Towards optimal non-rigid surface tracking

Martin Klaudiny, Chris Budd, and Adrian Hilton

Centre for Vision, Speech and Signal Processing, University of Surrey, UK
{m.klaudiny,chris.budd,a.hilton}@surrey.ac.uk

Abstract. This paper addresses the problem of optimal alignment of non-rigid surfaces from multi-view video observations to obtain a temporally consistent representation. Conventional non-rigid surface tracking performs frame-to-frame alignment which is subject to the accumulation of errors resulting in drift over time. Recently, non-sequential tracking approaches have been introduced which re-order the input data based on a dissimilarity measure. One or more input sequences are represented in a tree with reducing alignment path length. This limits drift and increases robustness to large non-rigid deformations. However, jumps may occur in the aligned mesh sequence where tree branches meet due to independent error accumulation. Optimisation of the tree for non-sequential tracking is proposed to minimise the errors in temporal consistency due to both the drift and jumps. A novel cluster tree enforces sequential tracking in local segments of the sequence while allowing global non-sequential traversal among these segments. This provides a mechanism to create a tree structure which reduces the number of jumps between branches and limits the length of branches. Comprehensive evaluation is performed on a variety of challenging non-rigid surfaces including faces, cloth and people. This demonstrates that the proposed cluster tree achieves better temporal consistency than the previous sequential and non-sequential tracking approaches. Quantitative ground-truth comparison on a synthetic facial performance shows reduced error with the cluster tree.

Keywords: dense motion capture, non-rigid surface alignment, non-sequential tracking, minimum spanning tree, cluster tree, dissimilarity

1 Introduction

Over the last decade, there has been an increasing research effort in spatio-temporal reconstruction of dynamic surfaces using multi-view video and/or depth acquisition. An important challenge is to transform the sequences of independent surface measurements at each frame into the aligned sequences with consistent temporal structure and correspondence. The problem of dense tracking for surfaces undergoing fast complex non-rigid motions over longer time periods has been tackled by a number of techniques. They can be divided into two broad groups according to the type of information they are primarily based on: image-based techniques work directly with multi-view video sequences; geometry-based techniques with a sequence of unregistered meshes reconstructed per frame.

Image-based techniques commonly estimate a scene flow [1] between pairs of frames based on image constraints from multiple views. Multi-view 2D optical flows combined with per-frame geometry of the surface yield a 3D motion field which deforms a template mesh throughout the sequence [2]. Pons et al.[3] use a variational formulation of matching image information across views and over time to directly compute the surface shape and its motion field in alternation. The shape and motion computation can also be joined into a single complex optimisation [4]. Carceroni and Kutulakos [5] propose more efficient 3D tracking of independent surface patches with their own shape and appearance properties. Neumann and Aloimonos [6] iteratively refine shape and motion of the multi-resolution subdivision surface model by optimisation of individual surface patches. The patches can also be associated with triangle fans of a mesh deformed over time [7, 8]. Their shape changes with the tracked mesh which improves alignment of their textures with changing surface appearance in multi-view videos. Accumulation of tracking errors is reduced by fixed patch textures from the reference frame.

Geometry-based techniques directly create a temporally consistent representation of unregistered surface geometries [9–11] or fit a prior shape model to the unregistered sequence [12]. Cagniart et al.[10] perform hierarchical matching of overlapping rigid surface patches to sequentially track a sequence of multi-view reconstructions. Wand et al.[9] propose so-called urshape representing the surface and optimise its time-varying deformation field to fit a point cloud sequence. The animation cartography approach [11] employs geometric feature tracking to map surface regions to the 2D embedding space and build up a map of the complete surface from partial observations. Existing image-based or geometry-based approaches process input data sequentially which results in error accumulation causing a drift of the tracked mesh or a complete failure if the frame-to-frame alignment cannot handle rapid non-rigid deformation of the surface.

Non-sequential methods for surface tracking have been proposed which reorder the input sequence to overcome the problems of drift and failure. Beeler et al. [13] identify similar frames across a sequence of facial performance and use them to anchor a sequential alignment of intermediate frames using multi-view optic flow. In contrast, Budd et al. [14] optimise the traversal among all frames of whole-body performance by introducing the use of a minimum spanning tree in shape similarity space to re-order the frame-to-frame alignment process. Non-sequential alignment has been extended to register multiple non-rigid mesh sequences [15, 14].

Non-sequential approaches reduce the drift and improve robustness to tracking failures compared to sequential approaches. However, the independent accumulation of errors along different alignment paths can lead to jumps in the resulting mesh sequence where different paths meet. This paper addresses the problem of optimising the tree structure for non-sequential tracking to balance between drift and jump errors. The proposed concept is generalised for any frame-to-frame alignment method and variety of non-rigid surfaces. This is demonstrated by extensive evaluation on challenging datasets of faces, cloth and people.

