Comparison of Skeleton and Non-Skeleton Shape Descriptors for 3D Video

Peng Huang*, Tony Tung†, Shohei Nobuhara†, Adrian Hilton*, Takashi Matsuyama†

*Centre for Vision Speech and Signal Processing, University of Surrey, UK
{p.huang,a.hilton}@surrey.ac.uk

†Graduate School of Informatics, Kyoto University, Japan
{tung,nob,tm}@vision.kuee.kyoto-u.ac.jp

Abstract

This paper presents a performance evaluation of shape similarity metrics for 3D video sequences of people with unknown temporal correspondence. Previous evaluation focuses on non-skeletal similarity metrics. Since the human motion is essentially an articulated motion, it is also interesting to investigate skeletal-based similarity metrics. In this paper, we extend previous evaluation of non-skeletal similarity metrics to skeletal-based similarity metrics. A quantitative evaluation based on the Receiver-Operator Characteristic (ROC) curve for the descriptors using a ground-truth dataset for synthetic motion sequences is presented. Reeb Graph, Multi-Dimensional Scaling and Shape Histograms are compared with and without a temporal filter. Reeb Graph and Shape Histograms achieve comparable ROC performance, both outperform MDS in the task of finding similar poses of the same person in 3D video. Finally, temporal Reeb Graph and Shape Histograms are applied to a public database of 3D video of people to identify optimal transitions and synthesise 3D character animation. Results demonstrate the accurate matching of surface shape and motion.

1. Introduction

Multiple-view reconstruction of human performance as a 3D video [15] has advanced to the stage of capturing both surface motion and surface dynamics of the body, clothing and hair during motion. Several potential applications have arisen from this, such as concatenating 3D video sequences to produce novel character animation [8], 3D video summarization and compression [7, 17]. These potential applications subsequently require solving the problem to identify frames with similar surface shape and motion including pose, clothing and hair in 3D video sequences. However, 3D video reconstruction results in an unstructured volumetric or mesh approximation of the surface shape at each time instance without temporal correspondence, which makes deriving a similarity metrics suitable for 3D video a challenging task.

Three dimensional (3D) shape similarity metrics have been widely investigated [4, 10, 16] as a means of effective and efficient object retrieval. Typically, shape descriptors are first extracted for each object, and the shape similarity between objects is then computed as a distance between their descriptors. Conventional shape descriptors focus on classifying static objects into different classes, for example, differentiating a chair from a tank. However, many shape descriptors are not sufficiently discriminative to distinguish different poses of a person, for example, distinguish a walk pose from a run. Previous work has investigated and evaluated several non-skeletal shape descriptors on 3D video, and extended them over time by applying a temporal filter [9], which has demonstrated improved ROC performance since both shape and motion similarity are considered.

Previous studies of similarity metrics for 3D video only evaluated spatial shape descriptors based on object surface or volume. In this paper we present a comparative evaluation of skeleton-based shape descriptors against previous spatial descriptors. The main contribution is a quantitative evaluation of 3D shape similarity in time-varying 3D mesh sequences using an articulated ground-truth dataset in which surface correspondence is predefined. A comparison is made between a non-skeletal shape descriptor Shape Histograms previously shown to give good recognition performance [9], a skeleton-based shape descriptor Reeb Graph [18] and a bending-free descriptor Multi-Dimensional Scaling [3]. Ground-truth evaluation is performed for synthetic sequences of 14 people performing 28 motions (Figure 5 and Table 1). Real sequences of 4 people/costumes each performing $6-8$ different motions from a public database [15] include a variety of loose and tight fitting clothing together with long sequences of com-

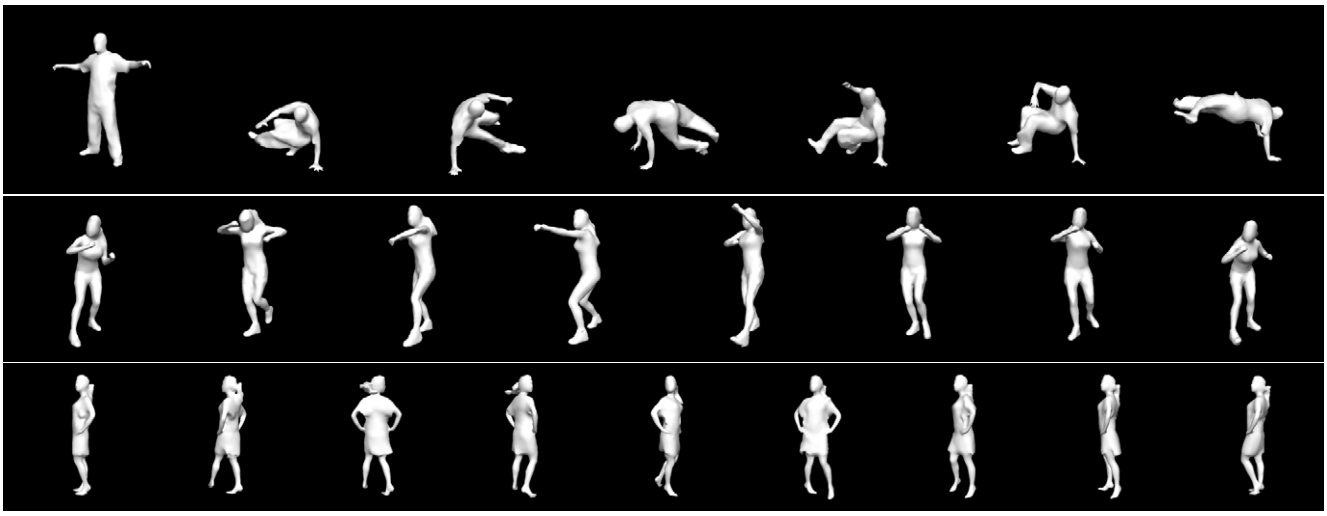


Figure 1. Real 3D video sequences. Top to bottom: JP "Free" dance motion (7 key-frames), Roxanne-Character1 "Hit" motion (8 key-frames) and Roxanne-Fashion2 "Twirl" motion (9 key-frames).

