High-detail 3D capture and non-sequential alignment of facial performance

Martin Klaudiny, Adrian Hilton
Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, United Kingdom
Email: m.klaudiny, a.hilton@surrey.ac.uk

***Abstract*—This paper presents a novel system for the 3D capture of facial performance using standard video and lighting equipment. The mesh of an actor's face is tracked non-sequentially throughout a performance using multi-view image sequences. The minimum spanning tree calculated in expression dissimilarity space defines the traversal of the sequences optimal with respect to error accumulation. A robust patch-based frame-to-frame surface alignment combined with the optimal traversal significantly reduces drift compared to previous sequential techniques. Multi-path temporal fusion resolves inconsistencies between different alignment paths and yields a final mesh sequence which is temporally consistent. The surface tracking framework is coupled with photometric stereo using colour lights which captures metrically correct skin geometry. High-detail UV normal maps corrected for shadow and bias artefacts augment the temporally consistent mesh sequence. Evaluation on challenging performances by several actors demonstrates the acquisition of subtle skin dynamics and minimal drift over long sequences. A quantitative comparison to a state-of-the-art system shows similar quality of temporal alignment.**

***Keywords*-facial performance capture, dense motion capture, non-sequential surface tracking, colour photometric stereo**

I. Introduction

The 3D capture of facial performance has become a necessary tool for delivering visual effects related to an actor's face or for creating new realistic characters. Two key aspects of creating a digital copy of the performance are temporal consistency and high fidelity. Full alignment of a 3D face model over time enables easy spatio-temporal manipulation and animation rig building. Capture of fine skin detail is essential for high-quality rendering of the face under altered light conditions.

Facial performance capture can be treated as a sequence of static reconstructions. Some current systems provide per-frame models with fine geometric details in real-time [1] and with additional reflectance information which allows relighting [2], [3]. However, the temporally unregistered representation can only be used to replay the performance and is not suitable for editing. Model-based methods implicitly obtain temporal consistency using a generic deformable model. The performance capture is then reduced to optimisation of a limited set of parameters which fits the model to image motion estimates [4] and/or per-frame geometries [5], [6]. The main limitation is imprecise capture of any expressions that are different from the dataset used for creating the deformable model.

The standard approach to temporal alignment is marker-based motion capture which tracks distinct markers placed on the face in video. 3D trajectories of the markers constrain deformation of a high-resolution facial mesh over time. Use of a template mesh provides coarse temporal consistency and does not reproduce skin dynamics [7]. The results can be enhanced by reconstructing time-varying detail using shape-from-shading [8] or photometric stereo [9].

Flow-based motion capture improves the accuracy of template mesh deformation by exploiting 2D optic flows calculated on the image sequences. The motion of mesh vertices between frames is optimised to conform to flow fields and depth maps computed at every frame. This approach has been demonstrated for single-camera [10] and multi-camera setups [11]. Alternatively, flow fields can be combined with a mesh sequence reconstructed by multi-view stereo [12] or structured light [13]. To improve the optic flow computation which often hindered by weak skin texture, some approaches use the rich geometric detail in normal maps obtained by photometric stereo instead of the original images [13], [14]. Due to sequential concatenation of frame-to-frame flows, tracking errors accumulate and cause a drift of the mesh on the face. Bradley et al. [12] correct the drift using additional optic flow estimation in the UV domain of the mesh after the initial deformation but the results are not satisfactory in regions undergoing fast motions.

Furukawa and Ponce [15] do not rely on 2D optic flow but track the mesh vertices directly in the 3D space using deformable surface patches. Fixed patch textures initialised in a reference frame alleviate the drift problem. However, a dense random pattern painted on the face is required to maintain the stability of the method. Majority of techniques track either a prepared template mesh [11], [13] or a reference mesh reconstructed at the first frame [15], [12]. Surface topology changes or missing information pose a problem for these methods. Popa et al. [16] address these with a hierarchical assembly of a globally consistent model from the whole sequence.

Sequential face tracking is unreliable over long and complex performances because of the accumulation of alignment errors. Non-sequential methods tackle this problem

by processing the sequence in a different order than the temporal order. Beeler et al. [17] identify similar frames across the performance and use them to anchor a sequential alignment of intermediate frames using multi-view optic flow. For a whole-body performance, Budd et al. [18] optimise the traversal among all frames of the performance by introducing a minimum spanning tree in shape similarity space to re-order the frame-to-frame alignment process.

From the perspective of performance fidelity, the model-based methods [6], [4], [5] are limited by generality of the deformable model which does not capture subtle, person-specific motions. The systems based solely on multi-view stereo [11], [15], [12] or structured-light reconstruction [6], [5] do not recover fine skin detail. The details can be introduced through high-resolution template obtained by e.g. laser scanning [8]. However, fine skin geometry is unaffected by deformation of the template on larger scale. In contrast, passive methods using diffuse white illumination estimate approximate shape of details from shading at every frame [8], [17]. The systems using photometric stereo based on gradient illumination [13], [14] obtain accurate facial normals and reflectance but require a complex light stage and high-speed image capture. In comparison, photometric stereo using a simple colour lights [1], [9] provides the same accuracy of normals using simple light setup. The drawback is the assumption of uniform surface chromaticity which can be circumvented by applying make-up to the face.

