# High Resolution Tracking of Non-Rigid Motion of Densely Sampled 3D 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
{yangwang,mogupta,swang,gu,samaras}@cs.sunysb.edu
[2] Mechanical Engineering Department, Stony Brook University
{song.zhang,peisen.huang}@sunysb.edu

**Abstract**

We present a novel 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 allows to track non-rigid deformations accurately 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.

**Index Terms**

Vision and Graphics, Face and Gesture, Registration, Motion Analysis and Tracking.

## I. 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 image sequences, e.g. [45], [6], [19], [29], [1], [8], [42], [21], [35], [22], [44], [41], [34]. Previous methods establishing 3D inter-frame correspondences for non-rigid motion largely fall into two categories: One depends on markers attached to the object [24], [28], [2] or on feature correspondences manually selected by the users [31]; the other calculates correspondences based on the geometry using a 3D deformable/morphable model [19], [3], [33], [40], [15], [22], [47], [12], [7], [46], [16], or other 3D shape registration algorithms such as [13], [5], [50]. 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. This paper presents an algorithmic framework which makes use of the elements of conformal geometry theory for the 3D facial expression tracking problem. 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 [37], [36], and space-time stereo [49], [14]. These systems can capture dense 3D data at a high frame rate. Recently, a high-resolution 3D expression data acquisition system was developed in [37] 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 tracking algorithms have been proposed recently for 3D facial expression data [49], [43]. Tracking methods based on optical flow estimation[49], [22] can be sensitive to noise for textureless regions. A hierarchical tracking framework for high resolution 3D dynamic expression data was presented in [43], 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 Free-Form Deformation (FFD). Furthermore, this 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 data-sets required for data driven applications in facial expression analysis and synthesis[7].

In this paper, we present a novel method for high resolution, non-rigid dense 3D point tracking. This proposed method is fully automatic, except for the initial fitting step on the first frame. (Automatic initial fitting can be achieved using the automated correspondence selection technique

[31] but it is outside the scope of this paper.) High quality dense point clouds of facial geometry moving at video speeds are acquired using a phase-shifting structured light ranging technique [37]. To use such data for the 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 [38], [32], [17], [18]. Harmonic maps were used in [48] 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. Similar idea was also used in [30] where user-defined feature sets are used to constrain the surface deformation. 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 [45]). 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. It is important to point out that there are other approaches proposed to perform surface matching based on Riemannian geometry, such as generalized multidimensional scaling (GMDS) [9], [11], where an isometry-invariant embedding is used to compute an intrinsic-geometric representation of the surface. Furthermore, combined the canonical parameterization, other features, such as the texture information, can also be incorporated into the representation to improve the matching performance [10].

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.

The theory of harmonic maps is based on conformal geometry theory [23], [39]; 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. 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. Because the harmonic mapping between two surfaces is computed by solving an elliptic P.D.E., the resulting map has a higher continuity than the boundary condition [38], [20]. This implies that the harmonic maps depend on the geometry in a continuous manner, and allow certain approximation scheme to handle boundary variation and occlusion as demonstrated in [48]. Furthermore, in order to reduce the inconsistency caused by the changing boundaries during the tracking process, we use the Neumann boundary condition as the *soft boundary constraint* to give the boundary condition a relatively lower weight, and use the interior feature constraints as the *hard constraints* to minimize the overall harmonic energy.

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 [43], [26], 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 II, we give an overview of harmonic mapping. Section III 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 IV. We conclude with a discussion and future directions in Section V.

## II. HARMONIC MAPPING

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 $M$ [38], [32], [17], [18]. Suppose we want to compute a harmonic map, $H : M \rightarrow D$, where $M$ is the domain manifold and $D$ is the target manifold, $H$ can be represented as two functions $(H_1, H_2)$, $H_i : M \rightarrow R$. More specifically, in our case, $M$ is the 3D face scan, $D$ is the unit disk on the plane $R^2$, and $(H_1, H_2)$ is the parametric coordinate $(u, v)$ in the unit disk $D$. Thereby the harmonic energy can be represented as

$$E(H) = \int_M |\nabla H_1|^2 + |\nabla H_2|^2,$$

By minimizing the harmonic energy, a harmonic map can be computed using the Euler-Lagrange differential equation for the energy functional, i.e. $\Delta H = 0$, where $\Delta$ is the Laplace-Beltrami operator [38], [32], [17], [18].

Since our source manifold $M$ is in the form of a *discrete* triangular mesh, we approximate the harmonic energy as [17], [48], [23],

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

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

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

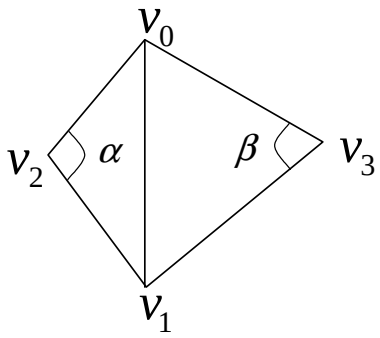



Fig. 1. Illustration of two conjuncted triangle faces: $\{v_0, v_1, v_2\}$ and $\{v_0, v_1, v_3\}$, where $\alpha$ and $\beta$ are the two angles used in Eqn. 2.

where $\{v_0, v_1, v_2\}$ and $\{v_0, v_1, v_3\}$ are two conjuncted triangular faces, and $\alpha$ and $\beta$ are the two angles against the edge $[v_0, v_1]$, as illustrated in Fig. 1.

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 $\Delta$ is the Laplace-Beltrami operator [38], [32], [17], [18]. This will lead to solving a sparse linear least-square system for the mapping $H$ of each vertex $v_i$ [17], [48], [23]. If the boundary condition is given,

$$H|_{\partial M} : \partial M \rightarrow \partial D, \tag{3}$$

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 I), 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 3. 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 [48], 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*. However, because the boundary is not reliable due to the boundary variation and occlusion, we give the boundary condition a relatively lower weight, in the absence of any strong features on the boundary, and a *soft boundary constraint* - the Neumann boundary condition - is employed to minimize the overall harmonic energy.

Intuitively, consider the manifold $M$ to be made of a sheet of rubber [17]. 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. According to Rado's Theorem [38], an arbitrary convex domain could be adopted to compute the harmonic mapping and the resulting map depends on the boundary in a continuous manner. However, this property does not hold for a concave domain in general. Therefore, in order to simplify the implementation, we use a unit disk as the target domain in our tracking method. Furthermore, based on Riemann Mapping Theorem on Conformal Geometry [38], we can compute the mapping from any simply connected surface to a disk domain. This provides the theoretical foundation for our tracking method to be applied to arbitrary simply connected surfaces and not limited to convex surfaces only. The Harmonic maps between the source surface and the target domain 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 mapping.
- 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*. Similar observations have also been used in other 3D face matching methods, such as GMDS-based methods proposed by Bronstein *et al*. [10], [11]. 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 [27], they are easy to implement.

## III. The Non-Rigid Tracking Algorithm

In this section, we present our novel method for high resolution, non-rigid dense 3D point tracking using harmonic maps. This proposed method is fully automatic, except for the initial fitting step on the first frame. (Automatic initial fitting can be achieved using the automated correspondence selection technique [31] but it is outside the scope of this paper.) 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.

The outline of the algorithm is given in Table I.

