Int J Comput Vis
DOI 10.1007/s11263-010-0319-9

Shape Similarity for 3D Video Sequences of People

Peng Huang · Adrian Hilton · Jonathan Starck

Abstract This paper presents a performance evaluation of shape similarity metrics for 3D video sequences of people with unknown temporal correspondence. Performance of similarity measures is compared by evaluating Receiver Operator Characteristics for classification against ground-truth for a comprehensive database of synthetic 3D video sequences comprising animations of fourteen people performing twenty-eight motions. Static shape similarity metrics shape distribution, spin image, shape histogram and spherical harmonics are evaluated using optimal parameter settings for each approach. Shape histograms with volume sampling are found to consistently give the best performance for different people and motions. Static shape similarity is extended over time to eliminate the temporal ambiguity. Time-filtering of the static shape similarity together with two novel shape-flow descriptors are evaluated against temporal ground-truth. This evaluation demonstrates that shape-flow with a multi-frame alignment of motion sequences achieves the best performance, is stable for different people and motions, and overcome the ambiguity in static shape similarity. Time-filtering of the static shape histogram similarity measure with a fixed window size achieves marginally lower performance for linear motions with the same computational cost as static shape descriptors. Performance of the temporal shape descriptors is validated for real 3D video sequence of nine actors performing a variety of movements. Time-filtered shape histograms are shown to reliably identify frames from 3D video sequences with similar shape and motion for people with loose clothing and complex motion.

Keywords Temporal shape similarity · 3D video · Surface motion capture · Human motion

1 Introduction

Three-dimensional (3D) shape matching has been widely investigated (Bustos et al. 2007; Del Bimbo and Pala 2006; Iyer et al. 2005; Tangelder and Veltkamp 2004) as a means of effective and efficient object retrieval. However, shape matching techniques typically only consider a single static shape and are designed to classify objects from different classes. In this paper we consider the problem of 3D shape matching in temporal sequences where the goal is to discriminate between the same object in different poses rather than different classes of objects.

Multiple view reconstruction of human performance as a 3D video sequence has received considerable interest over the past decade following the pioneering work of Kanade et al. (1997). This research has advanced to the stage of capturing detailed non-rigid dynamic surface shape of the body, clothing and hair during motion (Aguiar et al. 2008; Vlasic et al. 2008; Starck and Hilton 2007; Theobalt et al. 2007). Acquisition results in an unstructured volumetric or mesh approximation of the surface shape at each frame without temporal correspondence. Recent research has introduced data-driven animation synthesis where subsequences of captured motions are concatenated to construct highly-realistic animated content (Huang and Hilton 2009;

P. Huang (✉) · A. Hilton · J. Starck
Centre for Vision, Speech and Signal Processing (CVSSP),
University of Surrey, Guildford, GU2 7XH, UK
e-mail: p.huang@surrey.ac.uk

A. Hilton
e-mail: a.hilton@surrey.ac.uk

J. Starck
e-mail: j.starck@surrey.ac.uk

Published online: 05 February 2010

Starck and Hilton 2007; Xu et al. 2006; Starck et al. 2005). This requires a measure of temporal shape similarity to identify possible intra and inter sequence transitions which are suitable for concatenation without unnatural intermediate motion.

Previous research in concatenative synthesis of human motion has considered only the similarity in pose of the human skeleton (Arikan et al. 2003; Kovar et al. 2002; Lee et al. 2002). This does not account for surface shape deformations in clothing and hair. Surface similarity has been defined either manually (Starck and Hilton 2007; Starck et al. 2005) or through a shape descriptor (Xu et al. 2006). Similarity requires a shape descriptor that is sufficiently distinct to differentiate articulated pose and motion while tolerant to changes in surface topology for similar poses.

In this paper we review and compare current techniques from the shape retrieval literature for the problem of human surface shape similarity which we call static shape matching and extend them to the spatio-temporal domain which we call temporal shape matching. This paper extends the evaluation of static (Huang et al. 2007a) and temporal (Huang et al. 2007b) shape descriptors for 3D video sequences of people from individual sequences to a comprehensive corpus of both synthetic and real data for different people, motions and clothing. Shape descriptors are evaluated for matching shape and motion against known ground-truth on synthetic 3D video sequences for animated models of 14 people each performing 28 different motions giving a total of $40K$ frames. Comparison is made between local feature distribution techniques including: Shape Distribution (Osada et al. 2002), Spin Image (Johnson and Hebert 1999), Shape Histogram (Ankerst et al. 1999) and Spherical Harmonics (Kazhdan et al. 2003) assuming unknown correspondence. These techniques are extended to the spatio-temporal domain by applying a temporal filter and 4D shape-flow. Performance is evaluated by comparing the Receiver Operating Characteristic (ROC) showing the trade-off between correctly and incorrectly classified similarity. This comparison for a wide variety of people and movements validates the previous observation for a single person (Huang et al. 2007b) that the best performance for static shape matching is achieved by a volume-based shape histogram descriptor.

Novel temporal shape-flow descriptors are introduced extending the volume-based Shape Histogram descriptor to the temporal domain. Evaluation of the novel shape-flow descriptors against ground-truth demonstrates improved performance over previous static shape descriptors and time-filtered shape descriptors with consistent classification for 3D video sequences of different people with a wide variety of movement, body-shape and clothing. Temporal shape matching is demonstrated on real 3D video sequences of 9 people each performing 6–10 different motions from a public data base (Starck and Hilton 2007) with a total of $5K$ frames. Real sequences include a variety of loose and tight fitting clothing together with long sequences of complex motions from a street dancer (Fig. 1). Results demonstrate that the proposed temporal shape descriptor correctly identifies 3D video frames with similar shape and motion.

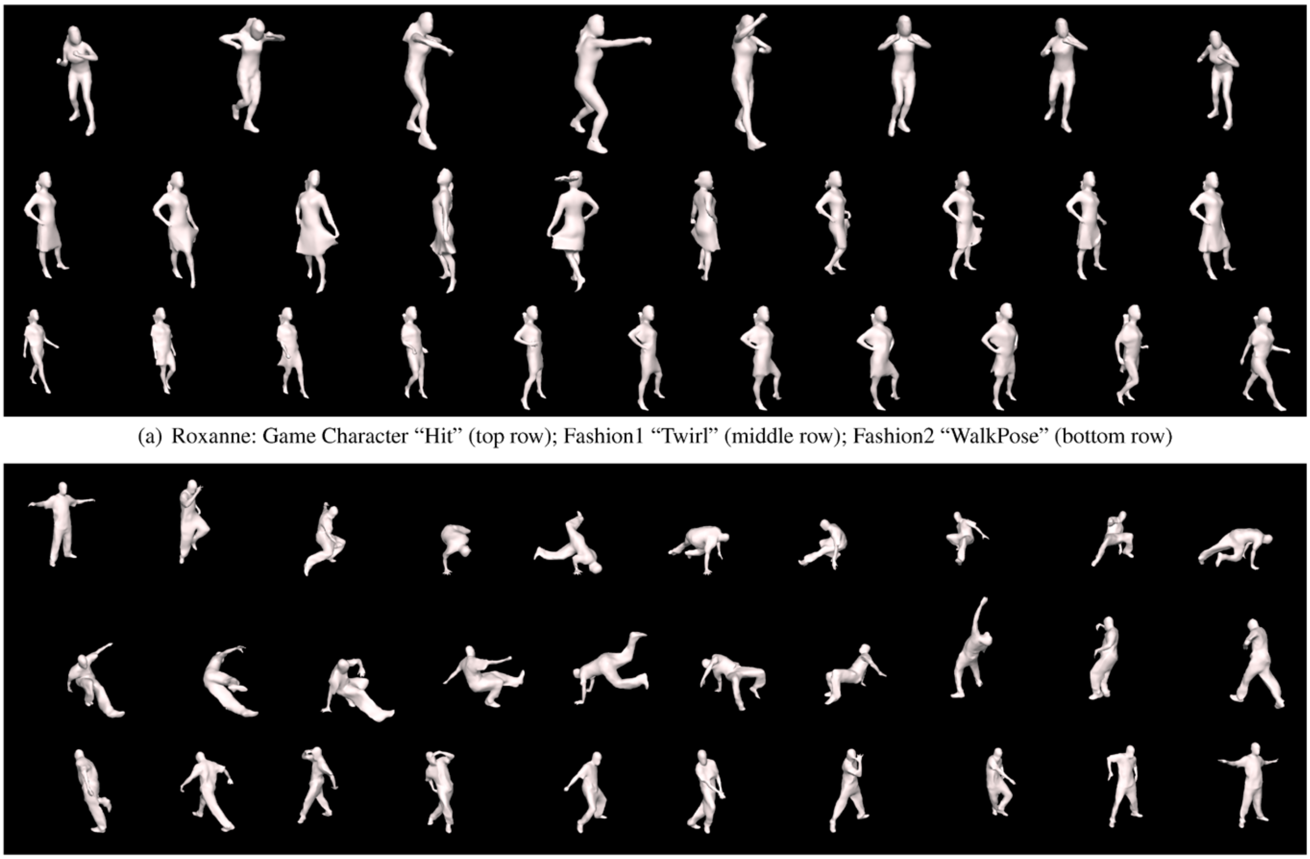
2 Related Work

The problem of shape similarity has been widely studied in the 3D shape retrieval literature. These descriptors aim to discriminate between rigid shapes for different object classes (book, mug, chair) and inter-class variations (cars, chairs). This paper focuses on shape descriptors to discriminate between instances from sequences of the same moving non-rigid object, a person, which differ in both shape and motion. The temporal shape descriptor extends previous approaches for measuring static shape similarity to temporal shape sequences. In this section we first review static shape matching techniques followed by approaches related to temporal matching of both human skeletal motion and shape sequences.

2.1 Static Shape Matching

Global features are used to characterise the overall shape of 3D models. Typical global features include: volume, surface area, moments, Fourier and Wavelet coefficients. Zhang and Chen (2001) propose an algorithm to efficiently calculate these global features of a 3D model directly from a surface mesh representation. Paquet (2000) provide three global feature-based descriptors for 3D shape matching, a cord-based descriptor, moment-based descriptor and wavelet-based descriptor. Corney et al. (2002) use three convex-hull based indices hull crumpliness, hull packing and hull compactness for coarsely filtering candidates prior to a more detailed analysis. Kazhdan et al. (2002) present a reflective symmetry descriptor that extracts the global symmetry information. Such global features are relatively simple to compute but do not provide discrimination at a local level.

Local features can give a more distinctive similarity measure. Shum et al. (1996) define similarity as the L_2 distance between the local curvature distribution over the mesh representation for two 3D objects. Zaharia and Preteux (2001) present the 3D Shape Spectrum Descriptor (3D SSD), which is defined as the distribution of a shape index over the entire mesh, to provide an intrinsic shape description of a 3D mesh. Chua and Jarvis (1997) provide a point signature to describe 3D free-form surfaces that is invariant to rotation and translation. Johnson and Hebert (1999) present a 3D shape-based object recognition system using Spin Images. These features provide local shape information to improve discrimination between similar shapes.



(a) Roxanne: Game Character “Hit” (top row); Fashion1 “Twirl” (middle row); Fashion2 “WalkPose” (bottom row)

(b) JP: “Free”

Fig. 1 Example frames from 3D video sequences of two people performing a variety of movements with different clothing

Local features are compared using a descriptor of the feature distribution. Osada et al. (2002) introduced a Shape Distribution as a signature to discriminate similar and dissimilar models. A Similarity Measure is computed as the difference between Shape Distributions, which is invariant to translation, rotation and tessellation of the 3D polygonal model. Ankerst et al. (1999) use a 3D Shape Histogram as a shape signature to classify a molecular database. Körtgen et al. (2003) attach a 3D Shape Context descriptor to each surface sample point. Shape Context was introduced by Belongie et al. (2002) for 2D matching. Ohbuchi et al. (2003) introduce two further shape descriptors, Angle Distance (AD) and Absolute Angle Distance (AAD) histograms for 3D matching.

Another popular approach is a transform-based representation which describes shapes in a transformation invariant manner. Kazhdan et al. (2003) propose Spherical Harmonic Descriptors that are invariant to rotation for 3D shape retrieval. However, the representation has a potential ambiguity problem. The frequency decomposition is performed independently in concentric spheres, such that two different shapes can have the same spherical harmonic representation. Novotni and Klein (2003) use 3D Zernike Descriptors for 3D shape retrieval, which is an extension of the Spherical Harmonic Representation. A set of descriptors is obtained that are orthonormal, complete and rotation invariant. However, 3D Zernike Descriptors suffer the same ambiguity problem.

