High Resolution Acquisition, Learning and Transfer of Dynamic 3-D Facial Expressions

Yang Wang[1], Xiaolei Huang[2], Chan-Su Lee[2], Song Zhang[3], Zhiguo Li[2],

Dimitris Samaras[1], Dimitris Metaxas[2], Ahmed Elgammal[2], Peisen Huang[3]

[1]Computer Science Department, State University of New York at Stony Brook, NY, USA
[2] Computer Science Department, Rutgers - the State University of New Jersey, NJ, USA
[3]Mechanical Engineering Department, State University of New York at Stony Brook, NY, USA

Abstract

Synthesis and re-targeting of facial expressions is central to facial animation and often involves significant manual work in order to achieve realistic expressions, due to the difficulty of capturing high quality dynamic expression data. In this paper we address fundamental issues regarding the use of high quality dense 3-D data samples undergoing motions at video speeds, e.g. human facial expressions. In order to utilize such data for motion analysis and re-targeting, correspondences must be established between data in different frames of the same faces as well as between different faces. We present a data driven approach that consists of four parts: 1) High speed, high accuracy capture of moving faces without the use of markers, 2) Very precise tracking of facial motion using a multi-resolution deformable mesh, 3) A unified low dimensional mapping of dynamic facial motion that can separate expression style*, and 4) Synthesis of novel expressions as a combination of expression styles. The accuracy and resolution of our method allows us to capture and track subtle expression details. The low dimensional representation of motion data in a unified embedding for all the subjects in the database allows for learning the most discriminating characteristics of each individual's expressions as that person's "expression style". Thus new expressions can be synthesized, either as dynamic morphing between individuals, or as expression transfer from a source face to a target face, as demonstrated in a series of experiments.*

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Computer Graphics]: Animation; I.3.5 [Computer Graphics]: Curve, surface, solid, and object representations; I.3.3 [Computer Graphics]: Digitizing and scanning; I.2.10 [Artificial intelligence]: Motion ; I.2.10 [Artificial intelligence]: Representations, data structures, and transforms; I.2.10 [Artificial intelligence]: Shape; I.2.6 [Artificial intelligence]: Concept learning

1. Introduction

Synthesis and re-targeting of facial expressions is central to facial animation and often involves significant manual work in order to achieve realistic expressions, due to the difficulty of capturing high quality expression data. Recent progress in dynamic 3-D scanning allows very accurate acquisition of dense point clouds of facial geometry moving at video speeds. In order to utilize such data for motion analysis and re-targeting, the question of correspondence must be addressed. Correspondences must be established between data of the same face in different frames, as well as between different faces. In this paper we present a data driven approach that consists of four parts: 1) High speed, high accuracy capture of moving faces, 2) Very precise tracking of facial motion by using a multi-resolution deformable mesh, 3) A unified low dimensional mapping of dynamic facial motion that can separate *expression style* and 4) Synthesis of novel expressions as a combination of expression styles.

Facial animation is an active area of research in computer graphics (see [PW96] for an overview of older work). In 2D facial animation, many advanced examples of talking faces have been produced with image-based methods [Bra, EGP], which are mainly focused on the mouth region. Small rotations are assumed in 2D methods and imaging conditions can only be those of the original video.

To allow 3-D animations, several techniques have been

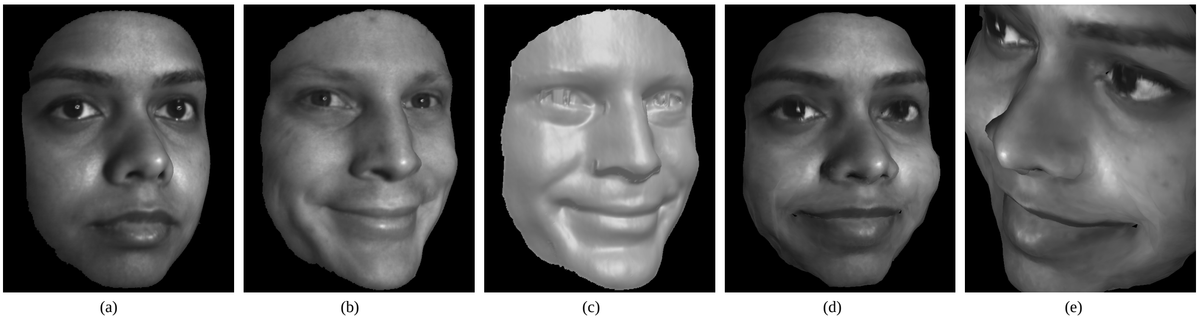

Figure 1: *(a) 3-D scan of Subject 1. (b) 3-D scan of Subject 2. (c) Untextured 3-D scan of Subject 2. (d) Subject 1 with synthetic smile transferred from Subject 2. (e) Detail of synthesized smile.*

developed to create photo-realistic face models from 2D images [PHL∗, BV]. Physics-based models are used to simulate the surface deformations caused by muscle forces [LTW, KHS03]. Mathematical approximation models include free form deformations [CHP, KMMTT], B-Spline surfaces [MNS88] and variational approaches [DMS]. Recently, both static 3-D scans of expressions [BV, BBPV] and time-sequences of 3-D motion [GGW∗, KG02] have been used to collect 3-D facial expressions. Expression cloning [NN] can produce facial animations by reusing existing motion data. Morph-based approaches [BN, PHL∗], geometry-based approaches [ZLGS03, JTDP03] and high level control mechanisms [BB02] generate photo-realistic facial expressions. Most current methods for capturing 3-D motion data either require the use of 100-200 markers (e.g. [GGW∗]) which then need to be manually removed from the images, or model fitting to multiple photographs. Using such methods the recovered geometry of the expressions is rather crude.

Recent technological advances in digital projection display, digital imaging, and personal computers, are making 3-D shape acquisition in real time increasingly available. Such ranging techniques include spacetime stereo [ZCS, DRR], and structured light [HHJC, RHHL]. In this paper we propose the use of high resolution dynamic 3-D shape data that capture very accurate geometry at speeds that exceed video frame rate. When scanning faces, our system returns an average of 75 thousand 3-D measurements per frame, at 40Hz frame rate, with an RMS of 0.05mm. Such quality of data allows for the capture of subtle expressions as well as the temporal study of facial expressions. A major contribution of this paper is the development of ways to parameterize such a high amount of data in order to make it easy to use while preserving the accuracy and visual quality that such data guarantees.