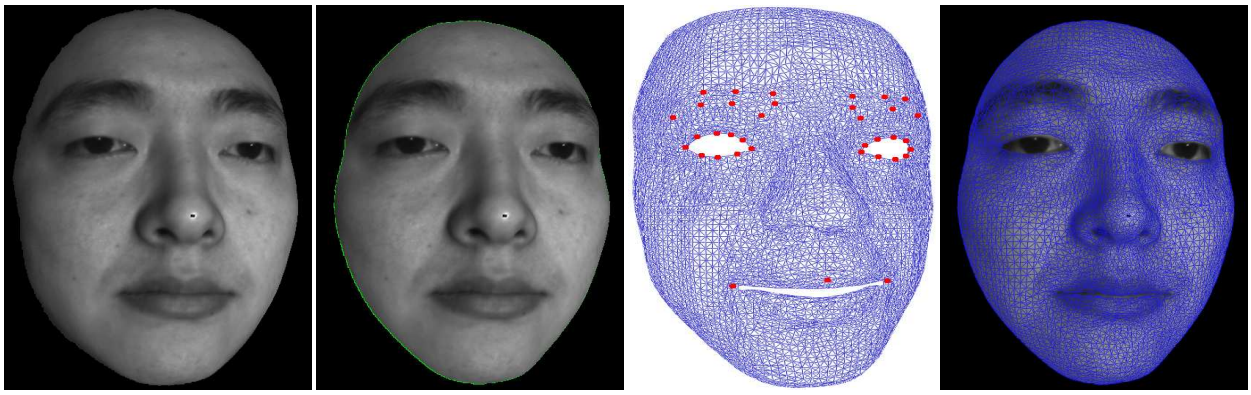
- **Data Preparation and Initialization**: Identify the boundary and fit a coarse generic face mesh model to the first frame using Free-Form Deformation (FFD).
- **Coarse Registration**:
  1) Rough alignment: Globally align the 3D face scans of the adjacent frames using the standard Iterative Closest Points (ICP) techniques.
  2) Since we have the boundary of the first frame, identify the boundary for all the subsequent frames automatically given the global alignment achieved with the previous frame. Calculate the initial harmonic maps onto 2D disks using the boundary condition.
  3) For coarse level registration between successive frames, introduce more constraints on the harmonic map using feature point correspondence constraints, where features are detected using standard methods like corner detection.
- **Iterative Refinement**:
  1) Iteratively augment the list of constraints for the harmonic map with the local feature correspondences obtained using optical flow methods.
  2) Repeat the previous step to progressively refine the harmonic map until the difference between the new source harmonic maps and the target harmonic maps recedes below a pre-defined threshold.
  3) Overlay the new source and the target harmonic map disks to establish dense registration, and hence recover the deformation parameters for the generic face mesh model between two consecutive frames.
  4) Continue this process over the whole sequence to achieve high resolution tracking.

TABLE I

THE OUTLINE OF OUR TRACKING ALGORITHM.

### *A. Data Preparation and Initialization*

The dynamic range sequences used in this paper are collected by a phase-shifting structured light ranging system [37]. When scanning faces, the real-time 3D shape acquisition system



(a) (b) (c) (d)

Fig. 2. **Initial Fitting for Tracking.** (a) The acquired 3D face scan data. (b) The 3D face data with identified boundary (marked in green). (c) The generic face model with manually selected feature points (marked as red dots). (Automatic initial fitting can be achieved using the automated correspondence selection technique [31] but it is outside the scope of this paper.) (d) The result of the initial fitting to a 3D face scan data.

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. [1] 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 [43], [25]. The FFD technique is employed only for fitting of the first frame, and not for subsequent tracking. Initial fitting is illustrated in Figure 2.

### *B. 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

[1]The RMS error is calculated using a planar board with a measurement area of $260 \times 244$mm [37].

scans globally. To start with, we manually mark and identify the boundary of the first frame. (See Figure 2) 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 [5]. 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.

### *C. 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 between $M_i$ and $M_{i+1}$. These correspondence constraints, along with the boundary condition define the map $R$ for the purpose of registration.

The relative ease of finding both texture and geometric 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 II. According to [48], 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 [48], associating the curvature information of vertices in $M_i$ to the corresponding ones in $D_i$. In practice, these feature corners usually have peak curvature value and can be easily detected by a pre-defined threshold. Figure 3 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 II, 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,

(a) (b) (c)

Fig. 3. **Harmonic Maps: Texture and Shape Images.** (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.

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 Figure 4 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.

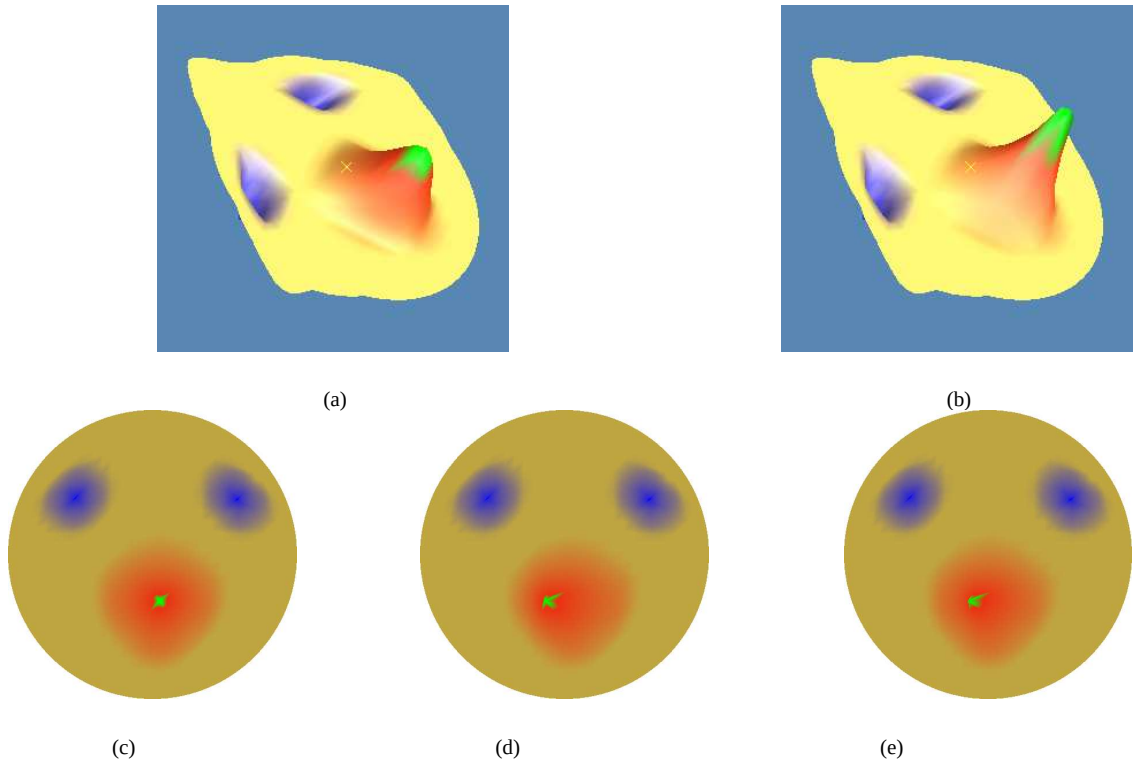



Fig. 4. **Illustration of Harmonic Map: A Synthetic Example.** (a) $S_o$: Initial configuration of surface (b) $S_t$: Surface after non-rigid deformation (c) $D_o$: Harmonic Map of $S_o$ with the hard boundary constraints only (d) $D_t$: Harmonic map of $S_t$ with the hard boundary constraints only (e) $D'_o$: 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 II), resulting in accurate registration.

### *D. 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 the new set of correspondences is dense enough to capture the facial deformation.

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 (40 Hz), 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 $\epsilon_L$. 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 Figure 5.

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.