In the CAD industry, the most common used graph-based representations are Boundary Representation (B-rep) and Constructive Solid Geometry (CSG). El-Mehalawi (2003) construct an attributed graph from a B-rep and measure similarity by using an inexact graph matching algorithm. Similarly, Mcwherter et al. (2001) compare models based on shape using information extracted from B-rep into Model Signature Graphs. However, these methods are limited to the CAD community, for example, in matching mechanical parts, and cannot apply to commonly used 3D mesh representations. Sundar et al. (2003) use a skeletal graph which encodes both the geometric and topological information in the surface to match and compare 3D models. Hilaga et al. (2001) propose a method based on Multi-resolutional Reeb Graphs (MRGs) to estimate a measure of similarity and correspondence between 3D shapes. The similarity is calculated with a coarse-to-fine strategy using the attributes of nodes in the MRG and topological consistency.

View-based methods represent objects by their image-plane projection. Chen et al. (2003) introduce the Light-Field Descriptor in which the appearance of an object is characterised by the projected appearance in a set of camera views. Similarity is computed by rotating the camera system surrounding each model until the highest overall similarity (cross-correlation) between the two models from all viewing angles is reached. The similarity between two 3D models is defined as the summation of the similarities across all the corresponding 2D images.

Bending-invariant techniques have been proposed to retrieve similar objects independent of changes in articulated pose. Elad and Kimmel (2003) present a method to construct a bending invariant signature for these models. They utilise the geodesic distance between surface points as an invariant to surface bending. A bending invariant surface is generated by transforming the geodesic distances between points into Euclidean ones (via an MDS procedure). They translate the problem of matching non-rigid objects in various postures into a simpler problem of matching rigid objects. Jain and Zhang (2007) present an approach to robust shape retrieval from databases containing articulated 3D models. Each shape is represented by the eigenvectors of a shape affinity matrix defining the geodesic surface distance between model points. This gives a spectral embedding which achieves normalisation against rigid-body transformations, uniform scaling, and shape articulation.

2.2 Temporal Shape Matching

The acquisition of temporal 3D surface sequences from multiple view video has received considerable interest over the past decade following the work of Kanade et al. (1997). Research has primarily focused on methods for multiple view reconstruction, structured representation and realistic rendering (Aguiar et al. 2008; Vlasic et al. 2008; Starck and Hilton 2003, 2007; Zitnick et al. 2004; Carranza et al. 2003). Advances in this field have led to the availability of 3D video data sets of actors performing multiple motions which support high-quality rendering with interactive viewpoint control. Reuse of captured 3D video sequences for concatenative synthesis of novel animations (Starck and Hilton 2007; Xu et al. 2006; Starck et al. 2005) requires temporal shape matching to identify transition points. Temporal shape matching for 3D video sequences has received limited investigation. Related work can be found in the literature on human motion recognition and concatenative animation from video or marker-based human motion capture.

Schödl et al. (2000) introduced video textures which identify transition points in a video sequence based on appearance similarity to produce an extended video sequence. Transition points are found by temporal matching of sub-sequences of the video to preserve the dynamics of the motion. In practice, such a sub-sequence match is achieved by time filtering of the frame-by-frame similarity matrix with a diagonal kernel. Similarity metrics based on 2D image differences cannot be directly extended to 3D time-varying surfaces.

A number of researchers have addressed the problem of temporal similarity for skeletal motion for concatenative synthesis. In the case of skeletal motion the temporal correspondence is known and similarity is evaluated from difference in joint angle or position together with their velocity and acceleration. Lee et al. (2002) modelled human skeletal motion data as a first-order Markov process and the probability of transitioning from one frame to another is estimated from a measure of similarity. The cost function is the sum of weighted differences of joint angles and joint velocities. The velocity term helps to preserve the dynamics of motion. Gleicher et al. (2003) developed methods to synthesize human motions from articulated motion sequences by piecing together existing motion clips. Transitions are located by matching the point clouds over two windows of frames. Each point cloud is formed by attaching markers to the skeletons representing the pose at each frame. Arikan et al. (2003) described a framework to synthesize motions from articulated data by assembling frames from a motion database. The distance between frames is measured as the squared distance between feature vectors extracted from the skeleton. Temporal similarity is achieved by including velocities and accelerations for every joint in the feature vectors. Similarity metrics on the skeletal motion cannot be directly applied to the surface sequences in 3D video.

In human motion recognition, volumetric analysis of video, where a sequence of images is treated as a 3D space-time volume, is widely used. Video features are then extracted: Bobick and Davis (2001) combine Motion-Energy Images (MEI) and Motion-History Images (MHI) as temporal motion templates for human movement recognition. Efros et al. (2003) propose a pixel-wise optical-flow motion descriptor which is measured in a figure-centric spatio-temporal volume for each person, to obtain a motion-to-motion similarity matrix and time-filter the frame-to-frame similarities. Gorelick et al. (2007) regard human actions as 3D shapes induced by the silhouettes in a space-time volume, extracting features such as local space-time salience, action dynamics, shape structure and orientation. Weinland et al. (2006) proposed a free-viewpoint representation for human action based on a multi-camera system using Motion History Volumes (MHV) where alignment and comparison are performed under a Fourier transform in cylindrical coordinates around the vertical axis. For further reading on motion recognition refer to Krüger et al. (2007) and on whole body human motion synthesis refer to Lee et al. (2002).

3 Shape Descriptors for 3D Video Sequences

In this section, we review static shape descriptors evaluated for similarity measurement in 3D video sequences of people and novel spatio-temporal shape descriptors. Static descriptors are presented first from the literature. The extension to temporal shape matching is then presented using a simple temporal filter and by extending the static descriptors to 4D shape-flow descriptors in the temporal domain.

3.1 Static Shape Descriptor

Evaluation of shape similarity is restricted here to local feature distributions. Global features provide only a coarse descriptor that is insufficient to distinguish similarity in a time varying sequence where an object can have the same global properties for a relatively large proportion of the time. In this section we briefly describe four widely used feature distribution descriptors previously introduced for general static shape matching which are evaluated for 3D video sequences.

3.1.1 Shape Distribution (Osada et al. 2002)

Shape Distribution (SD) provides a shape signature as a probability distribution of a shape function that measures some geometric properties of a 3D model. Typical shape functions are the angle, distance and area for randomised points on the model surface. Here we adopt the $D2$ measure, the distance between two random points on the surface, as proposed by Osada et al. (2002). Similarity is measured as the L_2 distance between the distribution $D2$ defined for two meshes. Figure 2 illustrates the Shape Distribution representation computed for a single frame of a 3D video sequence of a person. Given a 3D mesh representation the descriptor is constructed as follows:

Algorithm 1 (Shape Distribution)

1. Distance is iteratively measured between two random points on the surface.
2. A 1D histogram is created to count the number of point-pairs at different distances.
3. The final histogram is normalised.

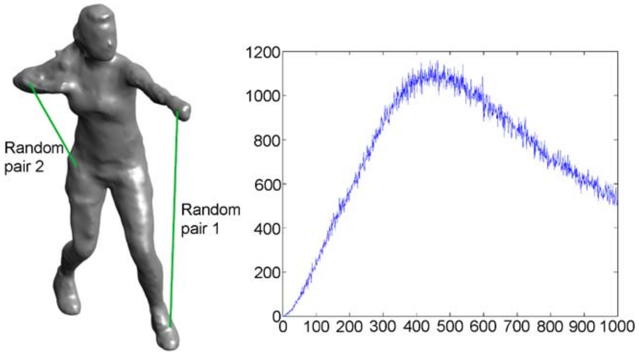


Fig. 2 Illustration of the Shape Distribution

3.1.2 Spin Image (Johnson and Hebert 1999)

A Spin Image (SI) is a 2D histogram which encodes the density of mesh vertices projected onto an object-centred space. Given a 3D surface mesh consisting of a set of oriented points corresponding to the mesh vertices, the histogram is constructed as follows:

Algorithm 2 (Spin Image)

1. An object-centred coordinate (α, β) is computed for each vertex according to the distance α along and the distance β from the principal axis of the object.
2. A 2D accumulator indexed by (α, β) is created and the accumulator is incremented for each vertex within the support of the spin image.
3. The final histogram is normalised.

The centre of mass and the first axis of the Principal Component Analysis (PCA) of the distribution of mesh vertices is used to define the object-centred coordinate system. Figure 3 illustrates the Spin Image for a single frame of a 3D video sequence, showing the histogram distribution resulting from a plane rotated about a vertical axis through the centroid of the shape.

3.1.3 Shape Histogram (Ankerst et al. 1999)

A Shape Histogram (SH) partitions the space containing an object into disjoint cells corresponding to the bins of a his-

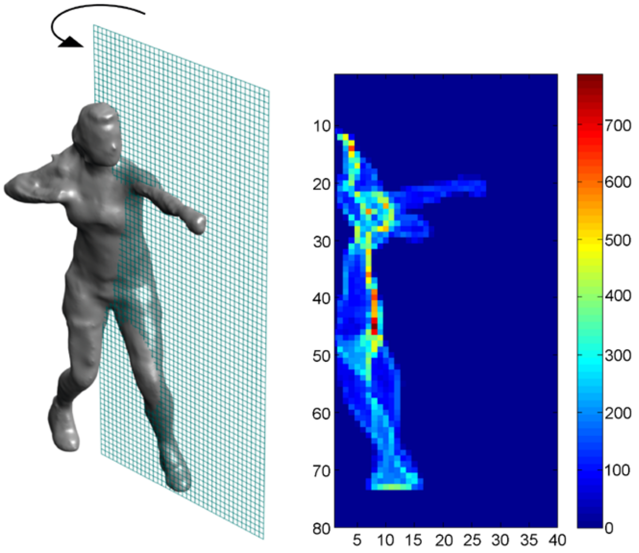


Fig. 3 Illustration of the Spin Image

togram. Given a 3D surface mesh, a volume sampling spherical histogram is constructed as follows:

Algorithm 3 (Shape Histogram)

1. A volumetric representation is constructed by rasterising the surface into a set of voxels that lie inside the model.
2. Space is transformed to a spherical coordinate system (r, ϕ, θ) around the centre of mass for the model.
3. A 3D spherical histogram is constructed, accumulating the voxels in the volume representation.
4. The final histogram is normalised.

The spherical coordinate histogram is compared invariant of rotation by testing similarity for all feasible rotations. Human models are assumed to have an upright direction and instead of rotating 3D mesh, we generate a fine histogram first, then shift the histogram with 1° resolution, and re-bin to a coarse histogram. A similarity measure is computed as the minimal of L_2 distance between the coarse histograms.

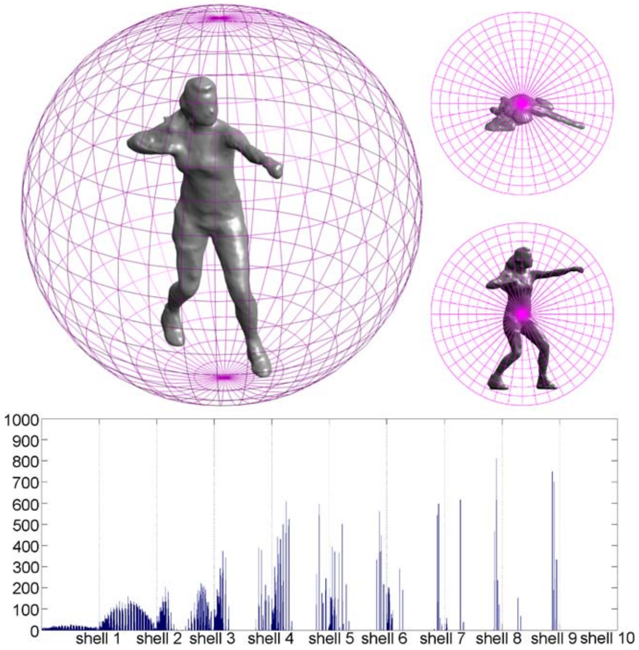


Fig. 4 Illustration of the Shape Histogram

The shape histogram representation with radial and spherical bins is illustrated in Fig. 4.

3.1.4 *Spherical Harmonics (Kazhdan et al. 2003)*

The Spherical Harmonic Representation (SHR) describes an object by a set of spherical basis functions. A descriptor is constructed by measuring the energy contained in different frequency bands, where the frequency components are rotation invariant. The descriptor is constructed as follows:

Algorithm 4 (Spherical Harmonics)

1. The volume of an object is divided into a set of concentric shells.
2. The frequency decomposition in each shell is computed.
3. The norm for each frequency component at each radius is concatenated into a 2D histogram indexed by radius and frequency.

The resolution of the shape descriptor is defined by the number of shells defining the radii (r) and the preserved bandwidth (bw) in the spherical harmonics. A similarity measure is computed as the L_2 distance between the histograms. Figure 5 illustrates the SHR functions and coefficient distribution for the single 3D video frame of a person.

