HAL
archives-ouvertes.fr

3D Human Video Retrieval: from Pose to Motion Matching

Rim Slama, Hazem Wannous, Mohamed Daoudi

▶ **To cite this version:**

Rim Slama, Hazem Wannous, Mohamed Daoudi. 3D Human Video Retrieval: from Pose to Motion Matching. Eurographics Workshop on 3D Object Retrieval, May 2013, Girona, Spain. hal-00829222

HAL Id: hal-00829222
https://hal.archives-ouvertes.fr/hal-00829222

Submitted on 2 Jun 2013

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 Video Retrieval: from Pose to Motion Matching

R. Slama[1,2], H. Wannous[1,2] and M. Daoudi[2,3]

[1]University of Lille 1, France
[2] LIFL (UMR 8022 Lille 1/CNRS)
[3]Institut Mines-Telecom/Telecom Lille1, France

Abstract

3D video retrieval is a challenging problem lying at the heart of many primary research areas in computer graphics and computer vision applications. In this paper, we present a new 3D human shape matching and motion retrieval framework. Our approach is formulated using Extremal Human Curve (EHC) descriptor extracted from the body surface and a local motion retrieval achieved after motion segmentation. Matching is performed by an efficient method which takes advantage of a compact EHC representation in open curve Shape Space and an elastic distance measure. Moreover, local 3D video retrieval is performed by dynamic time warping (DTW) algorithm in the feature space vectors. Experiments on both synthetic and real 3D human video sequences show that our approach provides an accurate shape similarity in video compared to the best state-of-the-art approaches. Finally, results on motion retrieval are promising and show the potential of this approach.

Categories and Subject Descriptors (according to ACM CCS): H.3.1 [Information storage and retrieval]: Content Analysis and Indexing—I.3.5 [Computer graphics]: Computational Geometry and Object Modeling —

1. Introduction

While human analysis in 2D image and video has received 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 3D static analysis of 3D human body, but still less on 3D human video analysis. Parallel to this, 3D video sequences of human motion is more and more available. In fact, their acquisition with a multiple view reconstruction systems or animation and synthesis approaches [CBK05] [dAST*08a] received a considerable interest over the past decade.

Most of the research topics on these 3D video recently focus mainly on performance, quality improvements and compression methods [TNM09] [dAST*08b]. Consequently, 3D videos are yet mainly only used for display. However, the acquisition of long sequences produces massive amounts of data which make the datasets difficult to handle: hence the need to develop efficient and effective segmentation and retrieval systems for managing the database and searching for relevant information quickly.

In this paper, we propose a novel descriptor for 3D human shape representation and 3D video matching. We then focus on the task of video segmentation and comparisons between motion segments for video retrieval, based on geodesic feature sets and elastic distance measure.

A 3D video of human motion, is considered to be composed of consecutive poses. As a first step of the retrieval pipeline, our geometric features called Extremal Human Curve (EHC) descriptor are extracted from body surface. Based on extremal features (4 end-effectors and head) and geodesics between each pair of them, our descriptor is invariant to rotation and scale. Every 3D frame will be represented by a collection of open curves whose comparison will be performed in a Riemannien Shape Space using an elastic metric.

For direct comparison of the video sequences, the motion segmentation can play an important role in the dynamic matching by splitting automatically the continuous 3D video data into meaningful segments that describe basic movements, called clips. Finally, we perform a dynamic time warping (DTW) between each pair of clips for every feature set using elastic distance measure. Based on this DTW-distances, we perform a ranking and obtain a ranked list of clips for each clip of the example dataset.

As key contributions in this paper:

- EHC: Use of this surface-based shape descriptor

[SWD13] to model 3D dynamic surface of human by a sequence of EHC, taking advantage of its structure which is invariant to isometric transformations.
- Characterization of the sequence as a collection of trajectories thanks to EHC representation in open curve Shape Space.
- Motion segmentation and retrieval using trajectories set and similarity metric using DTW in its feature space.

The outline of this paper is as follows. The next section discusses related works in the area of motion segmentation and retrieval. The extremal curves extraction and the elastic metric used for their comparison are presented in section III. In section IV, our framework used for motion segmentation and retrieval is presented. In section V, evaluation of our descriptor and experimental results for video segmentation and retrieval are performed. Finally, we conclude in section VI by summarizing our results and discussing issues for future works.

2. Related work

Few solutions related to 3D mesh video retrieval and shape similarity metrics have been found in the literature.