plex motions from a street dancer (Figure 1 and Table 1). Evaluation demonstrates comparable ROC performance on ground-truth synthetic sequences for Shape Histograms and Reeb Graphs. Real sequences also give comparable recognition performance enabling us the identification of frames with similar shape and motion for subjects with loose clothing and hair. However, more robustness is obtained for real sequences with changes in topology (due to reconstruction errors) using Shape Histograms.

2. Shape Descriptors

In this section, we present the shape descriptors evaluated together with their implementation details. The extension to temporal shape matching is then defined and computed as a temporal similarity metrics using a simple time-filter.

2.1. Shape Histogram

Ankerst *et al.* [1] introduce the 3D *Shape Histogram* (SH) as a shape signature to classify a molecular database. A 3D Shape Histogram is based on a partitioning of the space where an object resides, that is, a complete and disjoint decomposition into cells which correspond to the bins of the histogram. This approach is robust to topology changes, surface noise and subtle surface changes. However, the descriptor is not invariant to rotation, so a rotation-invariant comparison scheme is required. Here, we use a *rotated volumetric spherical Shape Histogram* (SHvr) which is reported as the best performer among several state-of-art shape descriptors both with and without time-filtering for the task of finding similar poses of the same actor in 3D video sequences [9]. Figure 2 illustrates a partition of the 3D space by SHvr.

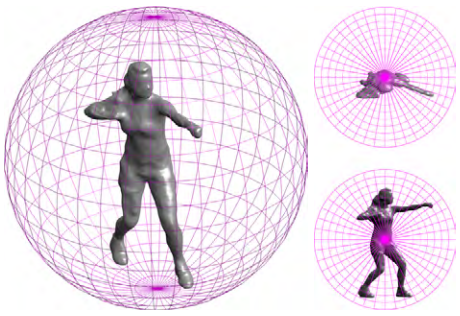


Figure 2. Illustration of SHvr partitioning the 3D space.

Shape Histogram construction. Given a 3D triangle mesh $M =< V, F >$, a volume sampling spherical histogram is constructed as follows:

1. A volumetric representation is constructed by first dividing the space into a $N_g \times N_g \times N_g$ voxel grid and then identifying the voxels which lie inside the 3D model. We denote $\mathbf{o} = (x, y, z)$ for 3D position of an occupied voxel centroid. This gives a set $O = \{\mathbf{o_n}\}, n = 0, ..., N_o - 1$, where N_o is the total number of occupied voxels, $N_o < N_g^3$.

2. Space in a Cartesian coordinate system is transformed into a Spherical coordinate system defined by the centre of mass of the model and a vertical axis. For each occupied voxel centroid $\mathbf{o} \in O$, the spherical coordinates $\mathbf{s} = (r, \theta, \phi)$ are calculated as follows,

$$(r, \theta, \phi) = (\sqrt{x^2 + y^2 + z^2}, \arccos \frac{z}{r}, \arctan \frac{y}{x}) \quad (1)$$

This gives a set of spherical coordinates of occupied voxel centroids $S = \{\mathbf{s_k}\}, k = 0, ..., N_o$.

3. A 3D spherical histogram $H(S) = [H_{\mathbf{l}}(S)]_{N_{\mathbf{l}}}$ is constructed, where $\mathbf{l} = [l_r, l_\theta, l_\phi]$, $N_{\mathbf{l}} = N_r \times N_\theta \times N_\phi$, accumulating the voxels in the volume representation,

$$H_{\mathbf{l}}(S) = \sum_{k=0}^{N_o} {}^{1} g(\mathbf{l}, k), g(\mathbf{l}k) = \begin{cases} 1 & \text{if } \mathbf{s_k} \text{ in } \mathbf{l}^{th} \text{ bin} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where the $\mathbf{l}^{th}$ bin is defined as a subspace $[l_r \cdot \Delta r, l_r \cdot \Delta r + \Delta r) \times [l_\theta \cdot \Delta\theta, l_\theta \cdot \Delta\theta + \Delta\theta) \times [l_\phi \cdot \Delta\phi, l_\phi \cdot \Delta\phi + \Delta\phi)$. $\Delta r, \Delta\theta, \Delta\phi$ denote the bin size for radius, inclination angle and azimuth angle respectively,

$$\Delta\{r, \theta, \phi\} = \frac{\{r, \theta, \phi\}_{up} \quad \{r, \theta, \phi\}_{low}}{N_{\{r,\theta,\phi\}}} \quad (3)$$

where we assume $(r_{up}, \theta_{up}, \phi_{up}) = (1.5m, \pi, 2\pi)$ and $(r_{low}, \theta_{low}, \phi_{low}) = (0, 0, 0)$ for a bounding sphere which certainly covers a human body. (N_r, N_θ, N_ϕ) are user-defined number of bins for each dimension in the spherical coordinate space.