3.1.5 *Static Shape Similarity for 3D Video*

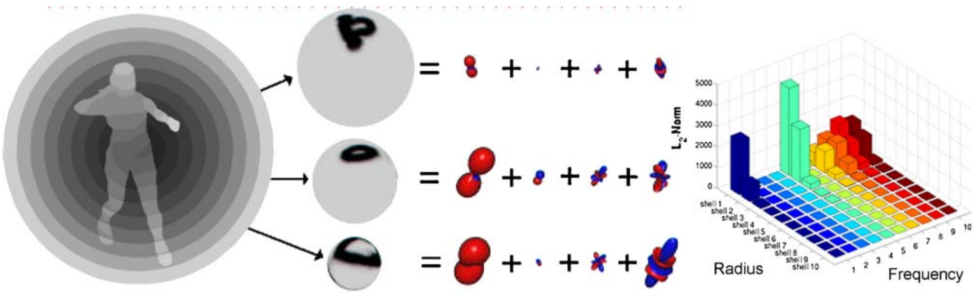
The shape descriptors presented in the previous section can be used to define a similarity measure. Given two individual frames x, y of 3D video sequences and their descriptors x', y', frame-to-frame similarity is defined as follows,

$$s(x, y) = d(x', y') \tag{1}$$

where function $d(x', y')$ computes the L_2 distance between descriptors x', y' for Shape Distribution, Spin Image and Spherical Harmonics,

$$d_{SD,SI,SHR}(x', y') = \|x' - y'\| \tag{2}$$

Fig. 5 Illustration of the Spherical Harmonic Representation



for Shape Histogram, let $x'(\theta)$ denote the histogram shifted with $\theta°$. Function $d(x'(\theta), y')$ computes the minimal distance by shifting the histogram with 1° resolution, i.e. $\theta = 0, 1, \ldots, 359$,

$$d_{SH}(x', y') = \min_{\theta} \|x'(\theta) - y'\| \tag{3}$$

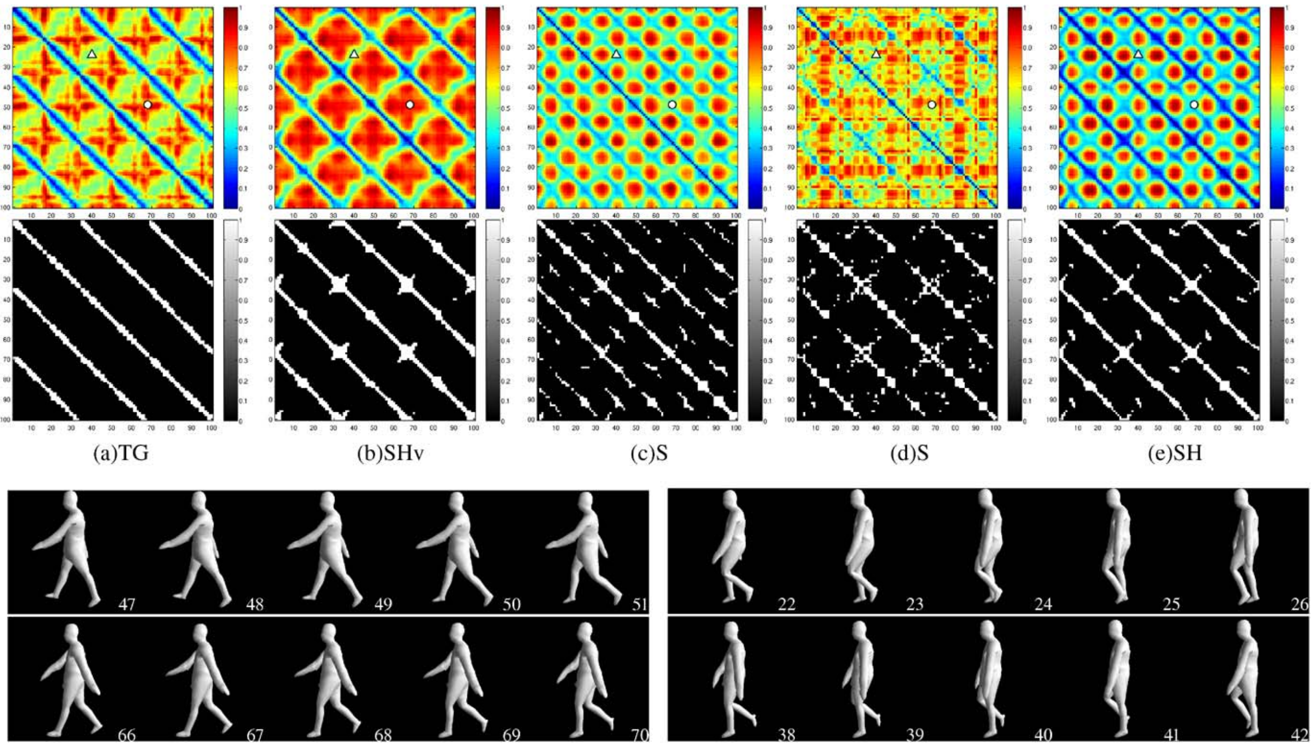
Given two sequences of 3D video $X = \{x_i\}$ and $Y = \{y_j\}$, the frame-to-frame similarity matrix S is defined as follows,

$$S(i, j) = s(x_i, y_j) \tag{4}$$

As illustrated in previous work (Cutler and Davis 2000) self-similarity is demonstrated for periodic motion. Figure 6(a) shows the self-similarity and classification evaluated using the known ground truth surface correspondence for shape and motion (Sect. 4). The periodic structure of the walking motion is illustrated by the diagonal lines of high similarity (dark blue). The self-similarity matrix for the four static shape descriptors computed independently frame-to-frame without known correspondence is shown in Fig. 6(b–e). The static shape similarity for all descriptors gives high similarity (dark blue) diagonal lines corresponding to the ground-truth periodic motion structure. Additional lines of high-similarity occur due to ambiguities in the static shape descriptor. Anti-diagonal lines of high-similarity occur with all static shape descriptors due to frames with similar shape but opposing motion such as the mid-point of the walk cycle (marked as a triangle) illustrated in Fig. 6(g). For the Shape Distribution and Spherical Harmonic similarity measures there is also a periodic line structure of high-similarity at twice the motion frequency in the diagonal direction due to mirror ambiguity where a shape and its mirror image have the same descriptor. An example of this (marked as a circle) is illustrated in Fig. 6(f) where frames 47–51 are dissimilar to frames 66–70 but similar to their mirror image.

3.2 Temporal Shape Descriptors

Extension of static shape descriptors to include temporal motion information is required to remove the ambiguities inherent in static shape descriptors for comparing 3D video sequences of similar shape. In this section we first extend static shape descriptors to the time domain by temporal filtering (Huang et al. 2007b) and introduce two novel shape-flow descriptors with global and local alignment of frames.



(a)TG (b)SHv (c)S (d)S (e)SH

(f) top row: Frames 47–51; bottom row: Frames 66–7

(g)toprow:Frames22–26;bottomrow:Frames38–4

Fig. 6 Static similarity measure for motion "Fast Walk" in a straight line compared with itself. Self-similarity and classification obtained by (**a**) Temporal Ground-Truth (TGT). Self-similarity and classification at $FPR = 5\%$ (Sect. 4.2) obtained by (**b**) the rotated volume-sampling Shape Histogram (SHvr); (**c**) Shape Distribution (SD); (**d**) Spin Image (SI); (**e**) Spherical Harmonics Representation (SHR). Example frames show (**f**) sub-sequences around frames 49 and 68 (centre) with "mirror ambiguity"; (**g**) sub-sequences around frames 24 and 40 (centre) with similar shape but different direction of motion for arms and legs

3.2.1 Time-Filtered Descriptors

Time information can be incorporated in a static 3D shape descriptor using a temporal filter (Huang et al. 2007b). Previously a similar strategy has been used to achieve motion-to-motion matching in video (Schödl et al. 2000; Efros et al. 2003). In practice, the time filter is applied to the frame-to-frame similarity matrix obtained from the static Shape Descriptors. Temporal shape similarity is obtained by convolving the static shape similarity with a time filter,

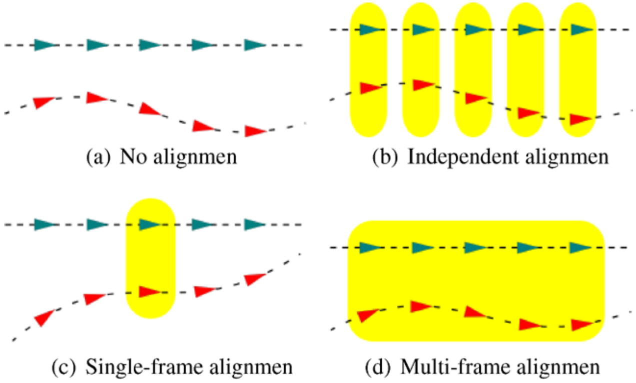
$$S_T(i,j) = S \otimes T(N_t) = \frac{1}{2N_t+1}\sum_{k=-N_t}^{N_t} S(i+k, j+k) \quad (5)$$

where S is the frame-to-frame similarity matrix and $T(N_t)$ is a time filter with window size $2N_t+1$, $T(N_t) = 1/(2N_t+1) * I$. The computational complexity of the time filtered static shape descriptor is the cost of computing the frame-to-frame static shape similarity together with a convolution of the resulting similarity matrix with the temporal filter. The cost is dominated by the cost of computing the static shape similarity with a relatively small additional cost of time filtering.

Time-filtering emphasises the diagonal structure of the similarity matrix and reduces minima in the anti-diagonal direction resulting from motion and mirror ambiguities in the static shape descriptor (Huang et al. 2007b). Figure 7(c–f) illustrates the effect of time-filtering with increasing temporal window size for each of the shape descriptors on a periodic walking motion. Comparison with the temporal ground-truth Fig. 6(a) shows that the incorrect shape similarity in the anti-diagonal direction which occur with static shape similarity is reduced. Performance evaluation of the time-filtered shape descriptors is presented in Sect. 4.5.1.

3.2.2 Shape-Flow Descriptors

Two new temporal shape descriptors are introduced to measure the change in shape for a surface in a sub-sequence corresponding to a given time window. Time filtering of the static shape similarity matrix breaks the temporal consistency in a motion as each static comparison is aligned independently on a frame-by-frame basis as illustrated in Fig. 8(b). The new descriptors consider not only the similarity between individual frames in a sub-sequence but also preserve the temporal changes using a sub-sequence alignment, referred to as a shape-flow descriptor.

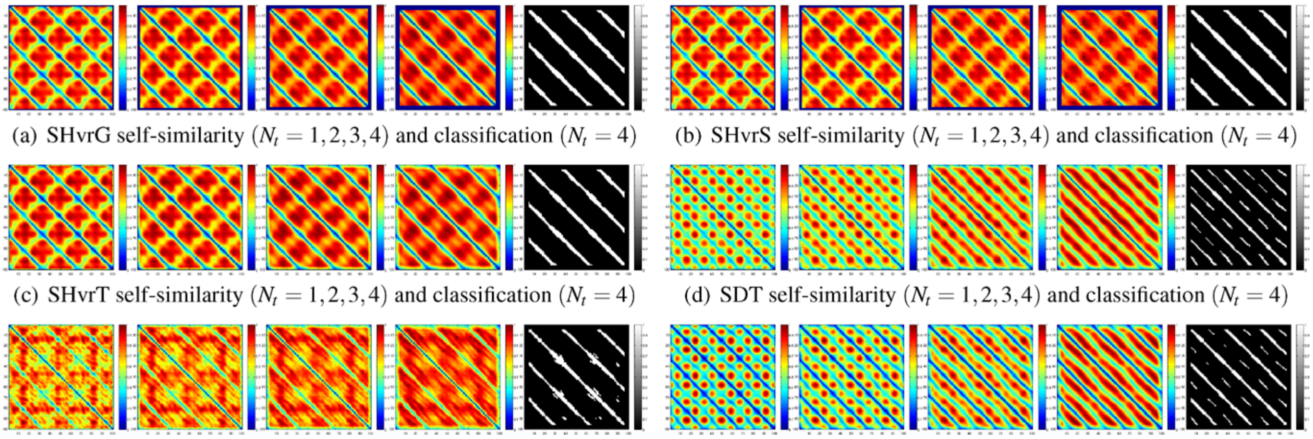

(a) No alignmen

(b) Independent alignmen

(c) Single-frame alignmen

(d) Multi-frame alignmen

Fig. 8 Shape-flow matching. Sequences (**a**) before applying any alignment; (**b**) after applying independent alignment for each frame used in static shape similarity; (**c**) after single-frame shape-flow matching; (**d**) after multi-frame shape-flow matching


(a) SHvrG self-similarity ($N_t = 1,2,3,4$) and classification ($N_t = 4$)

(b) SHvrS self-similarity ($N_t = 1,2,3,4$) and classification ($N_t = 4$)

(c) SHvrT self-similarity ($N_t = 1,2,3,4$) and classification ($N_t = 4$)

(d) SDT self-similarity ($N_t = 1,2,3,4$) and classification ($N_t = 4$)

(e) SIT self-similarity ($N_t = 1,2,3,4$) and classification ($N_t = 4$)

(f) SHRT self-similarity ($N_t = 1,2,3,4$) and classification ($N_t = 4$)

Fig. 7 Temporal similarity measure for motion "Fast Walk" in a straight line compared with itself. Self-similarity with window size 3, 5, 7, 9 and classification with window size 9 at $FPR = 5\%$ (Sect. 4.2) obtained by (**a**) multi-frame alignment volume-sampling spherical Shape Histogram (SHvrG); (**b**) single-frame alignment volume-sampling spherical Shape Histogram (SHvrS); (**c**) Temporal filtered volume-sampling spherical Shape Histogram (SHvrT); (**d**) temporal filtered Shape Distribution (SDT); (**e**) Temporal filtered Spin Image (SIT); (**f**) Temporal filtered Spherical Harmonics Representation (SHRT)

Performance evaluation of static shape descriptors (Huang et al. 2007a) extended in Sect. 4.4 demonstrates that the volume sampling 3D spherical histogram gives the best performance in classifying shape similarity. The 3D histogram is extended here to incorporate changes in shape using a 4D histogram in which each 3D spherical bin records the shape over a 1D time window. Similarity is again defined using the L_2 distance between coarse histograms after alignment.