## IV. Experimental Results and Error Analysis

In this section we provide experiments on real data and error analysis to measure the accuracy of our tracking algorithm. We performed tracking on four subjects performing various expressions for a total of twelve sequences of 250-300 frames each (at 30Hz). Each frame contains approximately 80K 3D points, whereas the generic face mesh contains 8K nodes. The accompanying video clips show tracking results on one male and one female face undergoing

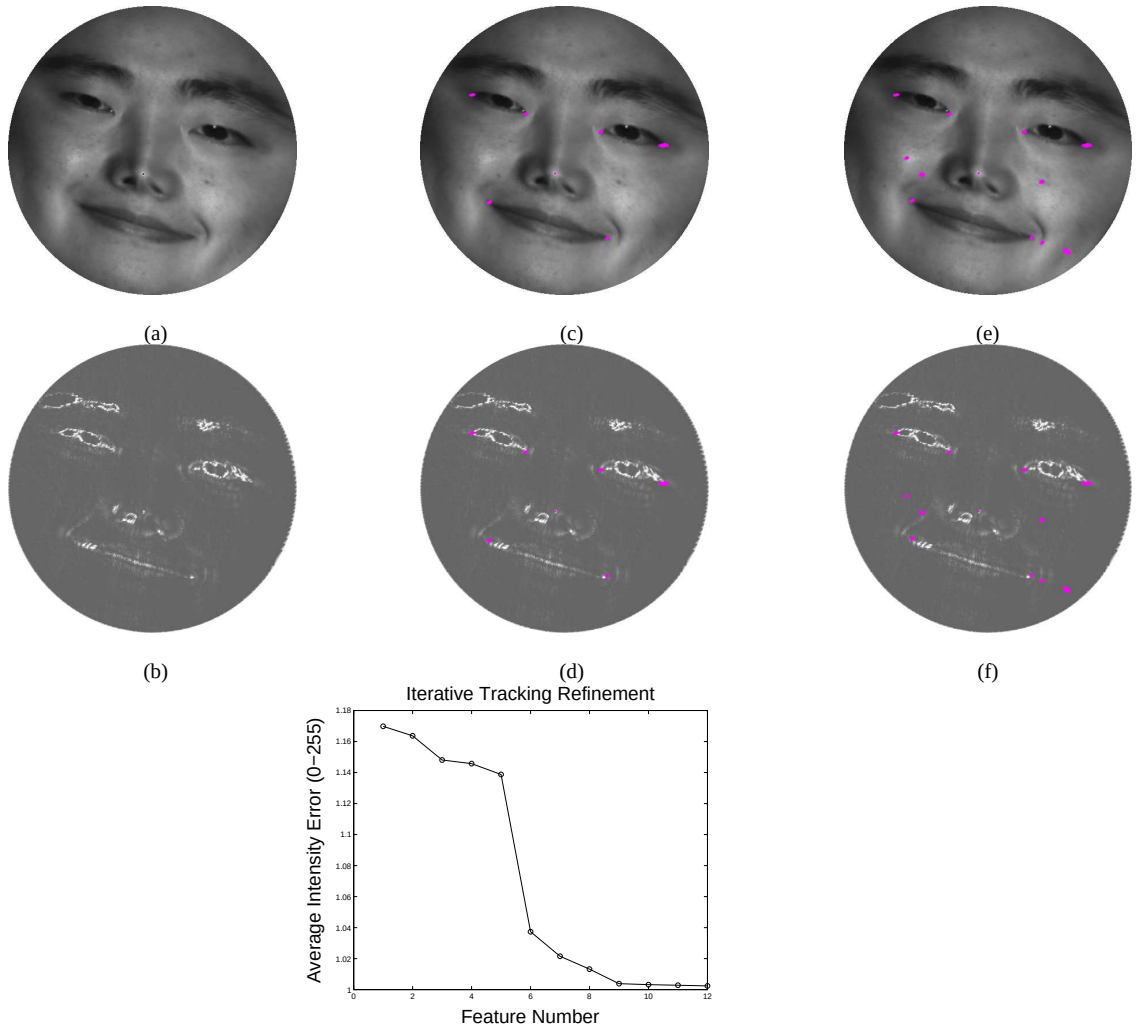
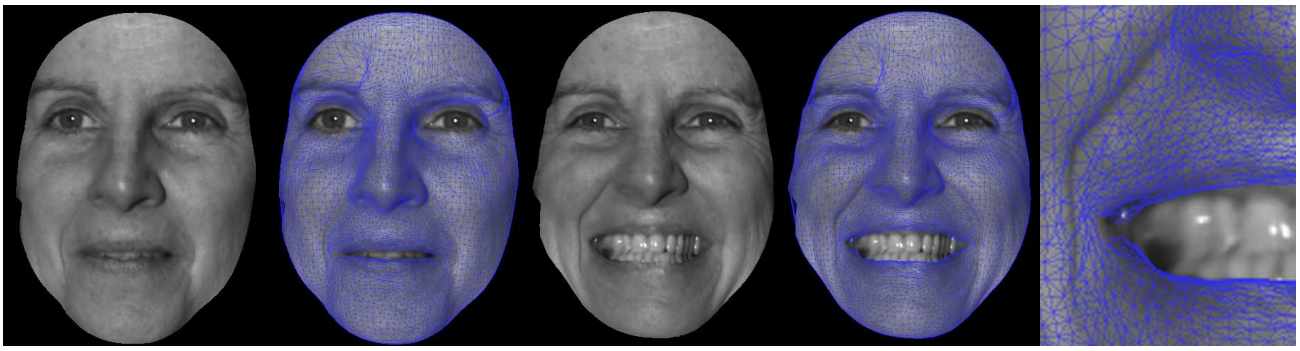



Fig. 5. **Tracking Algorithm: Iterative Refinement Step.** (a) and (b) The initial disk, $D_i$, with associated texture and curvature information respectively. $D_i$ is the harmonic map of $M_i$ (the source frame), with the boundary as the only constraint (as described in Section III). Similarly, $D_{i+1}$ 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 $\epsilon_L$. (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 magenta), 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 $\epsilon_L$. (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 $\epsilon_L$.



(a) (b) (c) (d) (e)

Fig. 6. **Snapshots from a Tracking Sequence of Subject A:** 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.

expressions of different intensity, 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. (See Figure 6)

### *A. Results*

Figures 7-10 show tracking results on two male and two female faces who were instructed to perform expressions of different intensity, which we described as: *Soft Affectionate Smile*, *Coy Flirtatious Smile*, and *Devious Smirk*. The sequences include opening and closing of the mouth (female subject), strongly asymmetric expressions and rigid head motion as well. As the results show, our method tracks very accurately even in the case of topology change and severe 'folding' of the data.

Figure 10 provides the tracking results for Subject A performing a *transition* expression, starting from a *Soft Affectionate Smile*, moving to a *Coy Flirtatious Smile*. The sequence is about $300$ frames long. The transition occurs around Frame $150$ (e). Frames $135 - 165$ (d-f) show a *blended* expression. We can observe that our method does well, even for *unusual* facial motion, arising, in this case, from a transition between two expressions.

