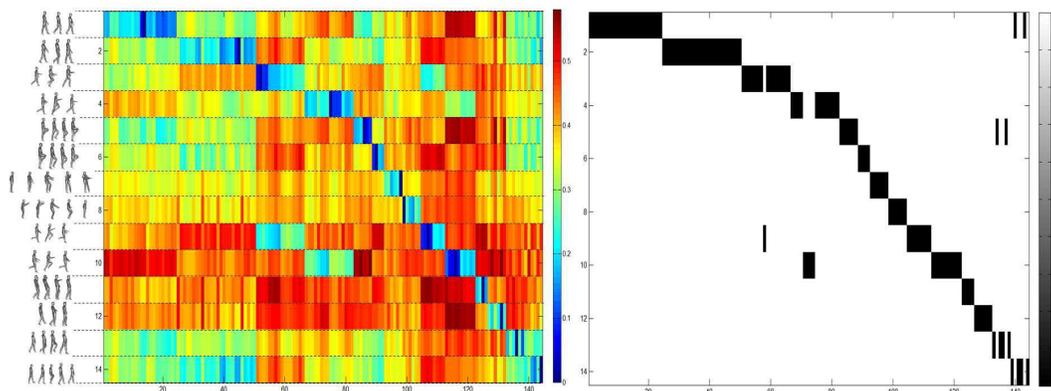


Figure 26: Similarity matrix and its binarization for template clip of each class against all clips in the dataset.

896 Retrieval performances obtained from this matrix for FT, ST and E-
 897 Measure are respectively 84.09%, 95.83% and 55.26%. In term of clip classi-
 898 fication, obtained accuracy is about 93.75%. The analysis of the result given
 899 by the binarized matrix shows that the most misclassified clips are those of

900 "fast run" class. In fact, they are assigned to class template representing
901 "sprint" motion class.

902 **8. Discussion**

903 The advantages of using EHC to represent human pose and motion in our
904 approach include: (1) invariance to affine transformation (2) possibility to
905 compute mean poses and mean clips (3) the use of well defined measure for
906 pose comparison in Reimannian manifold and (4) the use of well established
907 algorithm of DTW to align sequences taking benefits from the temporal
908 aspect of curves.

909 However, this representation has some limitations. Firstly, EHC depends
910 on the accuracy of extremities (head and limbs) extraction and on the defini-
911 tion of the path connecting end-points. In fact, the extraction of end-points
912 and extremal curves is based on the definition of geodesic distance between
913 each pair of curves. Thus, geodesic distances play an important role in our
914 geometric representation of the human body shape. However, they are sensi-
915 tive to significant topology changes as shown in Figure 27. In this figure, only
916 4 extremities are successfully detected and the left hand extremity is missed.
917 Thus, information about position of this hand is lost. In the future, other
918 strategies will be investigated for the extremities extraction step and shortest
919 path detection on the mesh by using diffusion or commute time distances as
920 presented by Elkhoury et al. [15] and Sun et al. [52].

921 Secondly, we note that our curve extraction can be sensitive to loose
922 clothes. For example, the mesh represented in Figure 27 shows a girl wearing
923 a skirt and the shape of the curve connecting her feet is different from the

924 same curve extracted on her mesh if she were wearing a trouser. This problem
925 will be even more critical if she wears a long skirt.

926 Thirdly, a prior knowledge on the direction of the posture of the human
927 body for the starting frame in video sequence is used to distinguish between
928 left/right hand and foot. Other feature matching algorithms, like Heat Kernel
929 Signature as proposed by Sun et al. [52] and Zheng et al. [53], could be used
930 in future work to correctly identify the right from the left side.

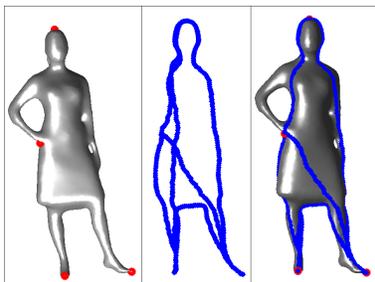


Figure 27: Example of failed extraction of EHC in presence of a topological change.