Histogram alignment is considered using two methods, either by finding the optimal alignment of the 3D descriptor for the centre frame of the temporal window, or by finding the optimal alignment of the entire sequence in the 4D temporal descriptor. We call the first method single-frame shape-flow matching and the second multi-frame shape-flow matching. Figure 8(c, d) illustrates the alignment in the time-filtered, single-frame and multi-frame shape-flow similarity. Single-frame shape-flow matching has the same computational complexity as static shape matching but may not find the optimal alignment for the whole sub-sequence. Multi-frame shape-flow matching is more robust but the computational cost is proportional to the time window size. Comparative performance evaluation of the temporal shape-flow descriptors is presented in Sect. 4.5.2.

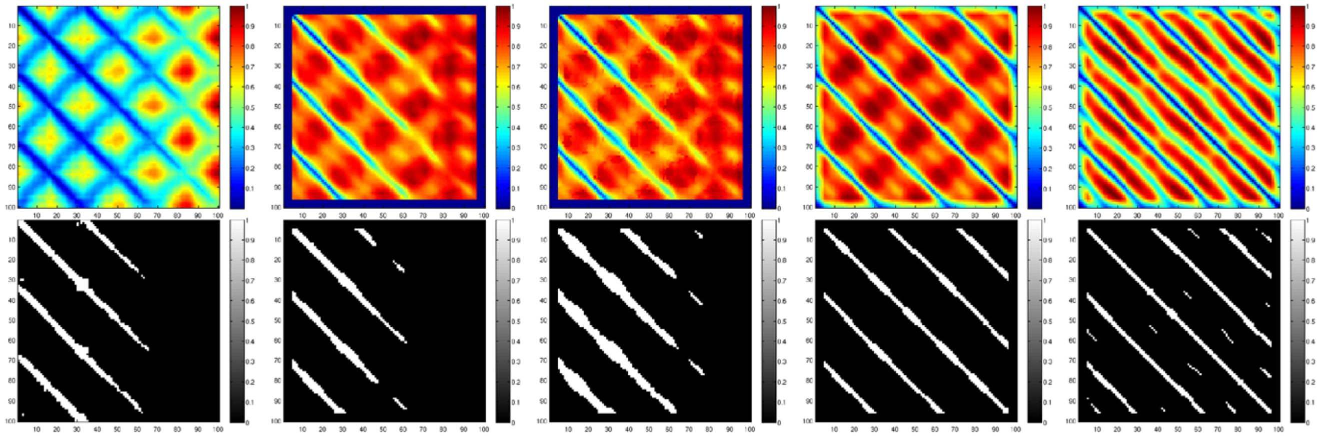
Optimal alignment is derived first by finding the translation that matches the centre of mass and then by shifting the histogram[1] to give the greatest similarity. Let x'_i and y'_j be 3D shape histograms of individual frames in two motions $X = \{x_i\}$ and $Y = \{y_j\}$, $d(x'_i(\theta), y'_j)$ computes the L_2 distance between $x'_i(\theta)$ with a shift θ about the vertical axis and y'_j with no shift. The similarity matrix for single-frame shape-flow matching $S_{SHvrS}(i, j)$ is defined as follows,

$$S_{SHvrS}(i, j) = \frac{\sum_{k=-N_t}^{k=N_t} \|x'_{i+k}(\mathrm{argmin}_\theta \|x'_i(\theta) - y'_j\|) - y'_{j+k}\|}{2N_t + 1} \tag{6}$$

For multi-frame shape-flow matching, the optimal rotation is found by searching for the rotation that minimises the distance between two sub-sequences and the similarity matrix $S_{SHvrG}(i, j)$ is computed as follows:

$$S_{SHvrG}(i, j) = \min_\theta \frac{\sum_{k=-N_t}^{k=N_t} \|x'_{i+k}(\theta) - y'_{j+k}\|}{2N_t + 1} \tag{7}$$

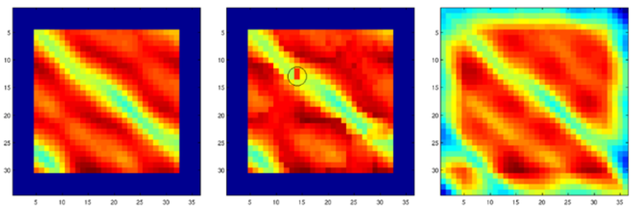
The effect of the shape-flow descriptors for a walking motion is illustrated in Figs. 7 and 9. For motion in a straight line Fig. 7(a–c) similar results are obtained for shape-flow with single and multiple frame alignment and the temporal filtering with independent alignment of frames. Comparison of the shape-flow descriptors (SHvrG, SHvrS) and the time-filtered descriptor (SHvrT) based on shape histograms with other time filtered descriptors (SDT, SIT, SHRT) Fig. 7(d–f) shows reduced temporal ambiguity with similarity and classification close to ground-truth 6(a). Distinction between the performance of shape-flow and time-filtered descriptors can be seen when the motion is not in a straight line. Figure 9 illustrates the cross-similarity between walking in a straight line and on a spiral for the ground-truth and temporal shape descriptors. Figure 9(d, e) show that time-filtered descriptors SHvrT and SHRT fail to correctly characterise



(a) TGT (b) SHvrG, N_t = 4 (c) SHvrS, N_t = 4 (d) SHvrT, N_t = 4 (e) SHRT, N_t = 4

Fig. 9 Temporal similarity measure for motion "Fast Walk" in a *straight line* compared with motion "Fast Walk" on a spiral. Cross-similarity and classification obtained by (**a**) Temporal ground-truth (TGT). Cross-similarity with window size 9 and classification at $FPR = 5\%$ (Sect. 4.2) obtained by (**a**) Temporal ground-truth (TGT); (**b**) Multi-frame alignment volume-sampling spherical Shape Histogram (SHvrG); (**c**) Single-frame alignment volume-sampling spherical Shape Histogram (SHvrS); (**d**) Temporal filtered volume-sampling spherical Shape Histogram (SHvrT); (**e**) Temporal filtered Spherical Harmonics Representation (SHRT)

[1] Shifting a SHvr descriptor in its θ bins is equivalent to rotating a mesh around the vertical axis but is more efficient (Sect. 3.1.3).



(a) SHvrG, N_t=4 (b) SHvrS, N_t=4 (c) SHvrT, N_t=4

Fig. 10 An example of the failure of SHvrS. Rachel's "Jog" compared to "Walk" with a fixed window size 9 ($N_t = 4$) using (**a**) multi-frame shape-flow SHvrG; (**b**) single-frame shape-flow SHvrS; (**c**) time-filtering shape histogram SHvrT

the change in similarity due to the non-linear motion path resulting in changes in direction between consecutive frames. Figure 9(b, c) shows that the shape-flow descriptors SHvrG with global and SHvrS with local alignment of frames produce similarity and classification which closely match the ground-truth Fig. 9(a). This illustrates a limitation of temporal filtering in correctly estimating similarity for non-linear motion paths and shows that shape-flow descriptors overcome this limitation. Quantitative performance evaluation is presented in Sect. 4.5. Figure 10 illustrates a limitation of shape-flow with local single frame alignment SHvrS versus global multiple frame alignment SHvrG for cross-similarity between real 3D video sequences of walk and jog motions. For shape-flow with multiple frame alignment SHvrG similarity Fig. 10(a) the diagonal structure is clearly visible. However, with single frame alignment Fig. 10(b) incorrect low-similarity scores occur on the diagonal (marked with a circle), this is due to failure of the single-frame alignment. Errors occur in SHvrS due to incorrect estimation of the alignment at the central frame. SHvrG is robust as optimal alignment is estimated for a sequence of frames.

4 Performance Evaluation

The performance of the shape descriptors is evaluated using a ground-truth dataset from simulated data. Temporal mesh sequences are constructed for different motions and the classification of correct and incorrect similarity is assessed using the Receiver-Operator Characteristic (ROC) curves for each technique. This evaluation extends previous comparison of shape descriptors for a single person performing eight motions (Huang et al. 2007a, 2007b) to a comprehensive dataset comprising models of 14 people each performing 28 motions. Optimal parameter settings for each shape descriptor are determined by evaluating the ROC for different parameter settings (Huang et al. 2007a). Similarity measures are then evaluated against temporal ground-truth to identify similar frames in the 3D video sequences.

4.1 Ground Truth

A simulated data-set is created using articulated character model for 14 people animation using motion capture sequences. Animated models of people with different body-shape and clothing were reconstructed from multiple view images (Starck and Hilton 2003). Each model has a single surface mesh with $1k$ vertices and $2k$ triangles. Models were animated using 28 motion capture sequences from the Santa Monica mocap archive for the following 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, rock n'roll, shoot. Each sequence comprised 100 frames giving a total of 39200 frames of synthetic 3D video with known ground-truth correspondence. Figure 11 shows 14 models and example frames of multiple motions for one model. Given the known correspondence rigid-body registration can be performed to align the frames for ground-truth assessment of similarity. The known correspondence is only used to compute the true ground-truth surface distance, and is not used in computing any of the shape similarity measures. Temporal Ground-Truth (TGT) which includes both the surface shape and motion is used to evaluate the performance of shape descriptors.

The ground-truth shape similarity between two surfaces is measured using the average distance between corresponding vertices. This characterises the frame-to-frame difference between the surfaces. Let X and Y be the set of mesh vertices for two surfaces, both have N vertices, if $d(x_i, y_i)$ denotes the Euclidean Distance between one vertex $x_i \in X$ and its corresponding vertex $y_i \in Y$, we calculate the average distance as follows:

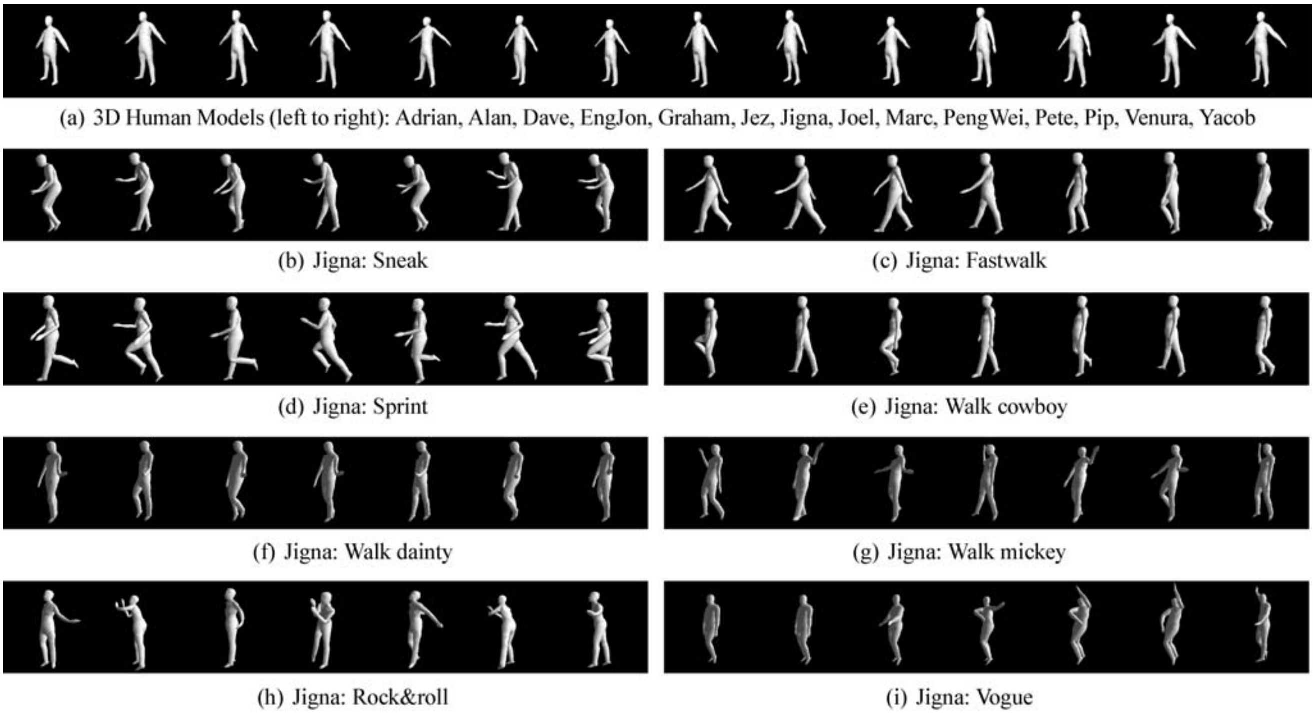
$$C_P(X, Y) = \frac{1}{N} \sum_i d(x_i, y_i) \tag{8}$$