### *B. Error Analysis*

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 at the

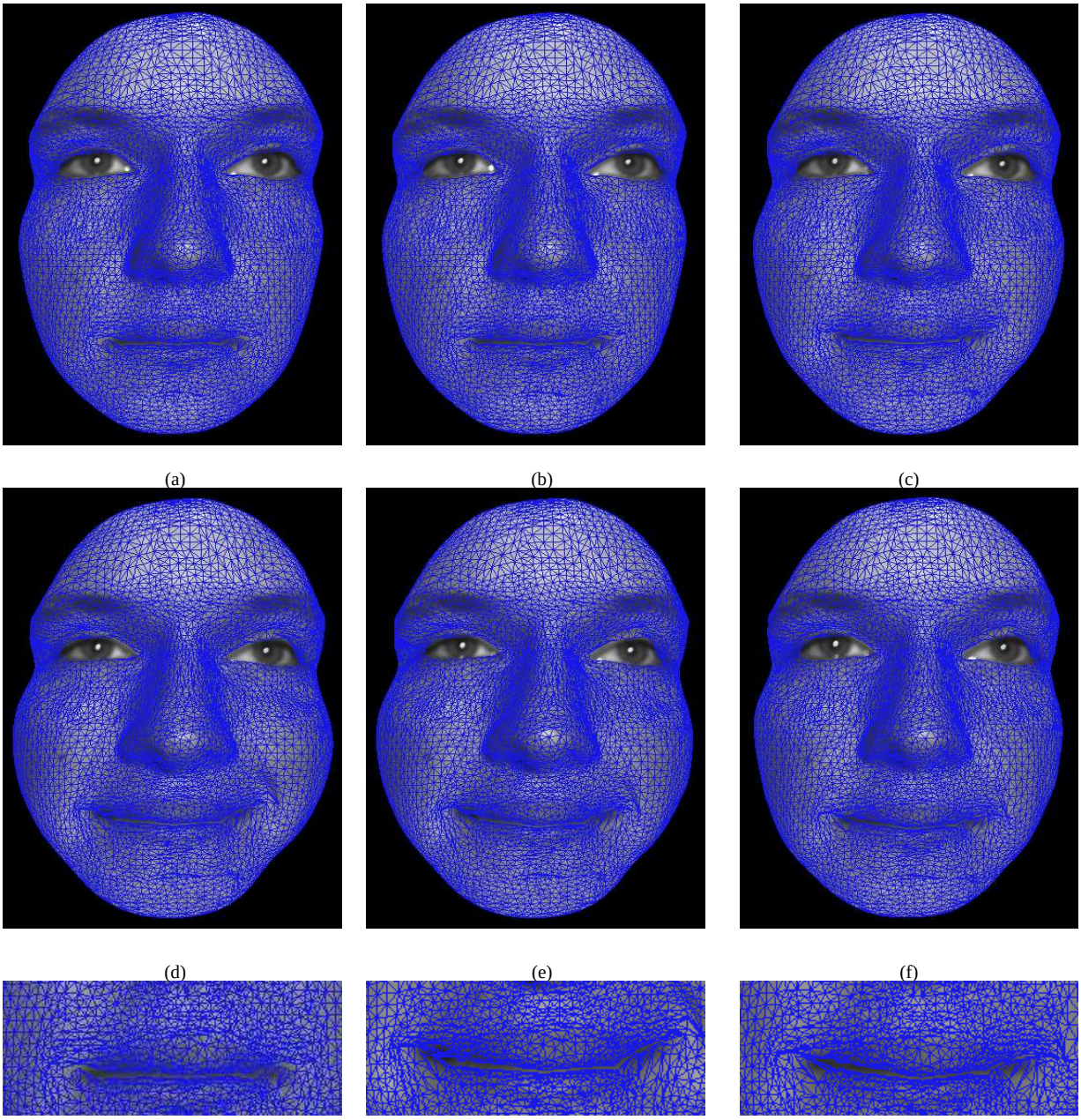



Fig. 7. (a-f) **Tracking Results for Subject C Performing a *Soft Affectionate Smile*.** (g-i) Close-Ups

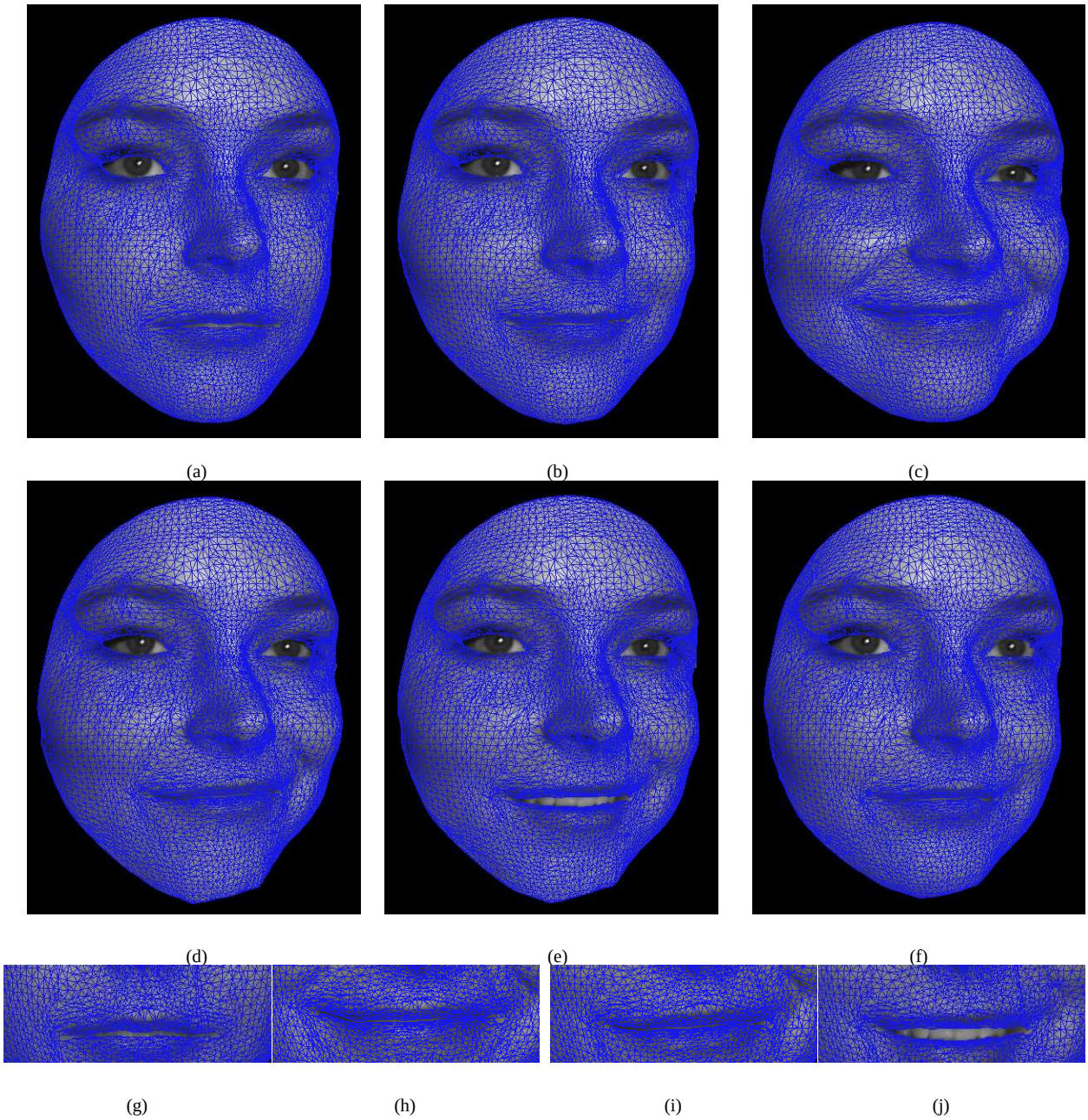



Fig. 8. (a-f) **Tracking Results for Subject B Performing a *Devious Smirk*.** (g-j) Close-Ups. **We can observe that our method does well even in the presence of asymmetry (i) and topology change (j) (opening of mouth)**