4. The final descriptor $SH(M)$ is a 3D histogram of the probability normalised by the total number of occupied voxels N_o,

$$SH(M) = \frac{H(S)}{N_o} \quad (4)$$

Similarity function. Since Shape Histograms are not rotation invariant, we have to define the similarity function to take rotation into account. The shape dissimilarity $s(p, q)$ between two 3D video frames p and q is defined as follows,

$$s(p, q) = \min_{\phi} |SH(M_p) \quad SH(\mathcal{R}(M_q, \phi))| \quad (5)$$

where $\mathcal{R}(M_q, \phi)$ denotes the 3D mesh M_q rotated by ϕ around the vertical axis. Here, we assume that human models have an upright direction, since we consider a human pose laying on the ground to be different from a standing pose, even though their shapes are similar since they cannot be concatenated seamlessly. In practice, instead of rotating the 3D mesh model M_p we first construct a high-resolution Shape Histogram $SH^*(M_q)$ and store it. Computing the minimal similarity against different ϕ requires shifting $SH^*(M_q)$ in dimension ϕ and re-binning back to $SH(M_q)$ for comparison. Since we only consider the rotation about the vertical axis, the bin size of high-resolution Shape Histogram is set to only increase the resolution in dimension ϕ, $(\Delta r^*, \Delta\theta^*, \Delta\phi^*) = (\Delta r, \Delta\theta, 1^\circ)$. Therefore, we can compute the minima by shifting the histogram with an array $\phi_n = [0, 1, .., 359]$,

$$s(p, q) = \min_{\phi_n} |SH(M_p) \quad \mathcal{B}(SH^*(M_q, \phi_n))| \quad (6)$$

where $\mathcal{B}(\cdot)$ denotes re-binning the high-resolution Shape Histogram $SH^*(\cdot)$ back to $SH(\cdot)$ and $SH^*(M_q, \phi_n)$ shifting $SH^*(M_q)$ with ϕ_n bins in the dimension of ϕ.

2.2. Multi-Dimensional Scaling

Schwartz *et al*. [12] introduced a representation of the intrinsic geometry of the cortical surface of the brain using *Multi-Dimensional Scaling* (MDS). Elad and Kimmel [5] extended the idea and proposed a non-rigid shape recognition method based on Euclidean embeddings. Here, we use a simplified version of the MDS based intrinsic dissimilarity measure proposed by Bronstein *et al*. [3]. MDS dissimilarity measures geometric (Hausdroff) difference between canonicalized meshes: (a) if the two input meshes have difference global topology, MDS dissimilarity captures the differences between the topology; (b) if the two input meshes share a global topology, MDS dissimilarity captures the differences on the geodesic distance, i.e., the differences of the amount of deformation. Figure 3 illustrates the canonicalization by MDS.

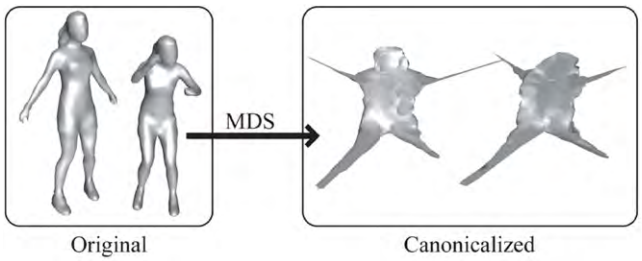


Figure 3. Canonicalization by MDS. MDS deforms extrinsically dissimilar but intrinsically similar meshes (left) to extrinsically and intrinsically similar forms (right).

MDS construction. The canonicalization is performed by finding a deformation which makes the geodesic distances between all pairs of the mesh vertices be equal to those of Euclidean distances. The best solution in terms of L_2 distortion is given by MDS [2]. Let D be an $N \times N$ matrix and its (i, j) element stores the geodesic distance between the i-th and j-th vertices. The vertex positions of the canonical shape are given as the first three eigenvectors of the double centred matrix of squared $\mathbf{D}$:

$$\tfrac{1}{2}\left(\mathbf{I} \quad \tfrac{1}{N}\mathbf{J}\right)\mathbf{D}^2\left(\mathbf{I} \quad \tfrac{1}{N}\mathbf{J}\right), \quad (7)$$

where $\mathbf{I}$ and $\mathbf{J}$ denote an $N \times N$ identity matrix and an $N \times N$ matrix whose elements are all 1. Bronstein *et al*. [2] have pointed out that other metric spaces can be used to obtain better embeddings, but we use Euclidean space for simplicity and efficiency.

Similarity function. The shape dissimilarity $s(p, q)$ between p and q are simply given as the Hausdorff distance of sets of vertices after MDS canonicalization,

$$s(p, q) = d_H(X, Y) \quad (8)$$

where X and Y denote sets of vertex positions of corresponding to canonical shapes and d_H computes Hausdorff distance between two sets of points.