2 Problem statement

Input is a sequence of measurements $\{O_t\}_{t=1}^N$ of a deforming surface for frames $\{t_1, ..., t_N\}$. It can consist of multiple segments from independent motions of the surface. Each measurement O_t consists of a set of images from multiple viewpoints I_t^c and a mesh G_t representing the current shape of the surface. The mesh sequence $\{G_t\}_{t=1}^N$ is temporally unregistered, thus each mesh $G_t = (\hat{X}_t, \hat{C}_t)$ has time-varying vertex positions $\hat{X}_t$ and time-varying connectivity $\hat{C}_t$. The required output is a *temporally consistent mesh sequence* $\{M_t\}_{t=1}^N$ where the vertex positions X_t of mesh M_t correspond to the same set of surface points in every frame t and the connectivity of vertices C is fixed throughout the sequence.

Conventionally, the output mesh sequence $\{M_t\}_{t=1}^N$ is obtained by *sequential tracking* which concatenates *frame-to-frame non-rigid alignment* between successive frames t_i, t_{i+1}. The frame-to-frame alignment estimates the correspondence between observations O_{t_i}, $O_{t_{i+1}}$. *Non-sequential tracking* processes the input sequence $\{O_t\}_{t=1}^N$ in an order different from the temporal order. The reordering of $\{O_t\}_{t=1}^N$ is guided by a measure which estimates difficulty of non-rigid alignment of measurements O_t between any two frames. Intuitively, the difficulty of transition between frame t_i and t_j is represented by the *dissimilarity* between respective measurements $d(O_{t_i}, O_{t_j})$. Given $d(O_{t_i}, O_{t_j})$ between all pairs of frames, paths to every frame are jointly optimised to have minimal length. This reduces accumulation of alignment errors when the tracking is performed along the paths.

The paths are represented by a *traversal tree* $T = (\mathcal{N}, \mathcal{E})$ which is a spanning tree with the nodes $\mathcal{N} = \{n_1, ..., n_N\}$ corresponding to all frames $\{t_1, ..., t_N\}$ (Figure 1). The edges $\mathcal{E} = \{(n_i, n_j), ...\}$ are directed and weighted by the dissimilarity $d(O_{t_i}, O_{t_j})$. The non-sequential nature of tracking using the traversal tree leads to the presence of *cuts* in the sequence at places where two different alignment paths meet (marked red in Figure 1). Independent accumulation of tracking errors along these paths can potentially manifest as glitches or jumps in the resulting sequence $\{M_t\}_{t=1}^N$. There is a trade-off between the minimisation of tracking path length and a large number of cuts. Longer paths lead to larger gradual drift but large amount of cuts introduce sudden glitches and jitter. The proposed method reflects this trade-off and allows calculation of the traversal tree which balances between these two kinds of artefacts.

The non-sequential traversal of the input sequence using T can be combined with any frame-to-frame surface tracking technique working with $\{O_t\}_{t=1}^N$. The dissimilarity measure d has to be proportional to the alignment error of the selected technique so it is valid for calculating T. However, d is designed as an approximate measure which is significantly easier to compute than direct alignment of the mesh M. Given T, a user needs to specify a shape and topology of the mesh $M_{t_r} = (X_{t_r}, C)$ for the root node n_r. M_{t_r} is subsequently tracked between the pairs of frames along the branches of T from n_r towards the leaves. The result is a temporally consistent mesh sequence $\{M_t\}_{t=1}^N$ which can span across multiple separate captures of the same surface.

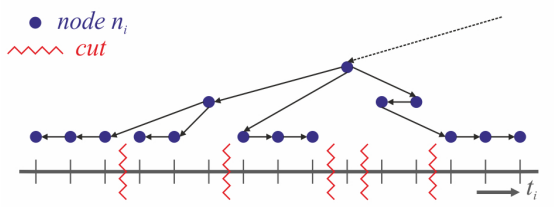

Fig. 1. Structure of a traversal tree T on the input frame sequence $\{t_1, ..., t_N\}$. The cuts separate adjacent frames which have different alignment paths along tree branches.

3 Minimum spanning tree

Non-sequential traversal of an input sequence based on the minimum spanning tree has been introduced by Budd et al.[14]. It is computed in a shape dissimilarity space and used for global alignment of multiple unregistered mesh sequences. This concept is generalised here for an arbitrary dissimilarity d between multimodal measurements O_t in every frame. The space of all possible pair-wise transitions between frames of the sequence is represented by a dissimilarity matrix D of size $N \times N$ where both rows and columns correspond to individual frames (Figure 2(a)). The elements $D(i,j) = d(O_{t_i}, O_{t_j})$ define a cost of alignment between frames t_i and t_j. The matrix is symmetric ($d(O_{t_i}, O_{t_j}) = d(O_{t_j}, O_{t_i})$) and has zero diagonal ($d(O_{t_i}, O_{t_i}) = 0$). The optimal traversal in this space can be found through graph formulation of the problem as suggested in [14].