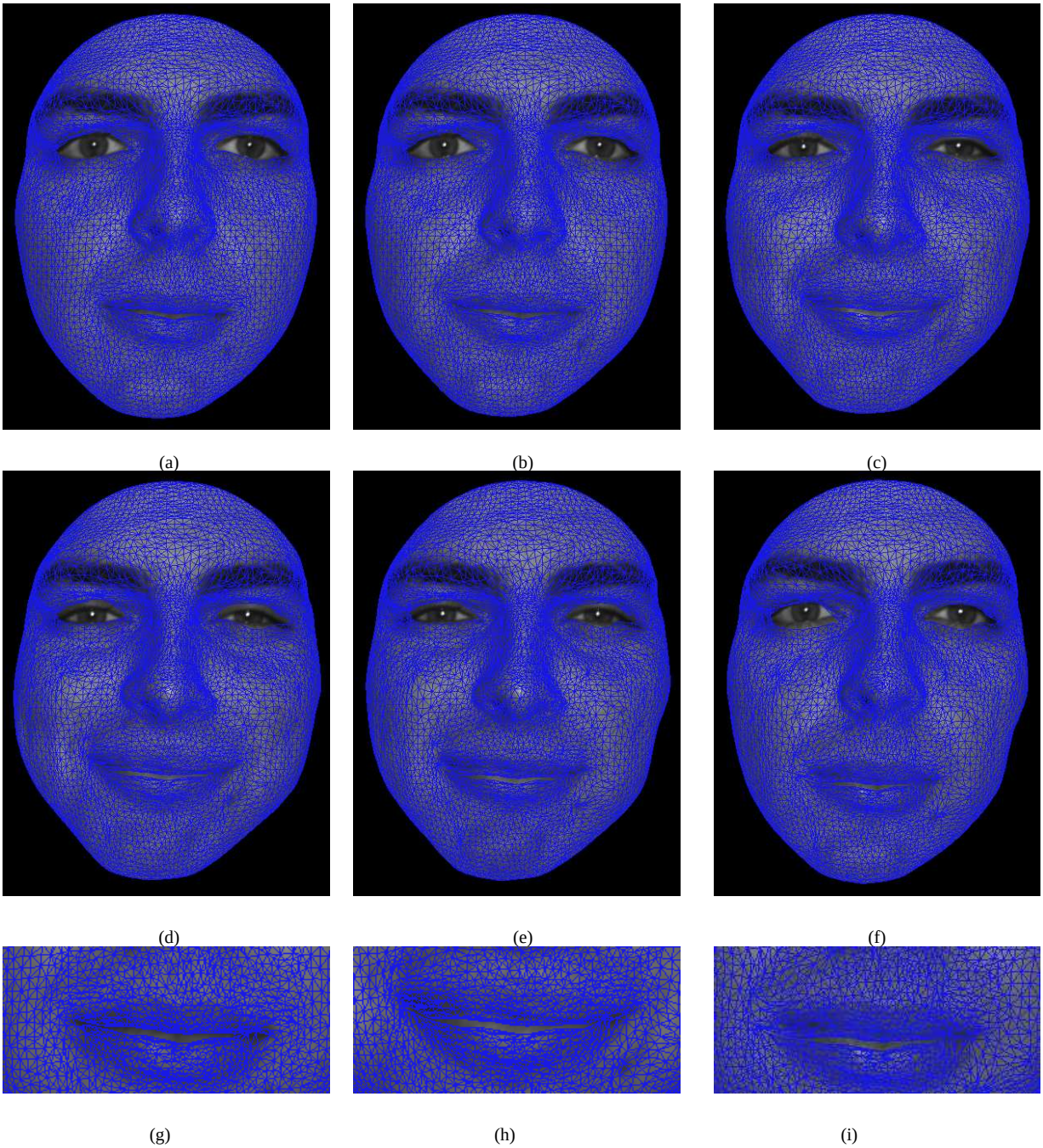



Fig. 9. (a-f) **Tracking Results for Subject D Performing a *Soft Affectionate Smile*.** (g-i) Close-Ups

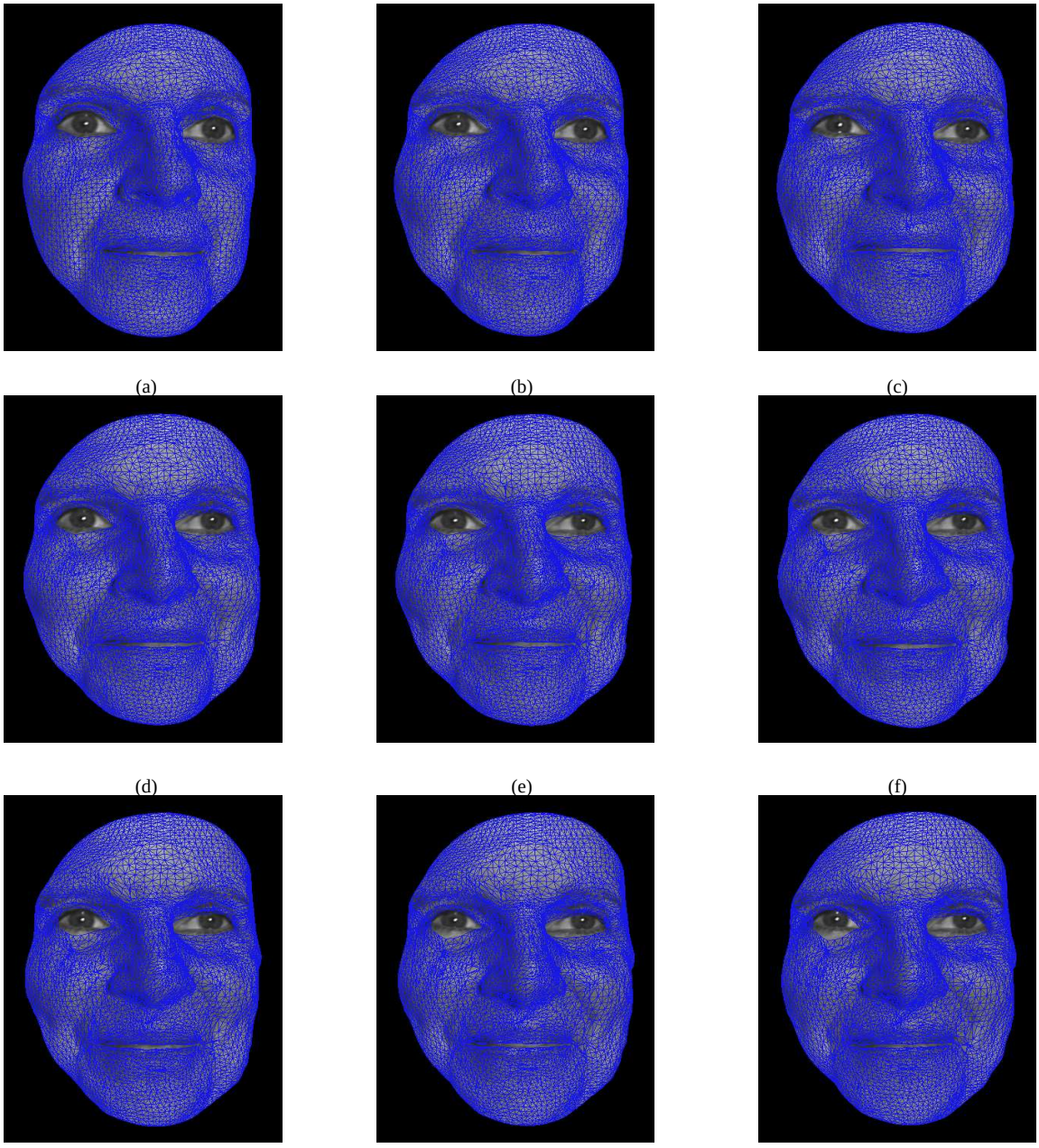
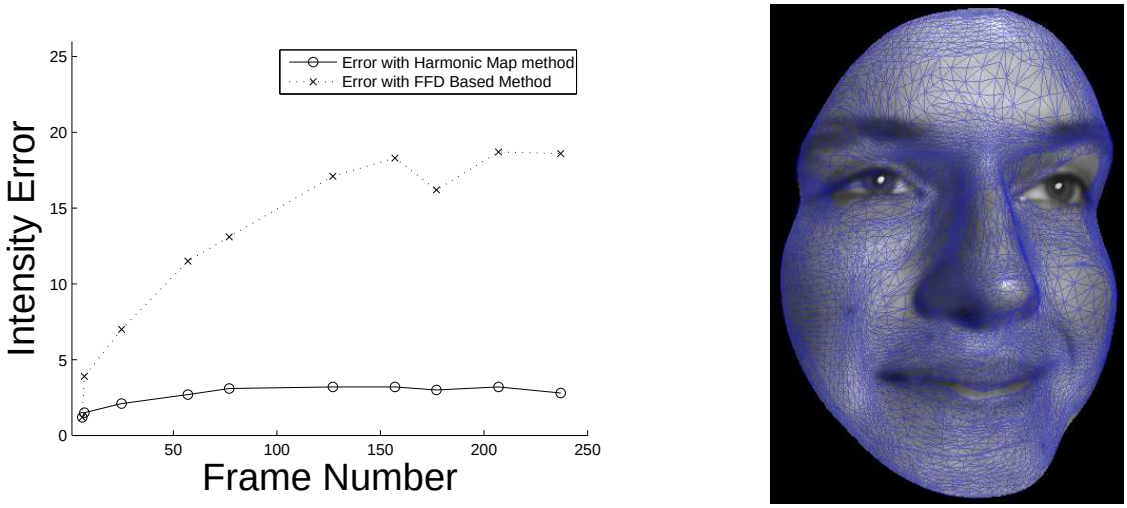