2.3. Reeb Graph

Reeb [11] first introduced the *Reeb Graph* (RG) as a high level 3D shape descriptor which represents both 3D mesh topology and shape using a graphical representation of surface properties. A Reeb Graph is built using a differentiable function μ defined on the model surface. The critical points of μ allow characterisation of the topology of the model. The surface is divided into regions according to μ values, and then a Reeb graph is obtained by first associating a node to each region and then linking the connected regions. Hilaga *et al*. [6] proposed a *Multi-resolution Reeb Graphs* (MRG) to estimate similarity and correspondence between 3D shapes. They choose normalised integral of geodesic distance as the continuous scalar function for rotation invariance and resistance against noise. The similarity is calculated with a coarse-to-fine strategy using the attributes of nodes in the MRG and topological consistency. Here, we use an *augmented Multi-resolution Reeb Graph* (aMRG) proposed by Tung and Schmitt [18]. A topological consistency criteria and geometric attributes are further introduced to the nodes in order to obtain better matching between nodes of graphs when comparing models. Figure 4 shows an example of Reeb Graph descriptor.

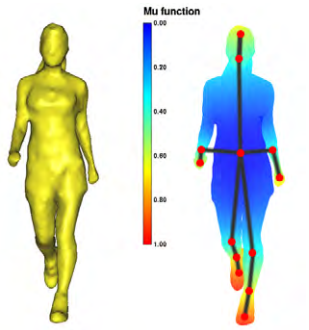


Figure 4. Illustration of Reeb Graph.

Reeb Graph construction. We assume 3D surface models approximated by compact 2-manifold meshes. Let S be a surface mesh. According to the Morse theory, the topology of the surface can be characterized using the critical points of a continuous function μ defined on S. The surface connectivities between the critical points can then be used to build the Reeb graph of μ, which is the quotient space defined by the following equivalence relation $\sim$: let $\mathbf{X} \in S$ and $\mathbf{Y} \in S$, then $\mathbf{X} \sim \mathbf{Y}$ if and only if: (1) $\mathbf{X}$ and $\mathbf{Y}$ belong to the same connected component of $\mu^{-1}(\mu(\mathbf{X}))$, and (2) $\mu(\mathbf{X}) = \mu(\mathbf{Y})$ [11]. In our framework, the Morse function μ is defined as in [6]:

$$\mu(\mathbf{v}) = \int_{\mathbf{p} \in S} g(\mathbf{v}, \mathbf{p}) dS \quad \text{and} \quad \mu_N(\mathbf{v}) = \frac{\mu - \mu_{\min}}{\mu_{\max} - \mu_{\min}} \tag{9}$$

where $g(\mathbf{v}, \mathbf{p})$ is the geodesic distance on S between $\mathbf{v}$ and $\mathbf{p}$, $\mu_N : S \rightarrow [0, 1]$ is the normalized function μ, and $\mu_{\min}$ and $\mu_{\max}$ are minimal and maximal values of μ respectively. As defined, μ_N is invariant to rotation, translation and scale transformations. The integral formulation provides robustness to local surface noise such as outlying vertices caused by reconstruction artefacts. Extremal values of μ_N return critical points corresponding to highly concave of convex regions of the surfaces. Thus, the Reeb Graph is constructed by: (1) partitioning the surface model into regular intervals based on μ_N values; (2) assigning a node to every region in each interval; (3) linking nodes of connected regions. The resolution level R of the Reeb graph relies on the number of intervals 2^R obtained by iterative subdivisions of $\mu_N \in [0, 1]$. Lower resolution Reeb graphs are then obtained by hierarchically merging intervals by pairs. Nodes are also linked to a unique parent-node from the lower graph resolution [6]. The multi-resolution Reeb graph is therefore represented as a set of Reeb graphs of various levels of resolution $r = 0...R$ extracted from the Reeb graph of the highest resolution R .

Similarity function. The shape dissimilarity $s(p, q)$ between p and q relies on the similarity evaluation of the $\mathcal{C}_r$ pairs of topologically consistent nodes for all level of resolution $r = 0$ to R:

$$\mathrm{s}(p, q) = \sum_{r=0}^{R} \sum_{(m,n) \in \mathcal{C}_r} \mathrm{sim}(m, n) \tag{10}$$

where $\mathrm{sim}(\cdot)$ measures the difference between two node features [18] by taking into account the embedded attributes, *i.e*. geometrical features such as surface local area and topology-based features such as graph connectivity of neighbouring nodes.

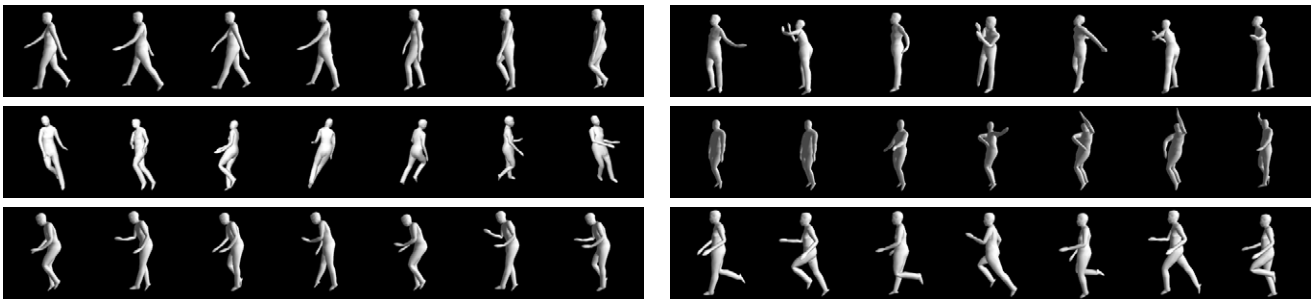
2.4. Similarity Metrics

The frame-to-frame similarity matrix is first obtained from the 3D Shape Descriptor. Given two 3D video sequences $P = \{p_i\}$ and $Q = \{q_j\}$, the similarity matrix $S := (s_{ij})_{N_p \times N_q}$ where $s_{ij} = s(p_i, q_j)$ measures the shape dissimilarity between p_i and q_j from equation 6, 8, 10 according to shape descriptors. Temporal similarity can be evaluated by applying a simple time filter. Here, we adopt an average weighted filter and the temporal similarity is evaluated as follows,

