High Resolution Tracking of Non-Rigid 3D Motion of Densely Sampled Data Using Harmonic Maps

Yang Wang[1], Mohit Gupta[1], Song Zhang[2], Sen Wang[1], Xianfeng Gu[1], Dimitris Samaras[1], Peisen Huang[2]
[1] Computer Science Department, Stony Brook University, Stony Brook, NY USA
{yangwang,mogupta,swang,gu,samaras}@cs.sunysb.edu

[2] Mechanical Engineering Department, Stony Brook University, Stony Brook, NY USA
{song.zhang,peisen.huang}@sunysb.edu

Abstract

We present a novel fully automatic method for high resolution, non-rigid dense 3D point tracking. High quality dense point clouds of non-rigid geometry moving at video speeds are acquired using a phase-shifting structured light ranging technique. To use such data for the temporal study of subtle motions such as those seen in facial expressions, an efficient non-rigid 3D motion tracking algorithm is needed to establish inter-frame correspondences. The novelty of this paper is the development of an algorithmic framework for 3D tracking that unifies tracking of intensity and geometric features, using harmonic maps with added feature correspondence constraints. While the previous uses of harmonic maps provided only global alignment, the proposed introduction of interior feature constraints guarantees that non-rigid deformations will be accurately tracked as well. The harmonic map between two topological disks is a diffeomorphism with minimal stretching energy and bounded angle distortion. The map is stable, insensitive to resolution changes and is robust to noise. Due to the strong implicit and explicit smoothness constraints imposed by the algorithm and the high-resolution data, the resulting registration/deformation field is smooth, continuous and gives dense one-to-one inter-frame correspondences. Our method is validated through a series of experiments demonstrating its accuracy and efficiency.

1. Introduction and Previous Work

Automatic tracking of non-rigid 3D motion is essential in many computer vision and graphics applications, especially dynamic facial expression analysis, such as facial expression recognition, classification, detection of emotional states, etc. In the literature, most non-rigid object tracking and registration algorithms utilize image data from 2D video sequences, e.g. [34, 4, 20, 1, 6, 31, 12, 13, 33, 30, 24]. Previous methods establishing 3D inter-frame correspondences for non-rigid motion largely fall into two categories: One depends on markers attached to the object [15, 19, 2] or on feature correspondences manually selected by the users [21]; the other calculates correspondences based on the geometry using a 3D deformable/morphable model [23, 29, 13, 36, 7, 5, 35, 9], or other 3D shape registration algorithms such as [3, 39]. In general, most of the existing methods rely on templates with relatively few degrees of freedom. While the recovered low dimensional configurations can often be used effectively in classification, they are hardly sufficient in many analysis applications, especially dynamic facial expression analysis, since many distinct characteristics of a person's expression lie in the subtle details such as the wrinkles and the furrows that are generated by highly local skin deformations. The major contribution of this paper is the use of the elements of conformal geometry theory for the 3D tracking problem, which to the best of our knowledge, has not been attempted before. Although our method was implemented in the context of facial expression tracking, it is general and could be applied to other classes of similarly deforming objects.

Recent technological advances in digital imaging, digital projection display and personal computers have made real time 3D shape acquisition increasingly more feasible. Such ranging techniques include structured light [26, 25], and space-time stereo [38, 8]. These systems can capture dense 3D data at a high frame rate. Recently, a high-resolution 3D expression data acquisition system was developed in [26] which captures highly accurate geometry at speeds that exceed regular video frame rate. Such high-quality data is very attractive for the analysis of facial expressions. However, since the dense data samples in these 3D face scans are not registered in object space, inter-frame correspondences can not be established, which makes the tracking of facial features, temporal study of facial expression dynamics and other analysis difficult. For this purpose, a number of track-

Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)
1550-5499/05 $20.00 © 2005 IEEE

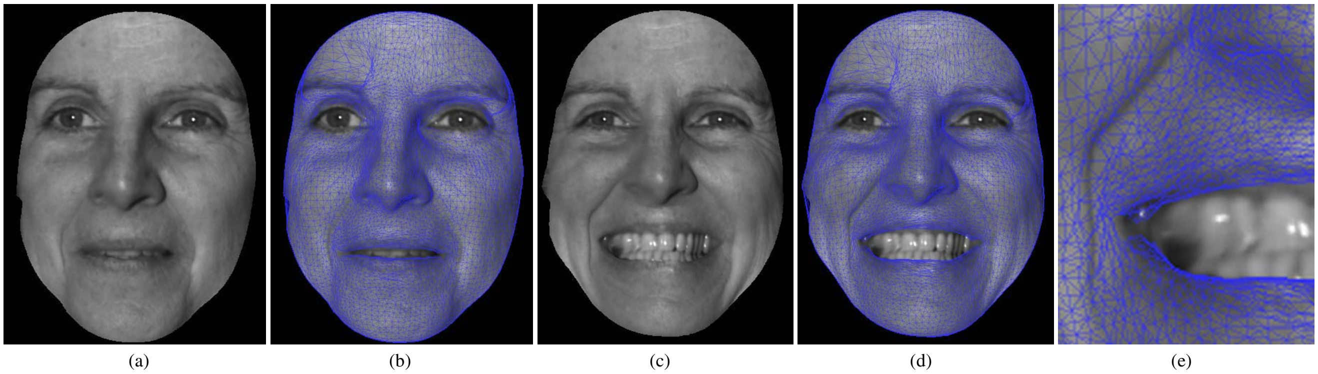


Figure 1. Snapshots from a tracking Sequence: a) Initial data frame b) Initial tracked frame c) Data at the expression peak d) Tracked data at the peak e) Close-up at the peak