Temporal Ground-Truth similarity $C_T(X, Y)$ between two frames is then defined by a combination of the shape similarity, $C_P(X, Y)$, and velocity similarity, $C_V(X, Y)$, as:

$$C_T(X, Y) = (1 - \alpha)C_P(X, Y) + \alpha C_V(X, Y) \tag{9}$$

$$C_V(X, Y) = \frac{1}{N} \sum_i d_v(x'_i, y'_i) \tag{10}$$

where $x'_i = x_i(t + 1) - x_i(t)$, $y'_i = y_i(t + 1) - y_i(t)$ are velocity vectors, $t + 1$, t denote the next and current frame, and $d_v(x'_i, y'_i) = |x'_i - y'_i|$ is the magnitude of vector difference between velocity vector x'_i and y'_i. Throughout the results presented in this work α is set as 0.5 to balance the shape and velocity similarity. For classification of frames as similar a threshold is set on the Temporal Ground-Truth similarity where the average distance $C_T(X, Y)$ falls below a fixed


(a) 3D Human Models (left to right): Adrian, Alan, Dave, EngJon, Graham, Jez, Jigna, Joel, Marc, PengWei, Pete, Pip, Venura, Yacob

(b) Jigna: Sneak

(c) Jigna: Fastwalk

(d) Jigna: Sprint

(e) Jigna: Walk cowboy

(f) Jigna: Walk dainty

(g) Jigna: Walk mickey

(h) Jigna: Rock&roll

(i) Jigna: Vogue

Fig. 11 Synthetic dataset. (**a**) 3D Human Models (*left* to *right*): Adrian, Alan, Dave, EngJon, Graham, Jez, Jigna, Joel, Marc, PengWei, Pete, Pip, Venura, Yacob. (**b–i**) Example frames from motion sequences of Jigna

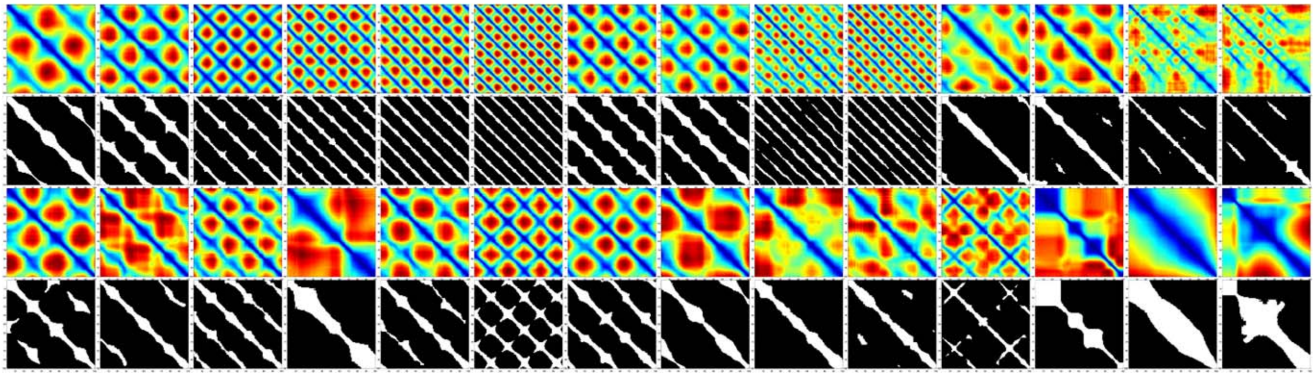


Fig. 12 Temporal ground truth self-similarity and classification of 28 motions from "Jigna" (from *top left* to *bottom right* in order): sneak, slow walk, fast walk, slow run, fast run, sprint, walk circle (*left* and *right*), run circle (*left* and *right*), walk turn (*left* and *right*), run turn (*left* and *right*), walk in styles (cool, cowboy, dainty, elderly, macho, march, mickey, sexy, tired, toddler) and complex motions (rock and roll, vogue dance, faint, shot arm)

predefined threshold τ_T. After normalisation of the self-similarity to the range [0, 1] the similarity threshold is set to $\tau_T = 0.3$ throughout this work which gives the ground-truth binary classification matrix $R_{TGT}(i, j)$ of frames i and j as similar or dissimilar. A ground-truth binary classification matrix $R_{TGT}(i, j) \in \{1, 0\}$ is then defined to classify frames as *similar* ($R_{TGT}(i, j) = 1$) if $C_T < \tau_T$ and *dissimilar* ($R_{TGT}(i, j) = 0$) otherwise. Inclusion of the surface motion in the ground-truth similarity together with the shape removes the ambiguity inherent in static single frame similarity measures. The lines of similarity in the diagonal direction indicate the periodic structure of the synthetic 3D video sequences. Figure 12 shows the similarity and ground-truth classification for the 28 motions with one of the models.

4.2 Evaluation Criterion

Performance of the shape descriptors 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.

$$TPR = \frac{ts}{ts + fd}; \qquad FPR = \frac{fs}{fs + td} \tag{11}$$

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 score for each shape descriptor is normalised to the range $S' \in [0, 1]$. A binary classification matrix for the shape descriptor $R_S^\tau(i, j) \in \{1, 0\}$ is then defined for a threshold τ to classify frames as *similar* ($R_S^\tau(i, j) = 1$) if $S' < \tau$ and *dissimilar* ($R_S^\tau(i, j) = 0$) otherwise. The classification $R_S^\tau(i, j)$ for a given τ is then compared to the ground-truth similarity classification $R_{TGT}(i, j)$ defined in Sect. 4.1. The number of true and false similarity classifications is then counted:

$$ts = \sum_{ij} \{R_{TGT}(i, j) \times R_S^\tau(i, j)\} \tag{12}$$

$$td = \sum_{ij} \{(1 - R_{TGT}(i, j)) \times (1 - R_S^\tau(i, j))\} \tag{13}$$

$$fs = \sum_{ij} \{(1 - R_{TGT}(i, j)) \times R_S^\tau(i, j)\} \tag{14}$$

$$fd = \sum_{ij} \{R_{TGT}(i, j) \times (1 - R_S^\tau(i, j))\} \tag{15}$$

The ROC performance for a given shape similarity measure is obtained by varying the threshold $\tau \in [0, 1]$ to obtain the true $TPR(\tau)$ and false $FPR(\tau)$ positive rates according to (11).

4.3 Parameter Setting for Static Shape Descriptors

Optimal parameter setting for each of the shape descriptors to match frames in 3D video sequences of people are determined by evaluating the ROC curve for a range of parameter settings (Huang et al. 2007a). This evaluation uses the known ground-truth similarity to evaluate the classification performance. Table 1 presents the optimal parameters for each shape descriptor. These parameter settings are used throughout the evaluation presented in this paper. Further details of the optimal parameter evaluation can be found in (Huang et al. 2007a).

Table 1 Optimal parameter settings for shape descriptors on 3D video sequences of people (Huang et al. 2007a)

Descriptor	Parameters
Shape distribution	No. of samples $N = 10^6$
Spin Image	No. of bins $Nb = Nb_\alpha = Nb_\beta = 40$
Shape Histogram	No. of radial bins $Ns = 10$
	No. of angular bins $Nb_\theta = 2Nb_\phi = 40$
Spherical Harmonics	No. of radial shells $N_s = 32$
	No. of harmonics $N_b = 16$

4.4 Evaluation of Static Shape Descriptors

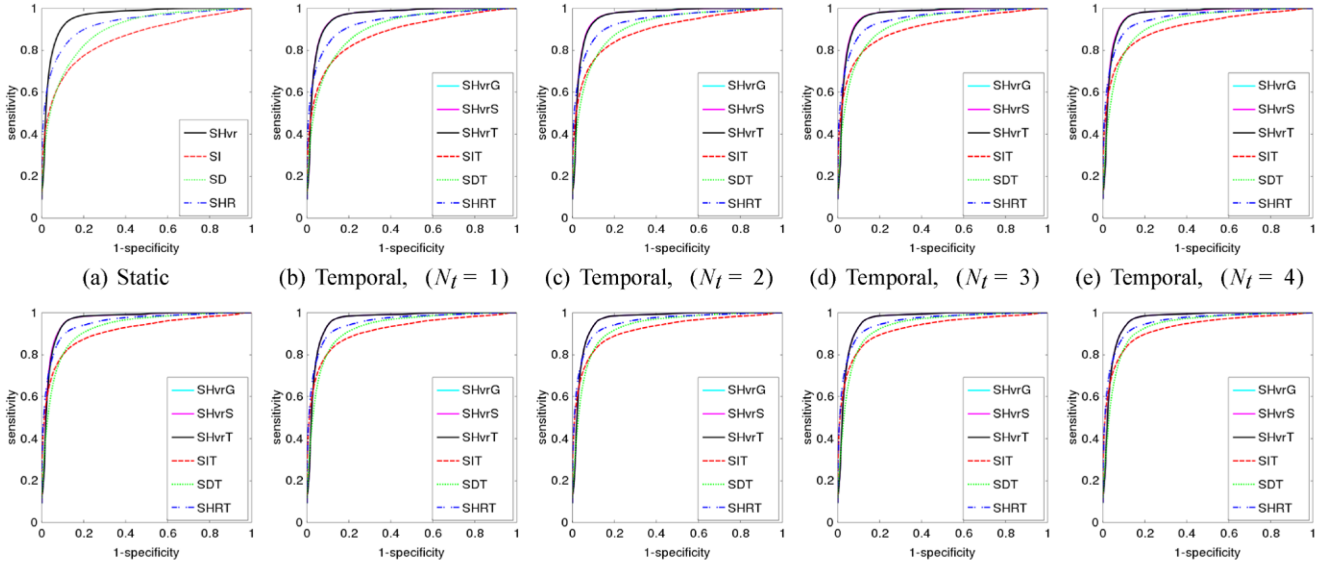
The performance of static shape descriptors for evaluation of static shape similarity was previously presented in (Huang et al. 2007a). In this paper we present an evaluation of static shape descriptors against the temporal ground-truth defined in Sect. 4.1 for an extended range of motions with different people. Figure 13(a) shows the combined ROC performance of static shape descriptors for the simulated dataset. Figure 17 presents the ROC curves for the 28 motions with 14 people for each of the shape descriptors. The ROC curves show that the volume-sampling shape-histogram descriptor (SHvr) achieves the highest performance among Shape Distribution (SD), Spin Image (SI), Spherical Harmonics Representation (SHR). SHvr consistently achieves the highest performance against ground-truth for all motions. The distribution of the curves for the 14 animated models of different people shows that the volume-sampling shape-histogram (SHvr) performs consistently with variation in size and clothing and outperforms other shape descriptors.

4.5 Evaluation of Temporal Shape Descriptors

In this section, we evaluate temporal shape descriptors defined in Sect. 3.2 against the Temporal Ground Truth. Optimal parameter settings for the shape descriptors given in Table 1 are used for evaluation of the temporal descriptors. Since the optimal temporal window size will depend on the rate of motion with a larger window size being required for slow motions, performance of each of the temporal shape descriptors is evaluated for each motion using the ROC curve with a range of temporal window size.