Some works have been addressed the problem of shape similarity for 3D video, and resolve the problem of video retrieval by matching frames and comparing correspondent ones using a specified metric. In [YA07], a modified shape distribution histogram has been employed as feature representation of 3D models. The similar motion retrieval is realized by Dynamic Programming matching using the feature vectors and Euclidian distance. In [KGH09], the problem of human action matching in outdoor sports broadcast environments is dressed. Shape histograms are constructed using a spherical coordinate system, and presenting the surface by an implicit function. Matching is achieved using Kullback Leibler divergence combined with a HMM.

The problem of 3D shape matching in temporal sequences, where the goal is to discriminate the same object in different poses, is addressed by [HHS10]. To do, classic shape histograms: shape distribution, spin image, shape histogram and spherical harmonics are used as static descriptors and extended to temporal ones using a time filter. A comparison of these shape descriptors combined with self-similarities has been made by Huang et al. [HHS10] and their experiments have showed that Shape Histogram gives the best performance for different people and motions. However, these similarity metrics evaluate only spatial shape descriptors and do not usually capture any geometrical information about the 3D human body pose and joint positions / orientations. This prevents its use in certain applications that require accurate estimation of the pose (and the joints in some cases) of the body parts. Tung et al. [TS05] proposed Multi-resolution Reeb-Graph as a skeleton-based-shape descriptor, where its evaluation has shown a competitive performance with spatial shape descriptors [HTN*10]. It was also used for video understanding as long as it is structured representation of the articulated structure [TM12]. However, in practice Reeb-Graph is sensitive to change in surface topology due to reconstruction error in real 3D video sequences.

Some other works have trends to accumulate static human shape or pose descriptors over time, or to capture the involvement of shape and pose changes in the sequence. Various representations of the body tracked in the time are used to deduce a motion vector in order to perform motion retrieval [HTTM11]. Such descriptors are: motion history volume (MHV), 3D optical flow, cylinder ellipsoid body model, skeletal and quadratic body model. More details about 3D video retrieval are recently presented in [DTP12].

In our approach, we propose to extend the use of EHC [SWD13] descriptor to model 3D video sequences of people in order to perform motion retrieval. For this purpose, a motion segmentation is performed on continuous sequence to split it into elementary action segments. These later present a human motion as a temporal sequence of poses, each characterized by EHC representation associated to human mesh. Elements of EHC representation are open curves in 3D space, which are viewed as point in shape space of open curves and hence each sequence will be represented by a trajectory on this shape space. Dynamic time warping is used to align different trajectories and it gives a similarity score between two local motions.

3. Human body shape and pose descriptor

We aim to present a body shape as a skeleton based shape representation. This skeleton will be extracted on the surface of the mesh by connecting extremal features located on the extremities of the body. The main idea behind the use of this representation is to analyze pose variation with elastic deformation of the body, using representative curves on the surface.

3.1. EHC descriptor

We chose to detect the body extremities as feature points resulting from the intersection of their two sets of local extrema, extracted by cross-analysis approach using geodesic based scalar functions defined over the body surface [TVD06]. Since it is based on geodesic distance evaluation, These extremities are stable and invariant to geometrical transformations and model pose (Figure 1). Now, let M be a body surface and $E = \{e_1, e_2, e_3, e_4, e_5\}$ a set of feature points on the body representing the output of feature points extraction. Let β denote the open curve on M which joints two feature points of M $\{e_i, e_j\}$. To obtain β, we seek for geodesic path P_{ij} between e_i and e_j. We repeat this step to extract extremal curves from the body surface ten times so that we do all possible paths between elements of E. As il-

lustrated on the right of Figure 1 the body is represented using these extremal curves $M \sim \bigcup \beta_{ij}$.

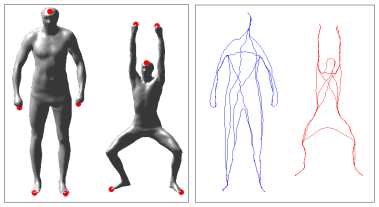


Figure 1: Feature points extracted from human body surface and correspondent extremal curves.

We have chosen to represent the body pose by a collection of curves for two reasons. Firstly, these curves connect limbs and give obviously a good representation of the body shape and pose, using a reduced representation of the mesh surface. Secondly, elastic analysis shapes of curves inside Shape Space is more efficient [JKSJ07]. However, to compare correspondent extremal curves we need a distance to evaluate how much the shape of the corresponding curves is similar. The distance we are going to use is called an elastic metric. It will be explained in more details in section 3.2.

3.2. Elastic distance

While human body is an elastic shape, its surface can be simply affected by a stretch (raising hand) or a bind (squatting). In order to analyze human curves independently to this elasticity, we need an elastic metric within a Shape Space framework.