A fully-connected undirected graph $G = (\mathcal{N}, \mathcal{D})$ is built from the matrix D. The nodes $\mathcal{N} = \{n_1, ..., n_N\}$ are associated with frames and interconnecting edges $(n_i, n_j) \in \mathcal{D}$ have the weight $D(i,j)$. A traversal visiting all frames is described by an undirected spanning tree $T'_s = (\mathcal{N}, \mathcal{E}')$ where $\mathcal{E}' \subset \mathcal{D}$. The optimal tree T'_{MST} is defined as the *minimum spanning tree* (MST) which minimises the total cost of pair-wise alignment given by d as outlined in Equation 1. This objective describes total non-rigid deformation of the surface which has to be overcome following the traversal tree, and is optimised by Prim's algorithm.

$$T'_{MST} = \underset{\forall T'_s \subset G}{\operatorname{argmin}} \left(\sum_{\forall (n_i, n_j) \in T'_s} D(i,j) \right) \tag{1}$$

The benefit of MST is that low-cost transitions are close to the root and the edges with larger d are pushed towards the leaves. This reduces the accumulation of errors along the branches and also limits the extent of a failure due to large inter-frame dissimilarity to the ends of branches. The drawback of MST is that it does not take into account the introduction of cuts and tends to temporally over-fragment the sequence. T'_{MST} then contains short off-shoots or re-shuffling of consecutive frames on a single branch as illustrated in the lower right corner of Figure 2(b). This happens mostly in slow-motion periods where T_{MST} over-fits to small changes in low range of d.

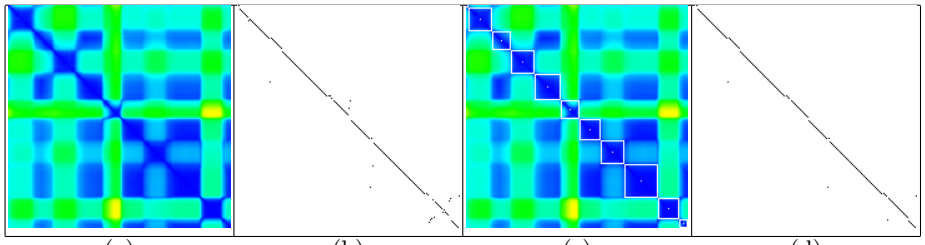

Fig. 2. Dissimilarity matrix D for a part of the dataset SyntheticFace (blue - low values, red - high values) (a). Traversal tree T_{MST} depicted in D (each directed edge (t_i, t_j) is marked black at respective location $D(i, j)$). T_{MST} is the directed T'_{MST} with optimal root by Equation 7 (b). Clustering S_β illustrated in D as white squares for individual clusters (c). Traversal tree T_β based on the clustering S_β (notice less fragmentation and longer sequential segments than in (b)) (d).

4 Cluster tree

To address shortcomings of MST the notion of temporal order of frames needs to be incorporated into the algorithm generating the traversal tree. MST is independent from the order of frames because the weight of edges in G does not change with re-ordering of the sequence $\{O_t\}_{t=1}^{N}$. A novel *cluster tree* is proposed which enforces sequential tracking locally to reduce the fragmentation of the sequence. The tree structure is still used to link the sequential segments together to obtain global non-sequential traversal of the sequence. The resulting tree shape is simpler with a smaller number of cuts which reduces the jumps/jitter in favour of relatively smooth sequential drift which is perceptually more acceptable.

4.1 Frame clustering

Intuitively, the segments traversed sequentially should contain little or no deformation of the surface, thus there is a minimal accumulation of errors. Clusters of similar successive frames form blocks with low d around the diagonal in the matrix D (Figure 2(a)). Ideally, large clusters should be generated in slow-motion segments and small clusters (even down to individual frames) in the segments with significant surface motion. The summarisation method by Huang et al. [16] is modified for the purpose of frame clustering. The clusters do not have any representative key-frames but all frames are compared to each other to measure overall intra-cluster consistency. This provides a more general clustering approach which suits our purpose better than grouping frames around a few distinct exemplars.

A sequence of frames $\{t_1, ..., t_N\}$ can be represented by a clustering $S = \{F_1, ..., F_L\}$ where a frame cluster $F_i(t_{ci}, \Delta t_i)$ is a set of successive frames $\{t_{ci} - \Delta t_i, ..., t_{ci} + \Delta t_i\}$. All L clusters have to cover together the whole sequence $F_1 \cup ... \cup F_L = \{t_1, ..., t_N\}$ and be pair-wise disjoint $F_i \cap F_j = \emptyset$. The inconsistency

of frames within cluster $A(F_i)$ is defined in Equation 2 as a sum of dissimilarities among them (the main difference to [16]).

$$A(F_i) = \frac{1}{2} \sum_{k=t_{ci}-\Delta t_i}^{t_{ci}+\Delta t_i} \left(\sum_{l=t_{ci}-\Delta t_i}^{t_{ci}+\Delta t_i} D(k,l) \right) \tag{2}$$

The clustering S is described by two costs: total intra-cluster inconsistency for all clusters and the number of clusters L. They are weighted against each other by the parameter $\beta \in < 0, 1 >$ to provide a combined cost which is minimised in Equation 3.

$$S_\beta = \underset{S}{\operatorname{argmin}} \left(\beta L + (1-\beta) \sum_{\forall F_i \in S} A(F_i) \right) \tag{3}$$

The optimal set of clusters S_β for the dissimilarity matrix D (Figure 2(c)) depends on β which influences granularity of the clustering. A value closer to 1 returns smaller number of large clusters while a value closer to 0 returns larger number of small clusters. For a given β Equation 3 is minimised through a graph-based formulation as in [16].