This paper describes a novel system for the 3D capture of facial performance using a practical acquisition setup without active illumination or high-speed recording. Non-sequential surface tracking based on the minimum spanning tree is introduced for the facial performance application. This optimal traversal through the sequence is combined with a frame-to-frame alignment method which is robust against weak texture and fast non-rigid motions. Alignment inconsistencies between different tree branches are resolved by multi-path temporal fusion and the result is a mesh sequence with accurate temporal alignment. Photometric stereo based on three colour lights is used to capture accurate skin geometry up to fine wrinkles and pores. High-detail UV normal maps are combined with the temporally consistent mesh sequence. Experimental results on challenging performances yield aligned 3D models with high-quality detail and minimal drift over long sequences.

II. System overview

The pipeline of the proposed facial performance capture system is illustrated in Figure 1. An actor is captured by two stereo camera pairs under red, green and blue light coming from approximately orthogonal directions.

3D reconstruction: The actor's face is firstly reconstructed at every frame using multi-view stereo. Disparity maps produced by stereo matching between each camera pair are merged into a single mesh by Poisson surface

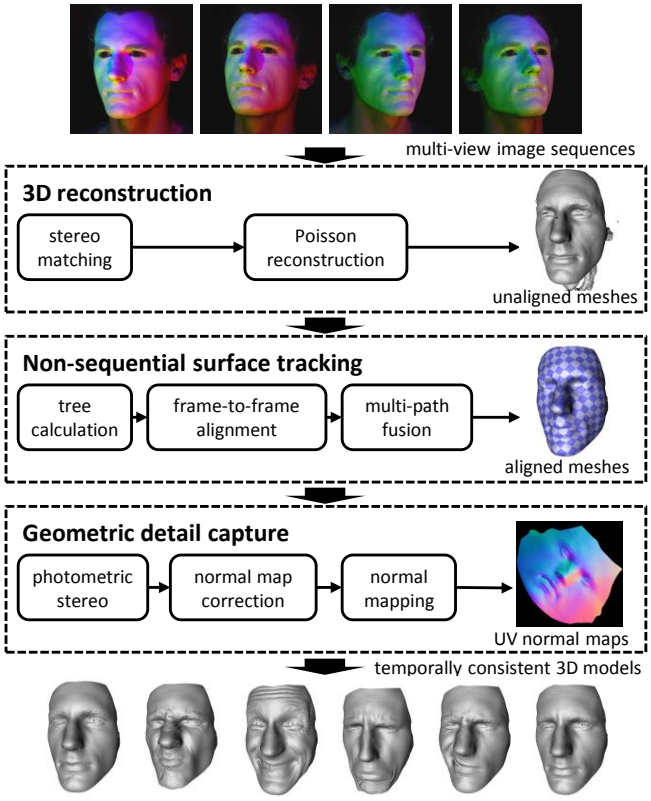

Figure 1. Diagram of the processing pipeline.

reconstruction. The resulting unaligned sequence of meshes constrains the subsequent surface tracking.

Non-sequential surface tracking: At first, a sparse set of facial features is sequentially tracked in the 3D space to establish a pair-wise dissimilarity among all frames. A non-sequential traversal of the sequence is calculated using the minimum spanning tree based on the dissimilarity of feature locations. A user needs to design a face mesh for the root frame of the traversal tree which is then tracked along tree branches to all frames. The tracking is performed using a surface appearance model consisting of textured 3D patches. A frame-to-frame alignment method combines temporal 3D matching of deformable patches with the weighted Laplacian mesh deformation. Alignment inconsistencies between adjacent frames on different tree branches are resolved by multi-path temporal fusion. The result is a temporally aligned mesh sequence with a fixed topology.

Geometric detail capture: Skin detail is captured by photometric stereo using colour lights with the aid of white uniform make-up on the actor's face. Normal maps of the face are obtained at every frame for one view in each stereo pair. Shadow artefacts and low-frequency bias present in the normals are corrected exploiting the base mesh. The normal maps from different views are combined into a single time-varying UV texture which covers the aligned mesh sequence.

The output of the pipeline is a temporally consistent 3D model of the performance capturing subtle dynamics of the face such as skin wrinkling, pore stretching, etc.

III. 3D RECONSTRUCTION

An actor's performance is recorded in several image sequences $\{I_t^c\}_{t=1}^N$ by multiple cameras. The shape of the face is reconstructed at each frame t from a set of images I_t^c (the index c denotes a camera). Initially, stereo matching based on graph cut [19] is performed separately on the images from each stereo pair. A single graph cut on a volumetric graph in the disparity space optimises an energy functional with the linear smoothness term and the data term based on normalised cross-correlation. The resulting disparity map is filtered according to the correlation score to remove unreliable regions. Filtered disparity maps from each stereo pair are converted into clouds of oriented points which are then fused into a single surface by Poisson surface reconstruction [20]. The final mesh G_t captures facial shape up to skin folds and larger wrinkles. Because the 3D reconstruction is performed independently at each frame, the number of vertices and connectivity of G_t vary between frames. Thus, the mesh sequence $\{G_t\}_{t=1}^N$ is temporally unaligned and serves as a shape prior for the subsequent surface tracking.

