2010

Single View Reconstruction for Human Face and Motion with Priors

Xianwang Wang
University of Kentucky, xianwangwang@gmail.com

Recommended Citation

Wang, Xianwang, "Single View Reconstruction for Human Face and Motion with Priors" (2010). *University of Kentucky Doctoral Dissertations*. 62.
https://uknowledge.uky.edu/gradschool_diss/62

ABSTRACT OF DISSERTATION

Xianwang Wang

The Graduate School
University of Kentucky
2010

Single View Reconstruction for Human Face and Motion with Priors

ABSTRACT OF DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Engineering at the
University of Kentucky

By
Xianwang Wang
Lexington, Kentucky

Director: Dr. Ruigang Yang, Associate Professor of Computer Science
Lexington, Kentucky 2010

ABSTRACT OF DISSERTATION

Single View Reconstruction for Human Face and Motion with Priors

Single view reconstruction is fundamentally an under-constrained problem. We aim to develop new approaches to model human face and motion with model priors that restrict the space of possible solutions. First, we develop a novel approach to recover the 3D shape from a single view image under challenging conditions, such as large variations in illumination and pose. The problem is addressed by employing the techniques of non-linear manifold embedding and alignment. Specifically, the local image models for each patch of facial images and the local surface models for each patch of 3D shape are learned using a non-linear dimensionality reduction technique, and the correspondences between these local models are then learned by a manifold alignment method. Local models successfully remove the dependency of large training databases for human face modeling. By combining the local shapes, the global shape of a face can be reconstructed directly from a single linear system of equations via least square.

Unfortunately, this learning-based approach cannot be successfully applied to the problem of human motion modeling due to the internal and external variations in single view video-based marker-less motion capture. Therefore, we introduce a new model-based approach for capturing human motion using a stream of depth images from a single depth sensor. While a depth sensor provides metric 3D information, using a single sensor, instead of a camera array, results in a view-dependent and incomplete measurement of object motion. We develop a novel two-stage template fitting algorithm that is invariant to subject size and view-point variations, and robust to occlusions. Starting from a known pose, our algorithm first estimates a body configuration through temporal registration, which is used to search the template motion database for a best match. The best match body configuration as well as its corresponding surface mesh model are deformed to fit the input depth map, filling in the part that is occluded from the input and compensating for differences in pose and body-size between the input image and the template. Our approach does not require any makers, user-interaction, or appearance-based tracking.

Experiments show that our approaches can achieve good modeling results for human face and motion, and are capable of dealing with variety of challenges in single view reconstruction, e.g., occlusion.

KEYWORDS: 3D reconstruction, marker-less motion capture, face modeling and analysis, motion modeling and analysis, non-linear dimensionality reduction

Author's signature: Xianwang Wang

Date: November 30, 2010

Single View Reconstruction for Human Face and Motion with Priors

By
Xianwang Wang

Director of Dissertation: Ruigang Yang

Director of Graduate Studies: Raphael A Finkel

Date: November 30, 2010

RULES FOR THE USE OF DISSERTATIONS

DISSERTATION

Xianwang Wang

The Graduate School
University of Kentucky
2010

Single View Reconstruction for Human Face and Motion with Priors

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Engineering at the
University of Kentucky

By
Xianwang Wang
Lexington, Kentucky

Director: Dr. Ruigang Yang, Associate Professor of Computer Science
Lexington, Kentucky 2010

Dedicated to my family.

ACKNOWLEDGMENTS

I am grateful to all those people who have helped me with this dissertation. Firstly, my sincere thanks to my advisor, Dr. Ruigang Yang for giving me an opportunity to do this work, developing my skills, and supporting me through the years. I have learned a great deal from working with him. Especially, I greatly benefited from his excellent advice and attempted to learn from his amazing writing skills. This work would never have happened without his help.

I would like to thank the members of my PhD committee for the comprehensive review of my thesis and the constructive discussion during the thesis defense. Thanks to Dr. Brent Seales, Dr. Melody Carswell, and Dr. Fuhua Cheng for the great discussions.

Many thanks to the current and former members of the Gravity Lab for the friendly atmosphere and for their help. Discussions with them regarding the work, were always interesting and inspiring. Special thanks to Mao Ye for the countless hours he spent in capturing the data and verifying the ideas during the project of motion capture.

Last but not least, sincere thanks my family for their love, patience and support, especially to my wife for her understanding and encouragement during the difficult stages of my thesis work, and also her endurance of the hard time, without which I will never be able to accomplish the work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

Human face and motion modeling has numerous applications in a variety of fields such as animation, surveillance, and human-computer interactions. Face modeling provides the application domain in Biometrics, such as authentication and security. Compared to other biometrics, i.e., finger, hand, and eye, faces have advantages in non-invasiveness and ease of use in [5]. The biggest application fields are entertainment and animation. Many animated films adopt face modeling to animate facial expressions, and a variety of characters in cinema and video games are realized through the transfer of captured motions to a particular shape. Motion modeling is also used for medical applications and athletic coaching. Specifically, one such application for athletes is the measurement of their range of motion and evaluation of their performance by comparison with standard motions. Such analysis helps prevent athletes from injuries and improves their performance.

Most current commercial systems address the 3-D face reconstruction problem by adding constraints in the form of adopting multiple views, projecting lasers or using structured light with patterns or special textures. These constraints and the requirement of special hardware reduce the operational flexibility of any such system.

Similarly, most marker or marker-less motion capture systems require a surrounding camera array to provide a complete observation of the motion [6] due to the complex nature of human motion. But marker-less motion capture from a single view has been and continues to be the ultimate grand challenge for motion modeling. The ease of operation and the reduced equipment cost of a single-camera system could eventually make motion capture a household routine, enabling many new applications that are currently blocked by the prohibitive cost and the cumbersome user interfaces of existing motion capture solutions.

Therefore, while the techniques of face modeling and motion capture using markers or special equipment are mature enough to be widely adopted in many practical applications, single view reconstruction (SVR) for human face and motion remains an active topic in both computer vision and computer graphics. Over the years, the research of SVR has been focused on using one regular video camera due to the cost-effectiveness and non-invasiveness. Unfortunately, human face and motion modeling from single-view video images is a hard problem due to the following challenges: (1) It is fundamentally an under-constrained problem, because the human face and body parts are rarely completely observable in any single given image. (2) The image-to-shape mapping is highly multi-modal, and a single function cannot model this inverse mapping, thus modeling the mapping requires a large amount of training data which is not often available. (3) The problem is further complicated by the complexity of the scene, such as lighting conditions, pose variation, cluttered background, and occlusion.

The objective of this thesis is therefore to develop approaches to overcoming the inherent ambiguities and challenges with model priors in our SVR problems. That is, for the problem of human face modeling, we aim to recover the 3D shape from a single image by using the local image or shape models learned from the training database, while for the problem of human motion modeling, we estimate the body configurations from a depth sequence from a known template model database.

1.1 Local Model-based Human Face Modeling

We incorporate learning techniques into 3D face reconstruction. To globally and completely capture the details and the underlying dynamics of both the images and the 3D shapes respectively, previous statistical learning techniques require a vast amount of training data to achieve accurate reconstruction. To overcome this limitation, we take advantage of the facts: 1) the local homogeneous surface follows the same deformation rule; 2) it is much easier to compensate illumination locally, and the local images and shapes have considerably smaller variance. 3) the local model can be captured from fewer examples due to more constraints in the local surface deformation. Therefore, we first divide the image and 3D shapes into overlapping patches and learn the local non-linear prior models by applying non-linear dimensionality reduction (DR) to each patch. A global shape can be recovered by encouraging its patches to conform to the local models. The non-linear approach of statistical learning We have developed overcomes the weakness of non-linear global models by learning local deformation models from manageable amounts of training examples. Therefore, We have developed an applicable and fully automated approach to recover 3D shapes from single face images with large variations in both pose and illumination, which is beyond previous state-of-the-art techniques.

More specifically, the ***technical contributions*** of our method include:

- We introduce a new parametrization of the face model. Rather than recording the absolute position of vertices, we record the per-triangle affine transformation between an individual model and a reference model. This parametrization is invariant to pose changes of 3D shapes and implicitly encodes the fact that the vertices cannot move independently off one another.

- We divide the image and 3D shapes into overlapping patches and apply the nonlinear DR method to each patch. Working on the patch level has two advantages over the whole face: it is easier to compensate illumination locally, and the images and shapes within a patch have considerably smaller variance [7]. Nonlinear DR methods have been shown to be more effective for deformation [8].

- A novel approach is developed to estimate the pose from a single input image by combining the unsupervised metric learning technique with the supervised metric learning techniques. Unlike previous DR-based pose estimation methods that treat illumination as a part of the pre-processing step, We have developed a unified framework that does not require any pre-processing for illumination normalization or correction.

- Instead of relying on explicit 2D-3D correspondences in the training database, we apply *manifold alignment* techniques to find the appropriate mapping between a 2D image and its corresponding 3D shape. This eliminates the need for tedious and manual labeling in the training database.

Using these novel components, our approach is able to deal with face images with both varying illumination and very large pose variation–up to 90° profile view

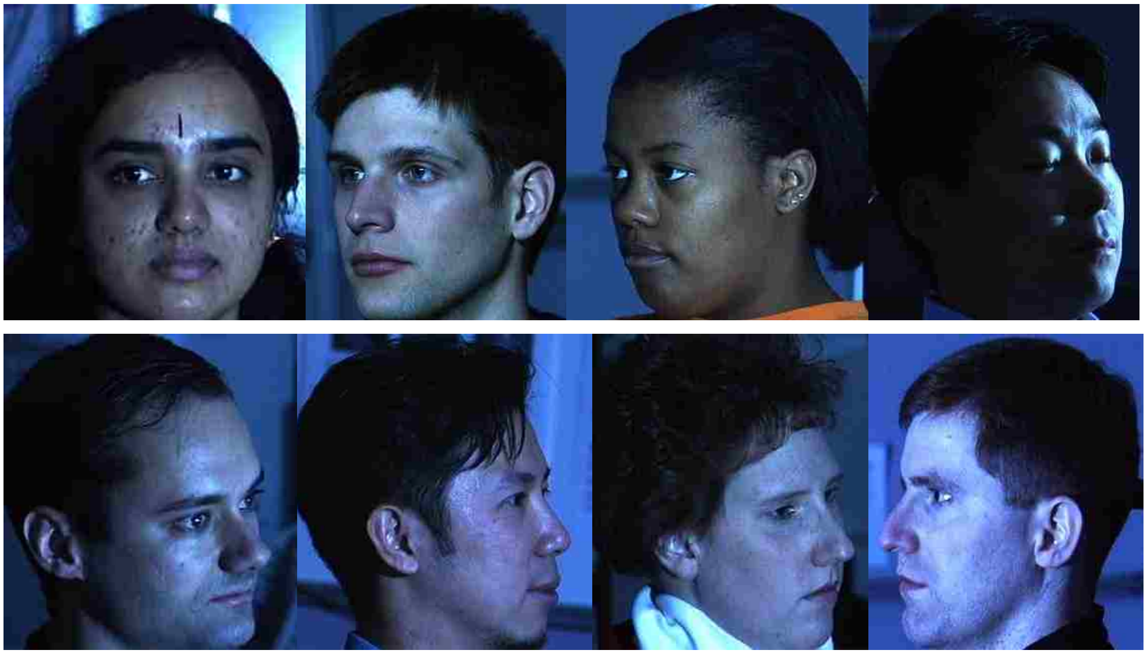


Figure 1.1: Sample input images for human face modeling. Notice the large variation in pose and illumination.

as shown in Figure 1.1, which we believe has not been demonstrated before. Note that the results are based on a training database without 2D-3D labeling or any illumination variations, making our method more accessible. Furthermore, the global reconstruction is achieved by solving a linear system in closed form. No iterative step is needed.

1.2 Template-based Human Motion Modeling

Compared to face modeling, human motion modeling has more challenges. To remove the challenges caused by the variations in lighting conditions and cluttered backgrounds, we present a single-view marker-less motion capture solution that uses a different cue, the scene depth. Depth is a fundamentally more stable cue than a 2D photograph. It not only provides metric 3D measurement, but also it is invariant

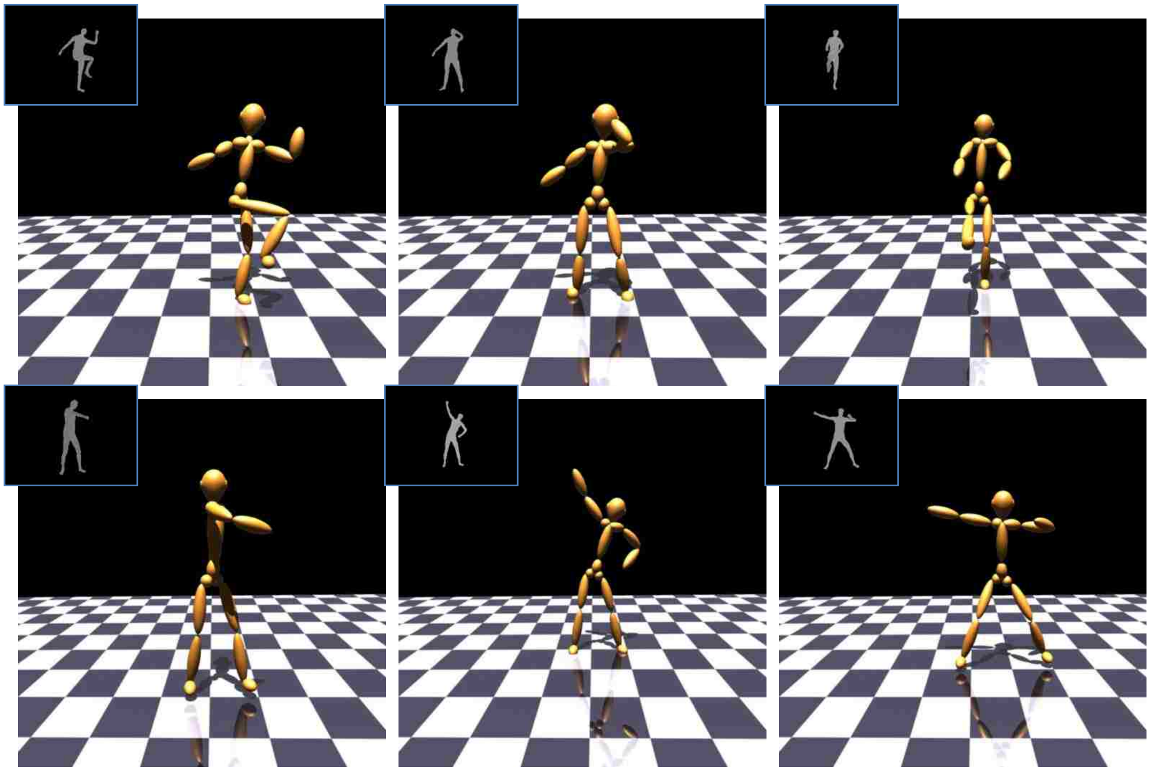


Figure 1.2: Sample input images for single person modeling. Notice the large variation in motion and occlusion.

of appearance, and the irrelevant information (such as the background) can be easily segmented. Our desire to use depth is also motivated by the recent availability of full-frame depth sensors, which provide more stable depth maps than typical passive stereo. Nevertheless, using depth for motion capture is not as simple as it appears at first glance. First, a depth sensor only generates a point cloud with noise and outliers; semantic information about which part corresponds to which joint must be extracted. Secondly, there exists large occlusion: at least 50% of the body is not observable in any single view.

We formulate motion capture from a single depth image sequence as a model fitting problem with a known template model database, from which we *extract semantics* and *fill the missing surface regions* so that the occluded body parts can be estimated. Our approach recovers the skeleton motion and the temporal deformation of the surface in an interleaved manner. We first estimate the body configuration (also called skeleton motion) of the current frame by registering the input (the current depth points) with the previous frame. The estimated body configuration is used to search for the skeleton motion in the template database most similar to the current input. We then reconstruct a complete 3D surface model by deforming the template surface model to fit the current depth points. The reconstructed 3D model ensures one-to-one vertex correspondence of reconstructed deformations through the entire motion sequence, which is very useful for numerous applications such as texturing and deformation transfer. The recovered full surface model in turn refines the accuracy of the body configuration, and reduces the problem of temporal tracking drift.