Fig. 10. **Tracking Results for Subject A Performing a *transition* expression, starting from a *Soft Affectionate Smile*, moving to a *Coy Flirtatious Smile***. The sequence is about 300 frames long. The transition occurs around Frame 150 (e). (d-f) (Frames 135, 150, 165 respectively) show a *blended* expression. We can observe that our method does well, even for *unusual* facial motion, arising, in this case, from a transition between two expressions.

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. The intensity values for the mesh nodes are calculated again for each subsequent frame of the sequence, as explained above. 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 [4] and local optimization techniques (FFD [25]) to track the same sequence. We present the comparison between the two techniques in Figure 11 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 for comparison, we have used the same set of feature constraints for our harmonic map based tracking method **as well as** for the FFD based method. 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.

Figure 12 depicts the error plots for different subjects performing various expressions, using the same error measure as in figure 11. The average intensity error is observed to be less than $0.03$ (on a scale of $0-1$), even at the peak of the expression, thus establishing the accuracy of our tracking method. **There is some *drift* however, as the intensity error doesn't come back to zero towards the end of the expression**.

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.




(a) (b)

Fig. 11. **Error Comparison between Our Method and an FFD Based Method.** (a) The plot of error between Our method and the FFD Based method (b) FFD breaks down while tracking large deformations. To ensure fairness for comparison, we have used the same set of feature constraints for our harmonic map based tracking method **as well as** for the FFD based method. **For the FFD based method, we can see clusters and folds developing near the selected features (corners of the eyes, mole on the cheek, corners of the lips etc.). Due to the local nature of the method, the rest of the points do not catch up (around lips, eyes), even though the feature correspondences match. We do not encounter such local minima problems with our harmonic map based method, since it uses global optimization**

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 $D_e$. Figure 13 (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. Figure 13 (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 [43], [26], 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 [43], [26] required up to 2 days per sequence with most of the time spent on tuning and parameter selection by the operator.

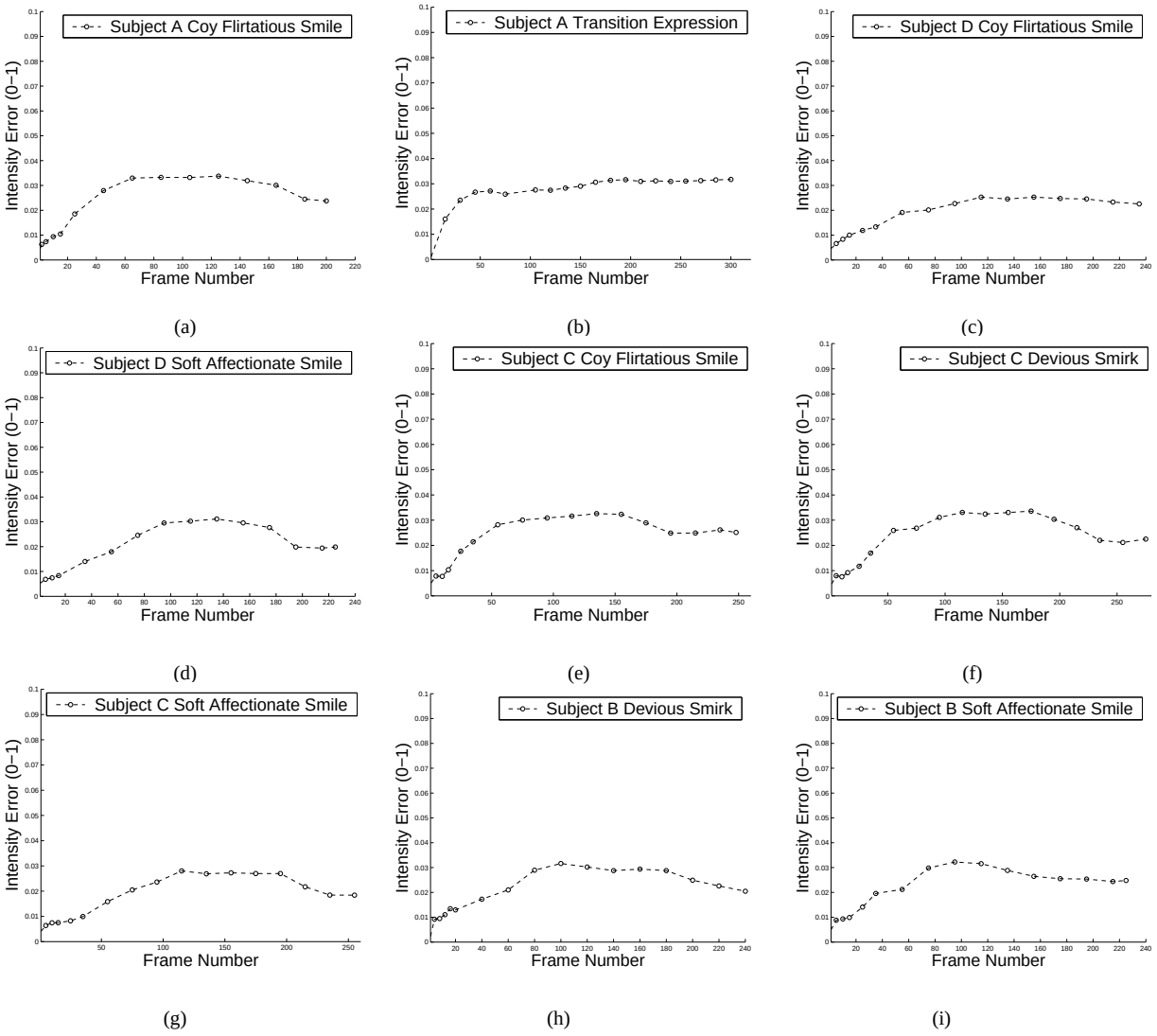
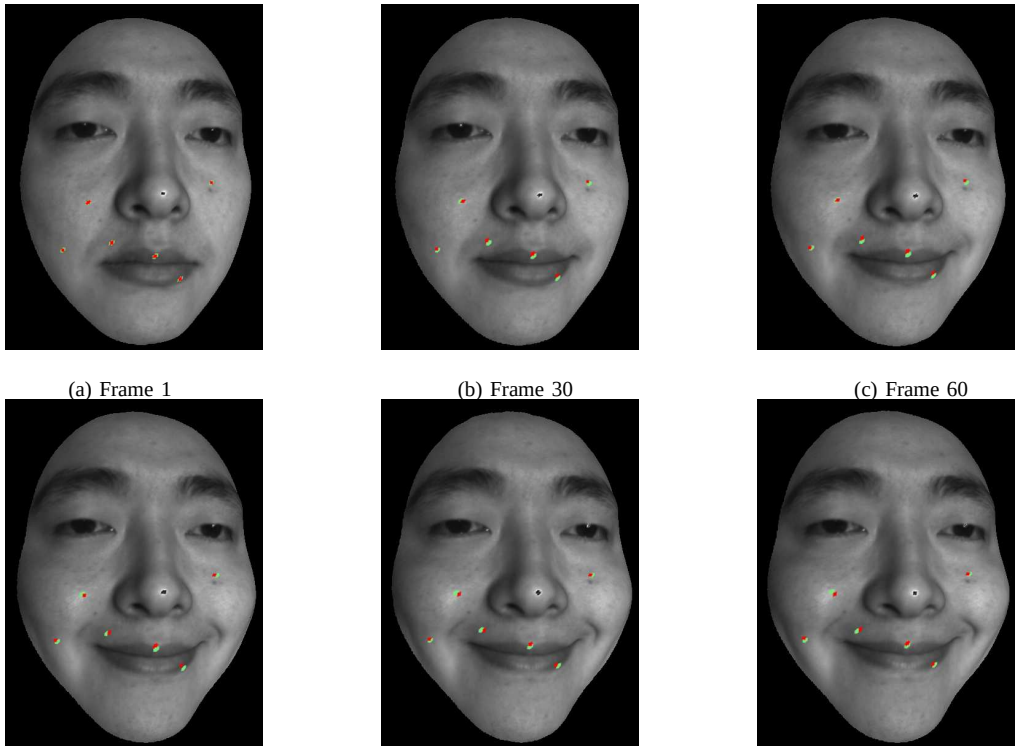
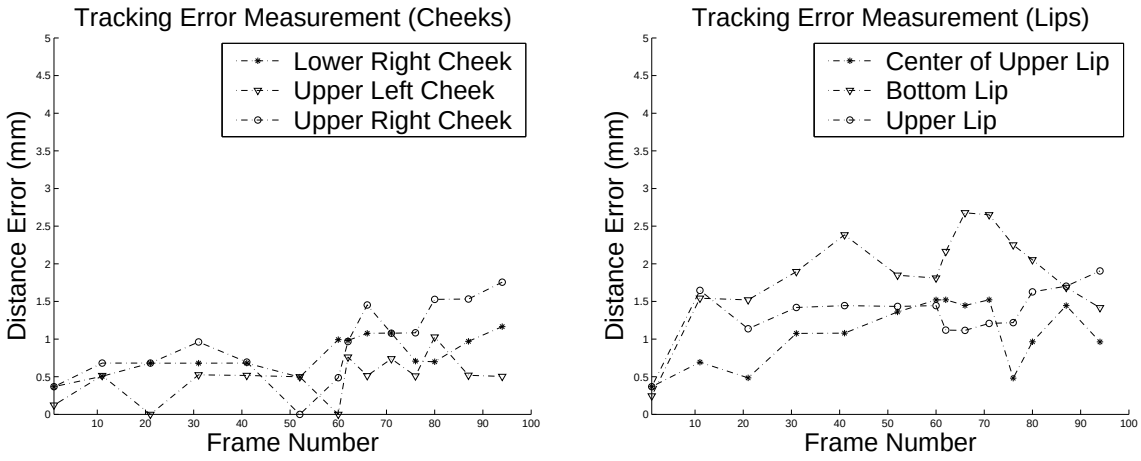