IV. NON-SEQUENTIAL SURFACE TRACKING

The input is a sequence of measurements $\{O_t\}_{t=1}^N$ of a face for frames $\{t_1, ..., t_N\}$. Without loss of generality it can contain multiple segments of separate performances of the same actor. Each measurement O_t consists of a set of images from multiple viewpoints I_t^c and a mesh G_t representing the current facial shape. 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 a *sequential tracking* which concatenates a *frame-to-frame alignment* between successive frames t_{i-1}, t_i. *Non-sequential tracking* introduced by Budd et al. [18] processes the input sequence $\{O_t\}_{t=1}^N$ in an order different from the temporal order. The re-ordering of $\{O_t\}_{t=1}^N$ is guided by *a dissimilarity* $d(O_{t_i}, O_{t_j})$ of measurements for any pair of frames t_i, t_j which indicates a difficulty of potential alignment between them. Optimal paths through the sequence to every frame can be jointly optimised based on d. The paths are represented by *a traversal tree* $T = (\mathcal{N}, \mathcal{E})$ which is the minimum spanning tree with the nodes $\mathcal{N} = \{n_1, ..., n_N\}$ corresponding to all frames $\{t_1, ..., t_N\}$ (Figure 2). The edges $\mathcal{E} = \{(n_i, n_j), ...\}$ are directed and weighted by the dissimilarity $d(O_{t_i}, O_{t_j})$.

Given T, a user needs to specify the shape and topology of the mesh $M_{t_r} = (X_{t_r}, C)$ for the root node n_r. The defined mesh is subsequently tracked along the branches of T from n_r towards the leaves using a frame-to-frame alignment technique. This reduces accumulation of alignment errors due to shorter paths in comparison to the sequential tracking, and therefore alleviates the common problem of drift over the actual surface.

The non-sequential nature of tracking leads to presence of *cuts* in the sequence at places where two different alignment paths meet in the pair of temporally adjacent frames (marked red in Figure 2). Error accumulation along different paths is independent and this can potentially manifest as glitches or jumps in the aligned mesh sequence. To combat this problem, the tracking is extended across the cuts from each side and multiple solutions for the neighbouring frames are blended into smooth transitions across the cuts. The final result is a temporally consistent mesh sequence $\{M_t\}_{t=1}^N$ which can span across multiple captures of the actor.

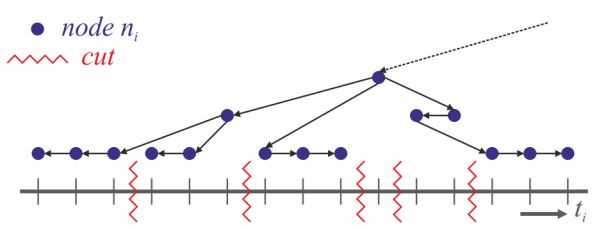

Figure 2. A traversal tree T on the frame sequence $\{t_1, ..., t_N\}$ in the temporal order. The cuts separate adjacent frames which have different alignment paths along tree branches.

A. Patch-based frame-to-frame alignment

The non-sequential traversal of the input sequence using T can be combined with any frame-to-frame alignment technique working with the measurements O_t. The approach based on 3D tracking of surface patches such as [15], [21] is used because of direct and robust 3D motion estimation. The techniques relying on multi-view 2D optic flows [11], [12] have to face artefacts around motion discontinuities, flow inconsistencies across the views and the overhead of their pre-computation.

A surface patch model introduced by Furukawa et al. [15] is built for the user-defined mesh M_{t_r} at the root frame t_r. Each mesh vertex has a patch associated with its adjacent triangle fan, so that any deformation of the mesh also changes the shape of the patch. The patch is covered by multiple textures sampled from individual views. These deformable patches are used to estimate the motion of their respective vertices between successive frames t_i, t_j along the alignment paths set by T. Matching the patch between the frames t_i and t_j is formulated as a two-fold problem. Multi-view patch texture sampled at the previous frame t_i is aligned with the images $I_{t_j}^c$ using full colour information [15]. Simultaneously, the patch position is constrained by the

unaligned geometry G_{t_j} [11]. The patch is rigidly translated in 3D space from its position at t_i to optimise both the image correspondence and the shape fitting.

The optimisation is performed for all patches in a cooperative manner [21]. Their positions are optimised by local random sampling and propagation of intermediate solutions among the neighbouring patches. These two phases are repeated in several iterations across the surface, so that all patches are converging jointly. This assumes that the adjacent patches move in a similar way, but no explicit smoothness is enforced which allows sharp motion discontinuities. The outcome is raw displacements for the mesh vertices between t_i, t_j. This optimisation scheme is robust against rapid complex deformations and weak texture of the surface. Therefore, it achieves better results for a face with the uniform make-up than the techniques based on independent gradient descent for each patch [15].

The estimated vertex displacements form a raw motion field which is regularised by a weighted Laplacian deformation to maintain the motion continuity across the surface [21]. The Laplacian deformation preserves a shape of M_{t_i} subject to the estimated vertex displacements as soft constraints. The constraints are weighted according to the patch matching errors to suppress outliers. The deformation of M_{t_i} to M_{t_j} is efficiently solved in a linear manner for the whole surface at once.