Experiments have showed that our method is accurate with positional error usually within 20mm of the ground truth. Figure 1.2 illustrates some examples of single-person motion recovery. The results in the accompanying video demonstrate that our method can correctly estimate motion configurations from a wide spectrum of scenes, including walking (Figure 1.3a), over-stretching (Figure 1.3b), kicking (Figure 1.3c), and swinging (Figure 1.3d). In addition to these basic motions, our approach can recover complex activities and details of motion models. For example, our approach can deal with challenging scenes, such as: dancing (Figure 1.3e), the partial occlusion, and total occlusion in body parts(Figure 1.3e, Figure 1.4a). Furthermore, our method is capable of handling extreme deformations as the subject moves into different poses and shapes. This is achieved automatically without any user interaction and careful

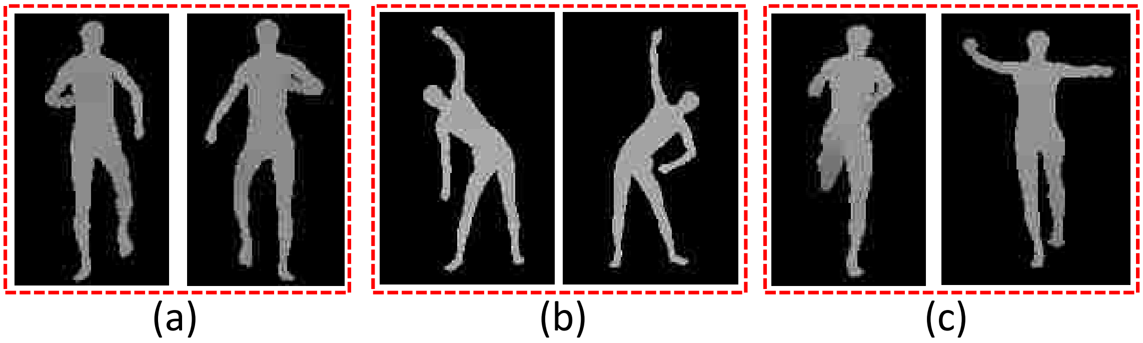
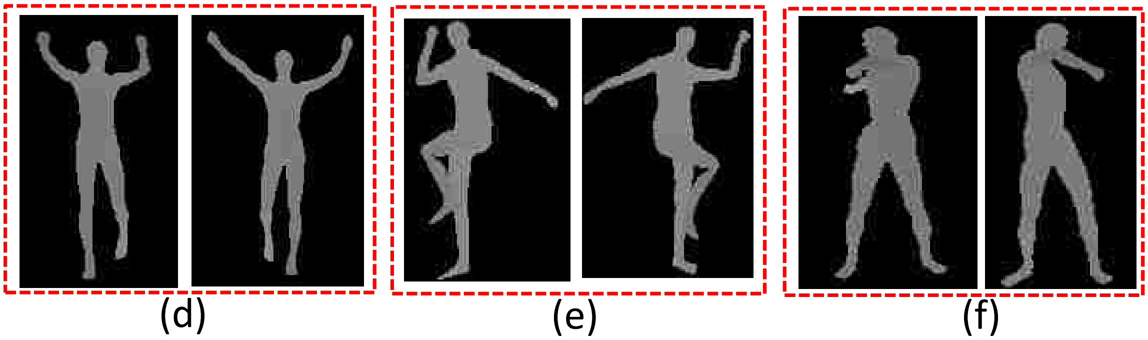



Figure 1.3: Some examples of input images to our approach. (a) walking; (b) stretching; (c) kicking; (d) swinging; (e)dancing; and (f) occlusion.

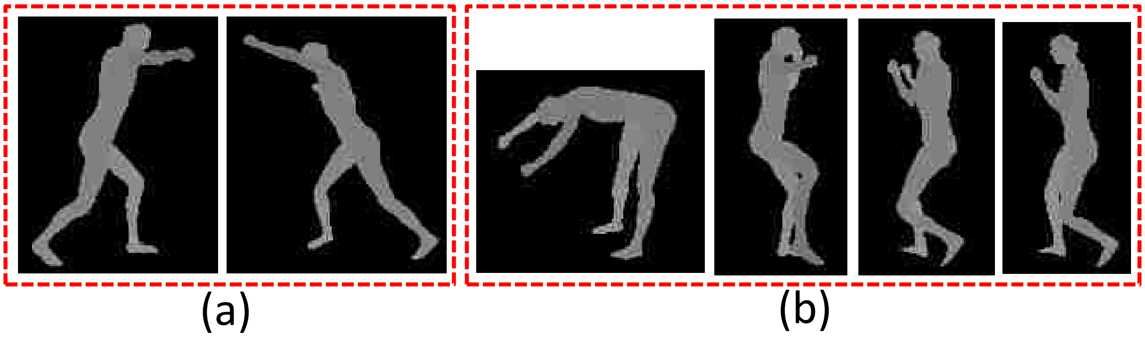


Figure 1.4: Occlusion examples. (a) partial occlusion and total occlusion; (b) extreme deformations and occlusions

placement of makers. In Figure 1.4b, the nature of the profile view occludes one side of the body, resulting in one leg (arm) of the subject hidden by the other leg (arm). This is significant as the template database has no similar pose, or any pose even close to the displayed pose. To demonstrate our method's robustness to occlusions, we even extended it to multiple-persons motion capture. As shown in Figure 1.5,

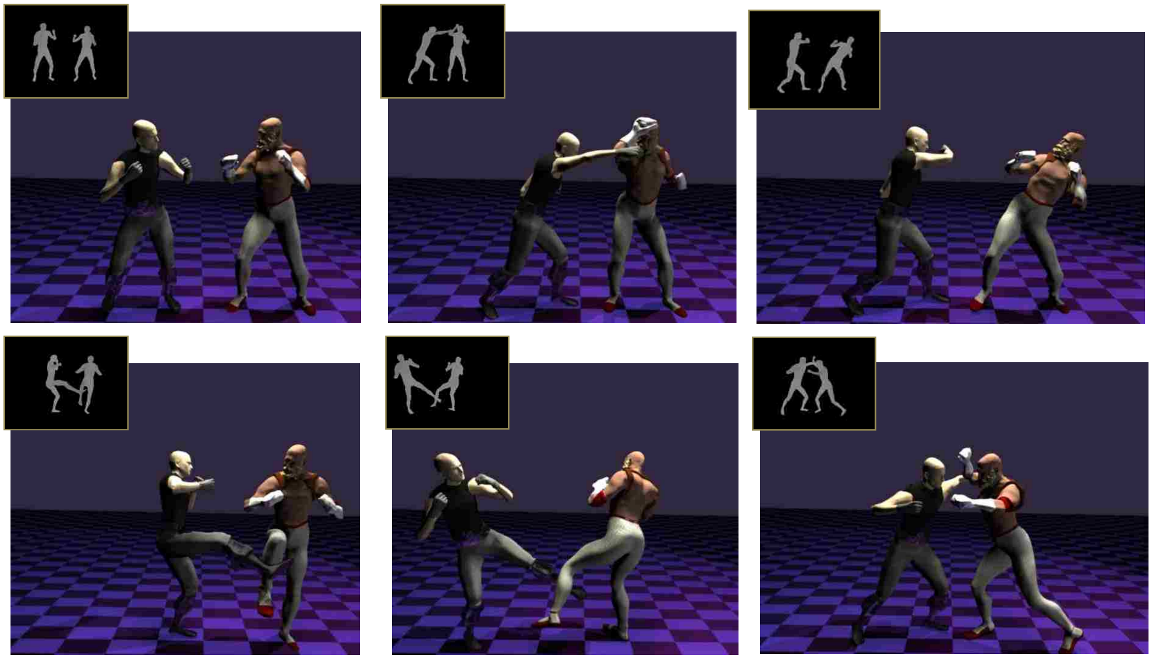


Figure 1.5: Multiple-person motion capture using a single depth video. The insets show the input depth maps, which has been segmented from the background and each other. The recovered body configurations are used to drive the two models.

not only is more than half of each subject occluded, the subject's silhouette is also destroyed. Correct object silhouette is a prerequisite for many video-based motion modeling approaches. Because we use the depth cue and the incorporation of a motion database, our approach can successfully handle such situations. To the best of our knowledge, this has never been demonstrated in multi-person cases.

In short, we present a novel approach to address these issues in current single-view 3D reconstruction of motion modeling. More specifically, the ***technical contributions*** of our method include:

- We present a robust framework of single view marker-less motion capture by adopting the depth cue and the template database. The framework is able to handle large or severe occlusion and deformation from depth images with large internal and external variations in a given scene, and achieves an average motion tracking accuracy of 20mm, which is much better than that of state-of-the-art approaches.

- Our algorithm is invariant of viewpoint and body-size through our two-step registration process. This eliminates the need of building a motion database covering all of the motions, viewpoints, and body sizes, which generally represents a very significant amount of data.

- We introduce a surface fitting technique to refine the accuracy of the body configuration, which significantly corrects the temporal tracking drift in single view motion capture. Furthermore, Temporal filtering is adopted to remove the jittering artifacts.

1.3 Thesis Outline

The rest of this thesis is organized as follows. Chapter 2 reviews the related work on single view reconstruction in the literature. In chapter 3, we present our local models to recover 3D shape from a single input image with large variations in pose, illumination and identity. Chapter 4 introduces our template-based modeling algorithm, which combines non-rigid registration and surface fitting techniques. The conclusion is made and the future directions of our research is discussed in Chapter 5.

Chapter 2 Related Work

Single view reconstruction from individual images is known to be highly ambiguous, because surfaces with very different shapes can generate very similar images under perspective projection. It is acknowledged to be an ill-posed problem unless the space of possible configuration is constrained [9]. Traditional shape-from-shading [10] and shape-from-texture [11] techniques only recover the Lambertian surfaces with known albedo, or surfaces with homogeneous texture patterns. Even given a calibrated perspective camera and a well-textured surface, the depth ambiguities cannot be resolved in individual images. The standard approach to overcoming these challenges is to introduce a deformable model and to recover the shape by optimizing an objective function that measures the fit of the model to the data. However, in practice this objective function is usually non-convex. Thus, to avoid being trapped in local minima, these methods require initial estimates that must be relatively close to the true shape. Geometric clues, such as silhouettes and normal maps are usually used to resolve the inherent ambiguities, but far from enough to obtain an unambiguous solution. Additional geometrical or topological priors, i.e., a specific class of objects, have been proposed in previous research to further constrain the problem. However, these methods make a strong assumption over the input scenes, e.g., planar outdoor architecture scenes and ground-vertical scenes. Such limitation prevents these methods from reconstructing surfaces with more complex and curved geometry like human faces and body motions.

To relax these constraints, approaches over the years endeavored to make this problem tractable by introducing prior models [12, 9, 13, 14] which are in the form of physics-based models, models reconstructed from non-rigid structure-from-motion algorithms [15, 16, 17], or models learned from statistical learning techniques [18, 19, 20, 12, 9, 13, 14]. Most methods still make restrictive assumptions about the object of interest that are sometimes hard to satisfy.

Physics-based models attempt to recover the shapes by introducing global models, which approximate intrinsic physical behaviors of a dynamical system in terms of the variables. Many variations of these models have been successfully applied to 2-D surface registration and 3-D surface modeling, e.g., under the form of superquadrics [21], triangulated surfaces [22], or thin-plate splines [23]. Physics-based models have shown their strong potential to solve SVR problems, due to excellence at fitting noisy image data and handling highly deformable 3D objects. However, these models reduce the number of degrees of freedom in surface deformation with linearity assumption through regularization terms or modal analysis [24, 22, 25] due to the high dimensionality of the problems. The simplification in the complexity of modeling prevents them from cases where there are large deformations present. Although more accurate and complex non-linear models [26, 27] have been proposed to deal with the non-linearity of deformation, the minimization of an image-based objective function with high complexity may have many local minima. Furthermore, the knowledge of physical properties of the surface is typically unknown, which introduces the challenges in the design of the object function.

Non-rigid structure from motion methods rely on tracking of feature points to simultaneously recover 3-D surface points and the deformation models [28,15,29,30,17] through image sequences. The advantage over other SVR approaches is that they require few priors. However, these methods suffer from two major drawbacks: 1) A sufficient number of feature points are required to be tracked throughout the whole sequence to learn both shape and motion, which limits their applicability. 2) Similar to physics-based model techniques, non-rigid structure from motion techniques are effective only when dealing with relatively small deformations, or smooth deformations, because they oversimplify the motion of surface by modeling deformations as either a linear combination of online learned basis vectors [16], constant basis vectors [15], or several piecewise rigid objects independently moving with respect to one another [17].

Both physics-based models and models reconstructed from non-rigid structure from motion techniques have difficultly in accurately capturing the non-linear physics of large deformations, due to the complexity of modeling the true physical properties of surfaces. Because of these limitations there has been an increase in statistical learning techniques over the years to model deformations. These techniques take advantage of training data in conjunction with dimensionality reduction techniques to learn low-dimensional models. Most of these models currently in use trace their roots to the early Active Appearance Models [31] in the 2-D case, followed by Morphable Models [18] and Active Shape Models [19]. These linear models can capture more of the true variability than modal analysis because they are learned from training examples. However, they have the same restriction of smooth constraints as before. Non-linear global models have also been demonstrated for surface deformations [32, 13,12,9]. Due to the many degrees of freedom of highly deformable surfaces, learning of these non-linear global models is tractable only when sufficient training data is available. Thus, the difficulty of building a database with enough examples has limited the spread of global non-linear model-based approaches. Furthermore, non-linear learning generally involves dealing with challenges of optimization of complex object functions that may be difficult to resolve because of non-convexity. Therefore, these models are typically designed to one specific kind of surface, such as that of a human face.

More sources of information may be used to overcome the ambiguities of single view reconstruction. Zhao and Chellappa [33] combine texture and shading cues to constrain the reconstruction problem. However, they make very restrictive assumptions on lighting conditions, which results in a method that lacks generality.

In the remainder of this chapter, we first present the related work of head pose estimation, which is one of the important components in human face modeling. Following that, we discuss the related work of our SVR problems in human face and motion modeling in more details.

2.1 Pose Estimation

In this section we review the research work in distance metric learning and head pose estimation approaches based on dimensionality reduction. Many methods have been developed. We discuss each one of them in the following sections.

Distance Metric Learning

Approaches in this category attempt to learn metrics that keep all the data points within the same classes close, while separating all the data points from different classes far apart. Xing *et al.* [34] formulate distance metric learning as a constrained convex programming problem and learn a global distance metric that minimizes the distance between the data pairs in the equivalence constraints subject to the constraint that the data pairs in the inequivalent constraints are well separated.

Local linear discriminative analysis [35] estimates a local distance metric using the local linear discriminant analysis. Relevant Components Analysis (RCA) [36] learns a full ranked Mahalanobis distance metric using equivalence constraints. The learned linear transformation can be used directly to compute distance between any two examples. Components Analysis (NCA) [37] maximizes the leave-one-out cross validation to learn a distance metric for KNN classifier. Large Margin Nearest Neighbor(LMNN) [38] extends NCA through a maximum frame work. Discriminative Component Analysis (DCA) and Kernel DCA [39] improve RCA by exploring negative constraints and capturing nonlinear relationships using contextual information. Essentially, RCA [36] and DCA [39] can be viewed as extensions of Linear Discriminant Analysis (LDA) [40] by exploiting the must-link constraints and cannot-link constraints. Local Fisher Discriminant Analysis (LFDA) [41] can be viewed as localized variant or an extension of LDA. It assigns greater weights to those connecting examples that are nearby. Kim et al. [42] provide an efficient incremental learning method for LDA by applying the concept of the sufficient spanning set approximation in each update step, i.e., for the between-class scatter matrix and the projected data matrix, as well as the total scatter matrix.

Globerson and Roweis [43] learn a Mahalanobis distance by constructing a convex optimization problem whose solution generates such a metric by trying to collapse all examples in the same class to a single point and push examples in other classes infinitely far away. Liu et al. [44] present an efficient algorithm to learn a Local Distance Metric (LDM) by employing eigenvector analysis and bound optimization from training data in a probabilistic framework that aims to optimize local compactness and local separability. The work [45] presents a Bayesian framework for distance metric learning that estimates a posterior distribution for the distance metric from labeled pairwise constraints. Schultz and Joachims [46] extend the support vector machine to distance metric learning by encoding the pairwise constraints into a set of linear inequalities. Unlike previous methods of the semi-supervised clustering approach, Locally Linear Metric Adaptation (LLMA) [47] performs nonlinear transformation globally but linear transformation locally.

Manifold Learning

The goal for approaches in this category is to learn a low-dimensional manifold in which most "intrinsic information" (e.g., distance) are preserved. Popular approaches include ISOMAP [48], Locally Linear Embedding (LLE) [49], and Laplacian Eigenmaps (LE) [50]. ISOMAP preserves the geodesic inter-point distances, LLE preserves the distance based on locally linear combination of neighborhood, and LE preserves the distance described by a weighted connected graph constructed from neighborhoods.

Hu *et al.* [51] first used ISOMAP to map video or image sequences of each individual into 2D embedded space. All the manifolds were further normalized into a unified embedding space, after each manifold was represented by an ellipse by employing an ellipse fitting method. The head pose angle was obtained by applying Radial Basis Function interpolation. This method only works on face images with temporal continuity and local linearity (e.g, video sequences), although good results have been shown in their experiment.

Raytchev *et al.* [52] apply the ISOMAP-based manifold learning technique for user-independent pose estimation and evaluate their method in comparison with the Linear Subspace and Locality Preserving Projections(LPP) [53].

Chen *et al.* [54] uses the face images of two specific head poses and estimates the head poses between them through classification-based nonlinear interpolation. This approach is based on the assumption that face images from multiple views lie on a manifold in the original image feature space.

Fu and Huang [55] present an appearance-based strategy for head pose estimation using supervised Graph Embedding(GE) analysis. The neighborhood weighted graph is first constructed in the sense of supervised LLE. The out-of-sample data points may be treated using the projection transformation solved in closed-form based on GE linearization. The K-nearest neighbor classification is then employed to estimate the head pose. Their method is successful with low pose estimation error. They consider face images with only pose variation, but not illumination, change in their experiment.

Table 2.1: performance comparison of different methods for head pose estimation

Method	Interval	Increment	Best result: error	Illumination (Yes/No)
ISOMAP [52]	$[-90° + 90°]$	15°	11°	No
Fisher manifold learning [54]	$[-10° + 10°]$		3°	No
BME with ISOMAP, LLE [56]	$[-90° + 90°]$	2°	3°	Yes$^{\diamondsuit}$
BME with LE [56]	$[-90° + 90°]$	2°	2°	Yes$^{\diamondsuit}$
LEA [55]	$[-90° + 90°]$	1°	2°	No

$^{\diamondsuit}$LoG filter is used.