Fig. 12. **Error Plots for Various Expressions Performed by Different Subjects.** Error Plots for : (a) Subject A Performing *Coy Flirtatious Smile*, (b) Subject A Performing the transition expression (from *Soft Affectionate* to *Coy Flirtatious Smile*), (c) Subject D Performing *Coy Flirtatious Smile*, (d) Subject D Performing *Soft Affectionate Smile*, (e) Subject C Performing *Coy Flirtatious Smile*, (f) Subject C Performing *Devious Smirk* (g) Subject C Performing *Soft Affectionate Smile*, (h) Subject B Performing *Devious Smirk*, (i) Subject B Performing *Soft Affectionate Smile*. **The average intensity error is observed to be less than** $0.03$ **(on a scale of** $0-1$**), even at the peak of the expression, thus establishing the accuracy of our tracking method.**



(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.

Fig. 13. **Error Measurement Using Markers.** 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)

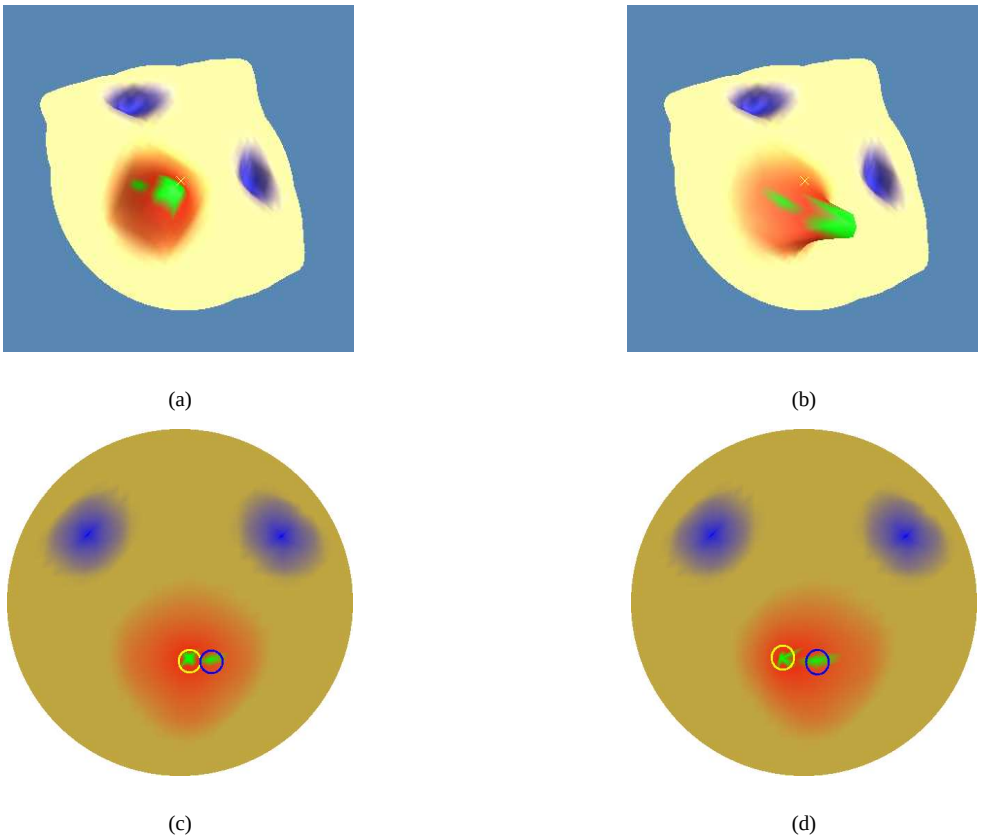



Fig. 14. **Potential Issue with the Method: A Synthetic Example.** In the presence of large deformations, ambiguities might arise while making feature correspondences, resulting in incorrect correspondence constraints. (a) Initial Surface. (b) Surface after undergoing large deformation. (c-d) Harmonic maps of initial and final surfaces respectively, with just the boundary constraint. We observe that feature 1 in (c) (circled in yellow) gets aligned with feature 2 in (d) (circled in blue), thus giving result to a correspondence mismatch. This, however, **does not pose any problems in our case** as we do not encounter such large deformations in real, high resolution (40fps) facial expression data.