4.5.1 Evaluation of Time Filtered Descriptors

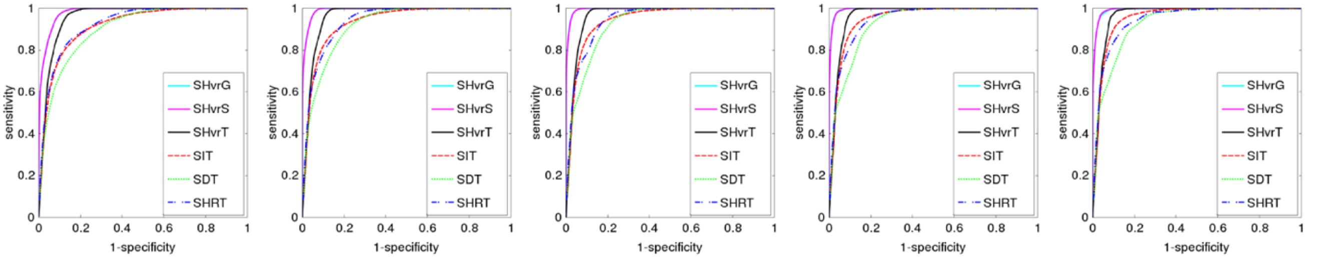
Combined ROC curves of the time-filtered descriptors on self-similarity against temporal ground truth across all people and motions in the simulated dataset with an increasing temporal window size are shown in Fig. 13(b–j). The performance of all descriptors increases compared to the equivalent static shape similarity in Fig. 13(a). The time-filtered volume-sampled shape histogram SHvrT gives the highest performance of all time-filtered shape descriptors against temporal ground-truth. This is expected as the volume-sampled shape histogram SHvr gives the best performance for the static shape descriptors and time-filtering reduces



(a) Static (b) Temporal, $(N_t = 1)$ (c) Temporal, $(N_t = 2)$ (d) Temporal, $(N_t = 3)$ (e) Temporal, $(N_t = 4)$

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

Fig. 13 Evaluation of ROC curves for static and temporal descriptors on self-similarity across 14 people each performing 28 motions



(a) Shape Flow $(N_t = 1)$ (b) Shape Flow $(N_t = 2)$ (c) Shape Flow $(N_t = 3)$ (d) Shape Flow $(N_t = 4)$ (e) Shape Flow $(N_t = 5)$

Fig. 14 Evaluation of ROC curves for temporal descriptors on cross-similarity "Fast Walk" in a straight line and on a spiral across 14 people

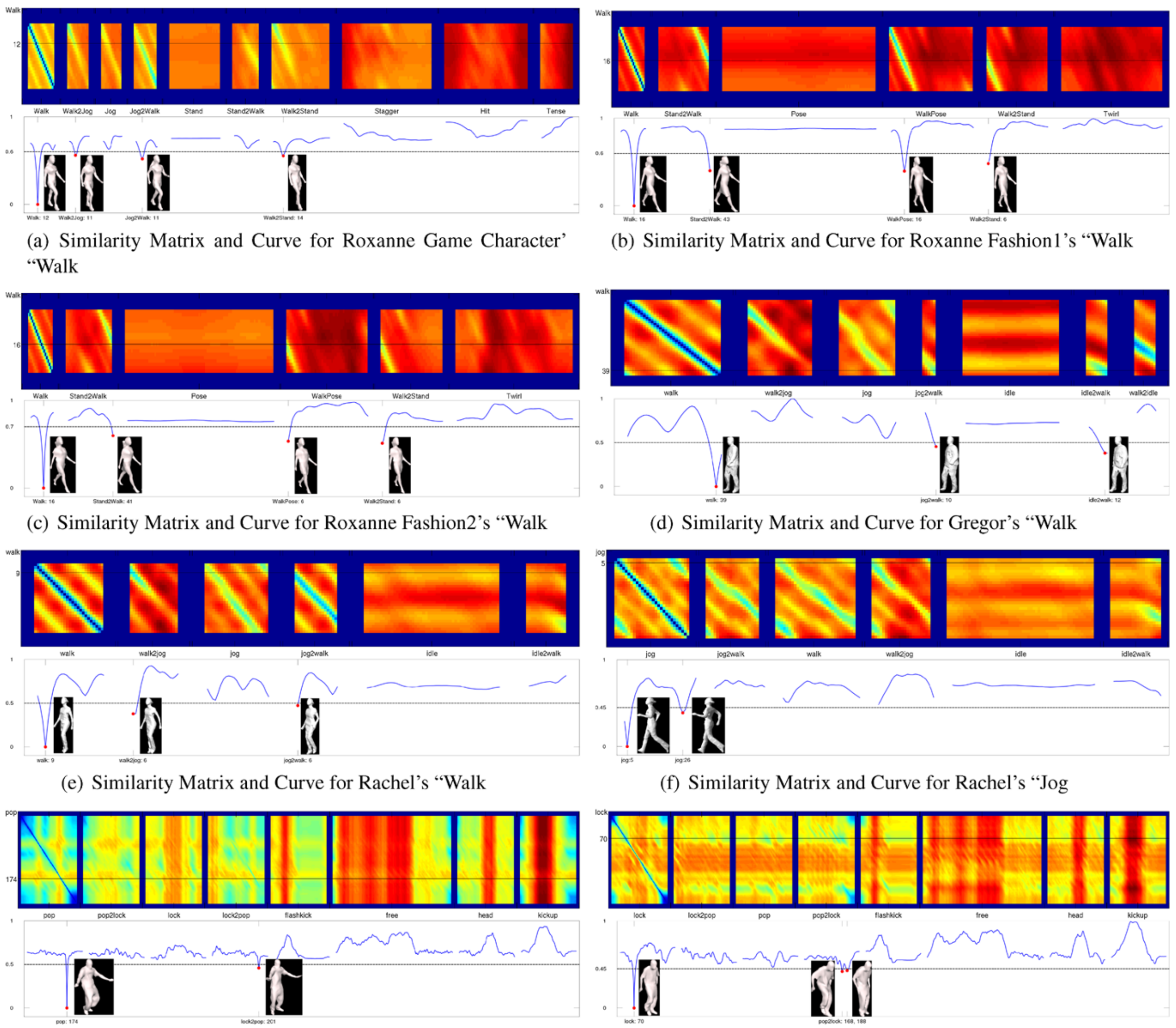
Table 2 Relative computational cost of temporal shape descriptors against window size for Roxanne's Game Character, motion "Hit". Relative cputime per frame

N_t	0	1	2	3	4	5	6
SHvrT	1.00	1.00	1.00	1.02	1.05	1.09	1.11
SHvrS	1.00	1.02	1.04	1.05	1.06	1.06	1.07
SHvrG	1.00	3.11	5.16	7.05	9.06	11.07	13.10

Table 3 Real 3D Video datasets for 9 actors and motions (transition motions are sequences which transition from one motion to another, i.e. walk to jog) N_s is the number of sequences and N_f the number of frames

Performer	Motions	N_s	N_f
JP	street dance: lock, pop, flash-kick, free-dance, head-spin, kickup + transitions	8	2300
Roxanne			
– Game Character	walk, jog, stand, stagger, hit, tense + transitions	10	442
– Fashion 1	walk, pose, twirl + transitions	6	491
– Fashion 2	walk, pose, twirl + transitions	6	435
Others			
– Adrian Gordon Gregor Rachel Jon Rafael Tony	idle, walk, jog, kick, punch + transitions	24	875
Total		54	4543

the temporal ambiguity increasing the classification accuracy. Comparison of the different shape descriptors with respect to window size also shows that the time-filtered shape histogram SHvrT is relatively insensitive to the change of window size. Figure 18 gives a more detailed comparison by presenting individual ROC curves for the 28 motions with 14 people for each of the time-filtered shape descriptors with a fixed temporal window size. This demonstrates the time-filtered volume-sampled shape histogram SHvrT



(a) Similarity Matrix and Curve for Roxanne Game Character' "Walk

(b) Similarity Matrix and Curve for Roxanne Fashion1's "Walk

(c) Similarity Matrix and Curve for Roxanne Fashion2's "Walk

(d) Similarity Matrix and Curve for Gregor's "Walk

(e) Similarity Matrix and Curve for Rachel's "Walk

(f) Similarity Matrix and Curve for Rachel's "Jog

(g) Similarity Matrix and Curve for JP's "Pop

(h) Similarity Matrix and Curve for JP's "Lock

Fig. 15 Intra-person similarity measure for Real Data. Similarity matrix, curve, example frames for (**a**) Roxanne Game Character's "Walk"; (**b**) Roxanne Fashion1's "Walk"; (**c**) Roxanne Fashion2's "Walk"; (**d**) Gregor's "Walk"; (**e**) Rachel's "Walk"; (**f**) Rachel's "Jog"; (**g**) JP's "Pop"; (**h**) JP's "Lock"

has low inter-person variance and consistently outperforms other time-filtered shape descriptors.

4.5.2 *Evaluation for Shape-Flow Descriptors*

Combined ROC curves of the shape-flow descriptors for classification of self-similarity against temporal ground truth across all people and motions in the simulated data set are presented in Fig. 13(b–j) with increasing temporal window size. Characteristics for the shape-flow descriptors SHvrG and SHvrS are superimposed with the time-filtered descriptor SHvrT showing that the performance is similar for straight line motions. Analysis of the detailed characteristics shows that in general the multi-frame shape-flow SHvrG achieves the highest performance with time-filtering SHvrT and single-frame shape flow SHvrS marginally lower. The difference between aggregate characteristics is lower than the variance for different people and motions as shown in Fig. 19. Figure 14 shows a case when shape-flow descriptor SHvrG and SHvrS achieve significantly higher performance than time filtering SHvrT for motion on a non-linear path. Shape-flow has significantly better performance than time-filtering for all window sizes. ROC curves for single and multiple frame frame shape-flow in Fig. 14 are super-

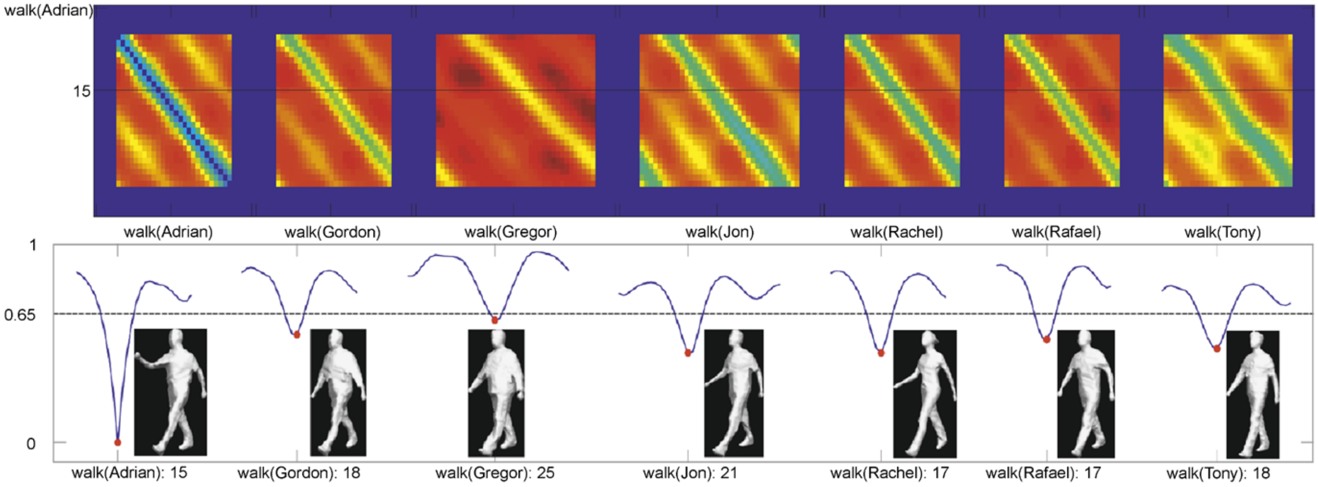


Fig. 16 Inter-person similarity measure for Real Data. Similarity matrix, curve, example frames for "Walk" across 7 people including Adrian, Gordon, Gregor, Rachel, Jon, Rafael and Tony

imposed indicating that performance of single and multiple frame alignment is comparable. Table 2 presents the relative computational cost for time-filtering and shape-flow measures with increasing window size. This shows that there is an order of magnitude increase in computational cost for the shape-flow descriptor with multiple frame alignment SHvrG, whereas single frame shape flow and time-filtering have a computational cost similar to the static shape descriptor ($N_t = 0$). In conclusion, the multi-frame shape-flow descriptor (SHvrG) overcomes the limitation of temporally filtered static shape descriptors (SHvrT, SIT, SDT, SHRT) for 3D video sequences with non-linear motion paths and is robust to errors in single frame alignment which occur with (SHvrS).