Balasubramanian *et al.* [56] propose Biased Manifold Embedding (BME) framework for head pose estimation. The pose information of given face image data is used to compute a biased neighborhood of each point in the feature space, before determining the low-dimensional embedding. The distances are defined as 0 between face images of the same pose angle. For data points from different pose angles, BME defines the distances by: $\tilde{D}(i,j) = \lambda_1 \cdot D(i,j) + \lambda_2 \cdot f(P(i,j)) \cdot g(D(i,j))$, where λ_1 and λ_2 are constants, $D(i,j)$ is the Euclidean distance between two data points x_i and x_j, $P(i,j)$ is the pose distance between data points x_i and x_j, f is any function of the pose distance, g is any function of the Euclidean distance between the data points, and $\tilde{D}(i,j)$ is the modified biased Euclidean distance. Many approaches [57, 58, 59, 60, 61] similar to BME try to modify the distance matrix between all the data points to improve the performance of the manifold learning techniques. They are all special cases of the BME framework as shown in Table 2.2. Balasubramanian *et al.* [62] present the unified view of all these approaches. BME uses a Generalized Regression Neural Network (GRNN) to learn the non-linear mapping for dealing with out-of-sample data points, and applies linear multivariate regression to estimate the pose. Essentially BME proposes a more general model to modify the distance matrix, which unifies the other supervised manifold learning approaches [57, 58]. The interested reader can refer to the discussion in [62] for details. Their experiment shows this method works well with person-independent face images with pose variation. BME uses Laplacian of Gaussian (LoG) to remove the illumination effects, but LoG representation is not sufficient for pose estimation under a wide variety of lighting conditions, in particular harsh lighting with shadows, as in our experiments.

The performance of representative DR-based approaches for head pose estimation are summarized in Table 2.1. It shows that LEA and BME are the current state-of-art techniques in pose estimation. However, none of them treats illumination variations in a principled way, most techniques simply do not discuss the effect of illumination.

2.2 Face Modeling

A classic method to recover 3D shape from a single image is Shape-from-Shading (SFS) [10, 63]. Direct application of SFS to face modeling has limited success since a face has large albedo variation and both concave and convex regions. Some SFS-based methods have been developed to improve shape recovery using specific domain constraints. The symmetric SFS method [33, 64] reconstructs the faces by exploiting the bilateral symmetry of faces. However, it is difficult to establish the point-wise correspondence between the symmetric parts. Prados *et al.* [65, 66] use a unique critical point over the face image to enforce convexity for shape recovery but all the parameters of the light source, the surface reflectance, and the camera have to be known.

Table 2.2: The unified view of supervised manifold learning techniques

Method	Setting
Balasubramanian *et al.* [62]	$\lambda_1 = 0,$ $\lambda_2 = 1,$ $f(P(i,j)) = \frac{\beta * \lvert P(i,j) \rvert}{\max_{m,n} P(m,n) - P(i,j)},$ $g(P(i,j)) = \begin{cases} D(i,j), & P(i) \neq P(j); \\ 0, & P(i) = P(j). \end{cases}$
Ridder et al. [57]	$\lambda_1 = 1,$ $\lambda_2 = \alpha \times \max(\Delta),$ $f(P(i,j)) = \Lambda,$ $g(P(i,j)) = 1.$
Li and Guo [58]	$\lambda_1 = 0,$ $\lambda_2 = 1,$ $f(P(i,j)) = 1,$ $g(P(i,j)) = \begin{cases} D(i,j), & P(i) \neq P(j); \\ \rho_i \times D(i,j), & P(i) = P(j). \end{cases}$
Vlachos *et al.* [59]	$\lambda_1 = 0,$ $\lambda_2 = 1,$ $f(P(i,j)) = 1,$ $g(P(i,j)) = \begin{cases} D(i,j), & P(i) \neq P(j); \\ \alpha \times D(i,j), & P(i) = P(j). \end{cases}$
Geng *et al.* [60]	$\lambda_1 = 0,$ $\lambda_2 = 1,$ $f(P(i,j)) = 1,$ $g(P(i,j)) = \begin{cases} \sqrt{\frac{e^{D^2(i,j)}}{\beta}}, & P(i) \neq P(j); \\ \sqrt{1 - \frac{e^{-D^2(i,j)}}{\beta}}, & P(i) = P(j). \end{cases}$
Zhao *et al.* [61]	$\lambda_1 = 0,$ $\lambda_2 = 1,$ $f(P(i,j)) = 1,$ $g(P(i,j)) = \begin{cases} \infty, & P(i) \neq P(j); \\ D(i,j), & P(i) = P(j). \end{cases}$

Kemelmacher and Basri [67] present an example-based SFS method for 3D shape recovery of a face from a single image using a single 3D reference model of a different person's face. To achieve a desired reconstruction, the method seeks the shape, albedo and lighting that best fit the input image while preserving the rough shape and albedo of the reference model. This method provides accurate reconstruction of new faces. However, it makes the assumption of Lambertian reflectance and rough alignment of the input image and the reference model. Similar methods include [68], where results from only frontal face images are demonstrated.

Statistical SFS methods [69,70,71] represent face shapes in the parametric eigenspace by applying PCA to a training set of 3D faces. [69] seeks the shape-coefficients by fitting the PCA model to satisfy the image irradiance constraints, while [70] recovers the shape by fitting the PCA model to image brightness data using constraints on the surface normal direction provided by Lambert's Law. Dovgard and Basri [71] reconstruct the shape by combining the geometric constraint [33] and the statistical constraints [69]. These methods are computationally expensive in the fitting procedure for minimizing the error between the rendered facial surface and the intensity of the input face. Thus, the optimization may not converge.

3D Morphable Model (3DMM) [72] developed by Blanz and Vetter is a well-known face reconstruction method. It applies to the images and shapes separately to derive the linear models. The 3D shape reconstruction is an optimization process which aims to minimize the difference between the rendered model image and the input image. However, 3DMM suffers from the same problem as Statistical SFS methods, i.e., long runtime and multiple local minima. An approach is presented to accelerate fitting procedure of 3DMM in [73].

Some learning-based methods have been developed for the shape reconstruction. Reiter *et al.* [1] recover the 3D shape from a NIR facial image by learning the canonical correlation analysis (CCA) mapping from near infrared (NIR) facial images to 3D shape, which are both transformed to vectors. Lei *et al.* [3] present an approach (Tensor+CCA) similar to [1], while the mapping is learned from the NIR tensor space to the 3D shapes. Castelan and Hancock [2] apply coupled statistical models (CSM) to recover surfaces from brightness images of faces. However, these statistical learning approaches can handle the shape recovery only from a frontal face image. Georghiades et al. [74] develop a generative method to handle pose and illumination variations for face recognition. The change of pose is limited to be less than +/- 30 degrees, while we can deal $+/-90°$.

2.3 Motion Modeling

It is beyond the scope of this thesis to discuss all the related work in motion capture. We refer the reader to [75, 76] for extensive surveys of this broad subject.

There are two main approaches to solve motion modeling problems, categorized as learning-based approaches and model-based approaches. Here, we only provide a brief review of these approaches in motion capture.

The learning-based approaches can be further classified into two categories: *discriminative* and *generative*. Discriminative learning approaches attempt to learn a direct mapping from image observations to motion model. The learned approaches vary in the form of linear or nonlinear regression [77], mixture of Bayesian expert [78, 79], linear or nonlinear embedding [80,81], and nearest neighbor search [82,83,84]. These methods have the advantage of fast computational speed and full automation. However, it is difficult to learn the inverse, multi-modal mappings between the image space and the space of body configurations. Typically a vast amount of high quality training data is needed to achieve good reconstruction.

Generative methods search the space of body configuration that minimizes the error defined between a projection of the human body and the image observation. They provide the flexibility in representing large classes of complex human motions, but their computational cost is very expensive, due to search in the high dimensional motion space. Moreover, they can be trapped in local minima because of non-linear optimization. Particle filtering may be helpful in solving this problem [85], but it does not scale well in the space of body configurations. Linear or nonlinear dimensionality reduction (DR) techniques have been used to reduce the dimensionality of the space of body configurations. Methods such as kernel component analysis (kPCA [86]), Laplacian Eigenmaps (LE [87]), Gaussian Process Latent Variable Models (GPLVM) and its variants (e.g. [88,89,90]), are employed to learn the low-dimensional embedding. Recently, some research work has been done to combine both generative and discriminative approaches to complement each other(e.g. [91]).

In contrast, model-based approaches rely on an explicitly known parametric human model to recover the skeletal motion by searching high dimensional configuration spaces. The searching methods are typically formulated deterministically as a non-linear optimization problem [92], or probabilistically as a maximum likelihood problem [4]. The model-based approaches require known initialization and an approximate dynamical model.

Grest et al. [92] use the combination of depth and silhouette information to establish correct correspondences in the presence of non-static background and people wearing normal clothing, and track the motion of a subject from a single view by non-linear least squares. As opposed to a more global approach, the local nature of these algorithms leads to suboptimal solutions. In addition, these algorithms have the susceptibility of losing track for long input sequences.

Pekelny and Gotsman [93] present an algorithm capable of recovering a full 3D surface geometry and dynamic skeleton of a deforming object from a sequence of depth images taken by a single depth video camera. The algorithm identifies and tracks the rigid components between frames, and reconstructs an articulated 3D model in a single pass over the data. The algorithm can track the skeleton of the subject over long sequences. However, one limitation of the algorithm is the assumption that a deforming subject is piecewise-rigid.

Another recent work [4] addresses the same problem as ours. It employs a generative model with a discriminative model that identifies the body part locations by data-driven procedure. While their algorithm can achieve real-time performance, the accuracy, reported as around 100mm, leaves something to be desired. On the other hand, our approach's accuracy is around 20mm. In addition, our approach can estimate the skeletal motion in cases of severe occlusion by using the complete model from the template database.

Chapter 3 Learning 3D Shape from a Single Facial Image via Non-linear Manifold Embedding and Alignment

Many algorithms have been developed to address the problem of human face modeling. Some of them can be thought of as an extended Shape-from-Shading approach, in which the 3D shape is optimized so that its rendering matches the input image (e.g., [69, 72, 71, 70, 33, 64, 66, 65, 67, 73]). Domain-specific constraints are typically added to reduce the solution space so that meaningful results can be obtained. While some very impressive results have been obtained, one of the biggest challenges of these methods is that the optimization could be trapped in a local minimum.

Another class of methods use machine learning techniques to reconstruct the 3D shape (e.g., [2,3,1]). These learning-based methods take advantage of the availability of prior training data, i.e., face images with corresponding shapes, from which the relationship of shapes and facial images can be inferred. The reconstruction quality depends heavily on the training data sets. Given the need for high-quality 3D models and accurate data labeling, obtaining or reproducing good results is always difficult. In addition, they suffer from the curse of dimensionality problem, i.e., the requirement of a vast amount of training data to achieve accurate reconstruction. As a result, most of these methods focus solely on frontal images taken under ambient (or fixed) illuminations to reduce the amount of training data needed.

In this work, our objective aims to recover the 3D shape from a single face image by overcoming the ambiguities. We investigate non-linear statistical learning techniques to learn the texture and shape models, and impose them as priors for the 3D human face modeling problem. Section 3.1 introduces the preprocessing of the training data, images of faces and the corresponding shapes. Section 3.2 describes the fundamentals of the Gaussian Process Latent Variable Model (GP-LVM) and its application for learning the local image models and the local surface models. The learning of correspondences between these models using a manifold alignment method is detailed in section 3.3. We discuss the problem of head pose estimation in Section 3.4. Section 3.5 presents the reconstruction procedure of the global shape by combining the learned local surface shapes. The experimental results and analytic analysis are shown in section 3.6.

3.1 Training Data Preprocessing

2D Image Preprocessing All the training facial images are first automatically aligned to the reference images I_i^r using the method in [94]. The index i denotes the pose variation, and r means that the image is considered as the reference image. We use different reference images for different poses. Estimating a 3D shape from a facial image with M pixels can be viewed as a generic non-linear M-dimensional regression problem. Even for small images, the number of dimensions is still too large. To overcome this dimensionality issue, We have developed local, low-dimensional estimation based on small image patches. For each facial image of a specific pose, we divide it into N_z overlapping $p \times q$ rectangular patches. This patch representation not only reduces the problem dimension, but also makes illumination correction easier. Instead of applying global and complex methods (such as [95]), we simply use local image normalization to correct non-uniform illumination or shading artifacts for each patch by:

$$J(x,y) = \frac{I(x,y) - m_I(x,y)}{\sigma_I(x,y)} \tag{3.1}$$

where $I(x,y)$ is the original image patch, $m_I(x,y)$ and $\sigma_I(x,y)$ are, respectively, the mean and the variance of $I(x,y)$, and $J(x,y)$ is the output image patch. After that histogram equalization is performed on $J(x,y)$. Figure 3.1 demonstrates the effectiveness of this approach. The corrected patches show little effect of lighting. Applying the same approach to an entire image is unlikely to be effective.

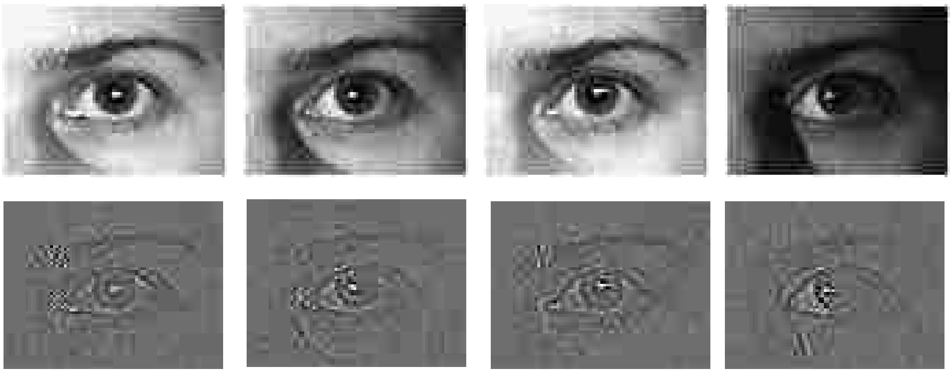


Figure 3.1: Local image patches before and after illumination correction.

As shown in Figure 3.2, after image subdivision and normalization, we construct the data $\{\mathbf{Y}_{i,j} = [\mathbf{y}_{i,j,1}, \cdots, \mathbf{y}_{i,j,k}, \cdots, \mathbf{y}_{i,j,N}],\ j = 1 \cdots N_z\}$, where $\mathbf{y}_{i,j,k}$ is the transformed column vector from the facial image region with pose i and patch index j of the k^{th} person, and N is the number of subjects.

3D Shape Preprocessing In the 3D shape preprocessing, we adopt the coherent point drift (CPD) algorithm. CPD is a probabilistic method for non-rigid registration of point sets; details can be found in [96]. For the sake of completeness we briefly introduce the CPD algorithm here.

CPD considers the alignment of two point sets as a probability density estimation problem. The first point set $X_{N\times D} = (x_1, \dots, x_N)$ represents the data points, while the second point set $Y_{M\times D} = (\mathbf{y}_1, \dots, \mathbf{y}_M)$ represents the Gaussian Mixture Model (GMM) centroids. CPD fits Y to X by maximizing the likelihood, or equivalently by minimizing the negative log-likelihood function as follows,

$$E(\theta, \sigma^2) = -\sum_{n=1}^{N} \log \sum_{m=1}^{M} P(m)p(\mathbf{x}|m) \tag{3.2}$$

where the GMM centroid locations are re-parameterized with a set of parameters θ; $p(\mathbf{x}|m) = \frac{1}{(2\pi\sigma^2)^{D/2}} e^{-\frac{1}{2}\|\frac{\mathbf{x}-\mathbf{y}_m}{\sigma}\|^2}$. We make the i.i.d. data assumption, and use equal isotropic covariances σ^2 and equal membership probabilities $P(m) = \frac{1}{M}$ for all GMM components ($m = 1, \dots, M$). Furthermore, the correspondence probability between two points $\mathbf{y}_m$ and $\mathbf{x}_n$ is considered as the posterior probability of the GMM centroid given the data point:$P(m|\mathbf{x}_n) = P(m)p(\mathbf{x}_n|m)/p(\mathbf{x}_n)$.

The Expectation Maximization (EM) algorithm [97] is used to find θ and σ. In the E-step, the posterior probability distribution $P^{old}(\mathbf{x}_n|m)$ of mixture components is computed using the "old" parameter values via Eq. 3.3,

$$P^{old}(\mathbf{x}_n|m) = \frac{e^{-\frac{1}{2}\|\frac{\mathbf{x}_n - \Gamma(\mathbf{y}_m, \theta^{old})}{\sigma^{old}}\|^2}}{\sum_{k=1}^{M} e^{-\frac{1}{2}\|\frac{\mathbf{x}_n - \Gamma(\mathbf{y}_m, \theta^{old})}{\sigma^{old}}\|^2}} \tag{3.3}$$

where $\Gamma(\mathbf{y}, \theta)$ is the transformation Γ applied to $\mathbf{y}$, and θ is a set of the transformation parameters.