$$s_{ij}^t = \frac{1}{2N_t + 1} \sum_{k=-N_t}^{N_t} s_{(i+k)(j+k)} \tag{11}$$

where the time filter is with a window size of $2N_t + 1$. The computational cost is dominated by the cost of computing the frame-to-frame shape similarity with a relatively small additional cost of time filtering using equation 11. Temporal filtering is a way of incorporating motion in the similarity measure [9].

(a) 3D models



(b) 3D models + MoCap

Figure 5. Synthetic dataset. (a) 14 models; (b) Jigna's "fast walk", "rock n'roll", "run circle left", "vogue dance", "sneak", "sprint".

Dataset	Motions	Frames
Synthetic: Adrian, Alan, Dave, EngJon, Graham, Jez, Jigna, Joel, Marc, PengWei, Pete, Pip, Venura, Yacob	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, rock n'roll, shoot.	39200 (100 per seq.)
Real: JP, Roxanne (Character1, Fashion1&2)	street dance (flashkick, free, head, kickup, lock, pop + transitions) standard movement (stand, walk, jog, pose, hit, twirl + transitions)	3592

Table 1. Synthetic and real 3D video sequences statistics. Synthetic (14 people $\times$ 28 motions) and Real (4 people/costumes $\times$ 6-8 motions).

3. Evaluation Methodology

The recognition performance of the shape descriptors is evaluated using a ground-truth dataset from synthetic 3D video sequences of people. The best performer is then used to identify similar frames in real 3D video sequences.

3.1. Ground-truth

A synthetic dataset is created using 14 articulated character model for people animation using 28 motion capture sequences (Table 1). Animated models of people with different gender, body-shape and clothing were reconstructed from multiple view images [14]. The height varies between about 1.6m to 1.9m. Each model has a single surface mesh with 1K vertices and 2K triangles. Figure 5 shows example frames of models and motions for one model.

3.2. Evaluation Criterion

Recognition performance is evaluated using the ROC curve, showing the true-positive rate (TPR) or *sensitivity* in correctly defining similarity against the false-positive rate (FPR) or *one-specificity* where similarity is incorrect,

$$TPR = \frac{ts}{ts + fd} \quad \text{and} \quad FPR = \frac{fs}{fs + td} \tag{12}$$

where ts denotes the number of true-similar predictions, fs the false similar, td true dissimilar and fd false dissimilar in comparing the predicted similarity between two frames to the ground-truth similarity. The similarity for each shape descriptor is normalised to the range $s'_{ij} \in [0, 1]$.

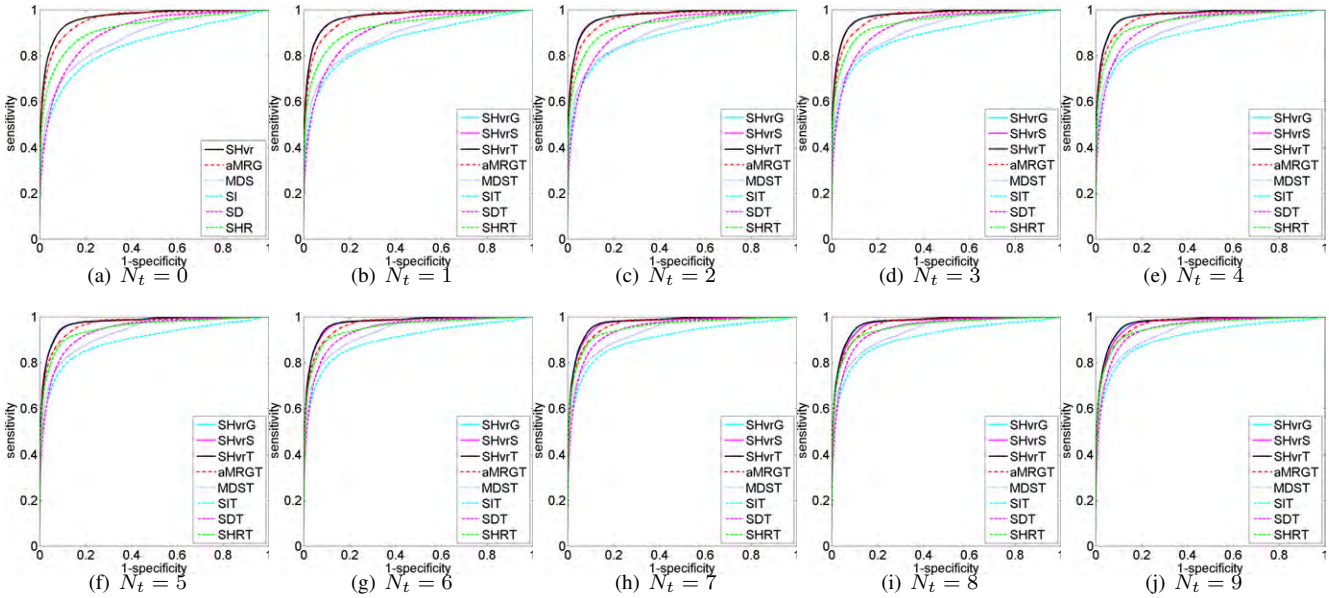
$$s'_{ij} = \frac{s_{ij} - s_{min}}{s_{max} - s_{min}} \tag{13}$$

where $s_{min} = 0$ and s_{max} is the maximal dissimilarity over all $s_{ij} \in S$ similarity matrix of the whole database. A binary classification matrix for the shape descriptor $C(\tau) = [c_{ij}(\tau)] \in \{1, 0\}$ is then defined

$$c_{ij}(\tau) = \begin{cases} 1 & \text{if } s'_{ij} < \tau \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