Let $\beta : I \to \mathbb{R}^3$, for $I = [0,1]$, represents an extremal curve obtained as described above. To analyse the shape of β, we shall represent it mathematically using a *square-root velocity function* (SRVF), denoted by $q(t) \doteq \dot{\beta}(t)/\sqrt{\|\dot{\beta}(t)\|}$. $q(t)$ is a special function introduced by [JKSJ07] that captures the shape of β and is particularly convenient for shape analysis.

The set of all unit-length curves in $\mathbb{R}^3$ is given by $\mathcal{C} = \{q : I \to \mathbb{R}^3 | \|q\| = 1\} \subset \mathbb{L}^2(I,\mathbb{R}^3)$. With the $\mathbb{L}^2$-metric on its tangent spaces, $\mathcal{C}$ becomes a Riemannian manifold. Since the elements of $\mathcal{C}$ have a unit $\mathbb{L}^2$ norm, $\mathcal{C}$ is a hypersphere in the Hilbert space $\mathbb{L}^2(I,\mathbb{R}^3)$. In order to compare the shapes of two extremal curves, we can compute the distance between them in $\mathcal{C}$ under the chosen metric. This distance is found to be the length of the minor arc connecting the two elements in $\mathcal{C}$. The geodesic length between any two points $q_1, q_2 \in \mathcal{C}$ is given by:

$$d_c(q_1,q_2) = cos^{-1}(\langle q_1, q_2 \rangle) \,, \tag{1}$$

and the geodesic path $\alpha : [0,1] \to \mathcal{C}$, is given by:

$$\alpha(\tau) = \frac{1}{\sin(\theta)} (\sin((1-\tau)\theta)q_1 + \sin(\theta\tau)q_2) \,,$$

where $\theta = d_c(q_1,q_2)$. In order to handle the variability due to re parametrization or rotation, we define orbits of the rotation group $SO(3)$ and the re-parametrization group Γ as equivalence classes in $\mathcal{C}$. We define the equivalent class containing q as: $[q] = \{\sqrt{\dot{\gamma}(t)} O q(\gamma(t)) | O \in SO(3), \ \gamma \in \Gamma\}$. The set of such equivalence class is called the shape space $\mathcal{S}$ of elastic curves [JKSJ07]. Two extremal curves with different elasticity or orientation are viewed as the same point on $\mathcal{S}$. We denote by d_s the geodesic distance between the corresponding equivalence classes $[q_1]$ and $[q_2]$ in shape space $\mathcal{S}$, and we denote by $q_2^*(t) = \sqrt{\dot{\gamma^*}(t)} O^* q_2(\gamma^*(t)))$ the optimal element of $[q_2]$, associated with the optimal rotation O^* and re-parametrization γ^* of the second curve, then

$$d_s([q_1],[q_2]) \doteq d_c(q_1, q_2^*) \,, \tag{2}$$

In practice, SVD is used to compute optimal rotation and the dynamic programming is performed for optimal parametrization.

3.3. Shape similarity

The elastic metric applied on extremal curve-based descriptors can be used to define a similarity measure. Given two 3D meshes x, y and their descriptors $x' = \{q_1^x, q_2^x, q_3^x, ..., q_N^x\}$ and $y' = \{q_1^y, q_2^y, q_3^y, ..., q_N^y\}$, the mesh-to-mesh similarity can be represented by the curve pairwise distances d:

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

where N is the number of curves used to describe the mesh. The average of curve distances between two descriptors can capture the similarity between their mesh poses. In case of change of shape in even one curve, the global distance will be affected and increase indicating that the poses are different. In Figure 2, a geodesic path between each corresponding two extremal curves, taken from two human bodies doing different poses, is computed in Shape Space. The evolution

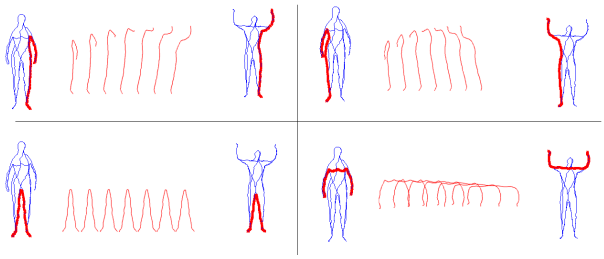


Figure 2: Geodesic path between extremal human curves of a neutral pose with raised hands.

of hand-foot curve between a pose with raised hand and another with hand down looks very natural under the elastic matching. Since we have geodesic paths denoting optimal deformations between individual curves, we can combine these deformations by an arithmetic distance to obtain full

deformations between two poses. Thanks to this global distance, we can compare human shape in different poses. For small deformation, the distance will be small and it is going to increase for models doing big different poses.