931 **9. Conclusion**

932 In this work, we have proposed a unified framework able to represent
933 human body shape with a pose descriptor, as well as a sequence of frames
934 with a specific representation. This framework relies on an Extremal Human
935 Curve descriptor (EHC), based on extremal features and geodesics between
936 each pair of them. This descriptor has the advantage of being a skeletal rep-
937 resentation, which is trackable over time. It is also an extremal descriptor
938 of the surface deformation which is composed by a collection of local open-

939 3D-curves. The representation of these curves and the comparison between
940 them are performed in the Riemannian shape space of open curves. By this
941 way, we have chosen to represent the pose of a mesh regardless to its rota-
942 tion, translation and scale. Convolved with a time filter to incorporate the
943 motion, it becomes a temporal descriptor for pose retrieval. The degree of
944 motion using feature vector, extracted from this descriptor, is then used for
945 splitting continuous sequence into elementary motion segments called clips.
946 Each clip describing an atomic movement is characterized by EHC repre-
947 sentation associated to human mesh. The open curves in 3D space, which
948 are the elements of EHC representation, are viewed as a point in the shape
949 space of open curves and hence each clip is represented by a trajectory on
950 this space. Dynamic time warping is used to align different trajectories and
951 give a similarity score between each two clips.

952 The quality of our descriptor regarding the recognition performance of
953 shape similarity in 3D video is analyzed and verified also by comparison
954 with other related recent techniques. Moreover, our approach achieves a
955 performance accuracy of 93.44% for video retrieval as second tier, which is
956 encouraging. Finally, we will investigate 3D human action recognition and
957 semantic activity analysis based on this framework.

958 **10. Acknowledgements**

959 This research program is supported partially by the Region Nord-Pas de
960 Calais (France).

961 **References**

- 962 [1] K. M. Cheung, S. Baker, T. Kanade, Shape-from-silhouette across time
963 part i: Theory and algorithms, in: International Journal of Computer
964 Vision, Vol. 62, 2005, pp. 221 – 247.
- 965 [2] E. De Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, S. Thrun,
966 Performance capture from sparse multi-view video, in: ACM SIG-
967 GRAPH, Vol. 27, ACM, 2008, pp. 1–10.
- 968 [3] T. Kanade, P. Rander, P. J. Narayanan, Virtualized reality: Construct-
969 ing virtual worlds from real scenes, in: IEEE MultiMedia, Vol. 4, 1997,
970 pp. 34–47.
- 971 [4] T. Tung, S. Nobuhara, T. Matsuyama, Complete multi-view reconstruc-
972 tion of dynamic scenes from probabilistic fusion of narrow and wide
973 baseline stereo, in: IEEE 12th International Conference on Computer
974 Vision (ICCV), 2009, pp. 1709 –1716.
- 975 [5] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, R. Szeliski, A compar-
976 ison and evaluation of multi-view stereo reconstruction algorithms, in:
977 IEEE International Conference on Computer Vision and Pattern Recog-
978 nition, Vol. 1 of CVPR '06, IEEE Computer Society, Washington, DC,
979 USA, 2006, pp. 519–528.
- 980 [6] A. E. Johnson, M. Hebert, Using spin images for efficient object recogni-
981 tion in cluttered 3d scenes, in: IEEE Transactions on Pattern Analysis
982 and Machine Intelligence, 1999, pp. 433–449.

- 983 [7] R. Osada, T. Funkhouser, B. Chazelle, D. Dobkin, Shape distributions,
984 in: ACM Transactions on Graphics, Vol. 21, 2002, pp. 807–832.
- 985 [8] M. Ankerst, G. Kastenmüller, H.-P. Kriegel, T. Seidl, 3d shape his-
986 tograms for similarity search and classification in spatial databases, in:
987 Proceedings of the 6th International Symposium on Advances in Spatial
988 Databases, SSD '99, 1999, pp. 207–226.
- 989 [9] P. Huang, A. Hilton, Shape-colour histograms for matching 3d video se-
990 quences, in: IEEE International Conference on Computer Vision Work-
991 shops (ICCV Workshops), 2009, pp. 1510 –1517.
- 992 [10] M. Körtgen, G. J. Park, M. Novotni, R. Klein, 3d shape matching with
993 3d shape contexts, in: The 7th Central European Seminar on Computer
994 Graphics, 2003.
- 995 [11] M. Kazhdan, T. Funkhouser, S. Rusinkiewicz, Rotation invariant spher-
996 ical harmonic representation of 3d shape descriptors, in: Proceedings
997 of the Eurographics/ACM SIGGRAPH symposium on Geometry pro-
998 cessing, SGP '03, Eurographics Association, Aire-la-Ville, Switzerland,
999 Switzerland, 2003, pp. 156–164.
- 1000 [12] P. Huang, A. Hilton, J. Starck, Shape similarity for 3d video sequences
1001 of people, in: International Journal of Computer Vision, Vol. 89, Kluwer
1002 Academic Publishers, Hingham, MA, USA, 2010, pp. 362–381.
- 1003 [13] T. Tung, T. Matsuyama, Topology dictionary for 3d video understand-
1004 ing, in: IEEE Transactions on Pattern Analysis and Machine Intelli-