The patch textures are updated at every frame along alignment paths. This adaptivity helps to cope with large changes of the skin appearance due to different facial expressions (wrinkling, stretching/shrinking of pores) and complex colour illumination. The approaches based on a fixed reference texture such as [15] struggle with this scenario. However, the per-frame update of the textures increases the risk of drift since the patch can gradually adapt to a different 3D point. In the case of sequential tracking over fast non-rigid motions, accumulation of alignment errors leads to the drift in spite of robustness of the frame-to-frame algorithm. Therefore, better results are produced if the alignment algorithm is combined with the non-sequential traversal of the sequence.

B. Traversal tree calculation

The dissimilarity measure d is designed as an approximate measure which is significantly easier to compute than the full frame-to-frame alignment of the mesh. It is derived from a sparse set of facial features which are manually selected, so that their motion represents well the overall motion of the face on a coarse level. The selected features are robustly tracked in $\{I_t^c\}_{t=1}^N$ by the linear predictor tracker [22] and their 3D trajectories are obtained by back-projection of 2D trajectories onto $\{G_t\}_{t=1}^N$. The dissimilarity $d(O_{t_i}, O_{t_j})$ represents an average Euclidean distance between 3D positions of the features at the frames t_i, t_j. Beforehand, the sets of positions are rigidly aligned by least-squares minimisation to discard a head pose change between t_i and t_j and allow a comparison of facial expressions only. The rigid alignment is used to initialise the non-rigid alignment algorithm to simplify the computation between t_i, t_j. The measure d is more appropriate for surface tracking than shape histogram comparison [18] or image correlation [17] which do not directly reflect motion of the surface.

The space of all possible pairwise 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 the individual frames (Figure 3(left)). The elements $D(i,j) = d(O_{t_i}, O_{t_j})$ are proportional to the difficulty of alignment between the frames t_i and t_j. The matrix is symmetrical $(d(O_{t_i}, O_{t_j}) = d(O_{t_j}, O_{t_i}))$ and has a zero diagonal $(d(O_{t_i}, O_{t_i}) = 0)$. The optimal traversal in this space can be found through the following graph formulation [18].

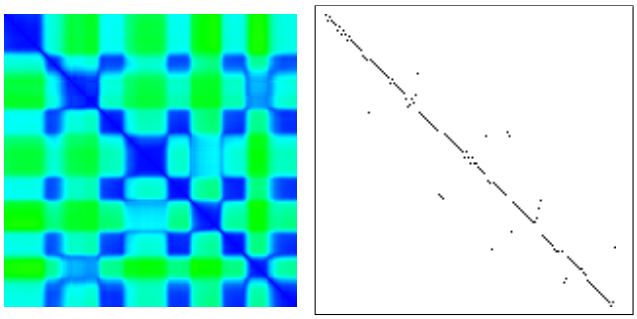

Figure 3. The dissimilarity matrix D for the dataset Actor1 (frames $150-299$; blue/red = low/high values) (left). The tree T depicted in the space of D where each directed edge (t_i, t_j) is marked black at its respective location $D(i,j)$ (right).

A fully-connected undirected graph $H = (\mathcal{N}, \mathcal{D})$ 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 weights $D(i,j)$. A traversal visiting all frames is described by a undirected spanning tree $T'_s = (\mathcal{N}, \mathcal{E}')$ where $\mathcal{E}' \subset \mathcal{D}$. The optimal tree T' is defined as the *minimum spanning tree* (MST) which minimises the total cost of pairwise alignment given by d:

$$T' = \underset{\forall T'_s \subset H}{\operatorname{argmin}} \left(\sum_{\forall (n_i, n_j) \in T'_s} D(i,j) \right). \qquad (1)$$

This objective describes total non-rigid deformation of the surface which has to be overcome following the traversal tree, and is optimised by the Prim's algorithm. The benefit of MST is that low-cost transitions are at the centre of tree 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 tree T' does not exactly describe a traversal of the sequence because it is undirected and does not have an explicitly defined root node. The root node n_r is selected optimally in contrast to [18], [17] to make the traversal as short as possible in terms of d for every frame. The length of weighted paths $n_l \rightarrow n_k$ from the candidate root node n_l to all other nodes n_k has to be minimal:

$$n_r = \underset{n_l \in \mathcal{N}}{\operatorname{argmin}} \left(\sum_{\forall n_k \in T'_\beta} \sum_{\forall (n_i, n_j) \in n_l \rightarrow n_k} D(i,j) \right). \quad (2)$$

The final traversal tree T (Figure 3(right)) is created from T' by setting the direction of the edges in $\mathcal{E}'$ according to a breadth-first search from n_r towards the leaf nodes.

C. Multi-path temporal fusion across tree branches

The drawback of non-sequential tracking based on a tree is the cuts between adjacent frames with different alignment paths (Figure 2) which can cause jumps in the resulting mesh sequence [18]. To ensure smooth transitions, the tree T is expanded with branches of a length m which extend the tracking across each cut in both directions. The expanded tree $\tilde{T} = (\tilde{\mathcal{N}}, \tilde{\mathcal{E}})$ includes T ($\mathcal{N} \subset \tilde{\mathcal{N}}, \mathcal{E} \subset \tilde{\mathcal{E}}$). New branches are added for every pair of adjacent frames (t_{i-1}, t_i) where $(\tilde{n}_{i-1}, \tilde{n}_i) \wedge (\tilde{n}_i, \tilde{n}_{i-1}) \notin \mathcal{E}$. A chain of new nodes with interconnecting edges is created for the frames $\{t_{i-m}, ..., t_{i-1}\}$ and linked to the original node $\tilde{n}_i \in \mathcal{N}$ (similarly for $\{t_i, ..., t_{i-1+m}\}$ and the node $\tilde{n}_{i-1}$).