The classification $c_{ij}(\tau)$ for a given τ is then compared to the ground-truth similarity classification c_{ij}^{GT} defined the same with that in [9]. The number of true and false similarity classifications, $ts(\tau), td(\tau), fs(\tau), fd(\tau)$ is then counted. The ROC performance for a given shape similarity measure is then obtained by varying the threshold $\tau \in [0, 1]$ to obtain the true $TPR(\tau)$ and false $FPR(\tau)$ positive rates according to equation 12.

Figure 6 presents combined ROC curves of descriptors with and without time-filtering for evaluating self-similarity against temporal ground-truth across all people and motions in the synthetic dataset. For comparison, previous evaluated non-skeletal shape descriptors are also included, *i.e.*



(a) $N_t = 0$ (b) $N_t = 1$ (c) $N_t = 2$ (d) $N_t = 3$ (e) $N_t = 4$

(f) $N_t = 5$ (g) $N_t = 6$ (h) $N_t = 7$ (i) $N_t = 8$ (j) $N_t = 9$

Figure 6. Evaluation of ROC curves for static and time-filtered descriptors on self-similarity across 14 people each performing 28 motions.

Shape Distribution (SD), Spin Image (SI), Spherical Harmonics Representation (SHR) and two Shape-flow descriptors, the global/local frame alignment Shape Histograms (SHvrG/SHvrS) [9]. The SHvr is set as the same optimal parameters in previous evaluation [9], *i.e.* $(N_r, N_\theta, N_\phi) = (5, 10, 20)$. The aMRG has been tuned as a global descriptor in order to characterise large temporal surface variation at the low resolution $r = 2$. The recognition performance of all descriptors with time-filtering increases as the temporal window size increases shown in Figure 6(b-j). They all show an improved recognition performance compared to the equivalent frame-to-frame shape similarity in Figure 6(a). This is because temporal filtering reduces the lines of similarity in the anti-diagonal direction which occur for similar shapes with different motions. The Shape Histograms (SHvr, SHvrT, SHvrG, SHvrS) and Reeb Graphs (aMRG, aMRGT) give similar characteristics achieving the best recognition performance of all shape descriptors against ground-truth. Both outperform MDS and MDST, this is expected as MDS is insensitive to any mesh deformation which maintains the geodesic distance. Comparison of the different shape descriptors with respect to window size also shows that the Shape Histograms and Reeb Graphs are relatively insensitive to the change of window size, $N_t = 0 \rightarrow 9$. SHvrG and SHvrT performs equally well while SHvrT slightly drops at a larger window size. This has been explained previously as that the SHvrS with a local-frame alignment is not as robust as the SHvrG with a global-frame alignment and although SHvrG is more discriminative with an additional computational cost than SHvrT, SHvrG only performs marginally better than SHvrT [9]. Finally, we compare the relative computational cost between Shape Histograms and Reeb Graphs. As a Shape Histogram descriptor counts the number of voxels occupied by the object in each bin, the computational complexity is $O(N_g^3)$. A Reeb Graph descriptor computes geodesic distance using Dijkstra's shortest path algorithm. The computational complexity is $O(Nlog(N))$ using a binary tree implementation, where N is the number of vertices. In practice, the voxel resolution is chosen much lower than the surface mesh resolution: $N_g << N$. Typically, for a 3D mesh with $N = 100K$ vertices, the voxelization resolution is $N_g = 200$. Thus $Nlog(N) = 500K < N_g^3 = 8000K$, and therefore a Reeb Graph is about 10 times more efficient than a Shape Histogram.

4. Similarity Measurement on Real Data

In this section we apply the SHvrT and aMRGT to captured 3D video sequences of people. Real 3D video sequences were reconstructed from multiple camera video capture available as a public research database [15]. These include a street dancer (JP) performing complex movements with baggy clothing, a performer (Roxanne) wearing 3 different costumes with shorts, a short-dress and a long-dress performing a standard set of movements (Table 1). Captured 3D video sequences are unstructured meshes with unknown temporal correspondence and time varying mesh connectivity, topology and geometry. Evaluation has been performed for all available sequences with the same resolution parameters used for synthetic evaluation. A different temporal window size is used for Roxanne ($N_t = 4$) and for JP ($N_t = 9$). Example results presented in Figure 7