**One potential issue with the method** is its inability to track large deformations, as illustrated in Figure 14, with the help of a *synthetic* example. We observe that in the presence of large deformations, ambiguities might arise while making feature correspondences, resulting in incorrect correspondence constraints. This, however, does not pose any problems in our case as we do not encounter such large deformations in real, high resolution (40fps) facial expression data. Since the motion is relatively small, correspondences can be established within a small neighborhood, thus preventing any ambiguous and hence incorrect correspondence constraints.

## V. Conclusions and Future Work

In this paper, we have presented a novel automatic method for high resolution, non-rigid dense 3D point tracking using harmonic maps. An important contribution of our tracking method is to reduce the non-rigid 3D tracking problem to a 2D image registration problem, where the feature correspondences are made on both texture and curvature images using standard techniques, such as corner detection and optical flow. 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.

## References


[1] Y. Akgul and C. Kambhamettu. Recovery and tracking of continuous 3d surfaces from stereo data using a deformable dual-mesh. In *IEEE International Conference on Computer Vision*, 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 Transactions on Graphics*, 22(3):587–594, 2003.

[3] S. Basu, N. Oliver, and A. Pentland. 3d lip shapes from video: a combined physical-statistical model. *Speech Commun.*, 26(1-2):131–148, 1998.

[4] S. S. Beauchemin and J. L. Barron. The computation of optical flow. *ACM Comput. Surv.*, 27(3):433–466, 1995.

[5] P. Besl and N. McKay. A method for registration of 3-d shapes. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 14(2), 1992.

[6] M.J. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *IEEE International Conference on Computer Vision*, pages 374–381, 1995.

[7] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, 2003.

[8] M. Brand and R. Bhotika. Flexible flow for 3d nonrigid tracking and shape recovery. In *IEEE Computer Vision and Pattern Recognition*, pages I:315–322, 2001.

[9] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. *Proc. National Academy of Sciences (PNAS)*, 103(5):1168–1172, 2006.

[10] A.M. Bronstein, M.M. Bronstein, and R. Kimmel. Three-dimensional face recognition. *International Journal of Computer Vision*, 64(1):5–30, 2005.

[11] A.M. Bronstein, M.M. Bronstein, and R. Kimmel. Efficient computation of isometry-invariant distances between surfaces. 28(5):1812–1836, 2006.

[12] 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.

[13] Y. Chen and G.G. Medioni. Object modeling by registration of multiple range images. In *IEEE Conf. on Robotics and Automation*, pages 2724–2729, 1991.

[14] J. Davis, R. Ramamoorthi, and S. Rusinkiewicz. Spacetime stereo: A unifying framework for depth from triangulation. In *IEEE Computer Vision and Pattern Recognition*, pages 359–366, 2003.

[15] D. DeCarlo and D. Metaxas. Adjusting shape parameters using model-based optical flow residuals. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(6):814–823, 2002.

[16] M. Dimitrijevic, S. Ilic, and P Fua. Accurate face models from uncalibrated and ill-lit video sequences. In *IEEE Computer Vision and Pattern Recognition*, pages II: 1034–1041, 2004.

[17] M. Eck, T. DeRose, T. Duchamp, H. Hoppe, M. Lounsbery, and W. Stuetzle. Multiresolution analysis of arbitrary meshes. In *ACM SIGGraph, Computer Graphics*, pages 173–182, 1995.

[18] J. Eells and J. H. Sampson. Harmonic mappings of riemannian manifolds. *Amer. J. Math.*, 86:109–160, 1964.

[19] I.A. Essa and A.P. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):757–763, 1997.

[20] Lawrence C. Evans. *Partial differential equations*. American Mathematical Society, 1998.

[21] S.B. Gokturk, J.Y. Bouguet, and R. Grzeszczuk. A data-driven model for monocular face tracking. In *IEEE International Conference on Computer Vision*, pages 701–708, 2001.

[22] S. K. Goldenstein, C. Vogler, and D. Metaxas. Statistical cue integration in dag deformable models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(7):801–813, 2003.

[23] X. Gu and S. Yau. Surface classification using conformal structures. In *IEEE International Conference on Computer Vision*, pages 701–708, 2003.

[24] B. Guenter, C. Grimm, D. Wood, H. Malvar, and F. Pighin. Making faces. In *ACM SIGGraph, Computer Graphics*, pages 55–66, 1998.

[25] X. Huang, N. Paragios, and D. Metaxas. Establishing local correspondences towards compact representations of anatomical structures. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 926–934, 2003.

[26] 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.

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

[28] Gregor A. Kalberer and Luc Van Gool. Face animation based on observed 3d speech dynamics. In *IEEE Conf. on Computer Animation*, 2001.

[29] J.J. Lien, T.K. Kanade, A.Z. Zlochower, J.F. Cohn, and C.C. Li. Subtly different facial expression recognition and expression intensity estimation. In *IEEE Computer Vision and Pattern Recognition*, pages 853–859, 1998.

[30] N. Litke, M. Droske, M. Rumpf, and P. Schröder. An image processing approach to surface matching. In *Eurographics Symposium on Geometry Processing*, pages 207–241, 2005.

[31] J-Y. Noh and U. Neumann. Expression cloning. In *ACM SIGGraph, Computer Graphics*, pages 277–288, 2001.

[32] B. O'Neill. *Elementary Differential Geometry*. Academic Press, 1997.

[33] F. Pighin, R. Szeliski, and D. Salesin. Resynthesizing facial animation through 3d model-based tracking. In *IEEE International Conference on Computer Vision*, pages 143–150, 1999.

[34] D. Ramanan and D.A. Forsyth. Finding and tracking people from the bottom up. In *IEEE Computer Vision and Pattern Recognition*, pages II: 467–474, 2003.

[35] J. Rittscher, A. Blake, and S.J. Roberts. Towards the automatic analysis of complex human body motions. *Image and Vision Computing*, 20(12):905–916, 2002.

[36] S. Rusinkiewicz, O. Hall-Holt, and Levoy Marc. Real-time 3d model acquisition. In *ACM SIGGraph, Computer Graphics*, volume 1281, pages 438 – 446, 2002.

[37] P.S. Huang 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.

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

[39] E. Sharon and D. Mumford. 2d-shape analysis using conformal mapping. In *IEEE Computer Vision and Pattern Recognition*, pages II: 350–357, 2004.

[40] H. Tao and T.S. Huang. Explanation-based facial motion tracking using a piecewise bezier volume deformation model. In *IEEE Computer Vision and Pattern Recognition*, pages I: 611–617, 1999.

[41] C. Tomasi, S. Petrov, and A. Sastry. 3d tracking = classification + interpolation. In *IEEE International Conference on Computer Vision*, pages 1441–1448, 2003.

[42] L. Torresani, D.B. Yang, E.J. Alexander, and C. Bregler. Tracking and modeling non-rigid objects with rank constraints. In *IEEE Computer Vision and Pattern Recognition*, pages I:493–500, 2001.

[43] 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, 2004.

[44] Z. Wen and T.S. Huang. Capturing subtle facial motions in 3d face tracking. In *IEEE International Conference on Computer Vision*, pages 1343–1350, 2003.

[45] A.P. Witkin, D. Terzopoulos, and M. Kass. Signal matching through scale space. *International Journal of Computer Vision*, 1(2):133–144, 1987.

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

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

[48] D. Zhang and M. Hebert. Harmonic maps and their applications in surface matching. In *IEEE Computer Vision and Pattern Recognition*, pages II: 524–530, 1999.

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

[50] Z. Zhang. Iterative point matching for registration of free-form curves and surfaces. *International Journal of Computer Vision*, 13(2):119–152, 1994.