After the expansion a frame can have multiple nodes associated with it which yield multiple solutions through different alignment paths. To combine them, every node $\tilde{n}_k \in \tilde{\mathcal{N}}$ is weighted by a coefficient

$$b_{\tilde{n}_k} = \left(1 / \sum_{\forall (\tilde{n}_i, \tilde{n}_j) \in \tilde{n}_r \rightarrow \tilde{n}_k} D(i,j) \right) \cdot \left(1 - \sum_{\forall (\tilde{n}_i, \tilde{n}_j) \in \tilde{n}_c \rightarrow \tilde{n}_k} \frac{1}{m+1} \right), \quad (3)$$

where the first term represents a tracking confidence from the root $\tilde{n}_r$ to $\tilde{n}_k$ as an inverse of accumulated dissimilarity. The second term linearly decreases the weight along the additional nodes where $\tilde{n}_c$ is the last node from the original T along the path $\tilde{n}_r \rightarrow \tilde{n}_k$ (equals 1 for $\tilde{n}_k \in \mathcal{N}$).

The final vertex positions X_t for the mesh M_t are blended from all candidate positions $X_{\tilde{n}_k}$ for the frame t as follows:

$$X_t = \frac{1}{\sum_{\tilde{n}_k \in t} b_{\tilde{n}_k}} \sum_{\tilde{n}_k \in t} b_{\tilde{n}_k} X_{\tilde{n}_k}. \quad (4)$$

Simple linear blending with the normalised coefficients $b_{\tilde{n}_k}$ has proved to be sufficient for short extensions across the cuts ($m = 3$). This produces the resulting mesh sequence $\{M_t\}_{t=1}^{N}$ which is temporally consistent.

V. Geometric detail capture

Photometric stereo based on colour lights [10] is used to capture fine skin geometry which is not present in the tracked mesh. This method calculates accurate surface normals from a single image of the face in contrast to other passive methods which approximate details from shading [8], [17]. It is also practical for dynamic face capture as shown by Hernandez et al. [1] because it does not require time-multiplexing light stage [13].

The main limitation is an assumption about uniform chromaticity of the surface which is not valid for skin. Therefore, uniform white make-up is applied on the actor's face which also fits better to the Lambertian reflection model assumed by the photometric stereo. Moreover, the make-up prevents sub-surface scattering of light in the skin which blurs the calculated normals. The application of make-up is not necessary for the technique but significantly improves the quality of normals as shown in [23], [9]. The calibration method presented in [9] is used for estimation of light directions and light-sensor interaction. High-detail normal maps of the face are computed at every frame for one view in each stereo pair.

A. Normal map correction

The normal maps obtained contain artefacts in shadow regions where one or more directional lights are occluded by other parts of the face [23]. The normals in these regions have a skewed orientation because of missing constrains from the occluded lights. Also, there is a weak low-frequency bias in the whole normal map due to errors in the photometric calibration [1]. These imperfections are corrected using the tracked mesh M_t. The correction procedure presented by Klaudiny et al. [9] is modified for this purpose.

The shadows cast by the individual R, G, B lights are segmented by thresholding a ratio between the respective component and the whole RGB vector (relies on a good spectral match between each colour light and camera sensor). The normals in the regions with a single light occluded are recalculated using 2 valid constraints from the other lights and a normal from the base mesh. This results in 3 additional normal layers to the original normal map. One extra layer is added for the regions with multiple lights occluded (or regions with very dark appearance) where the incorrect normals are replaced by the smooth mesh normals.

The low-frequency bias is eliminated by the first stage of the technique by Nehab et al. [24] which conforms overall orientation of the normals to a shape of the mesh. The bias correction is applied separately to the original normal map and each single-shadow layer in contrast to [9] because they are biased differently. This makes the corrected regions consistent with each other. All layers are fused together according to the initial shadow segmentation into the final normal map. Individual regions are linearly blended with each other over a 10-pixel range to avoid visual seams.

B. Normal mapping

The corrected normal maps are projected from their views onto the mesh M_t at every frame. Each half of the face is textured by a single normal map from the corresponding side view. The time-varying texture with normals is stored in a common UV domain where the fixed topology of M_t is unwrapped. This enables easy temporal editing of the surface normals where a modification in the UV domain is implicitly propagated throughout the whole sequence based on the temporal consistency of $\{M_t\}_{t=1}^{T}$. Rendering of the face is solely based on the normal map and the mesh just provides an underlying shape.

VI. Results

An actor's performance is recorded by 4 HD cameras at $25fps$ in HD-SDI uncompressed $4:4:4$ format for high colour fidelity (1920×1080 pixels). They are arranged into two vertical stereo pairs with narrow baseline on each side of the actor's face. All cameras are synchronised and fully calibrated with respect to a common coordinate system. Gamma correction is switched off to preserve a linear response of the sensors for the photometric stereo. Passive directional illumination is provided by three light projectors which illuminate the face from roughly orthogonal directions. The projected illumination is colour-filtered to spectrally differentiate the light sources (red, green, blue light). The spectral characteristics of the filters are approximately matched to the spectral sensitivities of the R, G, B sensors. Each illumination is then captured only by the respective sensor.