ing algorithms have been proposed recently for 3D facial expression data [38, 32]. Tracking methods based on optical flow estimation[38, 13] can be sensitive to noise for textureless regions. A hierarchical tracking framework for high resolution 3D dynamic expression data was presented in [32], using a deformable generic face model. However, it suffers from problems like*folding* and *clustering*, which are inherent to the methods employing local optimization techniques such as FFD. Also, the face model needs to be manually divided into several deformable regions, with associated shape and motion control parameters. This initial segmentation, along with the associated parameters has to be recovered statistically, requiring many experiments for each different expression of every subject. Although this might be acceptable for certain applications like motion capture for computer graphics, it requires prohibitive amounts of time and effort for processing of the large number of datasets required for data driven applications in facial expression analysis and synthesis[5].

In this paper, we present a novel fully automatic method for high resolution, non-rigid dense 3D point tracking. High quality dense point clouds of facial geometry moving at video speeds are acquired using a phase-shifting based structured light ranging technique [26]. To use such data for temporal study of the subtle dynamics in expressions, an efficient non-rigid 3D motion tracking algorithm is needed to establish inter-frame correspondences. In this paper, we propose such an algorithmic framework that uses a mathematical tool called harmonic maps [27, 22, 10, 11]. Harmonic maps were used in [37] to do surface matching, albeit focusing on rigid transformations. Given the source manifold M and the target manifold D, only the boundary condition $u|_{\partial M} : \partial M \rightarrow \partial D$ was used to constrain and uniquely determine the harmonic map $u : M \rightarrow D$. For applications like high resolution facial tracking though, we need to account for non-rigid deformations, with a high level of accuracy. To this end, we introduce additional feature correspondence constraints, in addition to the boundary constraint in our implementation of harmonic maps. We select a set of *motion-representative* feature corners (for example, for facial expression tracking, we select corners of eyes, lips, eye brows etc.) and establish inter-frame correspondences using commonly used techniques (for example, hierarchical matching used in [34]). We can then integrate these correspondence constraints with the boundary condition to calculate harmonic maps, which not only account for global rigid motion, but also subtle non-rigid deformations and hence achieve high accuracy registration and tracking.

The theory of harmonic maps is based on conformal geometry theory [14, 28]; the harmonic map between two topological disks is a diffeomorphism with minimal stretching energy and bounded angle distortion. Harmonic maps are invariant for the same source surface with different poses, thus making it possible to account for global rigid motion. Harmonic maps are highly continuous, stable and robust to noise. They depend on the geometry in a continuous manner. A very important property, which governs our registration and tracking algorithm is that the harmonic map is one-to-one. To register two frames, we align their respective harmonic maps as closely as possible by imposing the suitable boundary and feature constraints. *The motivation to do so is to establish a common parametric domain for the two surfaces*, which, coupled with the above mentioned property, allows to recover 3D registration between the two frames. In our case, the harmonic maps are diffeomorphisms, that is one to one and on-to, and hence lend themselves as a natural choice for surface parameterization in tracking applications.

As part of our framework, a deforming generic face model is employed to track the dense 3D data sequence moving at video speeds, with the harmonic maps guiding the deformation field. The harmonic maps are constrained, and hence driven by the feature correspondences established between adjacent frames using an *iterative scheme*; the feature correspondences are made on texture and curvature images using standard techniques, such as corner detection and optical flow. Most surface regions have strong features either in intensity or shape images. Our framework uses both simultaneously providing denser feature tracking. Harmonic maps, thus, help us to simplify a 3D surface registration problem to a 2D image matching problem. The resulting harmonic map provides dense registration between the face model and the target frame, thereby computing the motion vectors for the vertices of the generic face model. Our system can not only track global facial motion that is caused by muscle action, but also subtler expression details that are generated by highly local skin deformations. We have achieved high accuracy tracking results on facial expression sequences, which are comparable to those reported in [32, 17], using the same dense 3D data, while minimizing the amount of human labor required for preprocessing and initialization. The above mentioned level of accuracy, coupled with the *automatic nature* of our method, demonstrates the merits of our framework for the purpose of high resolution tracking of non-rigid 3D motion.

The remainder of the paper is organized as follows: In Section 2, we give an overview of harmonic mapping. Section 3 explains our tracking method in detail. We first describe the global alignment of 3D scans, followed by a description of the registration algorithm based on harmonic mapping and an iterative refinement scheme using local features. Experimental results are presented in Section 4, and we conclude with a discussion and future work in Section 5.