The samples returned by our system are not registered in object space and hence there is no guarantee of intra-frame correspondences, which would make tracking of facial features problematic. For this reason, we use a multi-resolution deformable face model. At the coarse level, we use a mesh with 1K nodes that is suitable for facial animation. The coarse mesh was first developed for robust face tracking in low quality 2-D images [GVM03] and we extend it to 3-D data. This method is fast, and the deformation parameters for each facial motion are few and intuitive. However it cannot capture accurately the large number of local deformations and expression details in our data, so we use it for a coarse-level initial tracking.

The highly local deformations and details in expressions are captured in a second level fitting process. For each frame of the range scan, the resulting mesh from the coarse-level tracking is used to initialize a subdivided refined mesh with 8K nodes. This finer mesh is registered to the frame based on the 3-D extension of a variational algorithm for non-rigid shape registration [HPM03]. This algorithm integrates an implicit shape representation [OS88] and the cubic B-spline based Free Form Deformations (FFD) model [SP, RSH∗99], and generates a registration/deformation field that is smooth, continuous and gives dense one-to-one correspondences.

Compared to other face model fitting techniques, such as the network of Radial Basis Functions (RBF) [NN] or mesh movement by blending nearest moving dots [GGW∗], our hierarchical tracking and fitting scheme reflects a more accurate model of facial motion. It can not only track global facial motion that is caused by muscle action (coarse level), but fit to subtler expression details that are generated by highly local skin deformations (fine-level). Past efforts to simulate facial muscle actions [KMMTT] did not always produce convincing facial expressions, due to the difficulty in simulating muscles. Instead, we solve the inverse problem by tracking real facial expressions using our hierarchical system, and replaying, synthesizing and re-targeting afterwards.

The availability of high quality dynamic expression data opens a number of research directions to the modeling of faces. Here we propose a new approach to the problem of facial expression transfer, i.e. the synthesis of novel facial expressions on new models based on the analysis of facial expressions captured from different subjects. Previously, researchers have used linear models (PCA [BV]) and variations such as bilinear models [TF00] and multilinear tensor models for facial expression analysis and synthesis [EGP, CDB02]. However, a major limitation of such models is that the dynamic facial expressions' visual manifolds are non-linear. Our approach is based on the use of a nonlinear dimensionality reduction framework [RS00, TdSL00] that allows us to find an improved representation of facial expressions and their related generative model, i.e. the mapping from a low dimensional manifold to the 3-D facial motion. This new approach allows us to synthesize new generative models that integrate the facial expression characteristics or *expression style* of different individuals. We can therefore capture the nonlinear aspects of an individual's facial expression and map them on a different individual. Another advantage of our approach is that it takes into account motions all over the face and not just around the mouth or eyes, thus obviating the need for explicit modeling of coarticulation effects and results in much more natural looking motions.

The first step of our algorithm finds a nonlinear manifold to represent the motion of an individual subject's facial motion from the tracked 3-D nodal motions. In order to be able to map the estimated motion manifolds from different individuals to a particular individual, the second step of our algorithm computes a warping transformation that places all the manifolds close in space. We term this new collection of individual manifolds the *unified* manifold. In this unified manifold we learn, in the third step of the algorithm, the mapping from each individual's manifold to the individual's 3-D motion. By analyzing the mapping functions based on the use of generalized radial basis functions we are able to determine the expression characteristics of each individual's facial motion. Finally, based on the learned mapping from the unified manifold to an individual's 3-D motion we can map the expression characteristics from one individual to another or from any combination of individuals to a single different individual. We demonstrate two types of synthesis results: morphing of geometry and expression between dynamic expression sequences of different individuals and expression transfer from a source face to a target face, without further changes in facial geometry.

In Section 2 of this paper we present our 3-D shape acquisition system, in Section 3 we present our high accuracy facial tracking method and in Section 4 we discuss our learning method for the separation of expression style from individual subjects. Expression synthesis results are presented in Section 5 and future work in Section 6.

2. Dynamic 3-D shape acquisition

The real-time 3-D shape acquisition system used in this research is a higher acquisition rate version of the system originally developed by Huang *et al*[HZC03]. It uses a single-chip DLP projector and a three-step sinusoidal phase-shifting algorithm [Mal92] to realize real-time 3-D shape acquisition. Figure 3 shows the schematic diagram of the developed system. A color fringe pattern, whose red, green, and blue components are the three phase-shifted fringe patterns, is generated by a PC. When this color fringe pattern is sent to a single-chip DLP video projector (Kodak DP900), the three color channels, or the three phase-shifted fringe patterns, are projected sequentially and repeatedly at a frequency of 80 Hz. Since gray-scale fringe patterns are more desirable, the color filters on the color wheel of the projector are removed. For image capture, two CCD cameras, positionally aligned with a beam splitter, are used, one color and one black-and-white (B/W). The color camera (Uniq Vision UC-930), with its exposure time set to one projection cycle (12.5ms), is used to capture a color 2D image for texture mapping (averaging the three sinusoidal phase-shifted fringe patterns with 120° phase shift cancels the fringes and produces a flat image of the object). The B/W camera (Dalsa CA-D6-0512W), which is a high-speed digital camera with a maximum frame rate of 262 fps, is synchronized with the projector to capture the three phase-shifted fringe images for 3-D shape reconstruction. Due to the limited frame rate of the camera, we are only able to capture the three phase-shifted fringe patterns in two projection cycles (25ms), thus achieving a frame rate of 40 Hz for 3-D shape acquisition. However, since the relationship between any two neighboring patterns is the same (with a phase shift of 120°), any newly captured fringe pattern can be combined with its two preceding patterns for 3-D shape reconstruction, thus achieving a real frame rate of up to 120 Hz for the current system setup. If a higher speed camera is used, this speed can be doubled to 240 Hz. On the other hand, if color texture mapping is required, the speed is lowered to 26 Hz due to the limited frame rate (maximum 30 fps) of the color camera used in the current system. The RMS of uncertainty of depth is 0.05 mm with a measurement area of 260 × 244 mm. Figure 2 shows the developed hardware system. This real-time 3-D shape acquisition system is described in detail in [ZH04].