Comparing correspondent curves requires the identification of end-points as head, right/left hand and right/left foot, which is a not affordable in practice. This requirement is important to perform the curve matching separately between models. In order to overcome this problem, our method takes advantage from morphology of the human body. In fact, the head end-point is comment point between shortest curves among all possible geodesics between the five end-points. Besides, identification of the couple of hand/foot as corresponding to the same side of the body is deduced from geodesic paths connecting right hand to left foot end-points or left hand to right foot end-points which is always the longest on the human body surface. For 3D video sequences, once the end-points are correctly detected from the starting frame in the video sequence, a simple algorithm of end-point tracking over time is performed.

4. Motion segmentation and 3D video retrieval

Based on our EHC representation of the shape model, it is possible to compare two video sequences by matching their correspondent extremal curves using the geodesic distance in shape space (Equation 2). However, a sequence of human action can be composed of several distinct actions, and each one can be repeated several times. Therefore, the motion segmentation can play an important role in the dynamic matching by dividing the whole 3D video data into small, meaningful and manageable elementary actions called clips. EHC descriptor will be employed to segment continuous sequences into clips.

4.1. Motion segmentation

Video segmentation has been studied for various applications: gesture recognition, motion synthesis and indexing, browsing and retrieval. Most of works on the 3D video segmentation use the motion capture data, and very few of them were applied to dynamic 3D mesh. One of them is presented in [XYA05], where a histogram of distance among vertices on 3D mesh was generated to perform the segmentation through threesholding step defined empirically. In [YA06], the motion segmentation has been automatically conducted by analyzing the degree of motion using modified shape distribution, but they make such assumption: actions are mainly Japanese dances and the sequence of motion is paused for a moment and consider such moments as segmentation points.

In our work, we consider the issue of segmenting a 3D video with unknown temporal correspondence, where the mesh can change both connectivity and topology. We propose an approach fully automatic to segment a 3D video efficiently without making neither thresholding step nor assumption in the motion's nature.

In motion segmentation, the purpose is to split automatically the continuous sequence into segments exhibit basic movements, called clips. As we need to extract meaningful clips, the segmentation is overly fine and can be considered as finding the alphabet of motion. For a meaningful segmentation, motion speed is an important factor. In fact, when human changes motion type or direction, the motion speed becomes small and this results in dips in velocity. We exploit this by finding the local minima for the change in type of motion and local maxima for the change in direction. The extrema detected on velocity curve should be selected as segment points. In Figure 3, the vector of motion degree for human walk is shown. We show frames detected in maxima (the actor changes the foot's direction) on the top of the plot, and frames belonging to the minima (the actor raise the other foot) on the bottom. In this work, we consider only change in type of motion as a meaningful clip. Like this, clips with slight variations and a small number of frames are avoided.

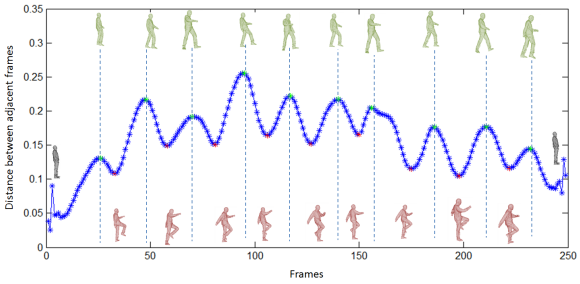


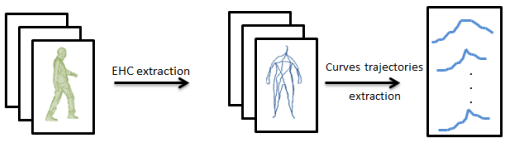
Figure 3: Detected extrema on extracted feature vector on a walk motion.

Note that optimum local minimum, that detect precise break points where the motion change, is selected in a predefined neighborhood. For this, we fix a size of window to test the efficiency of the local minimum in this condition. To calculate the speed variation, the degree of motion will be computed thanks to the distance between each two successive EHC in the sequence. The variations of a sequence are represented in vector of speed and a further smoothing filter is applied to obtain the final degree of motion vector.

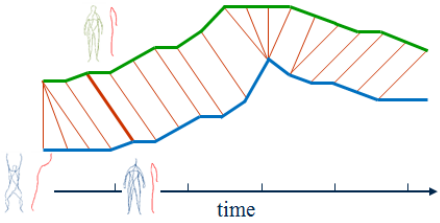
4.2. Clip matching

To seek for similar clips, example based queries are employed in a content-based retrieval context. Two motions can be considered similar even if there are changes in the shape of the actor and the speed of the action. This problem is similar to time-series retrieval where a distance metric is used to look for in a database the sequences whose distance to the query is below a threshold value. Each clip

is represented as a temporal sequence of human poses, characterized by EHC representation associated to shape model. Then, extremal curves are tracked in each sequence to characterize a trajectory of each curve in the shape space (Figure 4 (a)). Finally, the trajectories of each curve are matched and a similarity score is obtained. However, due to the variations in execution rates of the same clip, two trajectories do not necessarily have the same length. Therefore, a temporal alignment of these trajectories is crucial before computing the global similarity measure (Figure 4 (b)).