- 1005 gence, Vol. 34, IEEE Computer Society, Los Alamitos, CA, USA, 2012,
1006 pp. 1645–1657.
- 1007 [14] H. Tabia, M. Daoudi, J.-P. Vandeborre, O. Colot, A new 3d-matching
1008 method of nonrigid and partially similar models using curve analysis, in:
1009 IEEE Trans. Pattern Anal. Mach. Intell., Vol. 33, 2011, pp. 852–858.
- 1010 [15] R. El Khoury, J.-P. Vandeborre, M. Daoudi, Indexed heat curves for
1011 3d-model retrieval, in: 21st International Conference on Pattern Recog-
1012 nition (ICPR 2012), Tsukuba Science City, Japon, 2012, pp. 1964–1967.
- 1013 [16] S. Mahmoudi, M. Daoudi, A probabilistic approach for 3d shape retrieval
1014 by characteristic views, in: Pattern Recognition Letters, Vol. 28, 2007,
1015 pp. 1705 – 1718.
- 1016 [17] J. Sun, M. Ovsjanikov, L. Guibas, A concise and provably informative
1017 multi-scale signature based on heat diffusion, in: Proceedings of the
1018 Symposium on Geometry Processing, SGP '09, 2009, pp. 1383–1392.
- 1019 [18] Y. Lipman, T. Funkhouser, Mobius voting for surface correspondence,
1020 in: ACM SIGGRAPH 2009 Papers, Vol. 28 of SIGGRAPH '09, 2009,
1021 pp. 1–12.
- 1022 [19] M. Ovsjanikov, Q. Mérigot, F. Mémoli, L. J. Guibas, One point isometric
1023 matching with the heat kernel, in: Comput. Graph. Forum, Vol. 29,
1024 2010, pp. 1555–1564.
- 1025 [20] J. Lee, J. Chai, P. S. A. Reitsma, J. K. Hodgins, N. S. Pollard, Interac-
1026 tive control of avatars animated with human motion data, in: Proceed-

- 1027 ings of the 29th annual conference on Computer graphics and interactive
1028 techniques, ACM, New York, NY, USA, 2002, pp. 491–500.
- 1029 [21] P. Huang, T. Tung, H. Nobuhara, S. and Hilton, T. Matsuyama, Com-
1030 parison of skeleton and non-skeleton shape descriptors for 3d video, in:
1031 Proceedings of the Fifth International Symposium on 3D Data Process-
1032 ing, Visualization and Transmission (3DPVT'10), Pairs, France, 2010.
- 1033 [22] T. Yamasaki, K. Aizawa, Motion segmentation and retrieval for 3d video
1034 based on modified shape distribution, in: Journal on Applied Signal
1035 Processing EURASIP, Vol. 2007, Hindawi Publishing Corp., New York,
1036 NY, United States, 2007, pp. 211–211.
- 1037 [23] J. Wang, H. Zheng, View-robust action recognition based on temporal
1038 self-similarities and dynamic time warping, in: Computer Science and
1039 Automation Engineering (CSAE), 2012 IEEE International Conference
1040 on, Vol. 2, 2012, pp. 498–502.
- 1041 [24] L. Wang, L. Cheng, L. Wang, Elastic sequence correlation for human
1042 action analysis, in: Image Processing, IEEE Transactions on, Vol. 20,
1043 2011, pp. 1725–1738.
- 1044 [25] Y.-S. Tak, J. Kim, E. Hwang, Hierarchical querying scheme of human
1045 motions for smart home environment, in: Engineering Applications of
1046 Artificial Intelligence, Vol. 25, Pergamon Press, Inc., Tarrytown, NY,
1047 USA, 2012, pp. 1301–1312.
- 1048 [26] M. Mortara, G. Patane, Affine-invariant skeleton of 3d shapes, in: Pro-