In the M-step, the "new" parameter values are found by minimizing the expectation of the following complete negative log-likelihood function with respect to the "new" parameters, which is an upper bound of Eq. 3.2,

$$\begin{aligned} Q(\theta,\sigma^2) &= \tfrac{1}{2\sigma^2}\textstyle\sum_{n=1}^{N}\sum_{m=1}^{M} P^{old}(m|\mathbf{x}_n)\|\mathbf{x}_n-\Gamma(\mathbf{y}_m,\theta)\|^2 \\ &\qquad\qquad + \frac{N_{\mathbf{P}}D}{2}\log\sigma^2 \end{aligned} \tag{3.4}$$

where $N_{\mathbf{P}} = \Sigma_{n=1}^{N}\Sigma_{m=1}^{M}P^{old}(m|\mathbf{x}_n)$. The transformation Γ may have affine, rigid, and non-rigid forms.

Affine Registration For Affine point set registration, the transformation of the GMM centroid locations can be defined as $\Gamma(\mathbf{y}_m;\mathbf{B},\mathbf{t}) = \mathbf{B}\mathbf{y}_m+\mathbf{t}$, where $\mathbf{B}_{D\times D}$ is an affine matrix, and $\mathbf{t}_{D\times 1}$ is a translation vector. The objective function in Eq. 3.4 can be written as,

$$\begin{aligned} Q(\mathbf{B},\mathbf{t},\sigma^2) &= \tfrac{1}{2\sigma^2}\textstyle\sum_{m,n=1}^{M,N} P^{old}(m|\mathbf{x}_n)\|\mathbf{x}_n-(\mathbf{B}\mathbf{y}_m+\mathbf{t})\|^2 \\ &\qquad\qquad + \frac{N_{\mathbf{P}}D}{2}\log\sigma^2 \end{aligned} \tag{3.5}$$

Rigid Registration Rigid transformation is defined as $\Gamma(\mathbf{y}_m;\mathbf{R},\mathbf{t},s) = s\mathbf{R}\mathbf{y}_m+\mathbf{t}$, where $\mathbf{R}_{D\times D}$ is a rotation matrix, $\mathbf{t}_{D\times 1}$ is a translation vector, and s is a scaling parameter. Compared to the affine case, the rigid registration case is more complicated due to the constraints on $\mathbf{R}$. The work [98] discusses the closed form solution for the rotation matrix $\mathbf{R}$.

$$\begin{aligned} Q(\mathbf{R},\mathbf{t},s,\sigma^2) &= \tfrac{1}{2\sigma^2}\textstyle\sum_{m,n=1}^{M,N} P^{old}(m|\mathbf{x}_n)\|\mathbf{x}_n-(s\mathbf{R}\mathbf{y}_m+\mathbf{t})\|^2 \\ &\qquad + \frac{N_{\mathbf{P}}D}{2}\log\sigma^2, s.t.\mathbf{R}^T\mathbf{R}=\mathbf{I}, \det(\mathbf{R})=1 \end{aligned} \tag{3.6}$$

Non-rigid Registration The non-rigid registration is an ill-posed problem, because there is a broad class of transformations that align the two point sets. To deal with the problem, CPD defines non-rigid transform as the initial position plus a displacement function v, $\Gamma(\mathbf{y}_m; v) = \mathbf{y}_m + v(\mathbf{y}_m)$, and adds a regularization term to the negative log-likelihood function in Eq. 3.2,

$$E(v, \sigma^2) = -\sum_{n=1}^{N} \log \sum_{m=1}^{M} e^{-\frac{1}{2}\|\frac{\mathbf{x}_n - \Gamma(\mathbf{y}_m; v)}{\sigma}\|^2} + \frac{\lambda}{2}\phi(v) \tag{3.7}$$

where ϕ function represents the prior knowledge about the motion, which should be smooth. We define the regularization term as $\phi = \int_{R^d} |\tilde{v}(s)|^2 / \tilde{G}(s) ds$, where $\tilde{v}$ denotes the Fourier transform of the velocity and $\tilde{G}$ represents a symmetric low-pass filter. The displacement function v has the form of the radial basis function in Eq. 3.9, which can be estimated by minimizing the energy function in Eq. 3.7 using a variational calculus. Here, G is a Gaussian kernel. Therefore, we can write the upper bound of the function in Eq. 3.4 for non-rigid registration as,

$$\begin{aligned} Q(\mathbf{W}) &= \textstyle\sum_{n=1}^{N} \sum_{m=1}^{M} P^{old}(m|\mathbf{x}_n) \frac{\|\mathbf{x}_n - \mathbf{y}_m - \mathbf{G}(m,\cdot)\mathbf{W}\|^2}{\sigma^2} \\ &\quad + \tfrac{\lambda}{2} tr(\mathbf{W}^T \mathbf{G} \mathbf{W}) \end{aligned} \tag{3.8}$$

where $\mathbf{G}_{M \times M}(i,j) = e^{-\frac{1}{2\beta^2}\|\mathbf{y}_i - \mathbf{y}_j\|^2}$ is a square symmetric Gram matrix, and $\mathbf{W}_{M \times D} = (\mathbf{w}_1, \ldots, \mathbf{w}_M)^T$ is an unknown matrix of the Gaussian kernel weights in the displacement function v; $\mathbf{G}(m, \cdot)$ denotes the m^{th} row of $\mathbf{G}$.

$$v(\mathbf{z}) = \sum_{m=1}^{M} \mathbf{w}_m G(\mathbf{z} - \mathbf{y}_m) \tag{3.9}$$

Substituting Eq. 3.9 back into Eq. 3.7, we can rewrite the objective function in the new form as follows,

$$\begin{aligned} E(\mathbf{W}) &= -\sum_{n=1}^{N} \log \sum_{m=1}^{M} e^{-\frac{1}{2}\|\frac{\mathbf{x}_n-\mathbf{y}_m-\sum_{k=1}^{M}\mathbf{w}_m G(\mathbf{y}_k-\mathbf{y}_m)}{\sigma}\|^2} \\ &\quad +\frac{\lambda}{2} tr(\mathbf{W}^T\mathbf{G}\mathbf{W}) \end{aligned} \tag{3.10}$$

where $\mathbf{G}_{M\times M}(i,j) = e^{-\frac{1}{2\beta^2}\|\mathbf{y}_i-\mathbf{y}_j\|^2}$ is a square symmetric Gram matrix, and $\mathbf{W}_{M\times D} = (\mathbf{w}_1, \ldots, \mathbf{w}_M)^T$ is an unknown matrix of the Gaussian kernel weights in Eq. 3.9.

We select one 3D facial shape $\mathbf{M}_r$ as a reference model, and every other facial model $\mathbf{M}_h$, $h = 1 \ldots N$, is registered to $\mathbf{M}_r$ using CPD. After the registration, each facial shape has the same number of vertices and triangles (in our experiments, 2500 vertices and 4624 triangles for each facial shape), which provides us convenience for later processing. Then, we parameterize the 3D shape model with deformation transfer, which describes the shape transformation from the source ($\mathbf{M}_r$) to the target ($\mathbf{M}_h$) [99]. The source deformation is represented as a collection of affine transformations tabulated for each triangle of $\mathbf{M}_r$, e.g., $\mathbf{T}_h = [\mathbf{q}_1, \cdots, \mathbf{q}_m]^T$, where m is the number of triangles, and $\mathbf{q}_i$ denotes the affine transformation of the i^{th} triangle. However, the affine transformation cannot fully be determined with the three vertices of each triangle. A fourth vertex is added in the direction perpendicular to the triangle to resolve this issue [99]. It is computed as follows,

$$\begin{aligned} \mathbf{v}_4 &= \mathbf{v}_1 + (\mathbf{v}_2 - \mathbf{v}_1) \times (\mathbf{v}_3 - \mathbf{v}_1)/\sqrt{|\mathbf{v}_2 - \mathbf{v}_1| \times |\mathbf{v}_3 - \mathbf{v}_1|} \\ \tilde{\mathbf{v}}_4 &= \tilde{\mathbf{v}}_1 + (\tilde{\mathbf{v}}_2 - \tilde{\mathbf{v}}_1) \times (\tilde{\mathbf{v}}_3 - \tilde{\mathbf{v}}_1)/\sqrt{|\tilde{\mathbf{v}}_2 - \tilde{\mathbf{v}}_1| \times |\tilde{\mathbf{v}}_3 - \tilde{\mathbf{v}}_1|} \end{aligned} \tag{3.11}$$

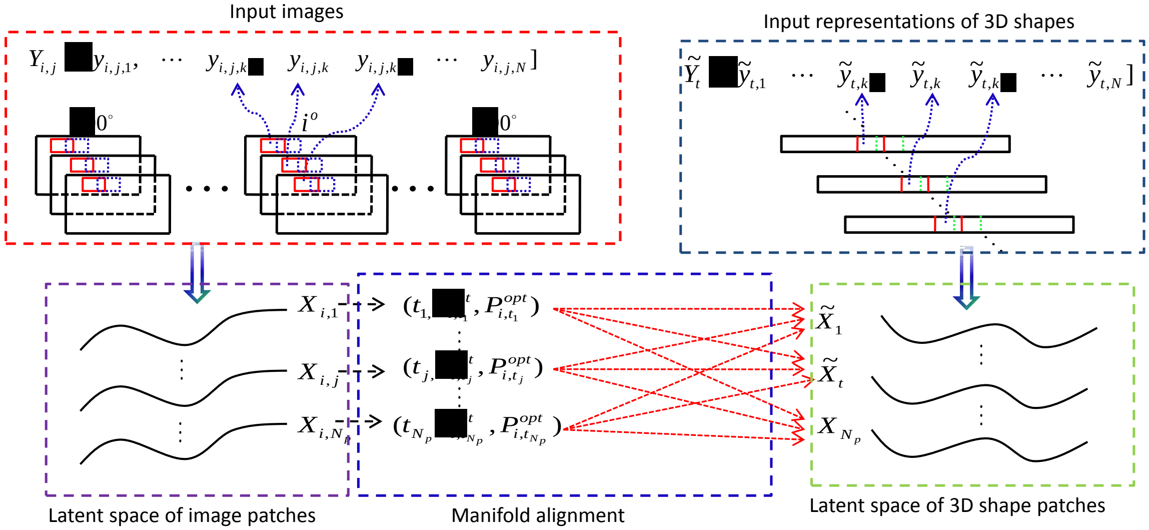

Input images
$Y_{i,j}$ $y_{i,j,1}$, ⋯ $y_{i,j,k}$ $y_{i,j,k}$ $y_{i,j,k}$ ⋯ $y_{i,j,N}$]
0°
i^o
0°
Input representations of 3D shapes
$\tilde{Y}_t$ $\tilde{y}_{t,1}$ ⋯ $\tilde{y}_{t,k}$ $\tilde{y}_{t,k}$ $\tilde{y}_{t,k}$ ⋯ $\tilde{y}_{t,N}$]
$X_{i,1}$
$X_{i,j}$
X_{i,N_p}
$(t_1, \; , P^{opt}_{i,t_1})$
$(t_j, \; , P^{opt}_{i,t_j})$
$(t_{N_p}, \; , P^{opt}_{i,t_{N_p}})$
$\tilde{X}_1$
$\tilde{X}_t$
X_{N_p}
Latent space of image patches
Manifold alignment
Latent space of 3D shape patches

Figure 3.2: Data flow chart of our algorithm. $\mathbf{Y}_{i,j}$ is constructed from the image regions of all training faces with pose i and patch index j, while $\widetilde{\mathbf{Y}}_j$ is from the representations of all 3D shapes with patch index j. $\mathbf{Y}_{i,j}$ and $\widetilde{\mathbf{Y}}_j$ are projected into the low dimensional space using GP-LVM and generate $\mathbf{X}_{i,j}$ and $\widetilde{\mathbf{X}}_j$. For each $\mathbf{X}_{i,j}$, its correspondence, $\widetilde{\mathbf{X}}_{t_j}$, is found as the one with the minimal alignment error by the manifold alignment algorithm.

where $\mathbf{v}_i$ and $\tilde{\mathbf{v}}_i, i \in 1 \dots 3$ are the original and deformed vertices of the triangle. The 3×3 matrix $\mathbf{q}$ can be computed by $\mathbf{q} = \tilde{\mathbf{V}}\mathbf{V}^{-1}$ where

$$\begin{aligned} \mathbf{V} &= [\mathbf{v}_2 - \mathbf{v}_1 \mathbf{v}_3 - \mathbf{v}_1 \mathbf{v}_4 - \mathbf{v}_1] \\ \tilde{\mathbf{V}} &= [\tilde{\mathbf{v}}_2 - \tilde{\mathbf{v}}_1 \tilde{\mathbf{v}}_3 - \tilde{\mathbf{v}}_1 \tilde{\mathbf{v}}_4 - \tilde{\mathbf{v}}_1] \end{aligned} \tag{3.12}$$

We also decompose $\mathbf{T}_h$ into N_z overlapping parts. With this, we construct the representations of 3D shapes $\widetilde{\mathbf{Y}}_j = [\widetilde{\mathbf{y}}_{j,1}, \cdots, \widetilde{\mathbf{y}}_{j,k}, \cdots, \widetilde{\mathbf{y}}_{j,N}]$, as shown in Figure 3.2, where $\widetilde{\mathbf{y}}_{j,k}$ is from the j^{th} patch of $\mathbf{T}_k$, corresponding to the facial image patches with patch index j of the k^{th} person.

3.2 The Local Image and Surface Models

In the previous sections, we explain how we gathered data as patches of facial images and 3D shapes. We now show how to learn the local image and surface models from such data. Generally, it is difficult to work with the data in the original high-dimensional space, since the number of training examples needed to fully cover the space of possible deformations grows exponentially with the number of dimensions. A large amount of research work on non-linear manifold embedding has been done to handle the *curse of dimensionality.* We adopt the Gaussian Process Latent Variable Model (GP-LVM) [100], which provides a good generalization from very small data sets using nonlinear models. An important characteristic of the GP-LVM is the reconstruction of a new point in the latent space with ease and accuracy. GP-LVM represents a Gaussian process (GP) mapping from the latent space $\mathbf{X}$ (low-dimensional embedding) to the data space $\mathbf{Y}$ (high-dimensional data set), where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]^T \in \Re^{N \times d}$ is the non-linear embedding matrix whose rows represent the corresponding positions in the latent space, $\mathbf{x}_i \in \Re^d$, and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N]^T \in \Re^{N \times D}$ is the data matrix in which each row is a single training sample, $\mathbf{y}_i \in \Re^D$. For a detailed discussion on GP and GP-LVM, see [100, 101]. Given a kernel function for the GP, $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)$, the likelihood of the data given the latent positions is

$$p(\mathbf{Y}|\mathbf{X}, \Theta) = \frac{1}{\sqrt{(2\pi)^{ND}|\mathbf{K}|^D}} \exp(-\frac{1}{2} tr(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^T)) \tag{3.13}$$

where $\mathbf{K}$ denotes the kernel matrix whose elements are defined by the kernel function $(\mathbf{K})_{i,j} = \mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)$, and Θ is the kernel hyper-parameters. In our experiments we use the form of the radial basis function (RBF) kernel, which controls the output variance, the RBF support width, the bias and the variance of the additive noise. GP-LVM learning consists of maximizing the posterior $p(\mathbf{X}, \Theta|\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{X}, \Theta)p(\mathbf{X})p(\Theta)$ with respect to the latent space $\mathbf{X}$, and the hyper-parameters Θ.

To reduce the computational complexity from an often prohibitive $O(N^3)$ to $O(Nk^2)$, where k is the number of points specified by the user, sparse approximation techniques were proposed [102] and were proven more accurate than simply using a subset of the data. All approximations involve augmenting the function values at the training points, $\mathbf{F} \in \Re^{N\times d}$, with $\mathbf{F} = [\mathbf{f}_1, \cdots, \mathbf{f}_N]^T$ and the function values at the test points, $\mathbf{F}_* \in \Re^{\infty\times d}$, by an additional set of variables, $\mathbf{X_u} \in \Re^{k\times d}$, called inducing variables. Learning the sparse GP-LVM involves maximizing with respect to $\mathbf{X}$, $\mathbf{X_u}$ and Θ the posterior

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{X_u}, \Theta)) = N(\mathbf{K_{f,u}}\mathbf{K}^{-1}_{\mathbf{u,u}}\mathbf{X_u}, \Lambda + \sigma^2\mathbf{I}) \tag{3.14}$$

where $\Lambda = diag[\mathbf{K_{f,f}} - \mathbf{K_{f,u}}\mathbf{K}^{-1}_{\mathbf{u,u}}\mathbf{K_{f,f}}\mathbf{K_{u,f}}]$ and $diag(A)$ is a diagonal matrix whose elements match the diagonal of A, $\mathbf{K_{f,u}}$ denotes the covariance function computed between $\mathbf{X}$ and $\mathbf{X_u}$, $\mathbf{K_{u,u}}$ is the kernel matrix for the elements of $\mathbf{X_u}$, $\mathbf{K_{f,f}}$ is the symmetric covariance between $\mathbf{X}$, and σ^2 is the noise variance.

Given a new test point $\mathbf{x}_*$, the predictive distribution of its high-dimensional position $\mathbf{y}_*$ can be obtained [100] by

$$p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{Y}, \mathbf{X_u}, \Theta) = N(\mu_*, \sigma^2_*) \tag{3.15}$$

where the mean and variance are

$$\mu_* = \mathbf{Y}^T\mathbf{K}^{-1}_{\mathbf{u,u}}\mathbf{K}_* \tag{3.16}$$

$$\sigma_*^2 = \mathbf{K}_{**} - \mathbf{K}_*^T \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{K}_* \tag{3.17}$$

where $\mathbf{K}_*$ is a vector with elements $\mathbf{K}(\mathbf{x}_*, \mathbf{x}_i)$ for latent positions $\mathbf{x}_i \in \mathbf{X_u}$, and $\mathbf{K}_{**} = \mathbf{K}(\mathbf{x}_*, \mathbf{x}_*)$.