(a) Motion characterization



(b) DTW between trajectories

Figure 4: Alignment process of different trajectories that model the motion.

In order to solve the temporal variation problem, we use DTW algorithm [Gio09]. This algorithm is used to find optimal non-linear warping function to match a given time-series with another one, while adhering to certain restrictions such as the monotonicity of the warping in the time domain. The optimization process is usually performed using dynamic programming approaches given a measure of similarity between the features of the two sequences at different time instants. The global accumulated costs along the path define a global distance between the query clip and the motion segments found in the database. Since DTW can operate with any measure of similarity between different temporal features, we adapt it to features that reside on Riemannian manifolds. Hence, we use the geodesic distance between different shape points $d_s(q_i, q_j)$, proposed in equation 2, as a distance function between the shape features at different time instants.

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

5. Experiments and discussions

To show the practical relevance of our method, we perform an experimental evaluations on several databases, and compare EHC descriptor performance, separately, to the most efficient descriptors of the state-of-the-art methods. We measure the efficacy of our descriptor to capture the shape similarity in 3D video sequences of different actors and actions from a public database. We evaluate this later against Temporal Shape Histogram [HHS10], Multi-resolution Reebgraph [HTN*10] and other classic shape descriptors, using provided ground truth. Finally, the performance of EHC descriptor to segment sequences and to retrieve clips is evaluated using a ground-truth dataset from simulated data. A real video sequence was also used to test the efficiency of our descriptor.

5.1. Shape similarity for 3D video sequences

The performance of EHC is evaluated against various shape similarity metrics for 3D video sequences of people with unknown temporal correspondence from the i3DPost dataset [SH03]. Performances of similarity measures are compared by evaluating Receiver Operator Characteristics (ROCs) for classification against ground-truth of a comprehensive data set of synthetic 3D video sequences consisting of animations of several people performing different motions. The similarity metric is represented by elastic measure values between each pair of models. The temporal ground truth similarity between two frames is defined as a combination of shape and velocity similarity as described in [HHS10]. In order to classify frames as similar or dissimilar a threshold is set on temporal ground truth similarity matrix. An example of self-similarity matrix computed using ground-truth descriptor, static and temporal descriptors is shown in Figure 5. This figure illustrates also the effect of time filtering with increasing temporal window size for ECD descriptors on a periodic walking motion.

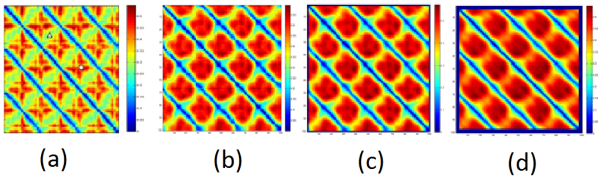


Figure 5: Similarity measure for "Fast Walk" motion in a straight line compared with itself. Coldest colors indicate most similar frames. (a) Temporal Ground-Truth (TGT), (b-d) Self-similarity matrix computed with TEHC with window size 3, 5 and 7 respectively.

Before all, we analyzed the performance of all possible combinations of curves on the shape similarity measurements and best combination is considered for all following experiments [SWD13]. We then compare our descriptor with the most popular descriptors. The comparison includes Shape Histograms (SHvr), Spin Image (SI), Shape Distribution (SD), and Spherical Harmonics Representation (SHR) using a time window of size 7. Results are shown in Figure 6 and observations resulting from the analysis of these results are the following: First, our descriptor outperforms

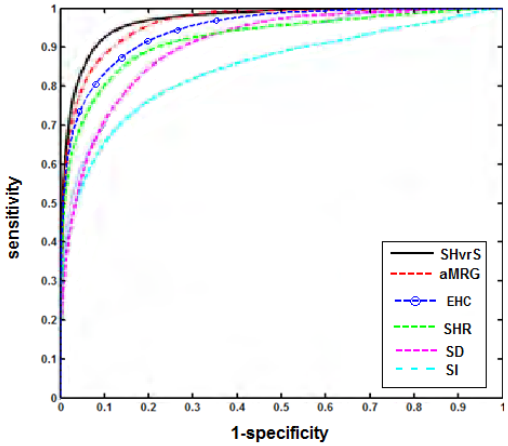


Figure 6: Evaluation of ROC curves. EHC is one of the top performers for shape retrieval in 3D video.