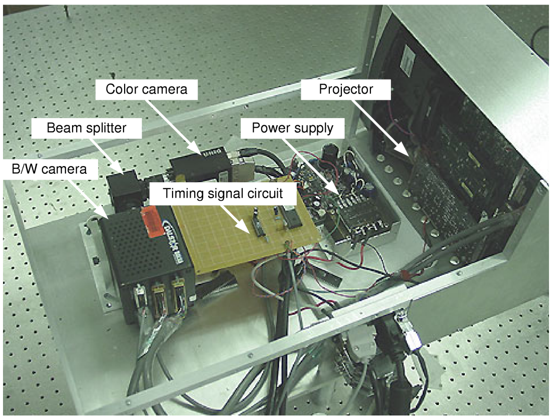

Figure 2: *Photograph of our real-time 3-D shape acquisition system (Box size: 24"×14"×10").*

3. Tracking facial expressions: capturing details and establishing correspondences

In order to utilize the acquired 3-D motion data, correspondences need to be established between data in different frames of the same face as well as between faces of different people. We adopt an approach in which we register the face scans of different actors before performing an expression with a generic face model. Then a new hierarchical framework is used to keep tracking the intra-frame deforma-

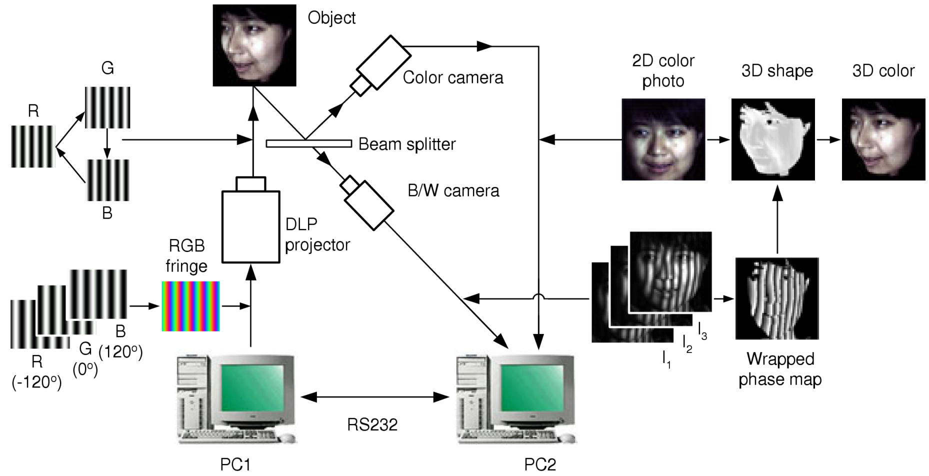

Figure 3: *Schematic diagram of our real-time 3-D shape acquisition system.*

tions of the face model points during an expression, providing a tight coupling between global and local deformations.

The generic face model has two resolutions: a coarse level with 1K nodes and a subdivided fine level with 8K nodes. We use the 8K node mesh for the initial fitting between the face model and an actor's face scan before an expression. Figure 4 demonstrates this initial fitting process. First, the face model and the 3-D scan data are roughly aligned by hand (Figure 4(b)). Then the 3-D extension of a variational non-rigid shape registration algorithm [HPM03] is used to register the face model with this range scan, achieving a complete surface match (Figure 4(c)). The algorithm is based on the integration of an implicit shape representation and cubic B-spline based Free Form Deformations (FFD). It represents shapes (e.g., the face model and the range scan) in an implicit form by embedding them in the space of distance functions of a Euclidean metric. A cubic B-spline based Free Form Deformation (FFD) model [SP, RSH*99] is then used to minimize a sum-of-squared-differences criterion between the embedding functions of a source (e.g., the model) and a target (e.g., the range scan) surface, and recover the FFD parameters that would map the source to the target. In this paper, in order to constrain the initial dense correspondences established by the registration algorithm, we define a small set of feature points on the face model (typically around 30, as in Figure 4(a)), then manually select their correspondences on the range data. These feature correspondences are incorporated as hard constraints during the optimization process of the registration algorithm (see [HZW*04] for details), establishing very good initial correspondences.

After the initial fitting, a hierarchical scheme is adopted to track the intra-frame deformations in an expression. At the coarse level, we use the 1K node face model and extend the deformable tracking system in [GVM03] to track 3-D dynamic range scans. The system divides the face model into several deformable regions whose shape and motion are controlled by a few parameters. Typically, for a smiling expression the face model is divided into 10 small regions with a total of 17 parameters. Because of the small parameter set, (which has the extra advantage of being intuitive to animators), this coarse-level tracking is very fast; however it can not capture highly local deformations and fine details in the expression. In order to estimate expression details, for each frame of the dynamic range data, we use the coarse-level tracking result to initialize the subdivided 8K node mesh at the higher level. Then this 8K node refined mesh is registered to the frame using the same variational non-rigid shape registration algorithm used for initial fitting. This hierarchical tracking/fitting protocol provides a tight coupling between global and local deformations, and results in efficient and very detailed fitting to the 3-D face scan data (see Figure 5). Based on our extensive experiments, intra-frame correspondences established by our system, especially between facial features, are highly accurate (Figure 5(c-d)), as expected due to a number of reasons. First, the dense correspondences in the tracking initialization process have high accuracy since we used manually selected facial feature correspondences as hard constraints. Second, due to the high acquisition speed, the intra-frame deformations in our range data are small, facilitating accurate and effective tracking. Third, both coarse-level tracking and fine-level fitting algorithms are sensitive to surface geometry. Facial features, such as the tip of the nose, corners of the eyes and mouth, have very distinctive geometry, and hence are tracked robustly. Last, but not least, our fine-level fitting/registration algorithm imposes very strong smoothness constraints. Using the implicit shape representation, the algorithm aligns the original surfaces as well as their clones, positioned coherently in the volumetric embedding space. The Free Form Deformations (FFD) model also enforces both implicit and explicit smoothness constraints. As a result, the established correspondences are one-to-one, coherent and globally consistent. More details on the mathematical formulation and experimental validation of our hierarchical tracking system can be found in [HZW*04].