The system has been evaluated on various performances of two actors. Due to dynamic nature of the results the reader is encouraged to watch supplementary videos[1]. Only snapshots from an example sequence Actor1 are presented in the paper. The dataset Actor1 is 299 frames long and provides a variety of exaggerated expressions changing at a fast pace. The typical set of features required for dissimilarity computation is located in the eye corners, on the eyebrows, around the nose and on the lip contour. In this case, 14 features has been landmarked and tracked in the 3D space throughout the sequence. The traversal tree has 51 branches starting at the root frame 271 (average branch length is ~ 25 frames). The temporally consistent mesh has 2689 vertices and 5248 faces and the associated UV normal map has a resolution 1500×1500 pixels. Both resolutions can be scaled up at the expense of longer computational time and larger data size. Complete processing for a single frame takes on average 5-6 minutes on 2.5 GHz processor (including overhead of multiple alignments for the frames around cuts). Single-threaded C++ code is used, but the processing stages related to a single view or a stereo pair are run in parallel.

[1]Supplementary material is available under: http://cvssp.org/projects/face3d/3dimpvt12/index.html

The example frames in Figure 4 illustrate correct reconstruction of a facial shape up to fine skin structure (pores, small wrinkles). The approach is able to handle extensive surface deformations in spite of the uniform white make-up and the colour illumination which alters the surface appearance over time. The mesh sequence is accurately aligned over time which is demonstrated by a fixed artificial texture locked down to the surface throughout the performance (Figure 4(middle)). Small drift occasionally appears in eyes and inside of mouth because of their view-dependent appearance. These regions have also inaccurate normals because of strong non-Lambertian properties and instability of the shadow segmentation as visible in Figure 4(bottom). Their changing orientation causes a noticeable flicker over time. The geometric detail is well corrected in the shadow regions. However, it appears slightly smoother than the rest of the face if the light comes from a similar direction as the original occluded light source.

To demonstrate the benefit of non-sequential tracking using MST, the dataset Actor1 is processed sequentially using the same frame-to-frame alignment algorithm. Figure 5 shows significantly worse temporal alignment in comparison to Figure 4(middle). Mesh distortions appear during fast expression changes and the tracking is not able to recover completely. Thus, the mesh gradually drifts over time around the eyes and the nose especially. In the supplementary material[1], there is also included a comparison with/without multi-path temporal fusion which illustrates elimination of sudden jumps at the places of cuts.

The proposed approach is compared to the state-of-the-art technique by Beeler et al. [17]. Because this technique estimates the geometric detail from the skin appearance under diffuse white illumination, it is not possible make direct comparison with our method requiring colour lights. However, the photometric stereo provides metrically correct normals in contrast to an approximation derived from skin appearance. The comparison is focused on temporal consistency of the mesh sequence. Beeler et al. track the face non-sequentially using a sparse set of anchor frames. This can be seen as an suboptimal traversal tree with direct transitions from the root frame to all anchor frames and sequential branches between the anchor frames.

The evaluation of both approaches is performed on the dataset Disney (courtesy of ETH Zurich/Disney Research [17]). An actor has been captured by 7 cameras at $46fps$ (1176×864 pixels). The sequence is 346 frames long and contains a moderately expressive speech. MST is computed using the dissimilarity based on 15 facial points. It has 44 branches starting at the root frame 216 (average branch length is ~ 26 frames). The tracked mesh has 20000 vertices and 39810 faces. The result for this dataset provided by Beeler et al. is sub-sampled to the same resolution to allow a direct quantitative comparison. The difference between the temporally consistent mesh sequences is expressed as an