demonstrates typical results with identification of frames with similar shape and motion. Figure 7(a) for the street dancer JP performing complex movements shows there is a lot of visible structure in the similarity matrix produced by SHvrT and aMRGT, and frames with similar pose and motion are also correctly identified. In Figures 7(b,c) for Roxanne similarity computed by SHvrT and aMRGT clearly identify the periodic structure of the walking motion and identifies frames with similar shape and motion even with the highly non-rigid movement of the loose dress and long-hair. Figure 7(d) shows a false case of similar pose identification by aMRGT where both SHvrT and aMRGT correctly identify a walk pose (Frame 39 of "Stand2Walk") as similar to the query frame (Frame 68 of "WalkPose"), however, aMRG also identify a stand pose with a hand touching the body (Frame 9 of "Stand2Walk") as similar, which is incorrect. This is because the Reeb Graph is intrinsically sensitive to surface topology changes which occur due to errors in reconstruction as it is based on geodesic distance. In practice to reduce sensitivity to topology changes a relative low-resolution Reeb Graph is used ($r = 2$) decreasing the resolution results in a erroneous similarity. This evaluation on real 3D video sequences demonstrates that the temporal similarity by both of the SHvrT and the aMRGT identify similar frames for complex movement and loose clothing while the SHvrT is more robust to global topology change.

5. Applications

Concatenative Human Motion Synthesis. 3D shape similarity metrics can be used to identify optimal transitions between 3D video sequences. Human motion synthesis is then performed by concatenating existing 3D video sequences to novel character animation. This example-based method is attractive as there is no loss of detail of from the original motion dynamics. The quality of synthesis is dependent on the smoothness of transitions. SHvrT and aMRGT accurately identify transitions which allow seamless concatenation of 3D video sequences. Example of synthesised motion by Huang *et al*. [8] are provided as support materials.

3D Video Summarization and Compression 3D shape similarity metrics can be used to extract key-frames for a 3D video. The key-frames can be regarded as a concise summarization. The basic idea is grouping similar frames and selecting representative ones by analysing self-similarity matrix. Huang *et al*. [7] present a method to optimise the trade-off between rate (number of key-frames) against distortion (deviation from the original sequence). Figure 1 shows an example of key-frame extraction from a 3D video by this method using self-similarity matrices produced by SHvrT. Similarly, 3D shape similarity metrics can also be used to reduce information redundancy in 3D video sequence for compression [17].

6. Conclusion

A comprehensive performance evaluation of three shape similarity metrics for 3D video sequences of people has been presented. Existing skeleton-based Reeb-Graph [18] which give good recognition performance for rigid shape retrieval and non-skeletal Shape Histograms [9] which have previously been shown to give the best recognition performance for time-varying non-rigid shape retrieval [9] together with a bending-free descriptor MDS [3] have been evaluated. Temporal shape similarity are presented to overcome the ambiguity in independent frame-to-frame comparison. Evaluation against a ground-truth synthetic 3D video dataset demonstrates that Shape Histograms and Reeb Graph consistently give the best recognition performance for different actors and motions. They are then applied in the evaluation of similarity measurement on real data. Intuitively skeleton-based Reeb Graph could be expected to outperform purely shape-based metrics, however, in practice comparative evaluation of skeletal Reeb Graph descriptors against non-skeletal Shape Histogram descriptors demonstrates similar recognition performance. The Reeb Graph has the advantage of a structured representation of the articulated structure. However, in practice it is sensitive to change in surface topology due to reconstruction error in real 3D video sequences. Shape Histograms as a simpler representation demonstrate to have comparable recognition performance and are robust to surface topology changes for real 3D video sequences. In future work, other skeleton-based descriptors such as Medial Axis/Surface [13] will also be considered.

7. Acknowledgement

This work was supported by EPSRC Grant EP/E001351 "Video-based Animation of People" and the JST-CREST project "Creation of Human-Harmonized Information Technology for Convivial Society".