2. Harmonic Mapping

An important contribution of our tracking method is to reduce the non-rigid 3D tracking problem to a 2D image registration problem, which has been extensively studied. We are dealing with 3D surfaces, but since they are manifolds, they have an inherent 2D structure which can be exploited to make the problem more tractable using harmonic maps.

A harmonic map $H : M \rightarrow D$ can be viewed as an embedding from a manifold M with disk topology to a planar graph D. A harmonic map is a critical point for the harmonic energy functional,

$$E(H) = \int_M |dH|^2 d\mu M,$$

and can be calculated by minimizing $E(H)$. The norm of the differential $|dH|$ is given by the metric on M and D, and $d\mu M$ is the measure on M [27, 22, 10, 11]. Since our source manifold M is in the form of a *discrete* triangular mesh, we approximate the harmonic energy as [10, 37, 14],

$$E(H) = \sum_{[v_0,v_1]} k_{[v_0,v_1]} |H(v_0) - H(v_1)|^2, \qquad (1)$$

where $[v_0, v_1]$ is an edge connecting two neighboring vertices v_0 and v_1, and $k_{[v_0,v_1]}$ is defined as

$$\frac{1}{2}\left(\frac{(v_0 - v_2) \cdot (v_1 - v_2)}{|(v_0 - v_2) \times (v_1 - v_2)|} + \frac{(v_0 - v_3) \cdot (v_1 - v_3)}{|(v_0 - v_3) \times (v_1 - v_3)|}\right),$$

where $\{v_0, v_1, v_2\}$ and $\{v_0, v_1, v_3\}$ are two conjuncted triangular faces.

By minimizing the harmonic energy, a harmonic map can be computed using the Euler-Lagrange differential equation for the energy functional, i.e. $\Delta E = 0$, where Δ is the Laplace-Beltrami operator [27, 22, 10, 11]. This will lead to solving a sparse linear least-square system for the mapping H of each vertex v_i [10, 37, 14]. If the boundary condition is given,

$$H|_{\partial M} : \partial M \rightarrow \partial D, \qquad (2)$$

then the solution exists and is unique.

For tracking purposes though, we need to align the two harmonic maps closely together (as explained in Section 1), and hence track interior non-rigid deformations as well. For this purpose, we also incorporate additional hard constraints to establish interior feature correspondences and to handle non-disk topologies (e.g., a 3D face scan with an open mouth). Suppose we have a point on an inner-boundary or an interior feature point v_i on the 3D mesh M, which should be mapped to a corresponding point w_i on the target 2D plane D. We can add it as a hard constraint $H(v_i) = w_i$ to the system from Equation 1 and 2. However, the resulting harmonic energy is expected to increase due to the additional hard constraints introduced. In order to reduce the energy to achieve a smoother mapping, we use the Neumann boundary condition, a soft constraint. This condition just constrains the boundary points of M to lie on the boundary of the $2D$ disk D, the exact positions being governed by the minimization of harmonic energy criteria. It is different from the fixed boundary condition used for surface matching [37], in which each boundary point on the $3D$ mesh M is mapped to a fixed point on the $2D$ disk, making it a hard constraint. In our method, all the interior feature correspondences on the face scans which can be reliably established are given the maximum weight, and hence are chosen as *hard constraints*. In the absence of any strong features on the boundary, the boundary condition is given a relatively lower weight, and hence the *soft boundary constraint* is employed to minimize the harmonic energy.

Intuitively, consider the manifold M to be made of a sheet of rubber [10]. The harmonic map with just the boundary constraint can be thought of as stretching the boundary of M over the boundary of the target $2D$ disk D. In this case, each point on the boundary of M is assigned a fixed location on the boundary of D, where it will be *nailed down*. The interior of the sheet then rearranges to minimize the stretching(or folding), thus minimizing the energy. Now, adding extra feature constraints is analogous to clamping down the rubber sheet at certain interior points. The harmonic map with added feature constraints acts like a clamped rubber sheet, rearranging around the nailed down interior points to achieve the most stable configuration. The points on the boundary of the rubber sheet M still remain on the boundary of D, though they are free to *slide* along it (Neumann boundary condition, a soft constraint) to help achieve the most stable configuration.

In our work, we compute harmonic maps between a surface undergoing non-rigid deformations (e.g. a human face) and a canonical unit disk on the plane. Harmonic maps have many merits which are valuable for tracking purposes:

- First, the harmonic map is computed through global optimization, and takes into account the overall surface topology. Thus it does not suffer from *local minima, folding, clustering*, which are common problems due to local optimization.

- Second, *the harmonic map is not sensitive to the resolution of the face surface, and to the noise on the surface.* Even if the data for the input surface is noisy, the result won't be affected significantly.

- Third, *the harmonic map doesn't require the surface to be smooth*. It can be accurately computed even when the surface includes sharp features.

- Forth, in our case, since the range of the map is a unit disk which is convex, the harmonic map exists, and is a *diffeomorphism*, namely, the map is one to one and on-to. So it can allow us to establish correspondences on 2D and recover 3D registration from the same.