4.2 Tree calculation

A non-sequential traversal can be computed on the sequence of clusters instead of the original frame sequence using MST as described in Section 3. The dissimilarity matrix D is collapsed to a cluster dissimilarity matrix D_F of size $L \times L$ where rows and columns correspond to the individual clusters from S_β. Equation 4 defines the dissimilarity $D_F(i,j)$ between the clusters F_i and F_j as the minimal cost of transition between the respective clusters in the full matrix D. A cluster pair (F_i, F_j) is then linked by the pair of frames (t_k, t_l) with minimal dissimilarity.

$$D_F(i,j) = min(D(k,l)) \qquad \forall t_k \in F_i, \forall t_l \in F_j \tag{4}$$

The matrix D_F is symmetric with zero diagonal elements as for D. A fully-connected graph $G_F = (\mathcal{N}_F, \mathcal{D}_F)$ with nodes corresponding to the clusters $\{F_1, ..., F_L\}$ is built from D_F. The minimum spanning tree $T'_F = (\mathcal{N}_F, \mathcal{E}'_F)$ among the clusters is computed as in Equation 1.

Afterwards, the tree among clusters T'_F needs to be transformed to a full spanning tree T'_β interconnecting all frames. The set of nodes $\mathcal{N}$ for T'_β is expanded to the full sequence of frames $\{t_1, ..., t_N\}$. The set of edges $\mathcal{E}'$ for T'_β firstly contains a sparse set of links $\mathcal{E}'_1$ interconnecting the original clusters which is derived from $\mathcal{E}'_F$ (Equation 5). Secondly, $\mathcal{E}'$ contains a set of edges $\mathcal{E}'_2$ linking the rest of the frames within the clusters to T'_β. Because of low intra-cluster dissimilarity of frames sequential traversal is enforced among them. Thus, $\mathcal{E}'_2$ defines chains of frames in temporal order for all clusters (Equation 6).

$$\mathcal{E}'_1 = \{(n_k, n_l) : (n_i, n_j) \in \mathcal{E}'_F, (F_i, F_j) \sim (t_k, t_l)\} \tag{5}$$

$$\mathcal{E}'_2 = \bigcup_{\forall F_i \in S_\beta} \{(n_k, n_l) : t_k, t_l \in F_i, |t_k - t_l| = 1\} \tag{6}$$

The construction of T'_β does not strictly create cuts at all boundaries between the clusters. Typically, the minimal transition between temporally adjacent clusters is the one linking the last frame of the first cluster to the first frame of the second cluster. Therefore, the algorithm has an option to chain together several neighbouring clusters into a single sequential segment if it is deemed optimal.

The tree T'_β does not exactly define a traversal of the input sequence because it is undirected and has no root node. The root node n_r has to be selected to set directions along the paths in T'_β. The selection is made by minimisation of Equation 7 which is derived from the criterion for a shortest path tree. The length of weighted paths $n_l \rightarrow n_k$ from a candidate root node n_l to all other nodes n_k has to be minimal.

$$n_r = \underset{n_l \in \mathcal{N}}{\operatorname{argmin}} \left(\sum_{\forall n_k \in T'_\beta} \sum_{\forall (n_i, n_j) \in n_l \rightarrow n_k} D(i,j) \right) \tag{7}$$

The final traversal tree T_β (Figure 2(d)) is created from T'_β by setting the direction of the edges in $\mathcal{E}'$ according to the expansion of breadth-first search from n_r towards the leaves.

The shape of T_β is influenced by the clustering parameter β. The granularity of clustering S_β influences a number of branches for T_β. The cluster tree T_0 for $\beta = 0$ is equivalent to T_{MST} because all clusters contain one frame. With increasing β trees become generally thinner with longer sequential branches. T_1 for $\beta = 1$ is equivalent to purely sequential traversal because a single cluster for the whole sequence is generated. The spectrum of possible cluster trees allows a selection of T_β which balances the trade-off between drift and jumps/jitter for a given dataset. However, the optimal value of β has to be manually tuned according to visual evaluation of the tracked mesh sequence.

5 Experiments

The proposed approach has been extensively tested under several different scenarios of deformable surfaces undergoing complex non-rigid motions. Table 1 summarises the datasets used which contain facial performances (SyntheticFace, Face, DisneyFace [13]), whole-body performances (StreetDance [17]) and cloth deformation (Garment). All datasets provide multi-view image sequences with camera calibration and an unregistered mesh sequence. The absence of ground-truth for real data is a common issue in dense surface tracking. To allow quantitative evaluation of the methods the dataset SyntheticFace is artificially created.