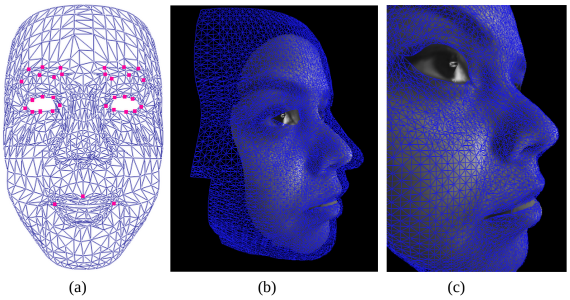

Figure 4: *(a) The generic face model with manually selected feature points. (b) The face model and the face scan data are roughly aligned. (c) The result of the initial fitting to a 3-D face scan data.*

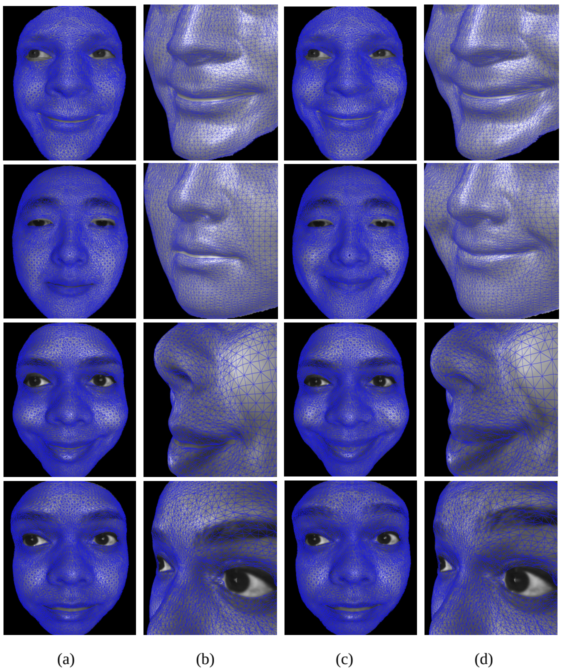

Figure 5: *Top Row: Snapshots of the* smile *expression of subject 1. Second Row: The* smile *expression of subject 2. Third Row: The* smile *expression of subject 3. Bottom Row: The* Raising eyebrow *expression of subject 3. Col.(a): Front view of frame 1. Col.(c): Front view of frame 2. First and second row of Col.(b,d) are rendered without texture - showing shape details; Third and fourth row of Col.(b,d) are rendered with texture - demonstrating accuracy of correspondences.*

4. Decomposable generative model for facial expressions

The question that we address is how to decompose three conceptually orthogonal factors: face geometry, facial expression content (i.e., the type of the expression e.g., smile), and expression style. For example, given several sequences of facial expressions, with different people performing the same expression, how to decompose the intrinsic face configuration through the expression (content) from the personalized style of the person performing the expression (style) and how to be able to cast such expression in a given style to a different face geometry. As a learning problem, we aim to learn a decomposable generative model that explicitly decomposes the following two factors given a facial expression:

- Content (face configuration): The intrinsic facial configuration through the motion as a function of time that is invariant to the person, i.e., the content characterizes the motion of the expression.
- Style (people) : Time-invariant person parameters that characterize the person performance of the expression.

If we consider a human facial expression as points in a high dimensional face configuration space, then, given the physical body constraints and the temporal constraints imposed by the expression being performed, it is expected that these points will lie in a low dimensional manifold. We can think of each expression performed by a certain person as a trajectory in the face configuration space, i.e., these points naturally lie on a one dimensional manifold characterizing the expression motion. Such a manifold might twist and self-intersect in such a high dimensional configuration space [BO, Bra]. The shape of such manifold for a certain expression (e.g., smile) is expected to be different from one person to another as different people's motion styles will follow different twists on such manifolds. This means that the manifold of the expression encodes both the content (e.g., smile) and also the personalized style.

Suppose that we can learn a unified, style-independent, embedded representation of the expression manifold in a low dimensional Euclidean embedding space, then we can learn a set of style-dependent mapping functions from the embedded representation to the face configurations space where each of these mapping functions represents a certain personalized style. If we can do this decomposition, then moving along the style-independent embedded manifold while choosing a certain style-dependent mapping function will generate an expression trajectory in the face configuration space. Of course, each person will have his own style-dependent mapping function. Therefore, we need to parameterize such a mapping function in order to decompose certain parameters that encapsulate the style. This way, we can have an embedded representation of the expression manifold and one mapping function that maps from the embedded representation to the face configuration given a parameter that describes the style. For example for a smile expression, moving along the manifold will generate a generic smile and changing the style parameter will stylize this smile.

Our approach is based on embedding the facial expression manifolds nonlinearly into a low dimensional space. Given such embedding, different manifolds corresponding to different people are normalized to achieve a unified embedding of the expression manifold. Given a unified embedding of the expression manifold, nonlinear mappings are learned from such embedding to the original space. Since the embedded manifolds are normalized, all the variations due to personalized style are expected to be represented in the nonlinear mapping space. Therefore, decomposing the nonlinear mapping coefficients would facilitate separation of

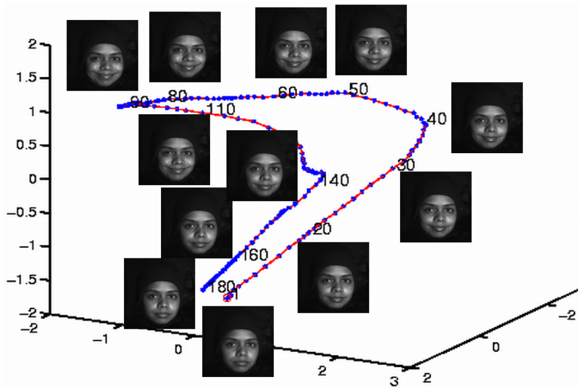

Figure 6: *Low dimensional representation of smile motion: An embedding of smile motion by LLE shows that smile motion can be well embedded in an one dimensional manifold located in 3-D Euclidean space. Manifold points for similar facial motions are located at nearby points in the manifold.*

the style parameters. A general framework for separation of style and content on nonlinear manifolds was introduced in detail in [EL]. We briefly describe the approach as adapted to facial expression analysis, in the next subsections.

4.1. Unified embedding of expression manifolds

Since expression manifolds are nonlinear, i.e., distances in the input space are nonlinearly related to distances along the manifold, PCA [BV], bilinear [TF00] and multilinear [VT] models will not be able to discover the underlying manifold and decompose orthogonal factors. Simply put, linear models will not be able to interpolate intermediate facial geometry and/or intermediate styles.