Given a new test point $\mathbf{y}_*$, its latent position can be inferred in the sparse GP-LVM by minimizing $-\ln p(\mathbf{y}_*, \mathbf{x}_* | \mathbf{Y}, \mathbf{X_u}, \Theta)$, up to an additive constant [8],

$$\ell(\mathbf{x}_*, \mathbf{y}_*) = \frac{\|\mathbf{y}_* - \mu(\mathbf{x}_*)\|^2}{2\sigma^2(\mathbf{x}_*)} + \frac{D}{2} \ln \sigma^2(\mathbf{x}_*) + \frac{1}{2}\|\mathbf{x}_*\|^2 \tag{3.18}$$

with the mean and variance given by

$$\mu(\mathbf{x}_*) = \mathbf{Y}^T \mathbf{K}_{\mathbf{f},\mathbf{u}}^T \mathbf{A}^{-1} \mathbf{K}_* \tag{3.19}$$

$$\sigma^2(\mathbf{x}_*) = \mathbf{K}_{**} - \mathbf{K}_*^T (\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} - \sigma^2 \mathbf{A}^{-1}) \mathbf{K}_* \tag{3.20}$$

where $\mathbf{A} = \sigma^2 \mathbf{K}_{\mathbf{u},\mathbf{u}} + \mathbf{K}_{\mathbf{u},\mathbf{f}} \mathbf{K}_{\mathbf{f},\mathbf{u}}$.

The local image model $\Theta_{i,j}^I$ and the low dimensional embedding $\mathbf{X}_{i,j} = [\mathbf{x}_{i,j,1}, \cdots, \mathbf{x}_{i,j,N}]$ are learned by the input image patches $\mathbf{Y}_{i,j}$, where $j = 1 \cdots N_z$. Similarly, we can get the local surface model Θ_t^S and the low-dimensional embedding $\widetilde{\mathbf{X}}_t = [\widetilde{\mathbf{x}}_{t,1}, \cdots, \widetilde{\mathbf{x}}_{t,N}]$ from the input 3D patches $\widetilde{\mathbf{Y}}_t$.

3.3 Learning the Correspondences

Previous methods of single-image 3D face modeling usually require explicit registration between the 2D images and the 3D models. Registration between different modality is a difficult problem. Typically this is done with user interaction. However, given our intention of dealing with both head pose and illumination variations, manually labeling all the images in both spaces is too time-consuming. Rather We have developed an automatic procedure to estimate the correspondences via manifold alignment with procrustes analysis [103].

More specifically, we have two collections of low-dimensional embeddings, 2D image patches $\{\mathbf{X}_{i,j}\}$ and 3D shape patches $\{\widetilde{\mathbf{X}}_t\}$. We estimate a transformation (i.e. procrustes analysis) to best align one data configuration ($\mathbf{X}_{i,j}$) to another ($\widetilde{\mathbf{X}}_t$). Each element of $\mathbf{X}_{i,j}$ and $\widetilde{\mathbf{X}}_t$ is first translated so that its centroid is at the origin, by

$$\begin{aligned} \mathbf{x}_{i,j,k} &= \mathbf{x}_{i,j,k} - \textstyle\sum_{k=1}^{N} \mathbf{x}_{i,j,k}/N, \quad j = 1 \cdots N_z \\ \widetilde{\mathbf{x}}_t &= \widetilde{\mathbf{x}}_{t,k} - \textstyle\sum_{k=1}^{N} \widetilde{\mathbf{x}}_{t,k}/N, \quad t = 1 \cdots N_z \end{aligned} \tag{3.21}$$

Then, we try to align $\mathbf{X}_{i,j}$ to all $\widetilde{\mathbf{X}}_t$. The alignment error of matching $\mathbf{X}_{i,j}$ and $\widetilde{\mathbf{X}}_t$ is defined by $\|\mathbf{X}_{i,j} - \lambda_{i,t}\widetilde{\mathbf{X}}_t\mathbf{P}_{i,t}\|_F$, where $\|\cdot\|_F$ denotes Frobenius norm, $\lambda_{i,t}$ is a re-scaling factor to either stretch or shrink $\widetilde{\mathbf{X}}_{i,t}$, and $\mathbf{P}_{i,t}$ is an orthonormal matrix, defining a rotation and possibly a reflection. We denote the correspondence of $\mathbf{X}_{i,j}$ as $\widetilde{\mathbf{X}}_{t_j}$ with patch index t_j, which has the minimal alignment error with $\mathbf{X}_{i,j}$. That is, the problem is simplified to find the patch index t_j of 3D shape representations, λ_{i,t_j}^{opt} and the transform $\mathbf{P}_{i,t_j}^{opt}$ such that

$$\{t_j, \lambda_{i,t_j}^{opt}, \mathbf{P}_{i,t_j}^{opt}\} = \arg \min_{t \in \{1 \cdots N_z\}, \lambda_{i,t}, \mathbf{P}_{i,t}} \|\mathbf{X}_{i,j} - \lambda_{i,t}\widetilde{\mathbf{X}}_t\mathbf{P}_{i,t}\|_F \tag{3.22}$$

For simplicity, we use $\mathbf{X}$, λ, $\mathbf{Y}$, and $\mathbf{Q}$ to represent $\mathbf{X}_{i,j}$, $\lambda_{i,t}$, $\widetilde{\mathbf{X}}_t$, and $\mathbf{P}_{i,t}$ respectively. Continuing on, the problem is formalized as:

$$\{\lambda_{opt}, \mathbf{Q}_{opt}\} = \arg \min_{\lambda, \mathbf{Q}} \|\mathbf{X} - \lambda\mathbf{Y}\mathbf{Q}\|_F \tag{3.23}$$

It can be written as,

$$\|\mathbf{X} - \lambda\mathbf{Y}\mathbf{Q}\|_F = trace(\mathbf{X}^T\mathbf{X}) + \lambda^2 \cdot trace(\mathbf{Y}^T\mathbf{Y}) - 2\lambda \cdot trace(\mathbf{Q}^T\mathbf{Y}^T\mathbf{X}) \tag{3.24}$$

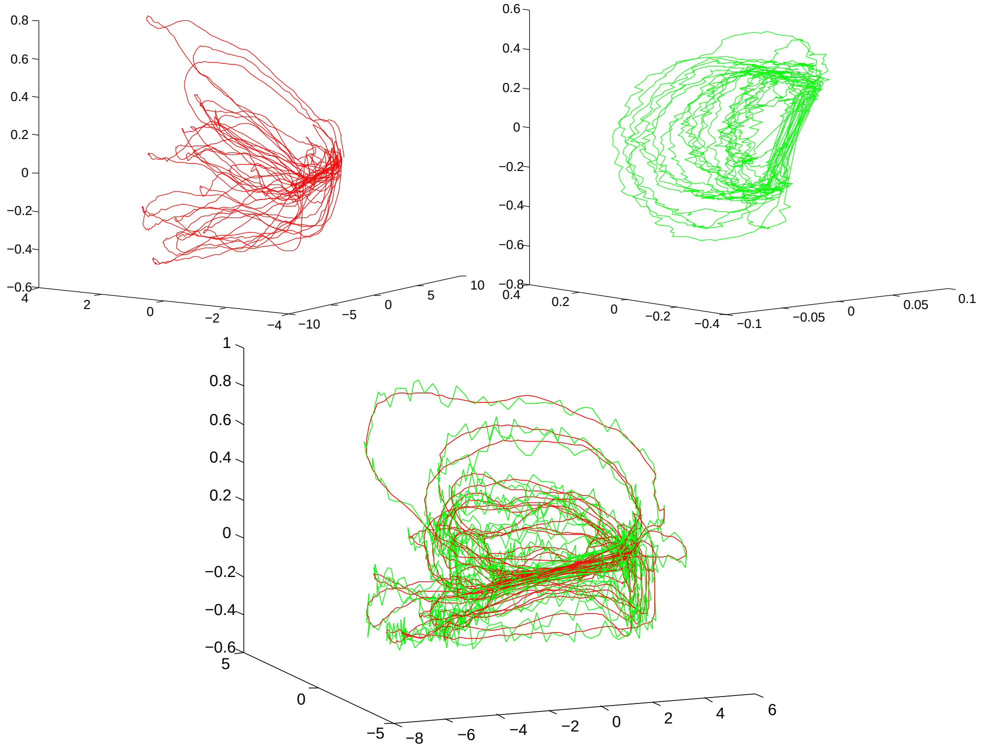


Figure 3.3: (Top) An example of low-dimensional embeddings of 2D image patches and 3D shape patches from different subjects with pose changes; They are in different coordinate systems. (Bottom) The two embeddings after alignment.

We can have $\lambda = trace(\mathbf{Q}^T\mathbf{Y}^T\mathbf{X})/trace(\mathbf{Y}^T\mathbf{Y})$ by differentiating with respect to λ. Thus, $\lambda^{opt}_{i,t_j} = trace(\Sigma)/trace(\widetilde{\mathbf{X}}^T_{t_j}\widetilde{\mathbf{X}}_{t_j})$. From 3.23 and 3.24, the minimization problem reduces to

$$\mathbf{Q}_{opt} = \arg\max_{\mathbf{Q}} trace(\mathbf{Q}^T\mathbf{Y}^T\mathbf{X})^2 \tag{3.25}$$

It is shown that $\mathbf{P}^{opt}_{i,t_j} = \mathbf{U}\mathbf{V}^T$ in [103], where $\mathbf{U}$, $\mathbf{V}$ and Σ are given by the SVD of $\widetilde{\mathbf{X}}^T_{t_j}\widetilde{\mathbf{X}}_{t_j}$, that is, $\mathbf{U}\Sigma\mathbf{V}^T = SVD(\widetilde{\mathbf{X}}^T_{t_j}\widetilde{\mathbf{X}}_{t_j})$.

Our method is based on the assumption that corresponding 2D and 3D embeddings have similar shape, yielding the minimum amount of registration error. Figure 3.3 shows that this assumption is likely to be valid.

Give a new point $\mathbf{x}_{i,j,*}$ in the embedding space of $\mathbf{X}_{i,j}$, the point $\widetilde{\mathbf{x}}_{t_j,*}$ corresponding to $\mathbf{x}_{i,j,*}$ can be computed by

$$\mathbf{x}_{i,j,*} = \lambda_{i,t_j}^{opt} \widetilde{\mathbf{x}}_{t_j,*} \mathbf{P}_{i,t_j}^{opt} \tag{3.26}$$

3.4 Head Pose Estimation

Human face analysis, due to its many applications from biometric authentication to human-computer interactions, is a very active topic in computer vision research. Head pose estimation is a central component for many of these applications. For example, face recognition systems require the capability of handling significant pose variations. Zhao *et al.* [104] shows that pose and illumination are the major factors affecting the performance of face recognition algorithms. The difference between two individuals' face images taken under the same lighting conditions is smaller than the difference between two face images of the same individual taken under varying lighting conditions. That is, image variation due to lighting changes is more significant than variation due to different personal identities [105]. While person-independent head pose estimation has been studied reasonably well in recent years [56, 62, 55], robust illumination-independent and person-independent head pose estimation remains a challenging problem. Here, we introduce a new approach to address this problem.

Generally, most human face analysis techniques use one of two main different approaches to circumvent the problem of illumination variation. The first approach uses pre-processing techniques, e.g., histogram equalization and Laplacian of Gaussian (LoG) transformed image [56], to modify the input image to a more suitable representation for pose estimation. However, the pre-processing cannot cope with illumination variation completely. As shown in [105], classical image representations such as edge maps as well as the image filtered with 2D Gabor-like functions are not sufficient for recognition purposes under a wide variety of lighting conditions. Our experiment demonstrates LoG representation cannot deal with pose estimation with varying illumination. On the other hand, the second approach tries to model objects of interest under all possible lighting conditions in the classification procedure. Our approach may be categorized as the second approach.

Related Work

There exist many methods for pose estimation from a single image (we exclude pose estimation from a video sequence in this context). These methods can be classified into five categories [62, 55]: shape-based geometric analysis methods [106], appearance-based methods [107,108], model-based methods [31], template-based methods [109,110], and dimensionality reduction based methods. In shape-based geometric analysis, the head pose is estimated through the geometric information that is defined by the configuration of facial landmarks or features. The main problem for this method is how to define geometric parameters for profile views of a face image. The pose estimation problem in the appearance-based methods may be viewed as a pattern classification problem. In addition to the weak generalization problem, most of these methods suffer from the problem of head pose estimation in a limited view range. In the model-based methods, the input image is fitted with the face model and a classifier such as a neural network is used to estimate the pose. The template-based method is based on nearest neighbor classification against texture templates. It is beyond the scope of this work to discuss all the related work in pose estimation. Instead, we will focus on the fifth category: dimensionality reduction (DR) techniques, which is the category our method belongs to.

Face images contain redundant information induced by pose, illumination, expression, occlusion, etc. Dimensionality reduction techniques are widely used to remove such data redundancy and find more compact feature representations. They are divided into two categories: linear dimensionality reduction and non-linear dimensionality reduction. The classical linear algorithms include Principle Component Analysis (PCA) and Multidimensional Scaling (MDS). PCA finds the subspace that best preserves the variance of the data, while MDS learns an explicit linear projective mapping that best preserves the inter-point distance. Linear methods cannot always reveal the intrinsic distribution of a given complex data set. ISOMAP [48], Locally Linear Embedding (LLE) [49], and Laplacian Eigenmaps (LE) [50] are the non-linear manifold learning techniques most often used in the last few years. More specifically, we denote $X = (x_1, \cdots, x_N) \in \Re^{D \times N}$ as the data matrix containing N data points in the original feature space, $Y = (y_1, \cdots, y_N) \in \Re^{d \times N}$ as the nonlinear embedding matrix, generally $d \leq D$, and W as the weight matrix. LLE preserves the distance based on locally linear combination of neighborhood. It computes the nonlinear embedding by minimizing the cost function $\phi(Y) = \sum_i \|y_i - \sum_{i=1}^{K} W_{ij} y_{ij}\| = \|Y^T M Y\|^2$, where $M = (I - W)^T (I - W)$, and $W = argmin \sum_i \|x_i - \sum_j w_{ij} x_j\|^2$. LLE preserves the distance described by a weighted connected graph constructed from neighborhood. It constructs a nonlinear mapping by solving the eigen problem $Ly = \lambda D y$, where D is a diagonal matrix whose entries are column sums of the weight matrix W, and

$L = D - W$ is the Laplacian matrix. ISOMAP preserves the geodesic inter-point distances. It first sets up neighbor relations for each point on the manifold and the neighbor relations are represented by a weighted graph G over the data points. The edge weight between x_i and x_j is assigned with the Euclidean distance $d_x(i, j)$. The pairwise geodesic distances $d_M(i, j)$ on the manifold are then estimated with the distance of the shortest path in the graph using Floyd's or Dijkstra's algorithm [111]. The low-dimension embedding is finally constructed by applying classical MDS to the geodesic distance matrix. These nonlinear dimensionality reduction techniques and their extensions have been applied to the head pose estimation problem [56]. Every DR approach is essentially a method to learn a distance metric that removes the data redundancy and leads to more compact feature representation. However, the problem of illumination variation is usually conceded or treated lightly as a pre-processing step. As we show in Figure 3.4, we aim to deal with very harsh illumination conditions. The lack of a DR-based pose estimation method under these difficult conditions is probably due to the fact that illumination changes are usually larger than that from different persons or small pose variations.

We present a novel approach to estimate the pose from a single input image by *combining both unsupervised metric learning techniques and supervised metric learning techniques.* Unlike previous DR-based pose estimation methods that treat illumination as a part of the pre-processing step, We have developed a *unified* framework that does not require any pre-processing for illumination normalization or correction. This is possible by applying a *learned* distance transformation after the use of nonlinear DR techniques. This is different from previous approaches in which the original

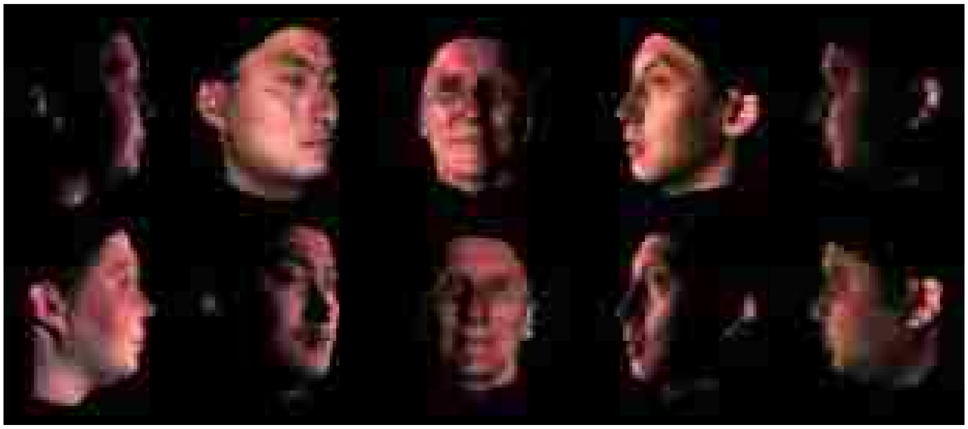


Figure 3.4: Sample input images for pose estimation. Notice the large variation in rotation and the harsh un-even illumination.

images are modified or filtered before applying nonlinear DR techniques. To the best of our knowledge this is the first time a DR-based method is capable of producing fairly accurate pose estimation (within a few degrees) under harsh illumination conditions as shown in Figure 3.4.