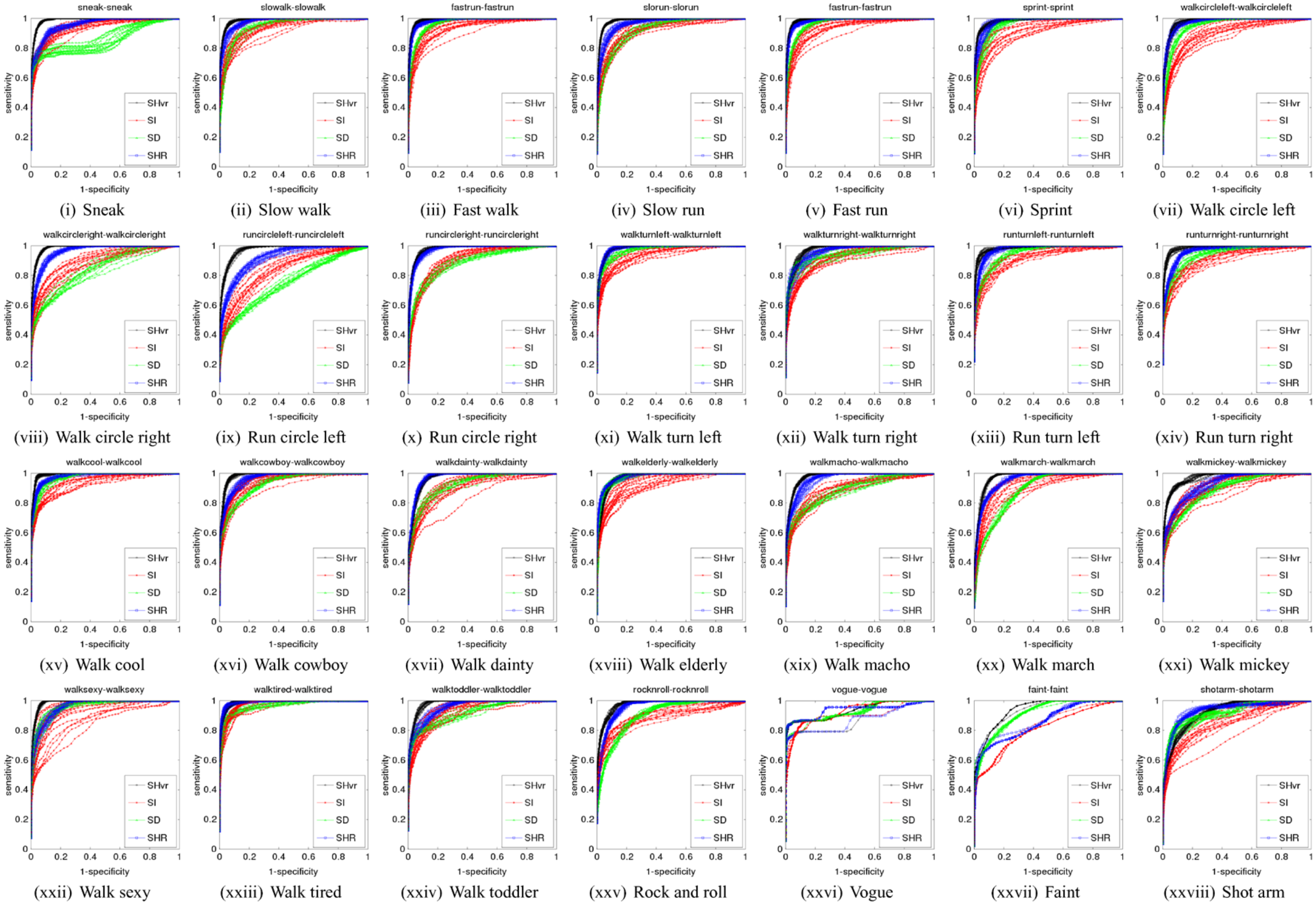
5 Similarity Measure on Real Data

In this section we apply the time-filtering shape histograms SHvrT to captured 3D video sequences of people. Real 3D video sequences were reconstructed from multiple camera video capture available as a public research database (Starck and Hilton 2007). The real 3D video sequences of nine people performing a variety of motion used in this evaluation are summarised in Table 3. 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 together with seven other actors performing a standard set of movements. Captured 3D video sequences are unstructured meshes with unknown temporal correspondence and time varying mesh connectivity, topology and geometry.

Time-filtering shape histograms SHvrT are used to evaluate intra-person similarity between the 3D video sequences of different motions for each performer/costume combination and the inter-person similarity for different performers performing the same motion. Evaluation has been performed for all available sequences. Example results are presented demonstrating typical results with identification of frames with similar shape and motion. SHvrT is applied to all sequences with the optimal resolution parameters (Table 1) and a temporal window size of 9 ($N_t = 4$).

Intra-person similarity across different motions for several performers together with an example similarity curve are presented in Fig. 15. The example matched frames for each performer show that the temporal similarity metric identifies frames of similar pose and motion across the different motions performed by each actor. In Figs. 15(b, c) for Roxanne the similarity clearly identifies 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. Figures 15(g, h) for the street dancer JP performing complex movements shows there is a lot of visible structure in the similarity matrix, frames with similar pose and motion are also correctly identified. This evaluation on real 3D video sequences demonstrates that the temporal similarity identifies similar frames and is robust to complex movement and loose clothing.

Inter-person similarity across several people each performing a walking motion together with an example similarity curve are shown in Fig. 16. The similarity measure correctly identifies frames with a similar shape and motion for each person. This illustrates that the temporal similarity measure can also be used to identify similar frames across different people.



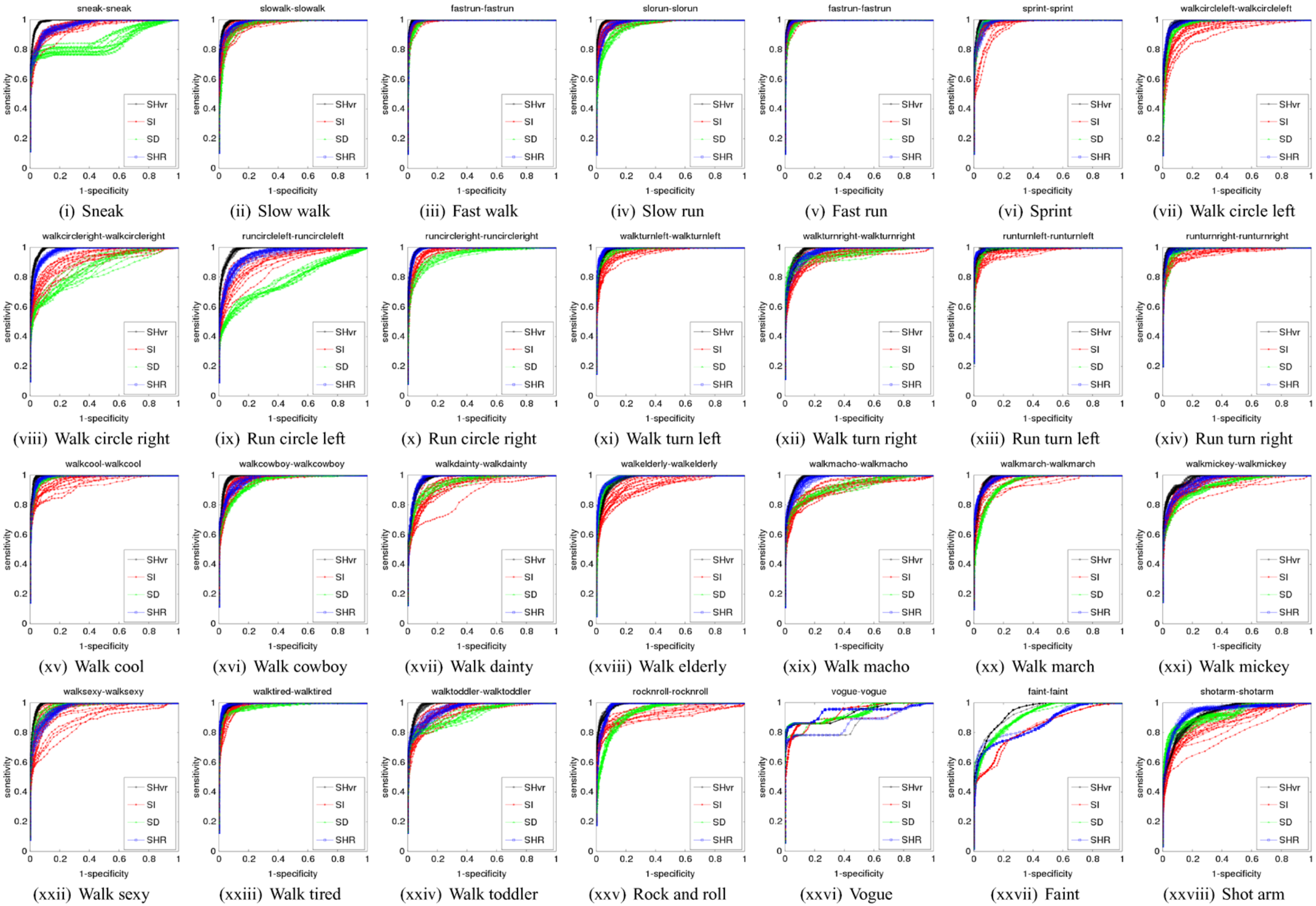
(i) Sneak (ii) Slow walk (iii) Fast walk (iv) Slow run (v) Fast run (vi) Sprint (vii) Walk circle left

(viii) Walk circle right (ix) Run circle left (x) Run circle right (xi) Walk turn left (xii) Walk turn right (xiii) Run turn left (xiv) Run turn right

(xv) Walk cool (xvi) Walk cowboy (xvii) Walk dainty (xviii) Walk elderly (xix) Walk macho (xx) Walk march (xxi) Walk mickey

(xxii) Walk sexy (xxiii) Walk tired (xxiv) Walk toddler (xxv) Rock and roll (xxvi) Vogue (xxvii) Faint (xxviii) Shot arm

Fig. 17 Evaluation of static shape descriptors against Temporal Ground Truth (TGT). ROC performance for 28 motions across 14 people

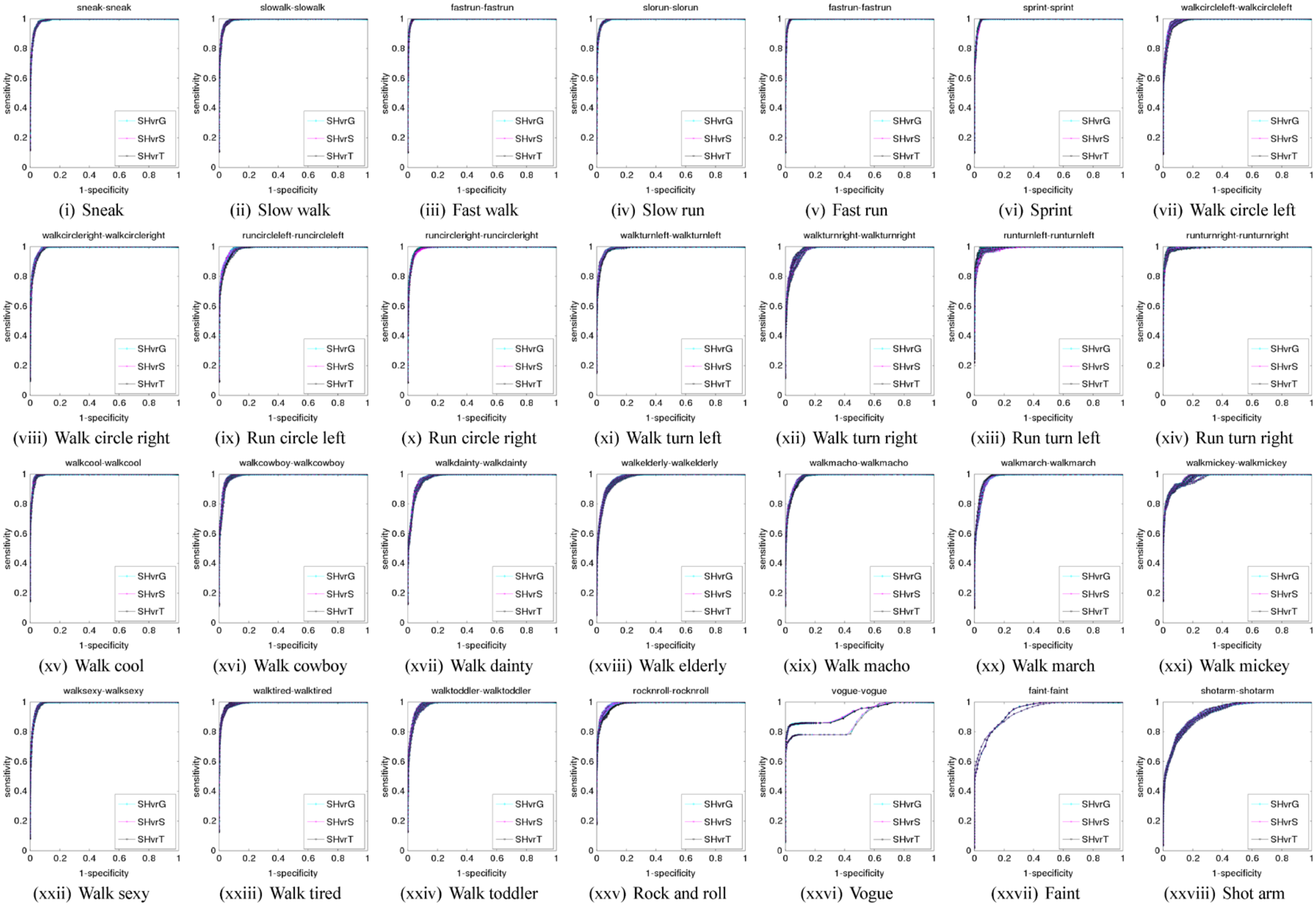

(i) Sneak (ii) Slow walk (iii) Fast walk (iv) Slow run (v) Fast run (vi) Sprint (vii) Walk circle left