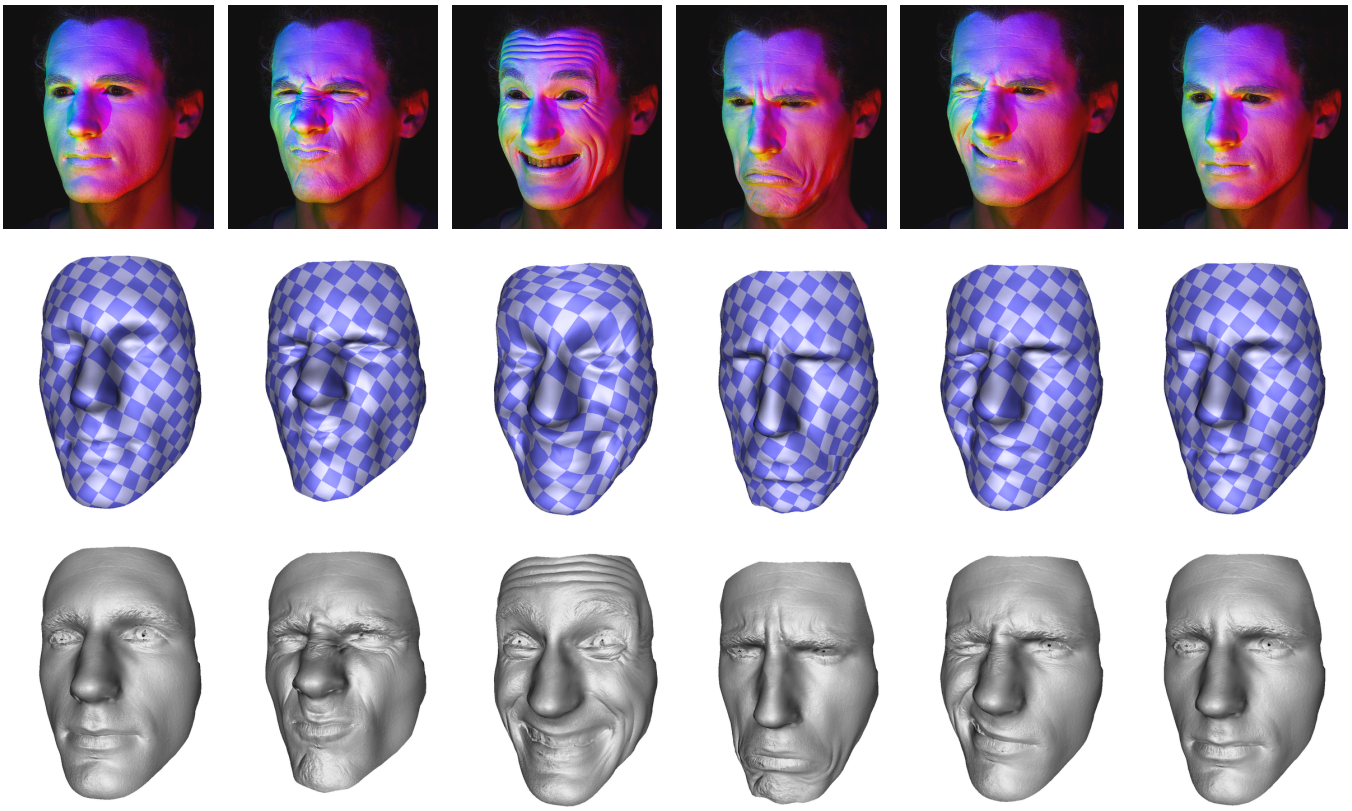

Figure 4. Snapshots from the temporally consistent 3D model for the dataset Actor1 - input images from one of the views (top), meshes rendered with a fixed UV texture (middle) and meshes rendered with a time-varying UV normal map (bottom). The results are from the frames 0, 40, 120, 165, 250, 299.

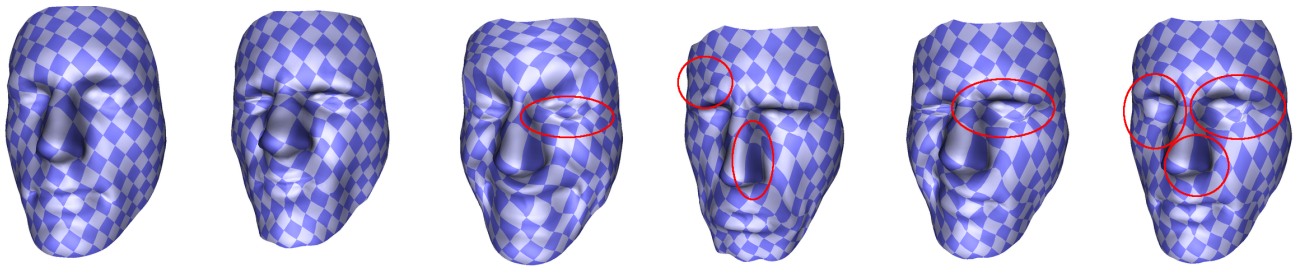

Figure 5. Snapshots from the sequential tracking of the dataset Actor1. Comparison to the non-sequential tracking in Figure 4 shows significant drift distorting the fixed texture (the most apparent distortions and drift are marked red).

Euclidean distance averaged across all mesh vertices for all frames: mean = $0.278mm$, standard deviation = $0.487mm$ (without multi-path temporal fusion). Spatial distribution of the difference is depicted in Figure 6 for selected frames. Note that the discrepancy may be due to the errors in either approach. Qualitatively, both techniques achieve accurate temporal alignment of comparable quality. The proposed approach suffers from a bit larger drift on inner lip in some situations.

VII. Conclusion

We have presented a facial performance capture system which combines robust non-sequential surface tracking with detail capture by photometric stereo with colour lights. The results demonstrate the ability of the system to acquire fine skin geometry and produce a stable temporally-consistent facial model comparable with state-of-the-art methods. This approach can be used for aligning multiple performances of the same actor. Also, there are no face-specific assumptions in this technique, thus it can be applied on other dynamic surfaces with similar shape and motion complexity as a face.

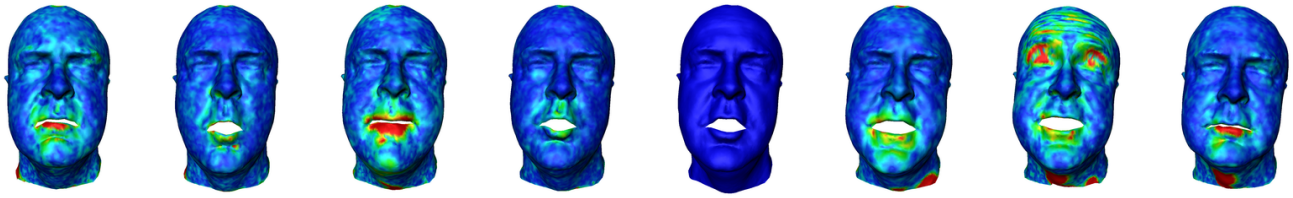

Figure 6. Comparison of the non-sequential tracking using MST and using anchor frames (Beeler et al. [17]) on the dataset Disney. An Euclidean distance between corresponding vertices is visualised across the face (blue = $0mm$, red = $2mm$). The compared meshes are from the frames: 48, 78, 114, 175, 216, 258, 333, 394.