Our Approach of Pose Estimation

In this section, we present our approach for pose estimation. We assume that there is a training face database, and each face image in the database is associated with a pose label. Our goal is to estimate the unknown pose label from an input face image that is not in the training database.

A good low-dimensional embedding for pose estimation should have the following properties: (1) Separation. The embedding from different poses are kept apart, and there is no overlap among them. Furthermore, the embeddings of different individuals with different illuminations but with the same pose should be close to each other, i.e., within a cluster. (2) Smoothness. The low-dimensional manifold should change

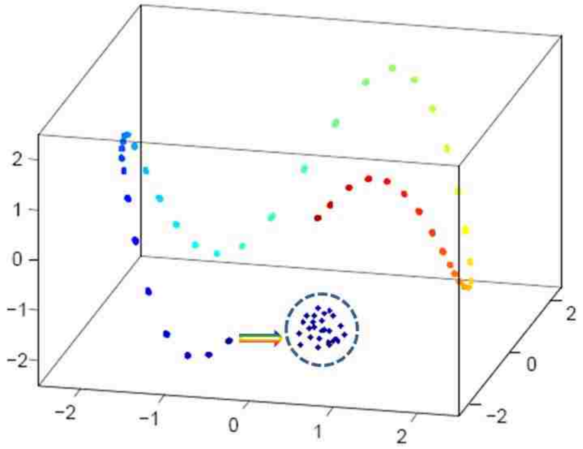


Figure 3.5: The ideal 3-dimensional embedding of 24 subjects' face images with only pose variation between [−90° + 90°] at 4° increments. The pose changes are represented by different colors.

smoothly according to the pose. Figure 3.5 shows an ideal 3-dimensional manifold embedding of 24 subjects with the same illumination while the pose angles vary from −90° to +90° with a granularity of 4° from our training database. In this figure, there are 46 clutters in total, each with a unique color corresponding to a specific pose angle. Within each clutter, there are 24 data points, which are the embeddings of 24 faces from the same pose.

The 3D embedding of ISOMAP and BME with ISOMAP are shown in Figure 3.6 and Figure 3.7. Figure 3.6 shows the ISOMAP embedding, in which 200 nearest neighbors (NN) are used. It maps face images of 24 subjects into 24 different pose manifolds. This is because ISOMAP cannot find the nearest neighbors of each point accurately when there are multiple individuals in the training set. To deal with identity variations, BME finds the right nearest neighbors for each data point by the given pose labels. However, when there are illumination variations, especially large illumination variations, BME cannot generate a good pose manifold either. As shown

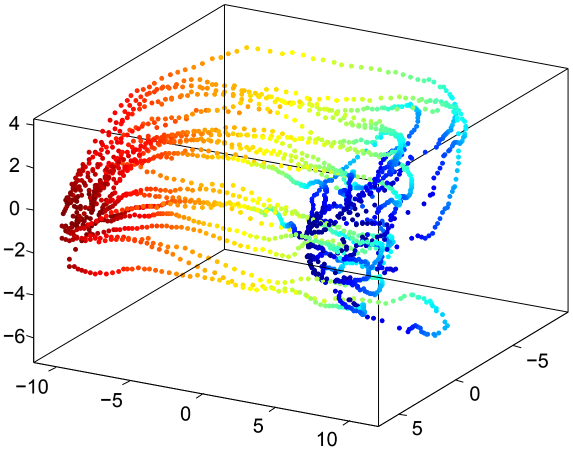


Figure 3.6: The 3-dimensional embedding of 24 subjects' face images with only pose variation between [−90° + 90°] at 4° increments, using ISOMAP embedding (NN = 200). The pose changes are represented by different colors.

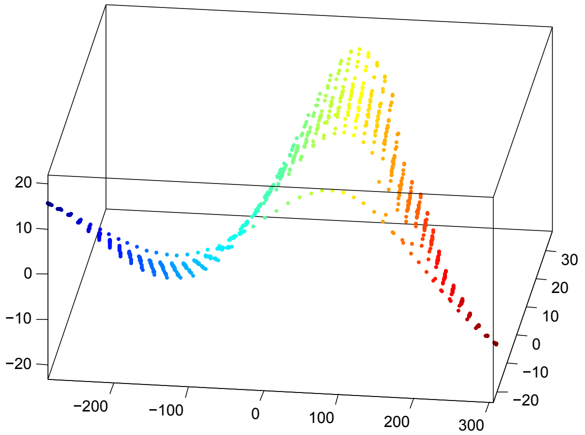


Figure 3.7: The 3-dimensional embedding of 24 subjects' face images with only pose variation between [−90° + 90°] at 4° increments, using BME (NN = 200). The pose changes are represented by different colors.

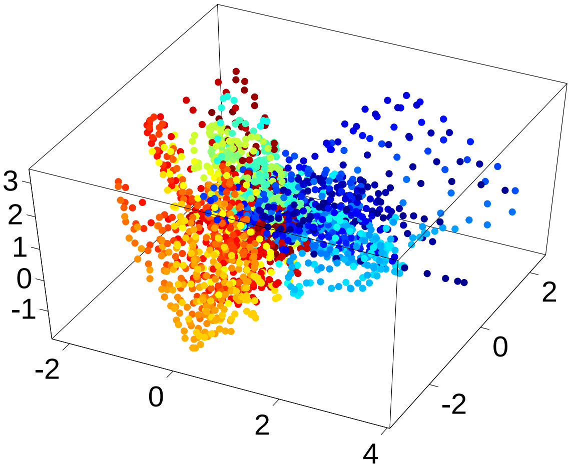


Figure 3.8: The 3-dimensional embedding of 10 subjects' face images with pose variation between $[-90° + 90°]$ at 4° increments and illumination changes from 0° to +45° at 5° increments, using ISOMAP embedding. The pose changes are represented by different colors.

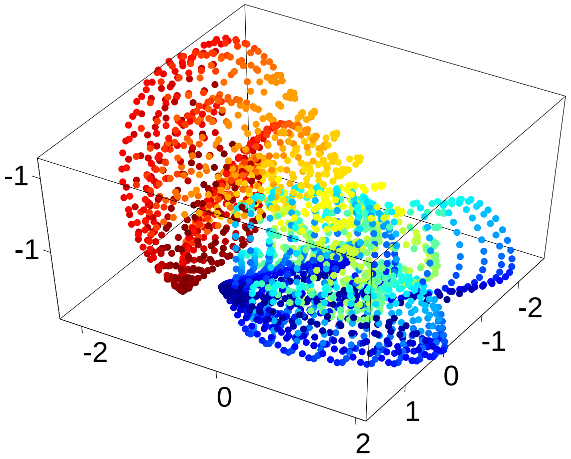


Figure 3.9: The 3-dimensional embedding of 10 subjects' face images with pose variation between $[-90° + 90°]$ at 4° increments and illumination changes from 0° to +45° at 5° increments, using biased ISOMAP embedding. The pose changes are represented by different colors.

in Figure 3.8 and 3.9, we generate two manifolds with ISOMAP and biased ISOMAP from 10 subjects with pose angles varying from −90° to +90° with a granularity of 4 and illumination changes from 0° to +45° at 5° increments. It is clear that there are many overlaps between pose angles in ISOMAP embedding (Figure 3.8), as well as in BME embedding (Figure 3.9). Hence, it is quite difficult to estimate poses from the manifolds generated by ISOMAP and BME. This is due to the following facts

- The computation of a nonlinear manifold relies on the distance between data points. For example, in order to compute a smooth pose manifold, the distance between face images under large illumination variation and the same pose should be small.
- BME only uses pose labels to find the right nearest neighbors. The distances between data points remain unchanged.
- The distortion caused by illumination variation is much larger than the distortion caused by the identity of individuals, which is apparent by comparing Figure 3.9 and Figure 3.7. Therefore, the distance between data points should be modified in order to obtain a smooth manifold. This brings forth the need to develop pose estimation techniques that can work well with face images from many different individuals having both large illumination changes and large pose changes.

To obtain a good low-dimensional embedding, we have developed an approach based on manifold learning techniques and supervised distance metric learning techniques for head pose estimation. We first construct the low-dimensional embedding using

ISOMAP (Figure 3.8). The low-dimensional embedding is then linearly mapped to the transformed feature space by modifying the distance between data points, using Local Fisher Discriminant Analysis (LFDA) [41] by pose labels. The combination of ISOMAP and Fisher Discriminant Analysis (FDA) was first proposed in the work [112]. In their work, each data point is represented by a feature vector, which is its geodesic distance from all other points. FDA is then applied to find an optimal projection direction for classification. The main difference between our approach and this extended ISOMAP is that we employ LFDA to refine the low-dimensional manifold and maintain pose class separation.

ISOMAP

The classical linear algorithms such as Principle Component Analysis (PCA) and Multidimensional Scaling (MDS) cannot always reveal the intrinsic distribution of a given complex data set. We therefore adopt ISOMAP for nonlinear dimensionality reduction [48]. The input is data matrix $X = (x_1, \cdots, x_N) \in \Re^{D \times N}$ containing N face images from the training data, where $x_i \in \Re^D (i = 1, 2, \ldots, N)$ are the D-dimensional samples of face images. In our case, D is equal to 1024, since we vectorized face images with the resolution 32×32. The output is the nonlinear embedding matrix $Y = (y_1, \cdots, y_N) \in \Re^{d \times N}$ of X, where $y_i \in \Re^d (i = 1, 2, \ldots, N)$ are the d-dimensional data points in the low-dimensional embedding. ISOMAP first determines the neighbor relationship on the manifold M based on the pairwise Euclidean distance $d_X(i, j)$ between pairs of face images x_i,x_j. These neighbor relations are represented as a weighted graph G over the data points, with the edges of weights $d_X(i, j)$ between neighboring points. The pairwise geodesic distance $d_M(i,j)$ on the manifold are then estimated with the distance of the shortest path $d_G(i,j)$ in the graph G using Floyd's or Dijkstra's algorithm. Classical MDS is finally applied to the matrix of graph distances $D_G = d_G(i,j)$ to construct d-dimensional embedding Y.

Local Fisher Discriminant Analysis

We use LFDA to learn the matrix $P_{LFDA} \in \Re^{d \times d}$ that transforms $y_i (i = 1, 2, \ldots, N)$ to $z_i (i = 1, 2, \ldots, N)$. $z_i \in \Re^d (i = 1, 2, \ldots, N)$ are the d-dimensional data points in the transformed feature space. $z_i = P_{LFDA}^T y_i$ in the same pose angle are kept close together, while z_i from different pose angles are well separated. LFDA evaluates within-class scatter and between-class scatter in a local manner by combining the idea of FDA and Locality-Preserving Projection (LPP) [53]. Here we briefly review the definition of FDA and LPP

Fisher Discriminant Analysis FDA considers maximizing the following objective:

$$J(W) = \frac{W^T S_b W}{W^T S_w W} \tag{3.27}$$

where S_b is the "between classes scatter matrix" and S_w is the "within classes scatter matrix". The definitions of the scatter matrices are:

$$S_b = \sum_c (\mu_c - \overline{\mathbf{x}})(\mu_c - \overline{\mathbf{x}})^T \tag{3.28}$$

$$S_w = \sum_c \sum_{i \in c} (\mathbf{x}_i - \mu_c)(\mathbf{x}_i - \mu_c)^T \tag{3.29}$$

where μ_c is the mean of samples in the class c and $\overline{\mathbf{x}}$ is the mean of all samples. It is known that W_{TDA} consists of the generalized eigenvectors associated to the generalized eigenvalues (in decreasing order) of the following eigenvalue problem:

$$S_b\varphi = \lambda S_w\varphi \tag{3.30}$$

Locality-Preserving Projection (LPP) Let A be the affinity matrix. All the elements of A are in $[0, 1]$. They will have smaller values if $\mathbf{x}_i$ and $\mathbf{x}_j$ are far apart. Several different methods [53,113] are proposed to define A. The minimization problem of finding the transformation matrix W_{LPP} is defined as follows,

$$\begin{array}{l} \arg\min\limits_{W} -\frac{1}{2}\sum\limits_{i,j} A_{i,j}\|W^T\mathbf{x}_i - W^T\mathbf{x}_j\|^2 \\ s.t.\ W^T\mathbf{X}\mathbf{D}\mathbf{X}^T W = \mathbf{I} \end{array} \tag{3.31}$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$, $\mathbf{I}$ is the identity matrix and $\mathbf{D}$ is a diagonal matrix; its entries are column sums of W, $\mathbf{D}_{i,i} = \sum_j W_{i,j}$. It is known that W_{LPP} consists of the generalized eigenvectors associated to the generalized eigenvalues (in decreasing order) of the following eigenvalue problem:

$$\mathbf{X}\mathbf{L}\mathbf{X}^T\varphi = \gamma\mathbf{X}\mathbf{D}\mathbf{X}^T\varphi \tag{3.32}$$

where $\mathbf{L} = \mathbf{D} - A$ is the Laplacian matrix.

The local within-class scatter matrix $\tilde{S}_B$ and the local between-class scatter matrix $\tilde{S}_W$ are defined as follows,

$$\tilde{S}_W = \frac{1}{2}\sum_{i,j=1}^{N} \tilde{W}_{i,j}^{(w)}(y_i - y_j)(y_i - y_j)^T \tag{3.33}$$

$$\tilde{S}_B = \frac{1}{2}\sum_{i,j=1}^{N} \tilde{W}_{i,j}^{(b)}(y_i - y_j)(y_i - y_j)^T \tag{3.34}$$

where

$$\tilde{W}_{i,j}^{(w)} = \begin{cases} W_{i,j}/N_c, & y_i \in c, y_j \in c \\ 0, & \text{otherwise} \end{cases} \tag{3.35}$$

$$\tilde{W}_{i,j}^{(b)} = \begin{cases} W_{i,j}(1/N - 1/N_c), & y_i \in c, y_j \in c \\ 1/N, & \text{otherwise} \end{cases} \tag{3.36}$$

,and $W_{i,j}$ is the affinity between y_i and y_j that is ranged in $[0, 1]$. The discussion about the definition of $W_{i,j}$ can be found in [53, 113]. LFDA considers the maximization of the following objective to find the transformation matrix P_{LFDA},

$$J(P) = \frac{\|P^T \tilde{S_B} P\|}{\|P^T \tilde{S_W} P\|} \tag{3.37}$$

Noticing that J is invariant with respect to scale, we can formulate the objective into the constrained optimization problem as follows:

$$\begin{aligned} &\min_P -\tfrac{1}{2} P^T \tilde{S_B} P \\ &s.t.\ P^T \tilde{S_W} P = I \end{aligned} \tag{3.38}$$

The lagrangian corresponding to this optimization problem is,

$$\ell = -\frac{1}{2} P^T \tilde{S_B} P + \frac{1}{2}(P^T \tilde{S_W} P - I) \tag{3.39}$$

Using the Karush-Kuhn-Tucker(KKT) conditions, the problem is transformed into the following generalized eigenvalue problem,

$$\tilde{S_B} \varphi = \lambda \tilde{S_W} \varphi \tag{3.40}$$

Then, the LFDA transformation matrix is defined by the solution as follows,

$$P_{LFDA} = [\varphi_1, \varphi_2, \cdots, \varphi_d] \tag{3.41}$$

where $\{\varphi_i\}_{i=1}^d$ are generalized eigenvectors corresponding to the generalized eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$.

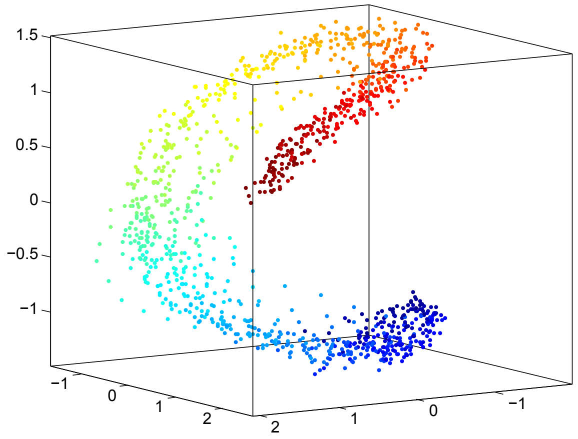


Figure 3.10: The 3-dimensional embedding of 24 subjects' face images with only pose variation from −90° to +90° at 4° increments using our method (NN =200). The pose changes are represented by different colors.

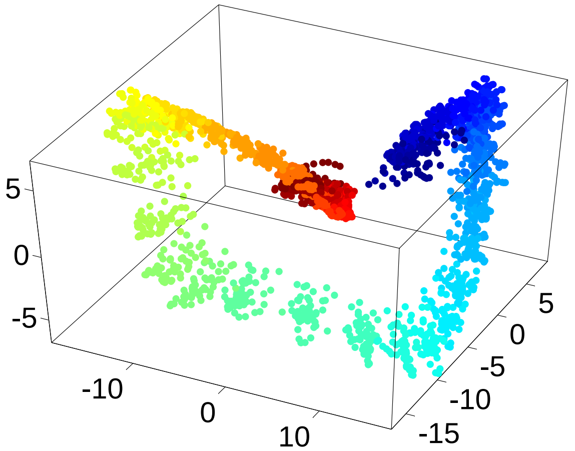


Figure 3.11: The 3-dimensional embedding of 10 subjects' face images with pose variation from −90° to +90° at 4° and illumination changes from 0° to +45° at 5° increments using our method (NN =200). The pose changes are represented by different colors.