classic shape descriptors (SI, SHR, SD) and shows competitive results with SHvrS and aMRG. Multiframe shape-flow matching required in SHvrS allows the descriptor to be more robust but the computational cost will increase by the size of selected time window. Second, EHC descriptor by its simple representation, demonstrates a comparable recognition performance to aMRG. It is efficient as the curve extraction is instantaneous and robust as the curve representation is invariant to elastic and geometric changes thanks to the use of the elastic metric. Third, the result analysis for each action shows that EHC gives a smooth rates that are stable and are not affected by the complexity of the motion [SWD13]. Such complex motions are rock and roll, vogue dance, faint, shot arm. However, this is not the case for SHvrS where performance recognition falls suddenly with complex motions as presented in figure 18 at [HHS10].

5.2. Motion segmentation

Plotting the distance between EHC representation of successive frames gives a very noisy curve. The break points from this curve do not define semantic clips and the extracting of minima leads to an over-segmentation of the sequence (see Fig7 (Top)). To obtain more significant local minima, we convolve the curve with a time-filter allowing to take into account the motion variation, not only between two successive frames but also in a time window. The motion degree after convolution is shown in Figure 7(Bottom). Break points are more precise and delimits significant clips corresponding to step change in the video sequence. The window size is defined empirically and fixed to 6 for all types of actions.

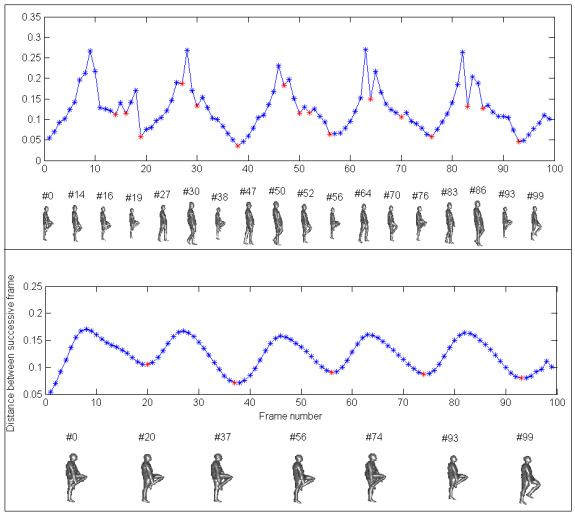


Figure 7: Speed curve smoothing.

In Figure 8, we show some results of motion segmentation on a slow and a fast walk. Although the walk speed increase, the action segmentation remains significant and does not change and corresponds to the step change of the actor. Segmentation for Rockn'roll dance motion is also illustrated in Figure 8(bottom). Thanks to the selection of local minima in a precise neighbourhood, only significant break points are selected.

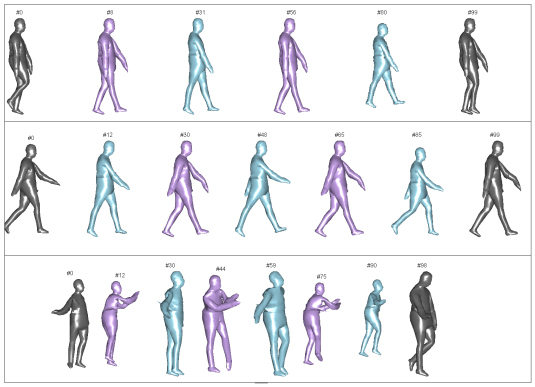


Figure 8: Segmentation results on various motion: (top) slow walk (middle) fast twalk (bottom) dance :Rockn'roll.

5.3. Motion retrieval

In the previous experiments, the temporal shape similarity performed by the state-of-the-art methods and compared to

our descriptor. In this experiment, we advocate the usage of the EHC representation and the motion segmentation for motion retrieval, where a query consists of a clip. As in a classical retrieval procedure, in response to a given query, our approach looks for in the benchmark database and returns an ordered list of responses called the ranked list. The evaluation of the algorithm is then transformed to the evaluation of the quality of the ranked list. For our experiments, we use 13 different actions from the i3DPost dataset [SH03], performed by two actors making a total of 26 actions. A motion segmentation stage is performed on these action sequences giving a total of 144 clips categorized into 14 classes. The action sequences consist mainly of different styles of walking, running and some dancing sequences. Classes grouped together present different styles of walking, running and dancing steps. For example, a step change in a walk may represent a class and groups similar clips done with different speed and in different trajectories. We notice that Right to Left change step is grouped in a different class than Left to Right change step.