We adapt the locally linear embedding (LLE) framework [RS00] to achieve a low dimensional manifold embedding for individual facial expressions that provides a good representation of facial motion. LLE finds the best embedding manifold by nonlinear dimensionality reduction given the assumption that each data point and its neighbors lie on a locally linear patch of the manifold, (details in [RS00]). Figure 6 shows the embedding of a smile motion to a 3-D Euclidean space. To optimally choose the neighborhood size we use an error criterion based on the reconstruction achieved by the fitted generative model in Equation 2.

A unified manifold embedding is achieved by warping sample manifold points in the embedding space. Unified embedding allows us to represent all the facial motion in one manifold. For each manifold, points are approximated by fitting a spline (with normalized parameters in the 0 to 1 range, since we assume similar starting and ending facial conditions across subjects). Correspondences are established by re-sampling the normalized spline at equal intervals. Given multiple manifolds a mean manifold is learned by warping each manifold using non-rigid transformation using an approach similar to [CR].

4.2. Learning a decomposable generative model

We aim to learn a generative model in the form

$$y_t^s = \gamma(x_t^c; a, b^s) \qquad (1)$$

where the observed face motion, y_t^s, at time t of style s is an instance driven from a generative model where the function $\gamma(\cdot)$ is a mapping function that maps intrinsic face configuration embedded coordinate x_t^c (content) at time t into the observation space given mapping parameters a and a style parameter b^s which is time invariant.

Given the unified embedding achieved in 4.1 we learn a set of style-dependent nonlinear mapping functions from the embedding space into the input space, i.e., functions $\gamma_s(x_t^c) : R^e \rightarrow R^d$ that map from embedding space with dimensionality e into the input space (observation) with dimensionality d for style class k. Since we consider nonlinear manifolds and the embedding is nonlinear, the use of nonlinear mapping is necessary. In particular we consider nonlinear mapping functions of the form

$$\gamma_s(x_t) = B^s \cdot \psi(x_t^c) \qquad (2)$$

where B^s is a $d \times N$ linear mapping and $\psi(\cdot) : R^e \rightarrow R^N$ is a nonlinear mapping where N radial basis functions can be used to model the manifold in the embedding space, i.e.,

$$\psi(\cdot) = [\psi_1(\cdot), \cdots, \psi_N(\cdot)]^T$$

We use generalized radial basis function (GRBF) interpolation [PG90] to the original sequence y_t^s by solving for multiple interpolants, i.e., $R^e \rightarrow R$ for each tracking feature. Thin-plate splines are used as the basis functions.

Since the embedded manifolds are normalized, all the variations due to personalized style are expected to be represented in the nonlinear mapping space. Therefore, decomposing the nonlinear coefficient, B^s, would facilitate separation of the style parameters. Given learned models in the form of equation 2 for each person, the style can be decomposed in the linear mapping coefficient space using a bilinear model in a way similar to [TF00, VT]. Therefore, input instance y_t can be written as an asymmetric bilinear model in the linear mapping space as

$$y_t = A \times_3 b^s \times_2 \psi(x_t^c) \qquad (3)$$

where A is a third order tensor (3-way array) with dimensionality $d \times N \times J$ and b^s is a style vector with dimensionality J and $\times_n$ denotes the mode-n tensor product. Given the role for style and content defined above, the previous equation can be written as

$$y_t = A \times_3 b^{people} \times_2 \psi(x_t^{configuration}) \qquad (4)$$

This decomposition can be achieved by arranging the mapping coefficients $B^1, \cdots, B^K$ for each person into a matrix form $\mathbf{B}$ and applying singular value decomposition as $\mathbf{B} = USV^T$, where the style vectors b^s are the rows of V.

Figure 8 shows an example of decomposed style vectors for three people. Each person's style vector shows a different dominant basis. New intermediate expression styles can be produced by linear combinations of people's style vectors. We can generate expressions in new styles by using any linear combination of the learned style vectors, plugging this new style as the vector b^s in equation 3 and change the variable x_t^c over time along the embedded manifold.

5. Experimental results

All our experiments run at interactive rates on a Pentium Xeon 3GHz dual processor platform. We present two experiments to verify the effectiveness of our algorithms. In both experiments, we analyze the expression style b^{s1}, b^{s2} of two persons. We then generate a new style vector b^{new} by linear interpolation of these two styles using a control parameter α as follows:

$$b^{new} = \alpha b^{s1} + (1-\alpha) b^{s2}, \qquad (5)$$

where, $\alpha = 0$ corresponds to expression style b^{s2} and $\alpha = 1$ corresponds to expression style b^{s1}. An expression style between b^{s2} and b^{s1} can be generated by varying the value of α between 0 to 1.

The capabilities of our approach are shown in the videos accompanying this paper. The first part of the video demonstrates the dynamic 3-D shape capture system and the dynamic multi-resolution tracking of facial expressions. The output of this part of our system are the 3-D nodal locations of the model over time. These 3-D model nodal locations are used for the dynamic morphing and expression transfer experiments described in the following sections. Standard texture mapping and shading techniques are used to render the tracking results. After the initial fitting on the first frame of a sequence, each vertex on the control mesh is assigned the color of the closest point from the 3-D scan data.

5.1. Dynamic morphing of expression and geometry

In the first experiment, we used the global 3-D locations of 8K model nodes after precise model fitting to the motion capture data as described in Sec. 3. These 3-D model nodal positions estimate a person's facial geometry and motion style for a given type of expression (content), see Sec. 4. Therefore, for our dynamic morphing applications, expression style is the combination of an individual's face geometric characteristics and motion characteristics. Our approach allows us to go beyond traditional morphing between two static faces, to dynamic morphing of geometry and expressions.

Figure 7 shows that each new facial expression combines geometric style (shape) as well as motion style when we combine two persons' style factors. Each column shows the generation of new motions through time (content) for a given person style. (fixed style). Each column corresponds to a different expression style as well as a new facial geometry. Columns (a) and (d) represent the motion styles of two persons and columns (b) and (c) represent intermediate morphing results, showing geometry variation according to style change along each row.

In addition, we can also generate expression variation with geometric morphing simultaneously by changing the style control parameter α between 0 and 1 over time. This corresponds to moving over time from the top left of Figure 7 to the bottom right of Figure 7. In the accompanying video sequences we show this exact case of dynamic morphing of expression and geometry with seamless transitions both in geometry and in expression.