- Fifth, the harmonic map is determined by the metric, not the embedding. This implies that the *harmonic map is invariant for the same face surface with different poses. Furthermore, if there is not too much stretching between two faces with different expressions, they will induce similar harmonic maps*. Because our dynamic range sequences are acquired at a high frame rate (40 Hz), we can assume that the local deformation between two adjacent frames is small.

Furthermore, harmonic maps are easy to compute and robust to numerical errors. By using a traditional finite element method [18], they are easy to implement.

3. The Non-Rigid Tracking Algorithm

In this section, we present our novel fully automatic method for high resolution, non-rigid dense 3D point tracking using harmonic maps. We first describe the global alignment of 3D scans, followed by a description of the registration algorithm based on harmonic mapping and an iterative refinement scheme using local features.

3.1. Data Preparation and Initialization

The dynamic range sequences used in this paper are collected by a phase-shifting based structured light ranging system [26]. When scanning faces, the real-time 3D shape acquisition system returns high quality dense point clouds of facial geometry with an average of 75 thousand 3D measurements per frame, at a 40Hz frame rate. The RMS (Root-Means-Squared) error of the 3D range data is about 0.05mm. Small holes around brows, eyes, nose, etc. are filled by a simple interpolation technique.

However, since the dense data samples in these 3D face scans are not registered in object space, inter-frame correspondences can not be established. Furthermore, the dense point clouds differ across the scans both in terms of the number of data samples as well as the relative positions of the samples on the surfaces. To solve these problems, a generic face model (a coarser face mesh) is fitted to the first 3D scan frame in the initialization step, by a variational Free-Form Deformation (FFD) shape registration method [32, 16]. The FFD technique is employed only for fitting of the first frame, and not for subsequent tracking.

3.2. Global Alignment and Boundary Identification

In the captured sequences, in addition to the non-rigid facial expression motion, there is also a certain amount of rigid head motion involved. To account for the latter, we align the $3D$ face scans globally. To start with, we manually mark and identify the boundary of the first frame. We can then apply the Iterative Closest Point(ICP) algorithm: for each sample on the *identified boundary of the first frame*, we find the closest sample on subsequent frames and apply a rigid body transformation that minimizes the distance between corresponding points [3]. Once we have the boundary of the initial frame and the rigid transformation, we can align the face scans globally and identify the boundaries of the subsequent frames.

3.3. Initial Coarse Registration

Once we have the global alignment, we want to capture the non-rigid deformation between two adjacent frames M_i and M_{i+1}. This inter-frame registration problem, resulting in a dense map $R : M_i \rightarrow M_{i+1}$, is solved by finding a coarse set of interior feature correspondences.

The relative ease of finding feature correspondences on 2D images as compared to 3D scans is the motivation for the next step of mapping M_i and M_{i+1}, to 2D disks D_i and D_{i+1} respectively, using the boundary constraint as described in Section 2. According to [37], the harmonic mapping is robust to boundary variation and occlusion. We define these mappings as H_i: $M_i \rightarrow D_i$ and H_{i+1}: $M_{i+1} \rightarrow D_{i+1}$. Following the disk mapping, we select a sparse set of easily detectable motion representative feature corners on the disks (for example, for facial expression tracking, we select corners of eyes, corners of lips, tip of the nose etc.) using texture and shape information. For the latter, we also adopted the idea of harmonic shape images as in [37], associating the curvature information of vertices in M_i to the corresponding ones in D_i. In practice, these

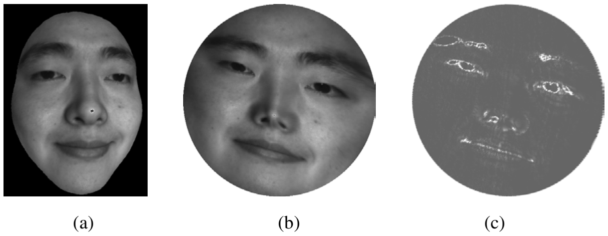


Figure 2. (a) The acquired 3D face scan data. (b) The resulting harmonic map onto a $2D$ disk with associated texture information. (c) The resulting harmonic map with associated curvature information, where brighter intensity signifies higher curvature.

feature corners usually have peak curvature value and can be easily detected by a pre-defined threshold. Fig. 2 shows an example of harmonic maps generated from one frame.

Once we have the set of correspondences on the $2D$ disks D_i and D_{i+1}, we can establish the correspondences on the $3D$ face scan M_i and the disk D_{i+1}, since the harmonic map H_i is one-to-one. Following this, as explained in Section 2, we augment the boundary constraint used to calculate H_i with these additional feature-correspondence constraints to define a new harmonic map $H'_i : M_i \rightarrow D'_i$.

As H'_i is driven by motion representative feature correspondences between the two frames, it captures the inter-frame non-rigid deformation at a coarse level. We can then overlay D'_i onto D_{i+1} to recover the inter-frame registration on 2D. Once again, we use the fact that the harmonic maps are one-to-one to calculate the dense map R required for registration of $3D$ frames. Harmonic maps, thus, help us simplify a 3D non-rigid tracking problem to a 2D image registration problem.