The similarity metric represented by elastic measure values between each pair of clips allows us to generate a confusion matrix for all classes of clips, in order to evaluate the recognition performance by computing dynamic retrieval measures thanks to a manually annotated ground truth. An example of the matrix representing the similarity evaluation score among clips in sequences performed by a female actress against the clips of sequences of actions performed by a male actor is showed in Figure 9. The coldest the color is, the more similar the two clips are.

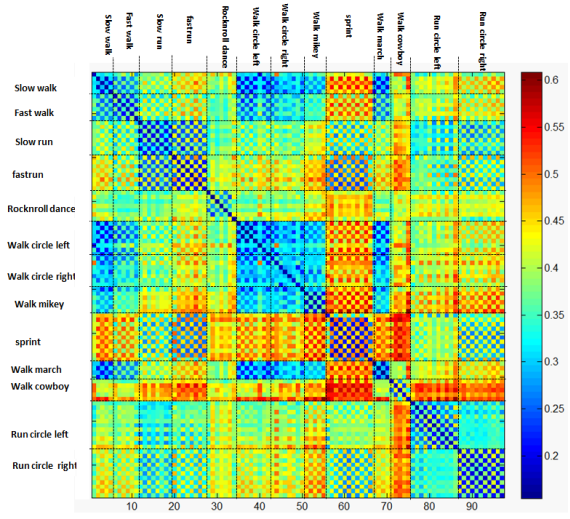


Figure 9: Similarity matrix evaluation between clips. The coldest the color is, the more similar the two clips are.

Thanks to the use of DTW, it is noticed that similarity score between same clips done in different speed is small (see Figure 9). The matching between the clip representing change in step in slow walk (25frames) and fast walk (18 frames)(Figure 8, top and middle row) is small.

Besides, our approach succeed to retrieve clips doing the same action in different ways. For example (see Figure 9), the walk circle clips can be matched with the clips of slow walk action done in a straight line. This can be explained by the use of the elastic metric to compare and match curve trajectories, which is independent to rotation. Although the actors performing the actions are different, it is observed that similar clips yield smaller similarity score. Like shown for Rocknroll dance action, when steps of the dance performed by different actors are correctly retrieved.

Retrieval performances are tested using the 28 actions performed by the two actors (Jigna and Adrien from the i3DPost dataset [SH03]). In the experiment, each clip from sequences are used as query. The clips from the segmented sequences present in the dataset are used as candidates. The query itself is not included in candidates. The used evaluation algorithm involves the evaluation measures used by information retrieval community (1st tier 2nd tier, NN and E-measure). It is demonstrated that 79.26% of similar motion clips are included in the first tier and more than 90% (93%) of clips are correctly retrieved in the second tier. Besides, accuracy of nearest neighbor is 99.1%. It is a rather good performance (Figure 10) considering that only such low-level feature as the EHC is utilized in the matching. The problem is that EHC is based on geodesics on 3D shapes, and our approach for retrieval is based on extremal curves trajectories in the sequence. However, extracted sequential curves that present the trajectory tend to change completely of path on the models while moving and thereby mislead the matching performed by DTW.

We also apply our retrieval approach to real captured 3D video sequences of people [VBMP08]. Self similarity example with an actor in a walking motion (walking in circular way) and its similarity curve are shown in Figure 11. The query clip is a Right-Left step change in the first position before doing a circle with the walk motion and retrieved clips are frames in the same class found later in the sequence when the actor is turning.

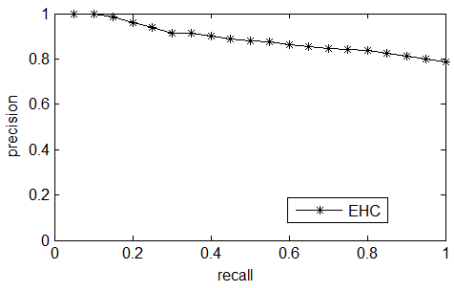


Figure 10: Recall/Precision curve for clip similarity.

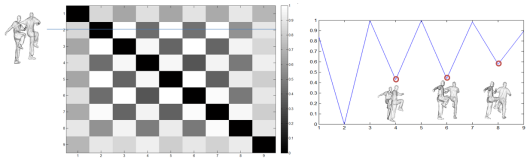


Figure 11: Experimental results for 3D video retrieval using motion of "walk in circle".