Two different frame-to-frame tracking techniques are used according to the nature of individual datasets. Image-oriented surface tracking (IOST) is used for the face and cloth datasets [8]. The dissimilarity measure d_{IOST} for IOST is derived from the 3D trajectories of a sparse set of strong features robustly tracked in $\{I_t^c\}_{t=1}^N$. Geometry-oriented surface tracking (GOST) is used for the whole-body performance [14]. The dissimilarity measure d_{GOST} for GOST is

based on comparison of G_t between frames using a shape histogram. Details of IOST, GOST and d-measures are given in the supplementary material [1].

Table 1. Description of datasets and frame-to-frame alignment methods used for their evaluation. StreetDance [17] and DisneyFace [13] are publicly available. $|X|$ denotes the number of vertices of the tracked mesh M.

Dataset	No. of cameras	Resolution	Fps	No. of frames	Method	$\|X\|$
SyntheticFace	4	800×950	25	355	IOST	2689
Face	4	1920×1080	25	355	IOST	2689
DisneyFace	7	1176×864	46	346	IOST	2700
Garment	4	1920×1080	25	320	IOST	425
StreetDance	8	1920×1080	25	1050	GOST	3484

The following traversals of the input sequence are compared across all datasets: the standard sequential traversal (represented by $\beta = 1$), the non-sequential traversal based on MST (represented by $\beta = 0$) and the non-sequential traversal based on cluster tree. Multiple traversal trees T_β are generated for the proposed cluster-based approach to explore the spectrum of possible tree shapes between the sequential traversal and MST. Figure 3 shows the number of clusters for the tested values of β across individual datasets. The aligned sequence $\{M_t\}_{t=1}^N$ is obtained by applying the respective frame-to-frame alignment algorithm along the branches of T_β. The temporal consistency of mesh sequences resulting from the individual T_β has been visually assessed from the perspective of gradual drift versus severity of jitter and rapid glitches (the best traversal tree is noted in Figure 3). Due to the visual nature of results the reader is encouraged to watch supplementary videos [1].

5.1 Synthetic facial performance

The dataset SyntheticFace is derived from the real performance Face to achieve realistic face motion. The aligned mesh sequence obtained for the dataset Face is temporally smoothed across cuts to remove jumps. This represents the ground-truth $\{M_t^{GT}\}_{t=1}^N$ which is textured with a fixed face texture to avoid introduction of any inconsistencies between appearance changes and underlying motion. The textured $\{M_t^{GT}\}_{t=1}^N$ is rendered into 4 virtual views to create $\{I_t^c\}_{t=1}^N$ and the ground-truth meshes serve as $\{G_t\}_{t=1}^N$. The dissimilarity d_{IOST} is computed from 3D trajectories of the vertices selected from $\{M_t^{GT}\}_{t=1}^N$. The initial mesh M_{t_r} is taken directly from $\{M_t^{GT}\}_{t=1}^N$ in the root frame, so that the resulting $\{M_t\}_{t=1}^N$ can be compared directly the ground-truth.

To be valid for tree computation, d_{IOST} needs to be proportional to the difficulty of frame-to-frame alignment observed by IOST technique. This is analysed

[1] Supplementary material including videos is available under:
http://cvssp.org/projects/face3d/eccv2012/index.html

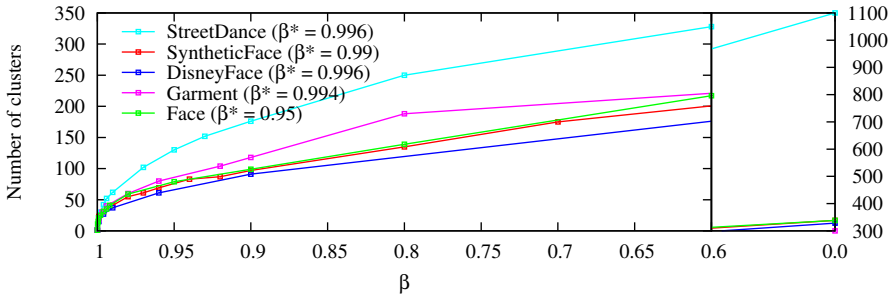

Fig. 3. Number of generated clusters for the tested values of β across the datasets. The amount of clusters increases from the sequential traversal ($\beta = 1$) towards MST ($\beta = 0$). β^* corresponds to the tree which gives the visually best tracking outcome.

by comparing the values of d_{IOST} with the tracking errors E_{IOST} reported by the alignment algorithm. The graph in Figure 4(right) aggregates pairs of (d_{IOST}, E_{IOST}) for every frame-to-frame transition across all traversals compared for SyntheticFace. The relationship has a scattered monotonically increasing trend. Low dissimilarities ($d_{IOST} < 0.4$) do not affect the quality of tracking and E_{IOST} linearly increases for higher values of d_{IOST}. The monotonic profile validates the use of d_{IOST} with IOST.