The algorithm is illustrated in Fig. 3 by considering the example of a synthetic surface S undergoing non-rigid deformation. S_o and S_t are the initial and final configurations respectively, and D_o and D_t are the corresponding harmonic maps with only the boundary constraint. We can notice that although D_o and D_t conform to each other around the boundary, the interior non-rigid deformation is still unaccounted for. Now, D'_o, a new harmonic map for S_o is calculated by mapping certain motion representative features on S_o to their corresponding positions on D_t, as described earlier. This is done in order to align the two maps D'_o and D_t as closely as possible, so that using the one-to-one property of harmonic maps, a dense registration between S_o and S_t can be recovered. As we can observe, D'_o and D_t are similar to each other even in the interior, thus providing accurate registration.

3.4. Iterative Refinement

The registration achieved from the previous step, although capable of capturing the coarse level facial deformation, is still insufficient to track subtle expressions. We adopt an iterative refinement scheme to improve the accuracy of the registration by progressively incorporating correspondence constraints of more local features. As part of this scheme, we keep on augmenting the set of sparse correspondences established in the previous step till it is dense enough to capture the facial deformation.

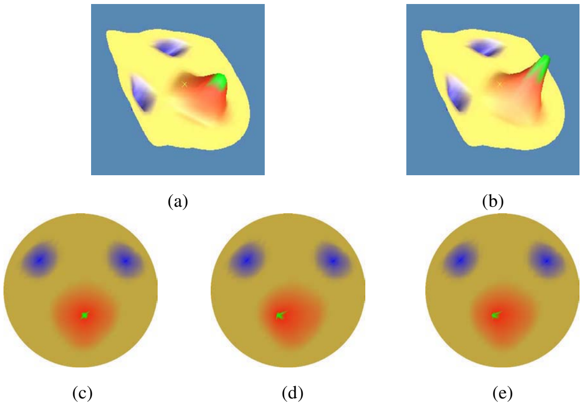


Figure 3. (a) S_o: Initial configuration of surface (b) S_t: Surface after non-rigid deformation (c, d) D_o, D_t: Harmonic Map of S_o and S_t respectively, with the hard boundary constraints only (e) D'_o: Harmonic map of S_o with the 'tip of the nose' as an additional feature-correspondence constraint. We can see that imposing correspondence constraints aligns D'_o and D_t better (as explained in Section 2), resulting in accurate registration.

In particular, we define the difference image Df_i for D'_i and D_{i+1} as $Df_i(u,v) = |D'_i(u,v) - D_{i+1}(u,v)|^2$. Using D'_i and D_{i+1} as calculated in the previous step, we find their difference image Df_i and identify the regions corresponding to significant differences. These regions indicate the areas on the face undergoing deformation, the motion of which has not been captured by the existing correspondence constraints. Because our dynamic range sequence is acquired at a high frame rate (40Hz), we can assume that the local deformation is relatively small, which allows us to apply standard 2D image registration methods within the difference regions. For high accuracy, we only consider areas with local features, which can be detected easily by applying a Laplacian filter to the image D_i and D_{i+1}.

A new D'_i is calculated by augmenting the set of correspondences with the new ones, which are kept if the new difference error between D'_i and D_{i+1} decreases, and discarded otherwise. We keep on iterating until the difference-error drops below the prescribed threshold ϵ_L. When we stop, as described in the previous subsection, we overlay D'_i on D_{i+1} to establish a dense set of correspondences, and hence recover inter-frame registration. This process is illustrated in Fig. 4.

We tackle the problem of *drifting*, a common issue in most tracking methods, in the following manner. During the initial fitting step, we identify some of the feature nodes on the mesh, like corners of the mouth etc. We then find the data points in M_i closest to these feature nodes, and constrain them to correspond to the respective features in the next data frame, i.e. M_{i+1}. Consequently, the distinct features on the face are always tracked correctly, thereby reducing the drift for other parts of the face.

Once we have the dense registration, we calculate the motion vectors for the vertices of the generic face mesh. For instance, to deform the generic face mesh from M_i to M_{i+1}, we localize each mesh vertex m_j inside a data triangle of M_i, followed by finding the corresponding data triangle of M_{i+1} and localizing m_j in M_{i+1} using bilinear interpolation. We continue this process for every frame, thereby calculating the motion vectors for the vertices of the generic face mesh across the whole sequence.

4. Experimental Results and Error Analysis

In this section we provide real data experiments and error analysis to measure the accuracy of our tracking algorithm. We performed tracking on three subjects performing various expressions for a total of ten sequences of 200-300 frames each (at 30Hz) Each frame contains approximately 80K 3D points, whereas the generic face mesh contains 8K nodes. The tracking results are available as video clips at $http : //www.cs.sunysb.edu/\tilde{}samaras/$, including opening and closing of the mouth (female subject) or strongly asymmetric smile (male subject). Our technique tracks very accurately even in the case of topology change and severe 'folding' of the data.