(viii) Walk circle right (ix) Run circle left (x) Run circle right (xi) Walk turn left (xii) Walk turn right (xiii) Run turn left (xiv) Run turn right

(xv) Walk cool (xvi) Walk cowboy (xvii) Walk dainty (xviii) Walk elderly (xix) Walk macho (xx) Walk march (xxi) Walk mickey

(xxii) Walk sexy (xxiii) Walk tired (xxiv) Walk toddler (xxv) Rock and roll (xxvi) Vogue (xxvii) Faint (xxviii) Shot arm

Fig. 18 Evaluation of temporal filtered shape descriptors with a fixed window size 5 ($N_t = 2$) against Temporal Ground Truth (TGT). ROC performance for 28 motions across 14 people



(i) Sneak (ii) Slow walk (iii) Fast walk (iv) Slow run (v) Fast run (vi) Sprint (vii) Walk circle left

(viii) Walk circle right (ix) Run circle left (x) Run circle right (xi) Walk turn left (xii) Walk turn right (xiii) Run turn left (xiv) Run turn right

(xv) Walk cool (xvi) Walk cowboy (xvii) Walk dainty (xviii) Walk elderly (xix) Walk macho (xx) Walk march (xxi) Walk mickey

(xxii) Walk sexy (xxiii) Walk tired (xxiv) Walk toddler (xxv) Rock and roll (xxvi) Vogue (xxvii) Faint (xxviii) Shot arm

Fig. 19 Evaluation of shape-flow descriptors with a fixed window size 5 ($N_t = 2$) against Temporal Ground Truth (TGT). ROC performance for 28 motions across 14 people

6 Conclusion

A comprehensive performance evaluation of shape similarity metrics for 3D video sequences of people has been presented. Existing static shape similarity metrics which give good performance for rigid shape retrieval have been evaluated: shape-distribution (Osada et al. 2002); spin-image (Johnson and Hebert 1999); shape-histogram (Ankerst et al. 1999); and spherical harmonics (Kazhdan et al. 2003). Temporal shape similarity metrics are presented to overcome the ambiguity in independent frame-to-frame comparison. Three approaches are evaluated based on extension of shape-histograms over time: time-filtering of the static shape similarity metric; and shape flow with single and multiple frame alignment.

Performance is evaluated using the Receiver Operator Characteristics for synthetic 3D video sequences with known ground-truth for animated models of 14 people each performing 28 motions giving a total of 39,200 frames. Evaluation of static shape similarity metrics demonstrates that shape-histograms with volume-sampling consistently gives the best performance for different actors and motions. However, all static shape similarity metrics are shown to exhibit temporal ambiguities in 3D video for frames with similar shape but different motion directions. Evaluation of temporal shape similarity metrics for a variety of synthetic motions demonstrates that multi-frame shape flow consistently gives the best performance for different people, motions and temporal window size. However, multi-frame shape flow has an order of magnitude increase in computational cost over time-filtering and single-frame shape flow. Time-filtered shape histograms are computationally efficient and give marginally lower performance for straight line motions but have significantly reduced performance for non-linear movements. Shape-flow with single-frame alignment achieves comparable performance to multi-frame shape flow, overcoming the limitations of time-filtered static similarity measures for 3D video sequences with non-linear paths, with a computational cost comparable to static shape similarity. However, single frame shape-flow may fail due to errors in alignment at the central frame whereas multi-frame shape-flow is robust.

Evaluation on real 3D video sequences for 9 people demonstrates that time-filtering shape histograms correctly identify frames with similar shape and motion for loose clothing (skirts), complex motions (street-dance) and between different people. Self-similarity also identifies the periodic structure in the motion such as walking and running even for sequences with loose clothing. Performance evaluation on a comprehensive set of real and ground-truth 3D video sequences of people shows that time-filtered shape-histograms are consistent for different people and movements giving a good trade-off between correct similarity and computational cost.

Acknowledgements This work was funded by the UK EPSRC Grant EP/E001351 *Video-based Animation Production*.

References

Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H. P., & Thrun, S. (2008). Performance capture from sparse multi-view video. *ACM Transactions on Graphics*, *27*(3), 1–10.

Ankerst, M., Kastenmüller, G., Kriegel, H. P., & Seidl, T. (1999). 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* (pp. 207–226). London: Springer.

Arikan, O., Forsyth, D. A., & O'Brien, J. F. (2003). Motion synthesis from annotations. *ACM Transactions on Graphics*, *22*(3), 402–408.

Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *24*(4), 509–522.

Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *23*(3), 257–267.

Bustos, B., Keim, D., Saupe, D., & Schreck, T. (2007). Content-based 3D object retrieval. Computer Graphics and Applications. *IEEE*, *27*(4), 22–27.

Carranza, J., Theobalt, C., Magnor, M. A., & Seidel, H. P. (2003). Free-viewpoint video of human actors. *ACM Transactions on Graphics*, *22*(3), 569–577.

Chen, D. Y., Ouhyoung, M., Tian, X. P., & Shen, Y. T. (2003). On visual similarity based 3D model retrieval. *Computer Graphics Forum (EUROGRAPHICS'03)*, *22*(3), 223–232.

Chua, C. S., & Jarvis, R. (1997). Point signatures: a new representation for 3D object recognition. *International Journal of Computer Vision*, *25*(1), 63–85.

Corney, J., Rea, H., Clark, D., Pritchard, J., Breaks, M., & Macleod, R. (2002). Coarse filters for shape matching. *Computer Graphics and Applications, IEEE*, *22*(3), 65–74.

Cutler, R., & Davis, L. S. (2000). Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *22*(8), 781–796.

Del Bimbo, A., & Pala, P. (2006). Content-based retrieval of 3D models. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *2*(1), 20–43.

Efros, A. A., Berg, A. C., Mori, G., & Malik, J. (2003). Recognizing action at a distance. In *ICCV '03: Proceedings of the ninth IEEE international conference on computer vision*. Washington: IEEE Computer Society.

El-Mehalawi, M. (2003). A database system of mechanical components based on geometric and topological similarity. part ii: indexing, retrieval, matching, and similarity assessment. *Computer-Aided Design*, *35*(1), 95–105.

Elad, A., & Kimmel, R. (2003). On bending invariant signatures for surfaces. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *25*(10), 1285–1295.

Gleicher, M., Joon, H., Lucas, S., & Jepsen, K. A. (2003). Snap-together motion: assembling run-time animation. *ACM Transactions on Graphics*, *22*, 181–188.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., & Basri, R. (2007). Actions as space-time shapes. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *29*(12), 2247–2253.

Hilaga, M., Shinagawa, Y., Kohmura, T., & Kunii, T. L. (2001). 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* (pp. 203–212). New York: ACM Press.

Huang, P., & Hilton, A. (2009). Human motion synthesis from 3D video. In *Proceedings of the 2009 conference on computer vision and pattern recognition (CVPR'09)* (pp. 1478–1485).

Huang, P., Starck, J., & Hilton, A. (2007a). A study of shape similarity for temporal surface sequences of people. In *3DIM '07: Proceedings of the sixth international conference on 3D digital imaging and modeling* (pp. 408–418). Washington: IEEE Computer Society.

Huang, P., Starck, J., & Hilton, A. (2007b). Temporal 3D shape matching. In *The fourth European conference on visual media production (CVMP'07)* (pp. 1–10).

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

Jain, V., & Zhang, H. (2007). A spectral approach to shape-based retrieval of articulated 3D models. *Computer-Aided Design*, *39*(5), 398–407.

Johnson, A. E., & Hebert, M. (1999). Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *21*(5), 433–449.

Kanade, T., Rander, P., & Narayanan, P. J. (1997). Virtualized reality: Constructing virtual worlds from real scenes. *IEEE MultiMedia*, *4*(1), 34–47.

Kazhdan, M., Chazelle, B., Dobkin, D. P., Finkelstein, A., & Funkhouser, T. A. (2002). A reflective symmetry descriptor. In *ECCV* (Vol. 2, pp. 642–656).

Kazhdan, M., Funkhouser, T., & Rusinkiewicz, S. (2003). Rotation invariant spherical harmonic representation of 3D shape descriptors. In *SGP '03: Proceedings of the 2003 Eurographics/ACM SIGGRAPH symposium on geometry processing* (pp. 156–164).

Körtgen, M., Park, G. J., Novotni, M., & Klein, R. (2003). 3D shape matching with 3D shape contexts. In *The 7th central European seminar on computer graphics*.

Kovar, L., Gleicher, M., & Pighin, F. (2002). Motion graphs. In *SIGGRAPH '02: Proceedings of the 29th annual conference on computer graphics and interactive techniques* (Vol. 21, pp. 473–482). New York: ACM Press.

Krüger, V., Kragic, D., Ude, A., & Geib, C. (2007). The meaning of action: A review on action recognition and mapping. *Advanced Robotics*, *21*(13), 1473–1501.

Lee, J., Chai, J., Reitsma, P. S. A., Hodgins, J. K., & Pollard, N. S. (2002). Interactive control of avatars animated with human motion data. *ACM Transactions on Graphics*, *21*(3), 491–500.

Mcwherter, D., Peabody, M., Regli, W. C., & Shokoufandeh, A. (2001). Solid model databases: Techniques and empirical results. *Journal of Computing and Information Science in Engineering*, *1*(4), 300–310.

Novotni, M., & Klein, R. (2003). 3D Zernike descriptors for content based shape retrieval. In *SM '03: Proceedings of the eighth ACM symposium on solid modeling and applications* (pp. 216–225). New York: ACM Press.

Ohbuchi, R., Minamitani, T., & Takei, T. (2003). Shape-similarity search of 3D models by using enhanced shape functions. In *Theory and practice of computer graphics, 2003 proceedings* (pp. 97–104).

Osada, R., Funkhouser, T., Chazelle, B., & Dobkin, D. (2002). Shape distributions. *ACM Transactions on Graphics*, *21*(4), 807–832.

Paquet, E. (2000). Description of shape information for 2D and 3D objects. *Signal Processing: Image Communication*, *16*, 103–122.

Schödl, A., Szeliski, R., Salesin, D. H., & Essa, I. (2000). Video textures. In *SIGGRAPH '00: Proceedings of the 27th annual conference on computer graphics and interactive techniques* (pp. 489–498). New York: ACM Press/Addison-Wesley.

Shum, H. Y., Hebert, M., & Ikeuchi, K. (1996). On 3D shape similarity. In *Proceedings of the 1996 conference on computer vision and pattern recognition (CVPR '96)* (pp. 526–531).

Starck, J., & Hilton, A. (2003). Model-based multiple view reconstruction of people. In *ICCV '03: Proceedings of the ninth international conference on computer vision* (pp. 915–922).

Starck, J., & Hilton, A. (2007). Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, *27*(3), 21–31.

Starck, J., Miller, G., & Hilton, A. (2005). Video-based character animation. In *SCA '05: Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on computer animation* (pp. 49–58). New York: ACM Press.

Sundar, H., Silver, D., Gagvani, N., & Dickinson, S. (2003). Skeleton based shape matching and retrieval. In *SMI '03: Proceedings of the shape modeling international 2003* (p. 130).

Tangelder, J. W. H., & Veltkamp, R. C. (2004). A survey of content based 3D shape retrieval methods. In *SMI '04: Proceedings of the shape modeling international 2004* (pp. 145–156). Washington: IEEE Computer Society.

Theobalt, C., Ahmed, N., Lensch, H., Magnor, M., & Seidel, H. P. (2007). Seeing people in different light-joint shape, motion, and reflectance capture. *IEEE Transactions on Visualization and Computer Graphics*, *13*(4), 663–674.

Vlasic, D., Baran, I., Matusik, W., & Popović, J. (2008). Articulated mesh animation from multi-view silhouettes. *ACM Transactions on Graphics*, *27*(3), 1–9.

Weinland, D., Ronfard, R., & Boyer, E. (2006). Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, *104*(2), 249–257.

Xu, J., Yamasaki, T., & Aizawa, K. (2006). Motion editing in 3D video database. In *3DPVT '06: Proceedings of the third international symposium on 3D data processing, visualization, and transmission* (pp. 472–479). Washington: IEEE Computer Society.

Zaharia, T., & Preteux, F. (2001). Three-dimensional shape-based retrieval within the mpeg-7 framework. In *Proceedings SPIE conference on nonlinear image processing and pattern analysis XII* (Vol. 4304, pp. 133–145).

Zhang, C., & Chen, T. (2001). Efficient feature extraction for 2D/3D objects in mesh representation. In *Image processing, 2001 proceedings 2001 international conference* (Vol. 3, pp. 935–938).

Zitnick, C. L., Kang, S. B., Uyttendaele, M., Winder, S., & Szeliski, R. (2004). High-quality video view interpolation using a layered representation. *ACM Transactions on Graphics*, *23*(3), 600–608.