We apply our method to the same data that we use in ISOMAP and BME methods to get the low-dimensional embedding, shown in Figure 3.11. Compared to the results of ISOMAP and BME in Figure 3.6 and Figure 3.7, our method is much better in clustering the face images in the same pose angle and separating the face images from different pose angles better than ISOMAP and BME methods in both the pose-only variation case and the pose+illumination variation case. For a new input face image, we can compute its low-dimensional embedding using the nonlinear mapping learned by Generalized Radial Neural Network (GRNN) [114], and then project it to the transformed feature space by applying the linear transformation. As the final step, its pose angle will be estimated by Relevance Vector Machine (RVM) [115]. The detailed algorithm procedure using ISOMAP and LFDA for pose estimation is shown in algorithm 1.

Experiment and Results of Head Pose Estimation

Data sets To evaluate the performance of our approach, we employed the 3D face dataset from [18]. The pose changes horizontally from $-90°$ to $+90°$ at 2 degree increments. The illumination varies from $0°$ to $+45°$ at $1°$ increments, as shown in Figure 3.12. Other public face databases such as FERET, the CMU-PIE database, Yale Face database, and MIT database, are not used in our experiment, because none of them provide a precise measure for pose and illumination angles; they also do not contain face images with a wide variety of illumination and pose changes [56]. To assess the robustness of our approach, we perform the experiment in two cases: pose estimation for face images without and with illumination variation. We compared

Algorithm 1 Pose Estimation Pipeline

```
1: Learning phase
   Input: the training face images $x_i (i = 1, 2, \cdots, N)$ and their corresponding labels
         of pose angles $\ell_i$.
   Output: the nonlinear mapping between $x_i$ and $y_i$, $P_{LFDA}$,
        the regression model for pose estimation.
   (a): Get the low-dimensional embedding $y_i$ using ISOMAP.
   (b): Learn the nonlinear mapping between $x_i$ and $y_i$ using GRNN.
   (c): Find the transformation matrix $P_{LFDA}$ from low-dimensional embedding $y_i$.
   (d): Learn the regression model between $y_i$ and its corresponding
        pose labels using RVM.

2: Testing phase
  Input: the test face images $x_i (i = N + 1, N + 2, \cdots, NN)$.
  Output: the pose labels for $x_i$.
   (a): Compute its low-dimensional embedding $y_i$ using learned nonlinear mapping.
   (b): Map $y_i$ to the transformed feature space using $z_i = P^T_{LFDA} y_i$.
   (c): Estimate the pose angle by applying the learned regression model on $z_i$
```

our method with the state-of-art pose estimation techniques, BME. The performance is analyzed with varying choices of the embedding dimensions (marked by "o" in the figures) and 200 neighbors.

Pose Estimation for Face Images without Illumination Variation

The experiment was performed over 24 subjects with pose angles varying from $-90°$ to $+90°$ at $2°$ increments and illumination of $22°$ using 8-fold cross-validation. We use 1911 face images from 21 subjects (91 images per subject) as the training data in each fold, and then use the 273 images from the other 3 subjects as testing data. The images were down-sampled to 32×32 resolution. The results of the experiment are shown in Figure 3.13. The red line indicates the performance of our method,

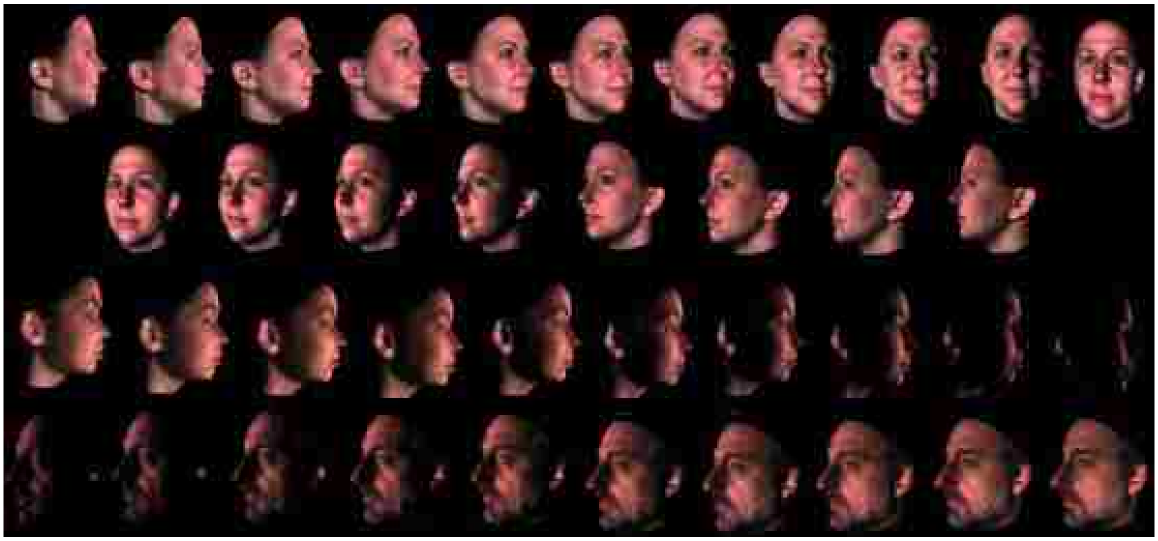


Figure 3.12: Samples of face images with varying pose and illumination from 3D face scans

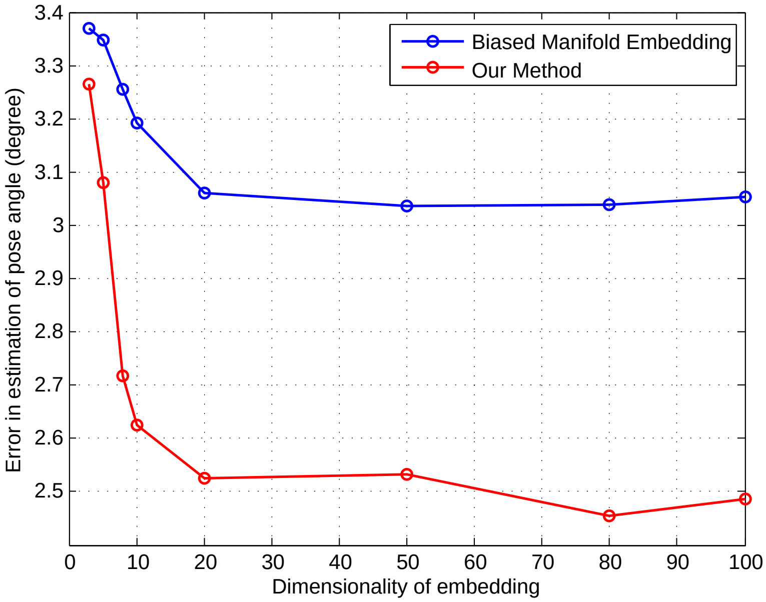


Figure 3.13: Pose estimation results comparison of our method against BME with ISOMAP for the face images without illumination variation (NN=200) in different dimensionality. The red line indicates the results of our method.

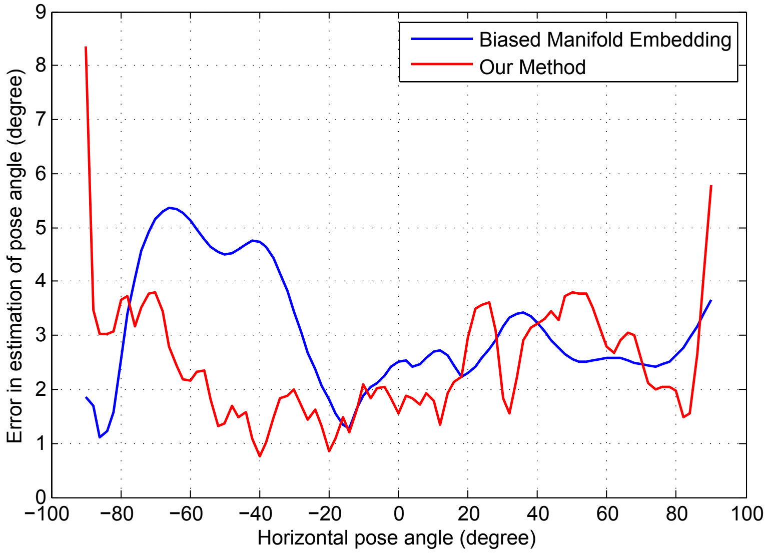


Figure 3.14: Distribution of the average error of our method against the BME framework in pose estimation for the face images without illumination variation. Each of the views is between [$-90°$, $+90°$] at 2° increments. The red line indicates the results of our method.

while the blue line shows the performance of the BME framework. The result shows the accuracy of our method is slightly better than that of the BME with ISOMAP. Figure 3.14 shows the head pose estimation error of our method against BME in each of the views in the pose angle interval.

Pose Estimation for Face Images with Illumination Variation

Due to memory limitation, this experiment was performed over 10 subjects, with pose angles varying from $-90°$ to $+90°$ at 4° increments and illumination variation from 0° to $-45°$ with a granularity of 5°. We use leave-one-out cross-validation (LOOCV), i.e., we sequentially take out the face images of one individual and use all the remaining images of the other individuals as training data. The images are also down-sampled to 32×32 resolution. The experiment results are shown in Figure 3.15. The red line

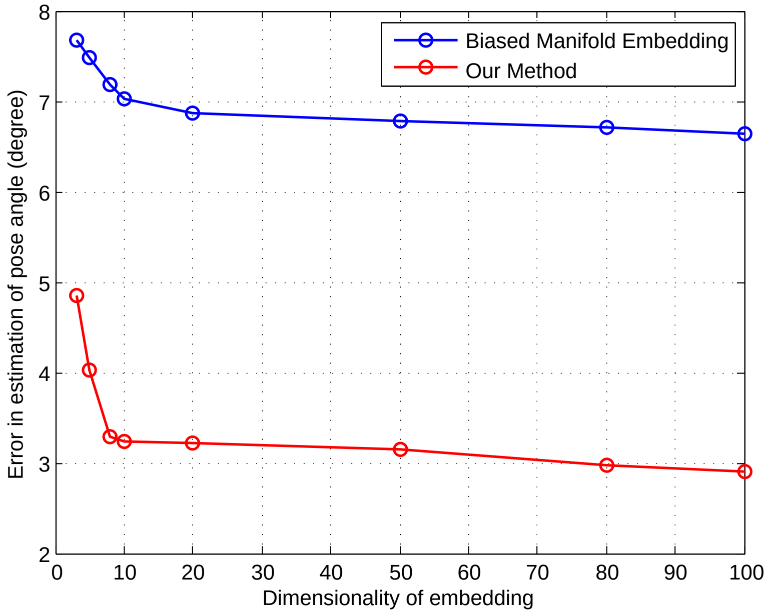


Figure 3.15: Pose estimation results comparison of our method against BME with ISOMAP for the face images with illumination variation. The red line indicates the results of our method.

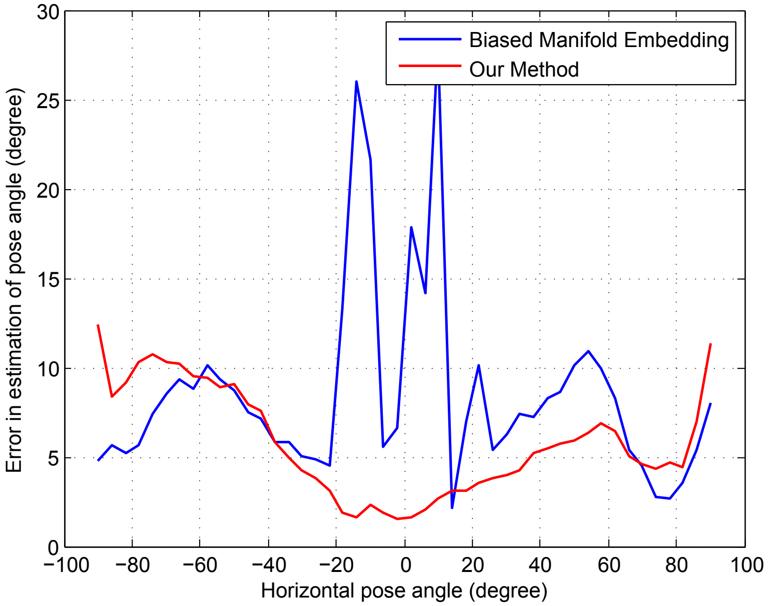


Figure 3.16: Distribution of the average error of our method against the BME framework in pose estimation for the face images illumination variation. each of the views is between [−90°, +90°] at 4° increments. The red line indicates the results of our method.

indicates the performance of our method, while the blue line shows the performance of the BME framework. It shows that our method significantly improves the head pose estimation performance compared to the BME framework. Figure 3.16 shows the head pose estimation error of our method against BME in each of the views in this pose angle interval (100-dimensional embedding) .

In both cases, our method performs well for the frontal and intermediate poses, but not for the profile (almost $-90°$ or $+90°$) views. This may be caused by the noisy features in the face images of these profile views, which results in some overlaps between the data points in the transformed feature space.

Comparing our results with those listed in Table 2.1, we can see that our method is comparable with the state-of-art methods for face images with only pose variation. When the illumination also varies, our method maintains high accuracy for most cases and performs better that BME.

In summary, we have developed an approach to illumination- and person-insensitive head pose estimation. We study the limitation of related approaches in pose estimation for face images with large illumination variation, and address the problem by combining ISOMAP and LFDA. We conduct several experiments to evaluate our approach. The experiment results demonstrate that our method is robust to variation in dimensionality of embedding, illumination and identity of individuals. Our method is easily extendable to other manifold learning techniques, such as LLE and LE, and supervised distance metric learning techniques like RCA, NCA, and LMNN. Furthermore, our method can be used to estimate the illumination direction by using illumination labels in LFDA.

3.5 Shape Recovery from a Single Image

From the training data set, local image models, surface models, and their correspondences are learned using manifold embedding and alignment techniques, as outlined in the learning phase of Algorithm 2. Now we can reconstruct the 3D shape of a new image. We first estimate the pose i of this image. We use Algorithm 1 in 3.4, which is robust to large illumination and pose variations. The facial image is then aligned to the reference facial image I_i^r with the estimated pose using the method in [94] and divided into N_z overlapping patches, $\mathbf{y}_1^*, \cdots, \mathbf{y}_{N_z}^*$. We have to correct the illumination of these patches before the shape recovery via Eq. 3.1. After illumination correction, the local shape for each patch can be estimated as outlined in the reconstruction phase of Algorithm 2.

Global Reconstruction The recovered representative of the local shapes, $\widetilde{\mathbf{y}}_1^*, \cdots, \widetilde{\mathbf{y}}_{N_z}^*$, need to be combined into the representative of a global shape, $\mathbf{s}^* = [\widetilde{\mathbf{y}}_1^* \cdots \widetilde{\mathbf{y}}_{N_z}^*]$. $\mathbf{s}^*$ can be considered as a collection of vectorized affine transformations of the triangles of the reference models $\mathbf{M}_r$. The problem that we need to solve here is to find the target shape $\mathbf{M}_u = \{\widetilde{\mathbf{v}}_1, \cdots, \widetilde{\mathbf{v}}_n\}$ to satisfy the constraints $\mathbf{s}^*$. For each target triangle of $\mathbf{M}_u$, and the affine transformation can be written as $\mathbf{T} = \widetilde{\mathbf{V}}\mathbf{V}^{-1}$ in terms of the original and deformed vertices. The elements of $\mathbf{V}^{-1}$ are coordinates of the known, original vertices of $\mathbf{M}_r$, while the elements of $\widetilde{\mathbf{V}}$ are coordinates of the unknown deformed vertices of $\mathbf{M}_u$. From this definition, we see that the elements of $\mathbf{T}$ are linear combinations of the coordinates of the unknown deformed vertices. Thus we can

Algorithm 2. Locally Estimating the Shape

```
I. Learning phase
  Input: A set of $N_z$ training examples, $(\mathbf{Y}_{i,1}, \widetilde{\mathbf{Y}}_1), \cdots, (\mathbf{Y}_{i,N_z}, \widetilde{\mathbf{Y}}_{N_z})$
  Output: the $N_z$ local models of image patches and shape patches, $\Theta^I_{i,1}, \cdots, \Theta^I_{i,N_z}$
      and $\Theta^S_1, \cdots, \Theta^S_{N_z}$; the correspondences (patch index) and the optimal mapping
      parameters between the models: $(t_1, \lambda^{opt}_{i,t_1}, \mathbf{P}^{opt}_{i,t_1}), \cdots, (t_{N_z}, \lambda^{opt}_{i,t_{N_z}}, \mathbf{P}^{opt}_{i,t_{N_z}})$
  for j = $1 \cdots N_z$
      Learn the local image and shape models, $\Theta^I_{i,j}$ and $\Theta^S_j$, and get the low
      dimensional embeddings of $\mathbf{Y}_{i,j}$ and $\widetilde{\mathbf{Y}}_j$, $\mathbf{X}_{i,j}$ and $\widetilde{\mathbf{X}}_j$ by maximizing the
      posterior of Eq. 3.14.
  end for
  for j = $1 \cdots N_z$
      Learn the correspondence of $t_j$ and the optimal mapping parameters: $\lambda^{opt}_{i,t_j}$,
      $\mathbf{P}^{opt}_{i,t_j}$, between $\mathbf{X}_{i,j}$ and $\widetilde{\mathbf{X}}_{i,t_j}$ via Eq. 3.21 and 3.22.
  end for
II. Reconstruction phase
  Input: the test facial image patches, $\mathbf{y}^*_1, \cdots, \mathbf{y}^*_{N_z}$
  Output: the recovered $N_z$ local shapes, $\widetilde{\mathbf{y}}^*_1, \cdots, \widetilde{\mathbf{y}}^*_{N_z}$
  for j = $1 \cdots N_z$
      Compute the low-dimensional embedding, $\mathbf{x}^*_j$, of $\mathbf{y}^*_j$ by minimizing the
      negative log likelihood of Eq. 3.18 with the learned local image model, $\Theta^I_{i,j}$.
      Map $\mathbf{x}^*_j$ into $\widetilde{\mathbf{x}}^*_j$ of low-dimensional space using the learned $\lambda^{opt}_{i,t_j}$ and $\mathbf{P}^{opt}_{i,t_j}$
      via Eq. 3.26.
      Recover the local shape of $\widetilde{\mathbf{x}}^*_j$, $\widetilde{\mathbf{y}}^*_j$, by computing the mean of posterior in
      Eq. 3.15 with the learned local shape model, $\Theta^S_{t_j}$.
  end for
```

formulate the problem as a minimization problem [99]:

$$\min_{\widetilde{\mathbf{v}}_1 \ldots \widetilde{\mathbf{v}}_n} \sum_{j=1}^{|M|} \|\mathbf{S}_j - \mathbf{T}_j\|^2_F \tag{3.42}$$

where $\mathbf{S}_j$ is the known source transformation, $|M|$ is the number of transformations in the constraint, and $\mathbf{T}_j$ is the unknown target transformation. Since the target transformations are defined in terms of the unknown deformed target vertices, the problem can be rewritten in the matrix form,

$$\min_{\widetilde{\mathbf{v}}_1 \ldots \widetilde{\mathbf{v}}_n} \|\mathbf{s}^* - \mathbf{A}\widetilde{\mathbf{x}}\|^2_2 \tag{3.43}$$

where $\widetilde{\mathbf{x}}$ is a vector of unknown deformed vertex coordinates, and A is a large, sparse matrix that relates $\widetilde{\mathbf{x}}$ to $\mathbf{s}^*$. Thus, all the vertices of the target shape $\mathbf{M}_u$ can be solved in the least-square sense.