- 1049 ceedings of the Shape Modeling International (SMI'02), SMI '02, IEEE
1050 Computer Society, Washington, DC, USA, 2002, pp. 245 – 252.
- 1051 [27] S. Katz, G. Leifman, A. Tal, Mesh segmentation using feature point
1052 and core extraction, in: *The Visual Computer*, Vol. 21, Springer-Verlag,
1053 2005, pp. 649–658.
- 1054 [28] J. Tierny, J.-P. Vandeborre, M. Daoudi, Invariant high level reeb graphs
1055 of 3d polygonal meshes, in: *International Symposium on 3D Data Pro-
1056 cessing, Visualization, and Transmission (3DPVT)*, IEEE Computer So-
1057 ciety, Los Alamitos, CA, USA, 2006, pp. 105–112.
- 1058 [29] F. Lazarus, A. Verroust, Level set diagrams of polyhedral objects, in:
1059 *Proceedings of the fifth ACM symposium on Solid modeling and appli-
1060 cations, SMA '99*, New York, NY, USA, 1999, pp. 130–140.
- 1061 [30] M. F. Abdelkader, W. Abd-Almageed, A. Srivastava, R. Chellappa,
1062 Silhouette-based gesture and action recognition via modeling trajecto-
1063 ries on riemannian shape manifolds, in: *Journal of Computer Vision and
1064 Image Understanding*, Vol. 115, Elsevier Science Inc., New York, NY,
1065 USA, 2011, pp. 439–455.
- 1066 [31] H. Drira, B. Ben Amor, A. Srivastava, M. Daoudi, R. Slama, 3d face
1067 recognition under expressions, occlusions, and pose variations, in: *IEEE
1068 Trans. Pattern Anal. Mach. Intell.*, Vol. 35, 2013, pp. 2270–2283.
- 1069 [32] S. Joshi, E. Klassen, A. Srivastava, I. Jermyn, A novel representation
1070 for riemannian analysis of elastic curves in \mathbb{R}^n , in: *IEEE Conference on
1071 Computer Vision and Pattern Recognition (CVPR '07)*, 2007, pp. 1 –7.

- 1072 [33] P. W. Michor, D. Mumford, Riemannian geometries on spaces of plane
1073 curves, in: J. Eur. Math. Soc., Vol. 8, 2006, pp. 1–48.
- 1074 [34] E. Klassen, A. Srivastava, W. Mio, S. Joshi, Analysis of planar shapes
1075 using geodesic paths on shape spaces, in: IEEE Pattern Anal. Mach.
1076 Intell., Vol. 26, 2004, pp. 372–383.
- 1077 [35] A. Yezzi, A. Mennucci, Conformal metrics and true "gradient flows"
1078 for curves, in: Proceedings of the Tenth IEEE International Conference
1079 on Computer Vision, ICCV, IEEE Computer Society, Washington, DC,
1080 USA, 2005, pp. 913–919.
- 1081 [36] A. Srivastava, E. Klassen, S. Joshi, I. Jermyn, Shape analysis of elastic
1082 curves in euclidean spaces, in: IEEE Transactions on Pattern Analysis
1083 and Machine Intelligence, Vol. 33, 2011, pp. 1415–1428.
- 1084 [37] I. Koprinska, S. Carrato, Temporal video segmentation: A survey, in:
1085 Signal Processing: Image Communication, 2001, pp. 477–500.
- 1086 [38] Y. Rui, P. Anandan, Segmenting visual actions based on spatio-temporal
1087 motion patterns, in: Proceedings. IEEE Conference on Computer Vision
1088 and Pattern Recognition, Vol. 1, 2000, pp. 111 –118 vol.1.
- 1089 [39] T.-S. Wang, H.-Y. Shum, Y.-Q. Xu, N.-N. Zheng, Unsupervised analy-
1090 sis of human gestures, in: Proceedings of the Second IEEE Pacific Rim
1091 Conference on Multimedia: Advances in Multimedia Information Pro-
1092 cessing, PCM '01, Springer-Verlag, London, UK, UK, 2001, pp. 174–181.
- 1093 [40] T. Shiratori, A. Nakazawa, K. Ikeuchi, Rhythmic motion analysis us-
1094 ing motion capture and musical information, in: Proceedings of IEEE