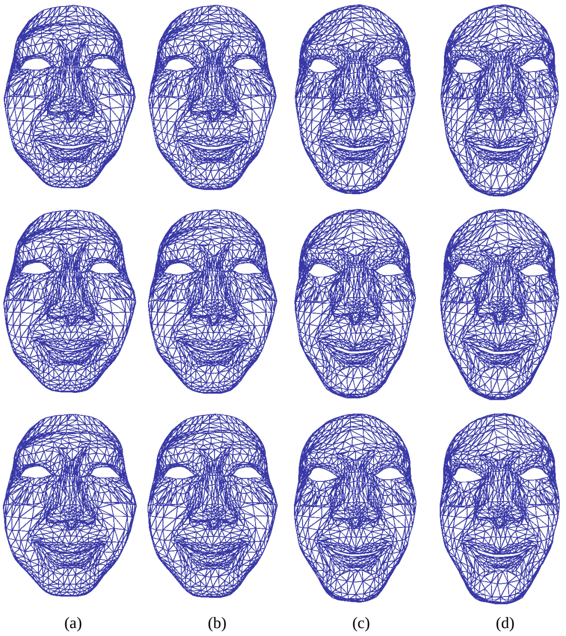

Figure 7: *Expression and geometry morphing. Col.(a): Second actor's original motion sequence, $\alpha = 0$. Col.(d): First actor's original motion sequence, $\alpha = 1$. Col.(c): Morphing result with $\alpha = 0.2$. Col.(d): Morphing result with $\alpha = 0.8$. From top to bottom, rows shows sampling of the generated motions at the 1st, 25th and 50th frame out of the 150 total frames.*

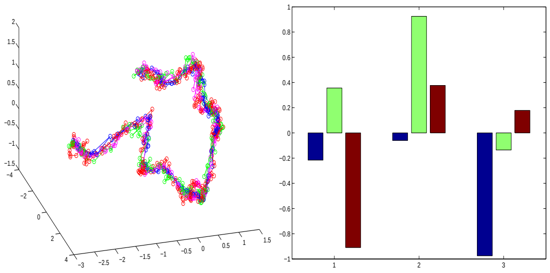

Figure 8: *Unified embedding of manifolds and motion style vectors for the geometry independent transfer of expression.*

5.2. Geometry independent expression transfer

In the second experiment, the goal is to synthesize and transfer only facial expressions to other individuals. This is achieved by analyzing facial expression styles independently of individual facial geometry. First, we define a base facial geometry for each individual in its object-centered reference frame. Then the facial motion in an expression is represented by the displacements between each frame geometry and the base geometry. This approach is possible as we can capture local deformations with consistent correspondences among different people as well as in different frames of the same person. We normalize the nodal points in each frame to eliminate variations in the displacements due to head size (a scaling operation) or head motion (face centering operation). In the normalization process, we subtract the mean value of all nodal points in each frame from the original nodal points and scale the whole mesh down to a unit size using the base model's face size. We choose the 3-D nodal positions of the face model in the first tracking frame as the base facial model since it is an actor's neutral face geometry.

In our experiments we analyzed the smiling expressions from three actors. Using our decomposable generative model, we analyzed the motion style factor for each person using the variation of the feature point location from the base geometry. Figure 8 shows the learned unified manifold (left) and the three individual motion style vectors (right). When we apply a motion style to a new actor we compute first the scale factor from the actor's base geometry. Therefore, we can transfer one actor's motion style to another actor regardless of her facial geometry. In Figure 9, we show two different base faces and the generation of new motion styles by combining different style factors. Each row shows four frames of the same motion style for two different base face geometries, demonstrating the effects of facial expression transfer to different base geometries. The fourth row for the first actor ($\alpha = 1$) and the first row of the second actor ($\alpha = 0$) show their original facial expressions. The second and third rows show new expressions styles by interpolation of style factors from rows 1 and 4. Figure 10 shows the mapping of the facial expression of previous two actors to new actor. The first row is the expression derived from the second actor, and the second row from the first actor.

6. Conclusions

In this paper we presented a system that accurately captures high speed, high resolution dynamic 3-D data and associated texture. We developed a multi-resolution method for intra-frame registration of freely deforming 3-D meshes and captured a small database of 3-D facial expressions from a few different subjects. We non-linearly projected our extremely high dimensional facial motion data onto low dimensional manifolds, which then were unified in a common embedding, which allows for the factorization of the most discriminating characteristics for each subject's expressions. Finally, synthesis of new facial motions was achieved through combined 3-D geometry and motion morphing, or through expression transfer.

Limitations of the method that will be addressed in future work include the absence of skin reflectance modeling for rendering under different illumination conditions as well as the absence of a specialized interior mouth and lip model to allow for large open mouth expressions. Since there are no coarticulation issues for the types of applications we examined in this paper, the incorporation of editing abilities of individual motion parameters that make sense to animators, into the global style analysis framework, should be straightforward.

7. Acknowledgments

The authors are grateful to S. Goldenstein and C. Vogler for their help with the tracking system[GVM03] and S. Venkataraman for initial face mesh work. Support for this research was provided by NSF grants ACI-0313184, ITR-0205671 and CMS-9900337 and NIH grant RR13995.

References

[BB02] BYUN M., BADLER N. I.: Facemote: qualitative parametric modifiers for facial animations. In *Symposium on Computer Animation* (2002), pp. 65–71.

[BBPV] BLANZ V., BASSO C., POGGIO T., VETTER T.: Reanimating faces in images and video. In *Eurographics'03*, pp. 641–650.

[BN] BEIER T., NEELY S.: Feature-based image metamorphosis. In *SIGGRAPH'92*, pp. 35–42.

[BO] BREGLER C., OMOHUNDRO S. M.: Nonlinear manifold learning for visual speech recognition. In *ICCV'95*, pp. 494–499.

[Bra] BRAND M.: Voice puppetry. In *SIGGRAPH'99*, pp. 21–28.

[BV] BLANZ V., VETTER T.: A morphable model for the synthesis of 3d faces. In *SIGGRAPH'99*, pp. 187–194.

[CDB02] CHUNANG E. S., DESHPANDE H., BERGLER C.: Facial expression space learning. In *Pacific Graphics* (2002), pp. 68–76.