A first error analysis is based on the difference in the intensity values of the nodes of the generic face mesh, between the initial and the subsequent frames. Initial intensity values to the mesh nodes are assigned after the initial fitting step, and are taken as the ground truth. The intensity value of each mesh node is calculated using bilinear interpolation of the intensities of the nearest 3D data points. If tracking was perfect, then the intensities of the nodes would change only due to shadowing and shading effects, which appear due to changing geometry. For comparison purposes, we use a traditional method based on optical flow estimation and local optimization techniques (FFD [16]) to track the same sequence. We present the comparison between the two techniques in Fig. 5 by plotting the averaged difference in intensities for the mesh vertices, where the difference for each frame is calculated with respect to the first frame. To ensure fairness, we have used the same set of feature constraints for both methods. We can see that our method does considerably better than the FFD based method, which fails to track large non-rigid motion and breaks down. The error, increases significantly as the sequence progresses for FFD whereas it remains relatively stable for our method, indicating minimal tracking drift issues.

Another measure that can be used to establish the accuracy of a tracking method is the displacement error of the mesh nodes from the ground truth. As part of our second experiment to calculate the error measure in terms of absolute displacements, we chose a set D_e of points spread uniformly over the data surface as test points, such that their motions form a representative subset of the motion vectors for all the vertices, i.e. the set of all the motion vectors is sampled sufficiently. To establish the ground truth, we attach markers on the face of the subject at locations given by the set D_e. The markers are for verification purpose only and are not used for tracking. In order to be detected, the diameter of each marker is about 3mm. For error analysis, we need to compare the ground truth against our tracking results, which requires identification of the corresponding set M_e of mesh nodes on the face model M. To this end, we register the first data frame with the face model $M($ about $16K$ nodes) during the initial fitting phase.

For each frame, we can calculate the tracking error by comparing the positions of the nodes in M_e to the ground truth, i.e. the positions of points in D_e. Fig. 6 (a-f) show the snap-shots of the tracking sequence at different instances; the green dots are the markers representing points in D_e and the red dots are the corresponding nodes in M_e, i.e. the tracking results. Fig. 6 (g-h) exhibit a comparative analysis of the tracking errors for different representative points. As we can see, the tracking error for most cases is around 1.5mm, which is low, given that the resolution of the 3D range scan data is about 0.5mm. The achieved accuracy of tracking is comparable to that reported in [32, 17], using the same dense 3D data. However overall processing time including initialization and parameter selection is approximately 6 hours per sequence on 2.2GHZ, 1GB PC (approximately 1 min per frame) spent mostly on harmonic map calculation and the method can be easily parallelized on a cluster. In comparison, the methods in [32, 17] required up to 2 days per sequence with most of the time spent on tuning and parameter selection by the operator.

5. Conclusions and Future Work

In this paper, we have presented a fully automatic method for high resolution, non-rigid dense 3D point tracking using harmonic maps. A deforming generic face model is employed to track the dense 3D data sequence moving at video speeds, with the harmonic maps guiding the deformation field. The harmonic maps are constrained, and hence driven by the correspondences established between adjacent frames using an iterative scheme; the features are detected using corner detection and other standard techniques on texture and curvature images. The resulting harmonic map provides dense registration between the face model and the target frame, thereby making available the motion vectors for the vertices of the generic face model. The use of harmonic maps, in this manner, reduces the problem of establishing correspondences in 3D, to that of 2D image registration, which is more tractable. We have achieved high accuracy tracking results on facial expression sequences, without manual intervention, demonstrating the merits of our algorithm for the purpose. In future work, we will exploit the knowledge of underlying facial muscle structure to impose more constraints on the tracking process, in order to further increase accuracy. We also plan to use the proposed framework for more applications like face recognition and dynamic expression recognition for dense 3D data.

6. Acknowledgements

We are grateful to Deborah Mayo for the help with the work presented in this paper. This work was partially supported by grants: NSF ACI-0313184, NSF CMS-9900337, NSF CAREER Award CCF-0448339, NIH RR13995, and DOJ 2004-DD-BX-1224.

References

[1] Y. Akgul and C. Kambhamettu. Recovery and tracking of continuous 3d surfaces from stereo data using a deformable dual-mesh. In *ICCV99*, pages 765–772, 1999.

[2] B. Allen, B. Curless, and Z. Popović. The space of human body shapes: Reconstruction and parameterization from range scans. *ACM Trans. Graph.*, 22(3):587–594, 2003.

[3] P. Besl and N. McKay. A method for registration of 3-d shapes. *PAMI*, 14(2), 1992.

[4] M. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *ICCV95*, pages 374–381, 1995.

[5] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *PAMI*, 25(9):1063–1074, 2003.

[6] M. Brand and R. Bhotika. Flexible flow for 3d nonrigid tracking and shape recovery. In *CVPR01*, pages I:315–322.

[7] J. Chai, J. Xiao, and J. Hodgins. Vision-based control of 3d facial animation. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 193–206, 2003.

[8] J. Davis, R. Ramamoorthi, and S. Rusinkiewicz. Spacetime stereo: A unifying framework for depth from triangulation. In *CVPR'03*, pages 359–366, 2003.

[9] M. Dimitrijevic, S. Ilic, and P. Fua. Accurate face models from uncalibrated and ill-lit video sequences. In *CVPR04*, pages II: 1034–1041, 2004.