6. Conclusion

In this work, body shape is firstly represented as a set of geodesic curves extracted from shape surface using extremal feature points and presented as open curves in shape space where they become invariant to translation, scale and elasticity. Then, an elastic metric is calculated between two shape models in order to estimate their similarity. We extended this descriptor to 3D video retrieval, where a motion segmentation is performed on continuous sequence to split it into elementary action segments called clips. These later are represented by a temporal trajectories of selected human curves on the open curve shape space. Video retrieval is then performed by matching the trajectories using DTW algorithm on the features that reside in Rienmanian manifolds and operate with the elastic metric defined in the shape space. Moreover, our approach achieves a performance accuracy of 93.44% for video retrieval as second tier, which is encouraging and shows the potential of this approach.

Finally, we would encourage future works to extend our approach to investigate more challenging applications like 3D human action modelling using HMM like approach.

References

[CBK05] CHEUNG K. M., BAKER S., KANADE T.: Shape-from-silhouette across time part i: Theory and algorithms. vol. 62, pp. 221 – 247. 1

[dAST*08a] DE AGUIAR E., STOLL C., THEOBALT C., AHMED N., SEIDEL H.-P., THRUN S.: Performance capture from sparse multi-view video. vol. 27, ACM, pp. 98:1–98:10. 1

[dAST*08b] DE AGUIAR E., STOLL C., THEOBALT C., AHMED N., SEIDEL H.-P., THRUN S.: Performance capture from sparse multi-view video. vol. 27, ACM, pp. 98:1–98:10. 1

[DTP12] DANELAKIS A., THEOHARIS T., PRATIKAKIS I.: 3d mesh video retrieval: A survey. In *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2012* (oct. 2012), pp. 1 –4. 2

[Gio09] GIORGINO T.: Computing and visualizing dynamic time warping alignments in r: The dtw package. vol. 31, pp. 1–24. 5

[HHS10] HUANG P., HILTON A., STARCK J.: Shape similarity for 3d video sequences of people. vol. 89, Kluwer Academic Publishers, pp. 362–381. 2, 5, 6

[HTN*10] HUANG P., TUNG T., NOBUHARA S., HILTON A., MATSUYAMA T.: Comparison of skeleton and non-skeleton shape descriptors for 3d video. In *Proceedings of the Fifth International Symposium on 3D Data Processing, Visualization and Transmission (3DPVTŠ10)* (Pairs, France, May 2010). 2, 5

[HTTM11] HOLTE M. B., TRAN C., TRIVEDI M. M., MOESLUND T. B.: Human action recognition using multiple views: a comparative perspective on recent developments. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding* (New York, NY, USA, 2011), J-HGBU '11, ACM, pp. 47–52. 2

[JKSJ07] JOSHI S., KLASSEN E., SRIVASTAVA A., JERMYN I.: A novel representation for riemannian analysis of elastic curves in rn. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on* (june 2007), pp. 1 –7. 3

[KGH09] KILNER J., GUILLEMAUT, HILTON A.: 3D Action Matching with Key-Pose Detection. In *Search in 3D and Video (S3DV)* (2009). 2

[SH03] STARCK J., HILTON A.: Model-based multiple view reconstruction of people. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on* (oct. 2003), pp. 915 – 922 vol.2. 5, 7

[SWD13] SLAMA R., WANNOUS H., DAOUDI M.: Extremal human curves: a new human body shape and pose descriptor. 2, 6

[TM12] TUNG T., MATSUYAMA T.: Topology dictionary for 3d video understanding. vol. 34, pp. 1645 –1657. 2

[TNM09] TUNG T., NOBUHARA S., MATSUYAMA T.: Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In *Computer Vision, 2009 IEEE 12th International Conference on* (29 2009-oct. 2 2009), pp. 1709 –1716. 1

[TS05] TUNG T., SCHMITT F.: The augmented multiresolution reeb graph approach for content-based retrieval of 3d shapes. pp. 91–120. 2

[TVD06] TIERNY J., VANDEBORRE J.-P., DAOUDI M.: Invariant high level reeb graphs of 3d polygonal meshes. vol. 0, IEEE Computer Society, pp. 105–112. 2

[VBMP08] VLASIC D., BARAN I., MATUSIK W., POPOVIC J.: Articulated mesh animation from multi-view silhouettes. 7

[XYA05] XU J., YAMASAKI T., AIZAWA K.: 3d video segmentation using point distance histograms. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on* (sept. 2005), vol. 1, pp. I – 701–4. 4

[YA06] YAMASAKI T., AIZAWA K.: Motion segmentation of 3d video using modified shape distribution. In *Multimedia and Expo, 2006 IEEE International Conference on* (july 2006), pp. 1909 –1912. 4

[YA07] YAMASAKI T., AIZAWA K.: Motion segmentation and retrieval for 3d video based on modified shape distribution. vol. 2007, p. 059535. 2