[CHP] CHADWICK J. E., HAUMANN D. R., PARENT R. E.: Layered construction for deformable animated characters. In *SIGGRAPH'89*.

[CR] CHUI H., RANGARAJAN A.: A new algorithm for non-rigid point matching. In *CVPR'00*, pp. 44–51.

[DMS] DECARLO D., METAXAS D., STONE M.: An anthropometric face model using variational techniques. In *SIGGRAPH'98*, pp. 67–74.

[DRR] DAVIS J., RAMAMOORTHI R., RUSINKIEWICZ S.: Spacetime stereo: A unifying framework for depth from triangulation. In *CVPR'03*, pp. 359–366.

[EGP] EZZAT T., GEIGER G., POGGIO T.: Trainable videorealistic speech animation. In *SIGGRAPH'02*, pp. 388–398.

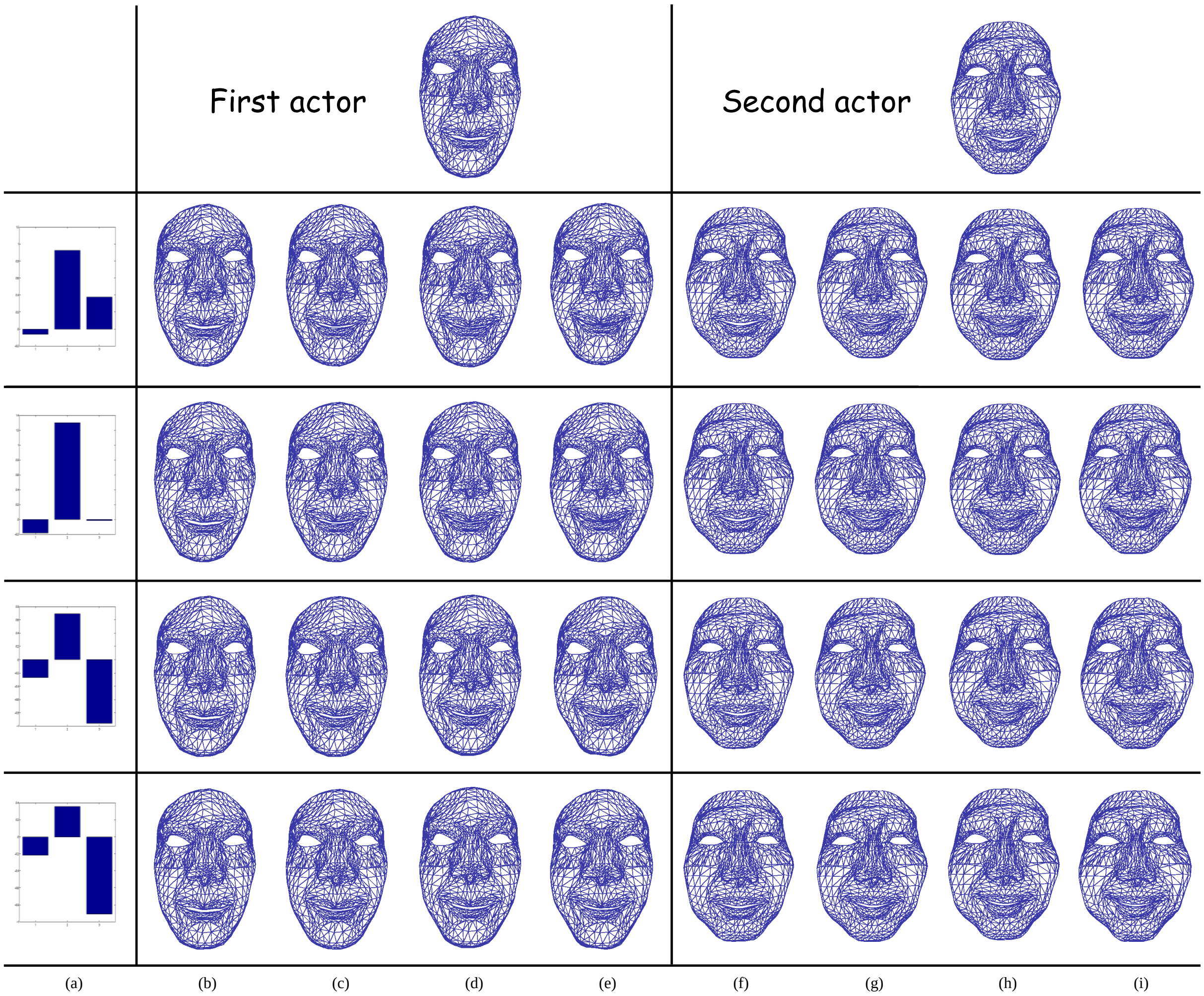

Figure 9: *Applying geometry-independent motion styles on two different base actors' geometries. Top: Two base actor facial geometries. Col. (a): style vectors. Col.(b-e): First actor. Col. (f-i): Second actor. Second row α = 0. Third row α = 0.3. Fourth row α = 0.8. Fifth row α = 1.0. Col. (b,f), (c,g), (d,h) and (e,i) correspond to motion images generated at frames 1, 15, 25 and 40 respectively, among 150 synthetic frames.*