ACKNOWLEDGEMENT

We would like to thank Alaleh Rashidnasab for being a test actor and T. Beeler et al. at ETH Zurich/Disney Research for releasing their dataset. Also, this work was partly supported by EU ICT project SCENE and EPSRC Visual Media Platform Grant.

REFERENCES

[1] G. Vogiatzis and C. Hernández, "Self-calibrated, Multi-spectral Photometric Stereo for 3D Face Capture," *IJCV*, vol. 97, no. 1, pp. 91–103, 2012.

[2] W.-C. Ma, T. Hawkins, P. Peers, C.-F. Chabert, M. Weiss, and P. Debevec, "Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination," in *Eurographics Symposium on Rendering*, 2007.

[3] G. Fyffe, T. Hawkins, C. Watts, W.-C. Ma, and P. Debevec, "Comprehensive Facial Performance Capture," *CGF*, vol. 30, pp. 425–434, 2011.

[4] D. Vlasic, M. Brand, H. Pfister, and J. Popović, "Face transfer with multilinear models," *ACM TOG*, vol. 24, pp. 426–433, 2005.

[5] T. Weise, H. Li, L. Van Gool, and M. Pauly, "Face/Off: live facial puppetry," in *SCA*, 2009.

[6] Y. Wang, X. Huang, C.-S. Lee, S. Zhang, Z. Li, D. Samaras, D. Metaxas, A. Elgammal, and P. Huang, "High Resolution Acquisition, Learning and Transfer of Dynamic 3-D Facial Expressions," *CGF*, vol. 23, pp. 677–686, 2004.

[7] I.-C. Lin and M. Ouhyoung, "Mirror MoCap: Automatic and efficient capture of dense 3D facial motion parameters from video," *The Visual Computer*, vol. 21, pp. 355–372, 2005.

[8] B. Bickel, M. Botsch, R. Angst, W. Matusik, M. Otaduy, H. Pfister, and M. Gross, "Multi-scale capture of facial geometry and motion," *ACM TOG*, vol. 26, no. 3, p. 33, 2007.

[9] M. Klaudiny, A. Hilton, and J. Edge, "High-detail 3d capture of facial performance," in *3DPVT*, 2010.

[10] G. J. Brostow, C. Hernandez, G. Vogiatzis, B. Stenger, and R. Cipolla, "Video Normals from Colored Lights," *TPAMI*, vol. 33, pp. 2104 – 2114, 2011.

[11] L. Zhang, N. Snavely, B. Curless, and S. M. Seitz, "Spacetime faces: high resolution capture for modeling and animation," *ACM TOG*, vol. 23, pp. 548–558, 2004.

[12] D. Bradley, W. Heidrich, T. Popa, and A. Sheffer, "High Resolution Passive Facial Performance Capture," in *SIGGRAPH*, 2010.

[13] W.-C. Ma, A. Jones, J.-Y. Chiang, T. Hawkins, S. Frederiksen, P. Peers, M. Vukovic, M. Ouhyoung, and P. Debevec, "Facial performance synthesis using deformation-driven polynomial displacement maps," *ACM TOG*, vol. 27, pp. 1–10, 2008.

[14] C. A. Wilson, A. Ghosh, P. Peers, J.-Y. Chiang, J. Busch, and P. Debevec, "Temporal upsampling of performance geometry using photometric alignment," *ACM TOG*, vol. 29, pp. 1–11, 2010.

[15] Y. Furukawa and J. Ponce, "Dense 3d motion capture for human faces," in *CVPR*, 2009.

[16] T. Popa, I. South-Dickinson, D. Bradley, A. Sheffer, and W. Heidrich, "Globally Consistent Space-Time Reconstruction," in *CGF*, 2010.

[17] T. Beeler, F. Hahn, D. Bradley, and B. Bickel, "High-quality passive facial performance capture using anchor frames," *ACM TOG*, vol. 1, no. 212, p. 1, 2011.

[18] C. Budd, P. Huang, M. Klaudiny, and A. Hilton, "Global non-rigid alignment of surface sequences," *International Journal of Computer Vision*, pp. 1–15, 2012.

[19] S. Roy, "Stereo without epipolar lines: A maximum-flow formulation," *IJCV*, vol. 34, pp. 147–161, 1999.

[20] M. Kazhdan and M. Bolitho, "Poisson surface reconstruction," in *SGP*, 2006.

[21] M. Klaudiny and A. Hilton, "Cooperative patch-based 3D surface tracking," in *CVMP*, 2011.

[22] E. J. Ong, Y. Lan, B. J. Theobald, R. Harvey, and R. Bowden, "Robust Facial Feature Tracking using Selected Multi-Resolution Linear Predictors," in *ICCV*, 2009.

[23] C. Hernández, G. Vogiatzis, and R. Cipolla, "Shadows in three-source photometric stereo," in *ECCV*, 2008.

[24] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi, "Efficiently combining positions and normals for precise 3d geometry," *ACM TOG*, vol. 24, pp. 536–543, 2005.