[10] M. Eck, T. DeRose, T. Duchamp, H. Hoppe, M. Lounsbery, and W. Stuetzle. Multiresolution analysis of arbitrary meshes. In *SIGGRAPH'95*, pages 173–182, 1995.

[11] J. Eells and J. H. Sampson. Harmonic mappings of riemannian manifolds. *Amer. J. Math.*, 86(109-160), 1964.

[12] S. Gokturk, J. Bouguet, and R. Grzeszczuk. A data-driven model for monocular face tracking. In *ICCV01*, 2001.

[13] S. K. Goldenstein, C. Vogler, and D. Metaxas. Statistical cue integration in dag deformable models. *PAMI*, 25(7).

[14] X. Gu and S. Yau. Surface classification using conformal structures. In *ICCV'03*, pages 701–708, 2003.

[15] B. Guenter, C. Grimm, D. Wood, H. Malvar, and F. Pighin. Making faces. In *SIGGRAPH'98*, pages 55–66, 1998.

[16] X. Huang, N. Paragios, and D. Metaxas. Establishing local correspondences towards compact representations of anatomical structures. In *MICCAI'03*, pages 926–934, 2003.

[17] X. Huang, S. Zhang, Y. Wang, D. Metaxas, and D. Samaras. A hierarchical framework for high resolution facial expression tracking. In *IEEE Workshop on Articulated and Nonrigid Motion*, 2004.

[18] T. Hughes. *The Finite Element Method*. Prentice-Hall, 1987.

[19] G. A. Kalberer and L. V. Gool. Face animation based on observed 3d speech dynamics. In *IEEE Conf. on Computer Animation*.

[20] J. Lien, T. Kanade, A. Zlochower, J. Cohn, and C. Li. Subtly different facial expression recognition and expression intensity estimation. In *CVPR'98*, pages 853–859.

[21] J.-Y. Noh and U. Neumann. Expression cloning. In *SIGGRAPH'01*, pages 277–288, 2001.

[22] B. O'Neill. *Elementary Differential Geometry*. 1997.

[23] F. Pighin, R. Szeliski, and D. Salesin. Resynthesizing facial animation through 3d model-based tracking. In *ICCV99*, pages 143–150, 1999.

[24] D. Ramanan and D. Forsyth. Finding and tracking people from the bottom up. In *CVPR03*, pages II: 467–474, 2003.

[25] S. Rusinkiewicz, O. Hall-Holt, and L. Marc. Real-time 3d model acquisition. In *SIGGRAPH'02*, pages 438–446.

[26] P. H. S. Zhang. High resolution, real time 3-d shape acquisition. In *IEEE Workshop on Real-time 3D Sensors and Their Use (joint with CVPR'04)*, 2004.

[27] R. Schoen and S. T. Yau. *Lectures on Harmonic Maps*. International Press, Harvard University, Cambridge MA, 1997.

[28] E. Sharon and D. Mumford. 2d-shape analysis using conformal mapping. In *CVPR04*, pages II: 350–357.

[29] H. Tao and T. Huang. Explanation-based facial motion tracking using a piecewise bezier volume deformation model. In *CVPR99*, pages I: 611–617, 1999.

[30] C. Tomasi, S. Petrov, and A. Sastry. 3d tracking = classification + interpolation. In *ICCV03*, pages 1441–1448, 2003.

[31] L. Torresani, D. Yang, E. Alexander, and C. Bregler. Tracking and modeling non-rigid objects with rank constraints. In *CVPR01*, pages I:493–500, 2001.

[32] 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. *Computer Graphics Forum*, 23(3):677–686.

[33] Z. Wen and T. Huang. Capturing subtle facial motions in 3d face tracking. In *ICCV03*, pages 1343–1350, 2003.

[34] A. Witkin, D. Terzopoulos, and M. Kass. Signal matching through scale space. *IJCV*, 1(2):133–144, 1987.

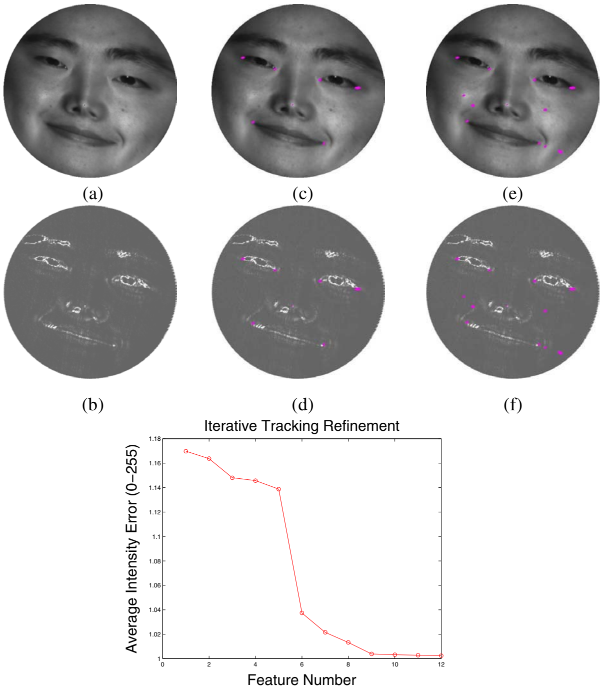
[35] J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-time combined 2d+3d active appearance models. In *CVPR04*, pages II: 535–542, 2004.