The ground-truth error of $\{M_t\}_{t=1}^N$ with respect to $\{M_t^{GT}\}_{t=1}^N$ is an average Euclidean distance of corresponding vertices across all frames $\{t_1, ..., t_N\}$. Figure 4(left) shows the graph of error for different β. The sequential tracking ($\beta = 1$) leads to the highest error due to accumulated drift. The profile for cluster trees demonstrates an improvement over MST ($\beta = 0$). In general, all non-sequential traversals achieve similar average imprecision $0.25 - 0.26mm$ per vertex which reflects the high quality of tracking. The ground-truth error reflects accumulation of the drift, however it does not quantify glitches due to the cuts. Despite this fact the graph of error correlates with visual assessment of the results and the cluster tree $T_{0.99}$ is selected as the best. The sequential result clearly suffers from significant mesh distortions built up during fast expression changes. The qualitative differences between $T_{0.99}$ and MST are fairly small because of the high-quality alignment achieved by IOST.

5.2 Facial performance

The dataset Face containing fast changes of facial expressions poses a problem for sequential tracking which results in mesh distortions in the most deforming eye and mouth regions. The fragmentation in MST does not show as visible jumps in most cases because IOST produces accurate alignments in spite of weak skin texture. The best $T_{0.95}$ yields accurate mesh sequence which improves over MST by eliminating several small glitches around the eyes and on the lips. The monotonic relationship of d_{IOST} and E_{IOST} shown in Figure 5(left) validates

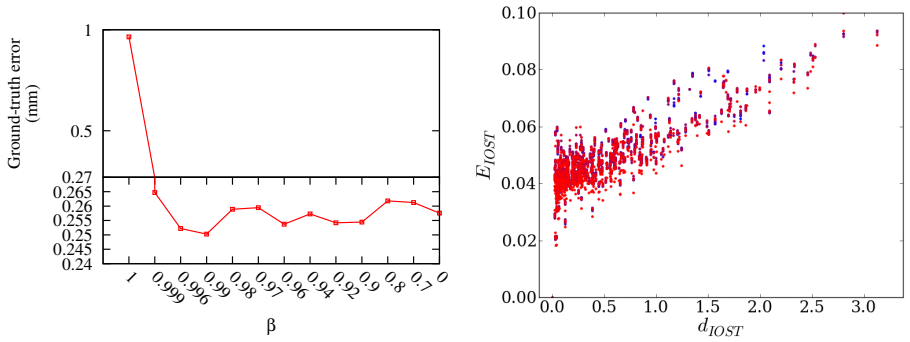

Fig. 4. Graph of the ground-truth error for SyntheticFace across different traversals given by β (left). The relationship of dissimilarity d_{IOST} and tracking error E_{IOST} for SyntheticFace (right). Colour scheme marks data samples from the sequential traversal (red) through $\beta = 1 \rightarrow 0$ to MST(blue).

d_{IOST} for IOST on real data as well. The tracking errors are generally higher than for SyntheticFace because of large changes in the face appearance during deformations.

The dataset DisneyFace contains moderately expressive speech which is tracked even by sequential traversal with small drift. Due to relatively low difficulty of the sequence the visual differences between MST and the best cluster tree $T_{0.996}$ amount to few noticeable glitches on the neck. Although the improvement by the cluster tree is relatively small (similarly for the dataset Face), it is significant because of the importance of accurate facial tracking for visual effects. Quantitative comparison has been performed on DisneyFace with the state-of-the-art non-sequential method for facial performance capture [13]. The difference to the temporally consistent mesh sequence released by Beeler et al.is calculated as for the dataset SyntheticFace with ground truth. The average vertex distance across all frames is $0.312mm$ with the standard deviation $0.357mm$ for the cluster tree $T_{0.996}$. Note that the difference may be due to the errors in either approach. Qualitatively, both techniques achieve comparable accuracy and temporal consistency.

5.3 Cloth

The dataset Garment contains fast free-form motions of a textured top on a subject's upper torso. Sequential alignment leads to fast degradation of the mesh at the beginning of sequence during rapid waving. Due to the partially repetitive motion pattern the number of branches of MST is excessive in some parts of the sequence. The increased presence of cuts causes many noticeable jumps. The cleaner structure of the cluster tree $T_{0.994}$ largely eliminates these artefacts apart from a few visible glitches at the peaks of complicated motions. The difference

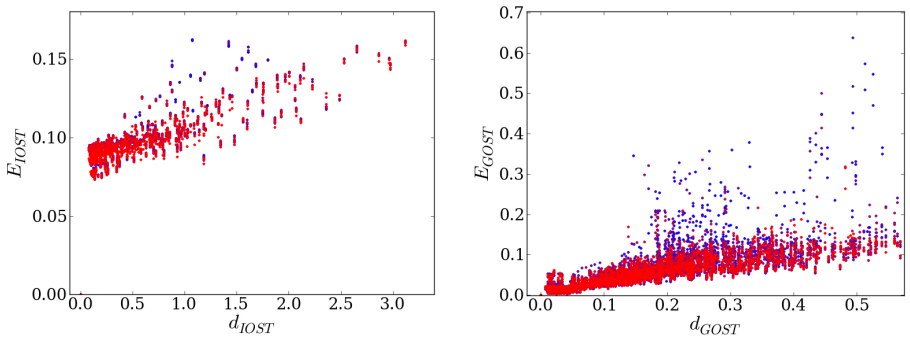

Fig. 5. The relationship of $d_{IOST} - E_{IOST}$ for Face (left) and $d_{GOST} - E_{GOST}$ for StreetDance (right).

between MST and the cluster tree is more apparent than for the face because of the more challenging surface deformations complicated by motion blur.