- 1095 International Conference on Multisensor Fusion and Integration for In-
1096 telligent Systems, 2003, pp. 89 – 94.
- 1097 [41] K. Kahol, P. Tripathi, S. Panchanathan, Automated gesture segmenta-
1098 tion from dance sequences, in: IEEE International Conference on Auto-
1099 matic Face and Gesture Recognition, 2004, pp. 883 – 888.
- 1100 [42] J. Xu, T. Yamasaki, K. Aizawa, 3d video segmentation using point dis-
1101 tance histograms, in: IEEE International Conference on Image Process-
1102 ing (ICIP), Vol. 1, 2005, pp. 701–704.
- 1103 [43] T. Yamasaki, K. Aizawa, Motion segmentation of 3d video using modi-
1104 fied shape distribution, in: IEEE International Conference on Multime-
1105 dia and Expo, 2006, pp. 1909 –1912.
- 1106 [44] P. Huang, A. Hilton, J. Starck, Automatic 3d video summarization: Key
1107 frame extraction from self-similarity, in: Fourth International Sympo-
1108 sium on 3D Data Processing, Visualization and Transmission, 2008, pp.
1109 1 – 8.
- 1110 [45] T. Giorgino, Computing and visualizing dynamic time warping align-
1111 ments in r: The dtw package, in: Journal of Statistical Softwar, Vol. 31,
1112 2009, p. 1–24.
- 1113 [46] R. Slama, H. Wannous, M. Daoudi, Extremal human curves: a new
1114 human body shape and pose descriptor, in: Accepted - 10th IEEE In-
1115 ternational Conference on Automatic Face and Gesture Recognition,
1116 Shanghai, China, 2013, pp. 1–6.

- 1117 [47] B. Allen, B. Curless, Z. Popović, The space of human body shapes:
1118 reconstruction and parameterization from range scans, in: ACM SIG-
1119 GRAPH, Vol. 22, ACM, New York, NY, USA, 2003, pp. 587–594.
- 1120 [48] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, H.-P. Seidel, A statistical
1121 model of human pose and body shape, in: P. Dutr’e, M. Stamminger
1122 (Eds.), Computer Graphics Forum (Proc. Eurographics 2008), Vol. 2,
1123 Munich, Germany: SMI:2002:Skeletonany, 2009, pp. 337–346.
- 1124 [49] D. Vlastic, I. Baran, W. Matusik, J. Popović, Articulated mesh animation
1125 from multi-view silhouettes, in: ACM Siggraph, ACM, New York, NY,
1126 USA, 2008, pp. 97:1–97:9.
- 1127 [50] J. Starck, A. Hilton, Surface capture for performance-based animation,
1128 in: Computer Graphics and Applications, Vol. 27, 2007, pp. 21–31.
- 1129 [51] T. Tung, F. Schmitt, The augmented multiresolution reeb graph ap-
1130 proach for content-based retrieval of 3d shapes, in: International Journal
1131 of Shape Modeling, Vol. 11, 2005, pp. 91–120.
- 1132 [52] J. Sun, M. Ovsjanikov, L. Guibas, A concise and provably informative
1133 multi-scale signature based on heat diffusion, in: Proceedings of the
1134 Symposium on Geometry Processing, SGP ’09, Eurographics Associa-
1135 tion, Aire-la-Ville, Switzerland, Switzerland, 2009, pp. 1383–1392.
- 1136 [53] Y. Zheng, C.-L. Tai, E. Zhang, P. Xu, Pairwise harmonics for shape
1137 analysis, Vol. 19, IEEE Computer Society, Los Alamitos, CA, USA,
1138 2013, pp. 1172–1184.