[36] A. Yezzi and S. Soatto. Deformotion: Deforming motion, shape average and the joint registration and approximation of structures in images. *IJCV*, 53(2):153–167, 2003.

[37] D. Zhang and M. Hebert. Harmonic maps and their applications in surface matching. In *CVPR99*, pages II: 524–530.

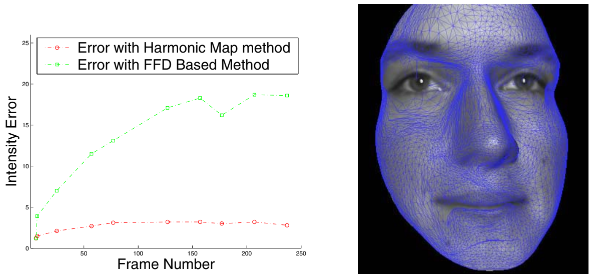
[38] L. Zhang, N. Snavely, B. Curless, and S. M. Seitz. Spacetime faces: high resolution capture for modeling and animation. *ACM Trans. Graph.*, 23(3):548–558, 2004.

[39] Z. Zhang. Iterative point matching for registration of freeform curves and surfaces. *IJCV*, 13(2):119–152, 1994.



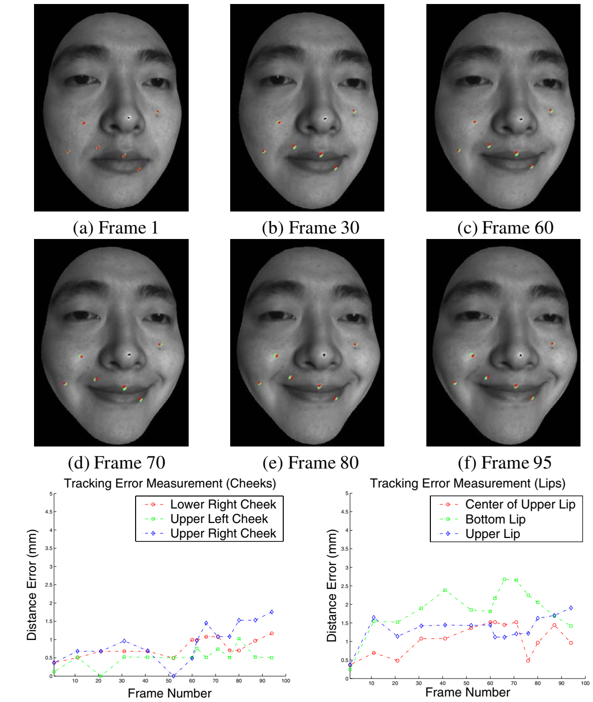
(g)

Figure 4. (a) and (b) The initial disk, D_i, with associated texture and curvature information respectively. D_i is the harmonic map of M_i (the source frame), with the boundary as the only constraint (as described in the previous sub-section). Similarly, D_{i+1} would be the harmonic map of M_{i+1}, the target frame. In order to register M_i and M_{i+1}, we iteratively augment the list of feature point constraints to obtain a progressively refined harmonic map of M_i, i.e. D'_i. We repeat the process until the difference-error between D'_i and D_{i+1} is less than ϵ_L. (c) and (d) are obtained by adding the feature corner constraints (the corners of the eyes, the tip of the nose, and the corners of the mouth.) for the calculation of the harmonic map. (e) and (f) are a further refinement, with additional local features (marked with blue), which are detected using optical flow, being added to the constraints list. In our experiments, we observe that typically $10 - 15$ feature correspondences place enough constraints on the harmonic map to reduce the error below the threshold ϵ_L. (g) plots the difference-error between D'_i and D_{i+1} against the number of feature constraints used to define the harmonic map (in addition to the boundary constraint). As is evident, the error recedes with the addition of new features, until it becomes less than the threshold ϵ_L.



(a) (b)

Figure 5. (a) The plot of error between Our method and the FFD Based method (b) FFD breaks down while tracking large deformations. We can see folds and clustering of nodes around the rim of the eyes and lips



(a) Frame 1 (b) Frame 30 (c) Frame 60

(d) Frame 70 (e) Frame 80 (f) Frame 95

(g) Error measurement around the cheeks(h) Error Measurement around the lips.

Figure 6. Error analysis on the tracking results of a smile expression sequence. An additional sequence with green markers attached to the face was acquired for error analysis; the green markers are attached for verification purposes only and are not used for tracking. (a-f) are the snap-shots of the tracking sequence at different instances, from neutral to the peak. The red dots illustrate the corresponding tracking results. (g,h) exhibit a comparative analysis of the tracking errors for different representative points, around the cheeks and the lips respectively. Since this is a smile sequence, error for points on the cheeks is expected to be relatively smaller than that for points on or near the lips, as is evident from (g) and (h)

1550-5499/05 $20.00 © 2005 **IEEE**