5.4 Whole-body performance

The subject performing break-dance moves in loose uniform clothing is captured in the dataset StreetDance [17]. The sequence is composited from 3 different performances (Free, KickUp and FlashKick) to demonstrate the ability of non-sequential approaches to align the data across separate motions. The monotonic trend in Figure 5(right) validates the dissimilarity measure d_{GOST} for the GOST technique. However, the graph is more scattered in comparison to Figure 5(left) which caused by a more challenging dataset and use of geometry-based alignment.

The sequential tracking gradually distorts the structure of the mesh but the result by MST does not suffer from this severe slippage on the real surface. However, the mesh jitters during static segments of the performance because of significant re-ordering of frames. The best cluster tree $T_{0.996}$ enforces sequential processing of these segments which leads to a more coherent alignment. Figure 6 shows quantitatively this improvement by means of average acceleration across all vertices. The peaks represent high acceleration related to fast changes of mesh motion manifested as the jitter. $T_{0.996}$ significantly reduces acceleration spikes in a slow-motion segment of StreetDance in comparison to MST. In addition, gross errors in the mesh shape (e.g. artificial connections between limbs) occur frequently for MST during complex movements such as back-flip. They are largely eliminated by $T_{0.996}$ for the price of increased local drift at the peaks of motion. However, this is perceptually more plausible than fast alternation between quite differently distorted meshes. Overall, there is a clear superiority of results by the cluster tree in comparison to MST.

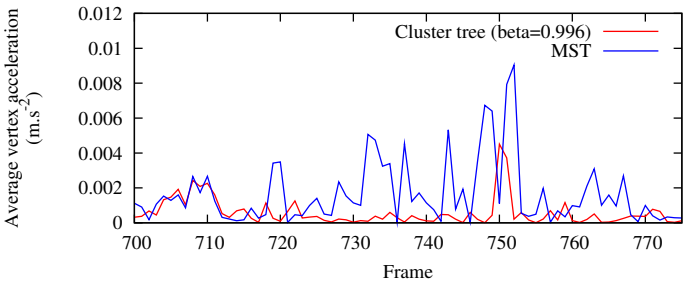

Fig. 6. Average vertex acceleration for MST and the best cluster tree $T_{0.996}$ for a segment in StreetDance where the subject stands still. The peaks representing fast change in motion correspond to high-frequency jitter in the aligned mesh sequence.

5.5 Discussion

The experimental results across different types of surfaces prove the existence of a trade-off between the accumulation of drift and the severity of glitches caused by cuts in temporal ordering. Perceptually, it is beneficial to increase the amount of local drift in the temporally consistent mesh sequence in exchange for the reduced amount of high-frequency jitter or glitches. Cluster trees provide a mechanism to balance this trade-off and therefore achieve results superior to fully sequential traversal or MST.

To analyse the trade-off between jumps and drift across the spectrum of trees, two quantitative measures representing each aspect are proposed. The measure SPL reflects the amount of potential drift in individual frames by a sum of path lengths between the root node and all other nodes (similar to Equation 7). The magnitude of potential glitch between adjacent frames separated by a cut is expressed as a sum of the non-overlapping parts of paths leading to them from the root node. The measure CUT is the total of these sums for all cuts created by the tree. Examples of SPL and CUT profiles across the tree spectrum are depicted in Figure 7 for the dataset SyntheticFace (graphs for the other datasets available online[1]). The trend of SPL across the datasets has a clear maximum for the completely accumulative sequential approach and generally decreases with some fluctuations towards MST. The measure CUT decreases from MST with a large amount of fragmentation towards the sequential traversal without any cuts. The middle range of both measures fluctuates because the different granularity of frame clustering given by β can lead to similar tree shapes. Some cluster trees have worse properties than MST in each measure but the majority of trees show an improvement in both. Intuitively, SPL and CUT should be combined into a single criterion which would express optimality of a tree with respect to the drift and jumps. This would enable automatic selection of the clustering parameter β defining a tree shape. However, any straightforward combination of the measures does not rank the trees consistently across different datasets, so that the order

correlates with visual assessment of the tracking results. A combined criterion defining the optimal traversal tree for sequences with different types of surface deformation is an open problem.

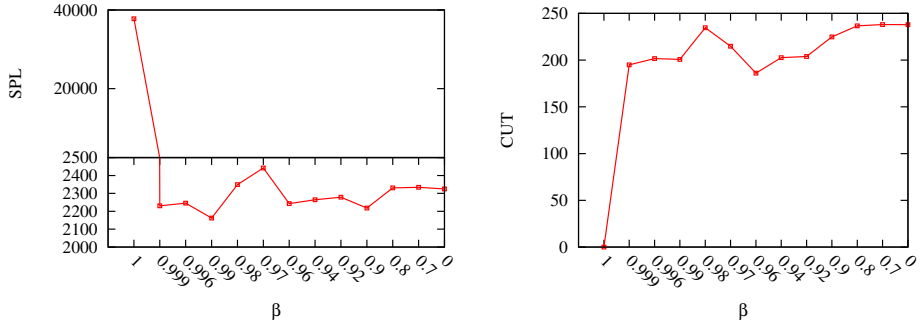

Fig. 7. SPL measure (left) and CUT measure (right) for SyntheticFace across different traversals given by β ($\beta = 1$ - sequential; $\beta = 0$ - MST).