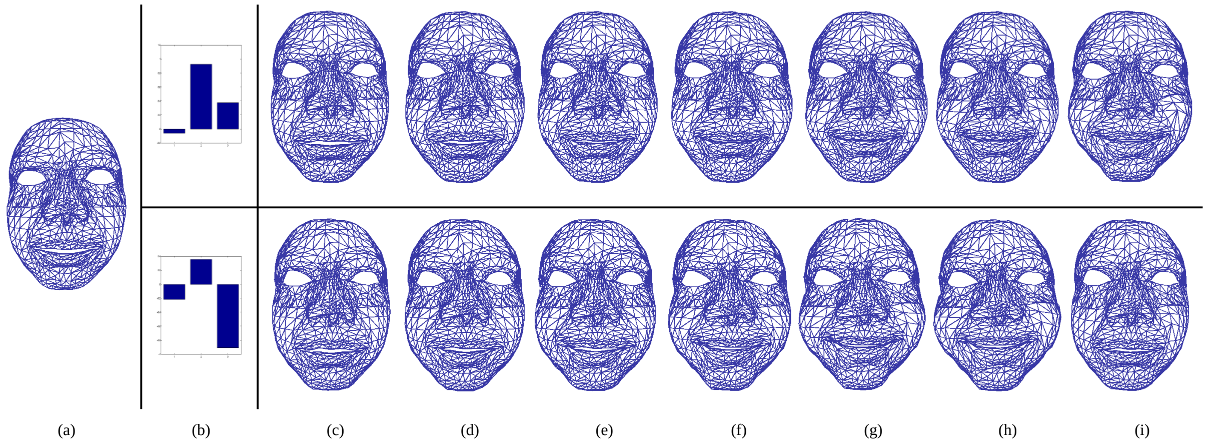

Figure 10: *Applying the motion styles of Figure 9 to a new geometry (different actor). Col. (a): New actor's base geometry. Col. (b): style vector for α = 0 (Second actor) and α = 1 (First actor). Col. (c-i): First row: α = 0. Second row: α=1. Col. (c-i) correspond to motion images generated at frames 1, 15, 25, 40, 60, 70, 130 respectively, among 150 synthetic frames.*

[EL] ELGAMMAL A., LEE C.-S.: Separating style and content on a nonlinear manifold. In *CVPR'04*.

[GGW*] GUENTER B., GRIMM C., WOOD D., MALVAR H., PIGHIN F.: Making faces. In *SIGGRAPH'98*, pp. 55–66.

[GVM03] GOLDENSTEIN S. K., VOGLER C., METAXAS D.: Statistical cue integration in dag deformable models. *PAMI 25*, 7 (2003), 801–813.

[HHJC] HUANG P. S., HU Q., JIN F., CHIANG F. P.: Color-encoded digital fringe projection technique for high-speed three-dimensional surface contouring. *Opt. Eng. 38(6)*, 1065–1071.

[HPM03] HUANG X., PARAGIOS N., METAXAS D.: Establishing local correspondences towards compact representations of anatomical structures. In *MICCAI'03* (2003), pp. 926–934.

[HZC03] HUANG P. S., ZHANG C., CHIANG F. P.: High-speed 3-d shape measurement based on digital fringe projection. *Opt. Eng. 42*, 1 (2003), 163–168.

[HZW*04] HUANG X., ZHANG S., WANG Y., METAXAS D., SAMARAS D.: A hierarchical framework for high resolution facial expression tracking. In *IEEE workshop on Articulated and Nonrigid Motion* (2004).

[JTDP03] JOSHI P., TIEN W. C., DESBRUN M., PIGHIN F.: Learning controls for blend shape based realistic facial animation. In *Symposium on Computer Animation* (2003), pp. 187–192.

[KG02] KALBERER G. A., GOOL L. V.: Realistic face animation for speech. *Intl. Journal of Visualization and Computer Animation* (2002).

[KHS03] KÄHLER K., HABER J., SEIDEL H.-P.: Reanimating the dead: reconstruction of expressive faces from skull data. *ACM Trans. Graph. 22*, 3 (2003), 554–561.

[KMMTT] KALRA P., MANGILI A., MAGNENAT-THALMANN N., THALMANN D.: Simulation of facial muscle actions based on rational free form deformations. In *Eurographics'92*, pp. 59–69.

[LTW] LEE Y., TERZOPOULOS D., WALTERS K.: Realistic modeling for facial animation. In *SIGGRAPH'95*, pp. 55–62.

[Mal92] MALACARA D. (Ed.): *Optical Shop Testing*. John Wiley and Songs, NY, 1992.

[MNS88] M. NAHAS H. H., SAINTOURENS M.: Animation of a b-spline figure. In *The Visual Computer* (1988), pp. 272–276.

[NN] NOH J.-Y., NEUMANN U.: Expression cloning. In *SIGGRAPH'01*, pp. 277–288.

[OS88] OSHER S., SETHIAN J.: Fronts propagating with curvature-dependent speed : Algorithms based on the Hamilton-Jacobi formulation. *Journal of Computational Physics 79* (1988), 12–49.

[PG90] POGGIO T., GIROSI F.: Networks for approximation and learning. *Proc. IEEE 78*, 9 (1990).

[PHL*] PIGHIN F., HECKER J., LISCHINSKI D., SZELISKI R., SALESIN D. H.: Synthesizing realistic facial expressions from photographs. In *SIGGRAPH'98*, pp. 75–84.

[PW96] PARKE F. I., WATERS K.: *Computer facial animation*. 1996.

[RHHL] RUSINKIEWICZ S., HALL-HOLT O., LEVOY M.: Real-time 3d model acquisition. In *SIGGRAPH'02*, pp. 438 – 446.

[RS00] ROWEIS S., SAUL L.: Nonlinear dimensionality reduction by locally linear embedding. *Science 290*, 5500 (2000), 2323–2326.

[RSH*99] RUECKERT D., SONODA L., HAYES C., HILL D., LEACH M., HAWKES D.: Nonrigid Registration Using Free-Form Deformations: Application to Breast MR Images. *IEEE Transactions on Medical Imaging 8* (1999), 712–721.

[SP] SEDERBERG T. W., PARRY S. R.: Free-form deformation of solid geometric models. In *SIGGRAPH'86*, pp. 151–160.

[TdSL00] TENENBAUM J. B., DE SILVA V., LANGFORD J. C.: A global geometric framework for nonlinear dimensionality reduction. *Science 290*, 5500 (2000), 2319–2323.

[TF00] TENENBAUM J. B., FREEMAN W. T.: Separating style and content with biliear models. *Neural Computation 12* (2000), 1247–1283.

[VT] VASILESCU M. A. O., TERZOPOULOS D.: Multilinear analysis of image ensembles: Tensorfaces. In *ECCV'02*, pp. 447–460.

[ZCS] ZHANG L., CURLESS B., SEITZ S. M.: Spacetime stereo: Shape recovery for dynamic scenes. In *CVPR'03*, pp. 367–374.

[ZH04] ZHANG S., HUANG P. S.: High-resolution, real-time 3-d shape acquisition. In *IEEE Workshop on Real-time 3D Sensors and Their Use (joint with CVPR04)* (2004).

[ZLGS03] ZHANG Q., LIU Z., GUO B., SHUM H.: Geometry-driven photorealistic facial expression synthesis. In *Symposium on Computer Animation* (2003), pp. 177–186.