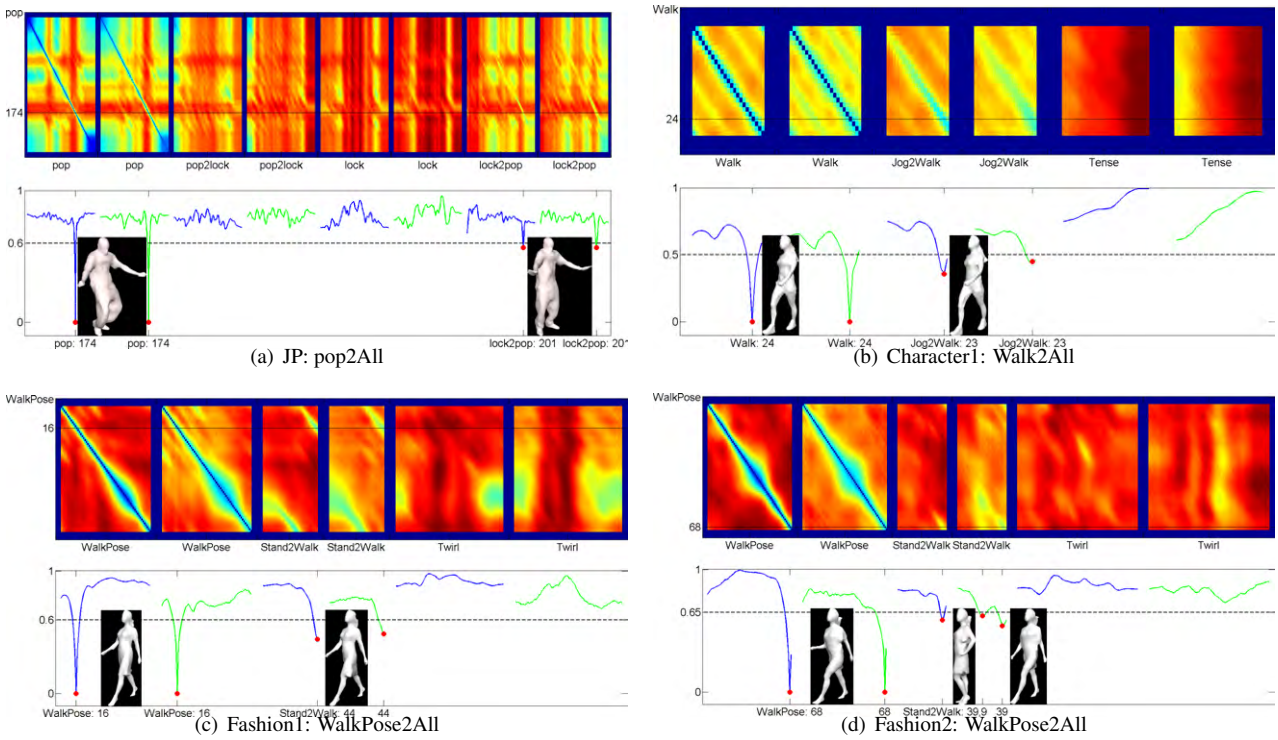
References

[1] M. Ankerst, G. Kastenmüller, H.-P. Kriegel, and T. Seidl. 3D shape histograms for similarity search and classification in spatial databases. In *SSD '99: Proceedings of the 6th International Symposium on Advances in Spatial Databases*, pages 207–226, London, UK, 1999. Springer-Verlag. 2

[2] I. Borg and P. J. F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, 2nd ed. edition, 2005. 3

[3] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Topology-invariant similarity of nonrigid shapes. *Int. J. Comput. Vision*, 81(3):281–301, 2009. 1, 3, 7

[4] B. Bustos, D. A. Keim, D. Saupe, T. Schreck, and D. V. Vranić. Feature-based similarity search in 3d object databases. *ACM Comput. Surv.*, 37(4):345–387, 2005. 1



(a) JP: pop2All (b) Character1: Walk2All

(c) Fashion1: WalkPose2All (d) Fashion2: WalkPose2All

Figure 7. Real data similarity measurement and a comparison between SHvrT and aMRGT. Similarity matrices produced by SHvrT (first) and aMRG (followed) are placed side by side. Similarity curves for selected query frame are shown below corresponding to similarity matrices (blue curve for SHvrT and green for aMRGT). Local minima after thresholding are marked as red points and retrieval frames by both SHvrT and aMRGT are shown near them.

[5] A. Elad and R. Kimmel. *Geometric methods in bio-medical image processing*, volume 2191. Springer, 2002. 3

[6] M. Hilaga, Y. Shinagawa, T. Kohmura, and T. L. Kunii. Topology matching for fully automatic similarity estimation of 3d shapes. In *SIGGRAPH '01: Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 203–212, New York, NY, USA, 2001. ACM. 4

[7] P. Huang, A. Hilton, and J. Starck. Automatic 3d video summarization: Key frame extraction from self-similarity. In *3DPVT '08: Proceedings of the Fourth International Symposium on 3D Data Processing, Visualization and Transmission*, Washington, DC, USA, 2008. IEEE Computer Society. 1, 7

[8] P. Huang, A. Hilton, and J. Starck. Human motion synthesis from 3d video. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1478–1485, June 2009. 1, 7

[9] P. Huang, A. Hilton, and J. Starck. Shape similarity for 3d video sequences of people. *International Journal of Computer Vision (IJCV) special issue on 3D Object Retrieval*, 2009. 1, 2, 4, 5, 6, 7

[10] N. Iyer, S. Jayanti, K. Lou, Y. Kalyanaraman, and K. Ramani. Three-dimensional shape searching: state-of-the-art review and future trends. *Computer-Aided Design*, 37(5):509–530, April 2005. 1

[11] G. Reeb. Sur les points singuliers d'une forme de Pfaff complétement intégrable ou d'une fonction numérique. *Comptes Rendus de L'Académie ses Sciences*, pages 847–849, 1946. 4

[12] E. Schwartz, A. Shaw, and E. Wolfson. A numerical solution to the generalized mapmaker's problem: flattening nonconvex polyhedral surfaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(9):1005–1008, Sep 1989. 3

[13] K. Siddiqi, J. Zhang, D. Macrini, A. Shokoufandeh, S. Bouix, and S. Dickinson. Retrieving articulated 3-d models using medial surfaces. *Mach. Vision Appl.*, 19(4):261–275, 2008. 7

[14] J. Starck and A. Hilton. Model-based multiple view reconstruction of people. *ICCV '03: Proceedings of the Ninth International Conference on Computer Vision*, pages 915–922, 2003. 5

[15] J. Starck and A. Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 27(3):21–31, 2007. 1, 6

[16] J. W. H. Tangelder and R. C. Veltkamp. A survey of content based 3d shape retrieval methods. In *SMI '04: Proceedings of the Shape Modeling International 2004*, pages 145–156, Washington, DC, USA, 2004. IEEE Computer Society. 1

[17] T. Tung and T. Matsuyama. Topology dictionary with markov model for 3d video content-based skimming and description. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 469–476, June 2009. 1, 7

[18] T. Tung and F. Schmitt. The augmented multiresolution reeb graph approach for content-based retrieval of 3d shapes. *International Journal of Shape Modeling (IJSM)*, 11(1):91–120, June 2005. 1, 4, 7