Even with the single criterion reflecting sequential drift versus non-sequential jumps the selected tree is optimal only with respect to the dissimilarity d used. Because it is an approximate measure, the relationship to the actual difficulty of frame-to-frame tracking is not likely to be perfectly linear. This is indicated by the graphs between d and the tracking error E (Figures 4(right),5(left) for IOST and Figure 5(right) for GOST) where the correlation is monotonic but non-linear. This trend validates the use of the chosen measures for guiding the tracking. However, the non-linearity can bias the tree shape away from the ideal result (such as excessive branching due to over-fitting in low range of d which does not influence much the quality of tracking). The consequences of the non-ideal relationship can be alleviated by tuning of the tree shape through β. Even with a perfect dissimilarity the problem of distributing alignment errors across the sequence remains and needs to be optimised by the cluster tree.

6 Conclusion

This paper proposes a cluster tree to non-sequential tracking of non-rigid surface sequences which balances accumulation of errors in frame-to-frame alignment against jumps due to re-ordering of the data. The approach is generalised for any type of non-rigid surface tracked by an arbitrary frame-to-frame method. Evaluation is performed on a variety of datasets including facial, whole-body performances and deformation of cloth. Results demonstrate qualitatively and quantitatively improved temporal alignment against previous sequential and non-sequential minimum-spanning tree approaches.

Acknowledgement: This work was partly supported by EU ICT project SCENE and EPSRC Visual Media Platform Grant.

References

1. Vedula, S., Baker, S., Rander, P., Collins, R., Kanade, T.: Three-dimensional scene flow. TPAMI **27** (2005) 475–480
2. Zhang, L., Snavely, N., Curless, B., Seitz, S.M.: Spacetime faces: high resolution capture for modeling and animation. ACM TOG **23** (2004) 548–558
3. Pons, J.P., Keriven, R., Faugeras, O.: Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. IJCV **72** (2007) 179–193
4. Courchay, J., Pons, J.P., Monasse, P., Keriven, R.: Dense and accurate spatio-temporal multi-view stereovision. In: ACCV. Volume 5995., Springer (2009) 11–22
5. Carceroni, R., Kutulakos, K.: Multi-view scene capture by surfel sampling: From video streams to non-rigid 3d motion, shape and reflectance. IJCV **49** (2002) 175–214
6. Neumann, J., Aloimonos, Y.: Spatio-temporal stereo using multi-resolution subdivision surfaces. IJCV **47** (2002) 181–193
7. Furukawa, Y., Ponce, J.: Dense 3d motion capture from synchronized video streams. In: CVPR, IEEE (2008) 1–8
8. Klaudiny, M., Hilton, A.: Cooperative patch-based 3D surface tracking. In: Conference for Visual Media Production, IEEE Computer society (2011) 67–76
9. Wand, M., Adams, B., Ovsjanikov, M., Berner, A., Bokeloh, M., Jenke, P., Guibas, L., Seidel, H., Schilling, A.: Efficient reconstruction of nonrigid shape and motion from real-time 3D scanner data. ACM TOG **28** (2009) 15:1–15:15
10. Cagniart, C., Boyer, E.: Free-form mesh tracking: A patch-based approach. In: CVPR, IEEE (2010) 1339–1346
11. Tevs, A., Berner, A., Wand, M., Ihrke, I., Bokeloh, M., Kerber, J., Seidel, H.P.: Animation Cartography - Intrinsic Reconstruction of Shape and Motion. ACM TOG **31** (2011) 12:1–12:15
12. Vlasic, D., Baran, I., Matusik, W., Popović, J.: Articulated mesh animation from multi-view silhouettes. ACM TOG **27** (2008) 97:1–97:9
13. Beeler, T., Hahn, F., Bradley, D., Bickel, B.: High-quality passive facial performance capture using anchor frames. ACM TOG **30** (2011) 75:1–75:10
14. Budd, C., Huang, P., Klaudiny, M., Hilton, A.: Global non-rigid alignment of surface sequences. IJCV (2012) DOI: 10.1007/s11263–012–0553–4
15. Huang, P., Budd, C., Hilton, A.: Global temporal registration of multiple non-rigid surface sequences. In: CVPR, IEEE (2011) 3473–3480
16. Huang, P., Hilton, A., Starck, J.: Automatic 3d video summarization: Key frame extraction from self-similarity. In: 3DPVT, IEEE Computer Society (2008)
17. Starck, J., Hilton, A.: Surface capture for performance-based animation. Computer Graphics and Applications **27** (2007) 21–31