3.6 Experimental Results

Data Sets To evaluate the performance of our approach, we employed two data sets in our experiments. The first one is a 3D face scans database [72], which contains shapes and textures of 120 real faces obtained with a laser scanner. We generate the synthetic facial images from them with pose and illumination changes. The pose changes horizontally from −90° to +90° at 5 degree increments. The illumination varies horizontally from −45° to +45° with a granularity of 5. The resolution of facial images is 256×256. Notice that the images provided in this database are not identical to the real albedos of the faces, due to noticeable effects of the lighting conditions. Our second dataset is the CMU-PIE database [116], which contains 68 individuals with 9 horizontal and 3 vertical pose variations and 21 illumination variations.

Among these images, we use 2052 synthetic facial images as the training data set. They correspond to 108 subjects under 19 pose variations. The illumination condition is fixed at a natural (ambient) setting. To learn the local image and surface models, we use 60 inducing variables, and the latent dimension $d = 8$.

Experiments Our first experiment shows the effectiveness of our patch-based method for illumination variations. Note that our training database contains no sample under changing illumination. Figure 3.17 shows a comparison without and with illumination normalization. We use $m_I(x, y) = \sigma_I(x, y) = 2$ for synthetic images and $m_I(x, y) = \sigma_I(x, y) = 0.5$ for real images in Eq. 3.1 to correct the illumination variation, and divided all the images into N_z overlapping patches with size 7×7. The value of N_z depends on the pose of images, e.g., 252 patches for the frontal faces in our experiments.

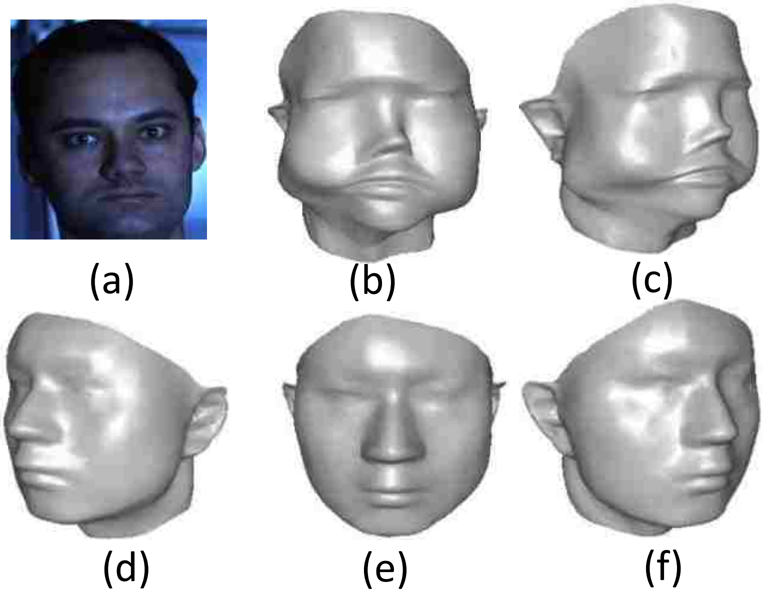


Figure 3.17: Shape recovery from a single frontal image w/o local illumination normalization. (a) the input frontal images with illumination; (b,c) different views of the reconstruction result without illumination normalization; (d,e,f) different views of the reconstruction result with illumination normalization.

Synthetic Inputs We use the images and shapes from the remaining 12 subjects in the first database as the testing data to run a controlled experiment. Our method is used to recover their shapes from the synthetic images. This experiment allows us to show comparisons of our reconstructions to the ground truth shapes. The quantitative accuracy of reconstruction can be defined as [3]:

$$\varepsilon = \frac{1}{n} \sum_{i=1}^{n} |D_r(i) - D_t(i)| \tag{3.44}$$

where D_r is the recovered shape and D_t is the ground truth shape, and n is the number of vertices in the shape. Figure 3.18 shows a few results. For comparison we show the reconstructed shapes and the ground truth, and plot the alignment of the reconstructed shapes (in gray) with the ground truth shapes (in blue). It can be seen that our algorithm can obtain accurate reconstructions in spite of illumination and pose variations. The reconstructed error in each pose is shown in Figure 3.20, which shows that our algorithm is fairly insensitive to pose variations and achieves the same level of accuracy as the methods [3, 2, 1] in all poses. The recovery accuracy for the frontal facial images in our method is slightly better than that of those methods. In addition, our method can handle illumination and pose changes.

Real Inputs We apply our method to several real images from the CMU database using the same training data set. Note that the real input images are not in the training database, and have the big difference in illumination. The reconstructed results are shown in Figure 3.19.

3.7 Conclusion

In summary, we have developed an approach to illumination- and person-insensitive head pose estimation, and a novel approach to the shape recovery from a single side-view image. For the problem of pose estimation, we studied the limitation of related approaches in pose estimation for face images with large illumination variation, and

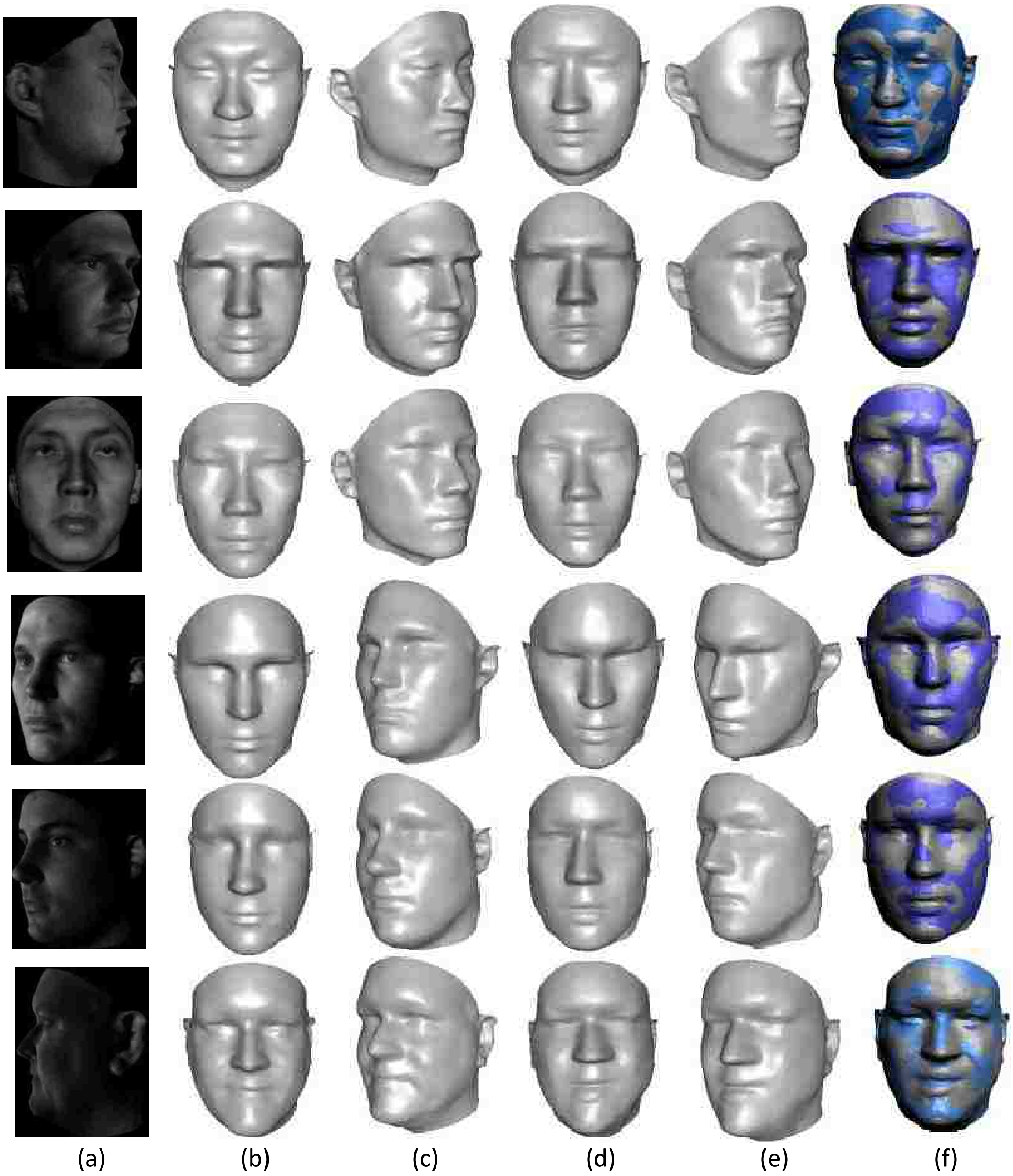


Figure 3.18: Results of shape recovery for the synthetic facial images. (a) the input image rendered from the 3D face scan database; (b,c) different views of the ground truth shape; (d,e) the frontal view and side view of the recovered 3D shape; (f) the aligned image of the ground truth shape (in blue) and the recovered shape (in gray), which is used for measuring the reconstruction accuracy.

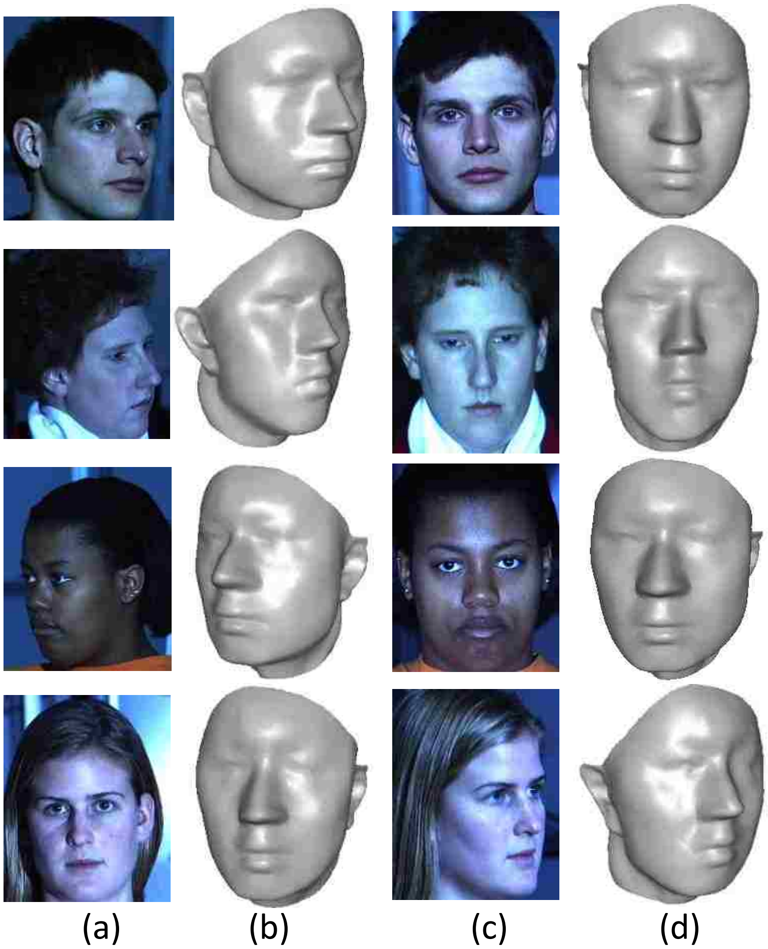


Figure 3.19: Results of shape recovery for the real images. (a) the input image from the CMU database; (b) the reconstruction from (a) using our approach; (c) The image of the same person as (a) in a different pose that was not used for the reconstruction; (d) the profile view of the reconstruction corresponding to the pose in (c).

addressed the problem by combining ISOMAP and LFDA. We conducted several experiments to evaluate our approach. The experiment results demonstrate that our method is robust to variation in dimensionality of embedding, illumination and identity of individuals. For the problem of shape recovery, we study the limitation of related approaches in shape recovery for facial images with illumination and pose variations and address the problem using non-linear embedding and alignment. We conduct experiments to evaluate our approach by comparing the reconstructed results

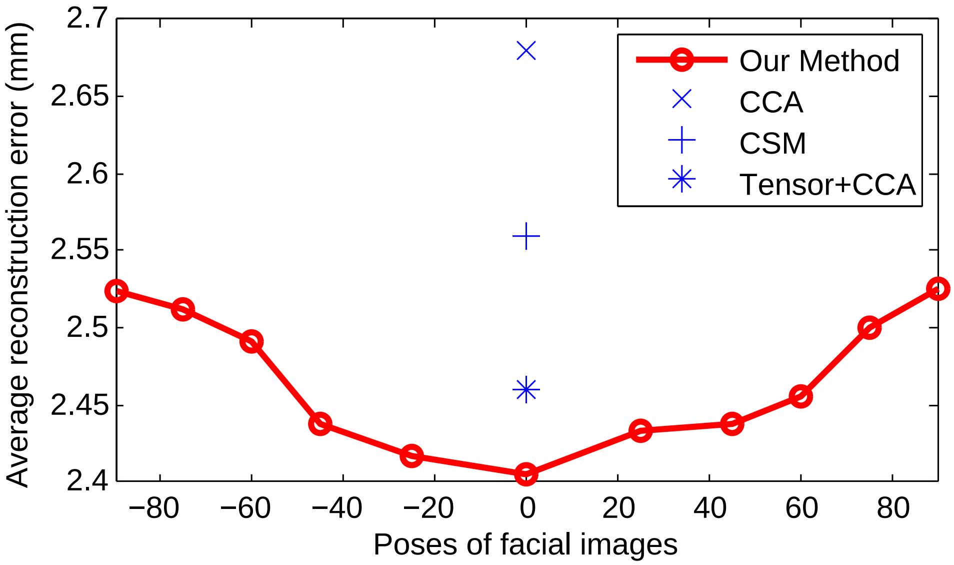


Figure 3.20: 3D reconstruction error vs. pose variation in our method. For comparison we also show the best reconstruction errors of CCA [1], CSM [2] and Tensor+CCA [3]. These methods can only deal with frontal images of faces without illumination variation. Note that these methods measure the reconstruction error with the specific training data set and testing data set.

to ground truth shapes and by applying the method to various real images. The experimental results demonstrate that our method of shape recovery is also robust to variation in pose, illumination and identity of individuals. Looking into the future, we would like to further evaluate the performance of our approach with more appropriate real training data. In addition, we plan to extend our approach to reconstruct the shapes of other objects, such as the human body.

Chapter 4 Accurate and Robust Skeletal Motion Capture from a Single Depth Sensor

Single view marker-less motion capture remains an open problem, even after many years of research. The main challenges come from high variability of the internal variations in human appearance, differences of movement across individuals, and external variations of scenes, such as lighting conditions, cluttered background, poor image resolution, non-rigidity of tissue and clothing, and partial occlusion including self-occlusion. The previous approaches of face modeling cannot be directly applicable for single view motion modeling, because the approaches of statistical learning rely on large training sets to learn the low-dimensional models. It is difficult to build a huge image database that covers the space with internal variations in human appearance, differences of movement style, and external variations of scenes. We decrease difficulty by adopting the depth cue from a depth sensor to remove the variation in appearance, illumination, and background. In addition, depth information can provide us metric 3D measurement. We therefore formulate motion capture from a single depth image sequence as a model fitting problem with a known template model database. The database contains 3D surface models of motion sequences from a single subject, which is much smaller compared to one used in statistical learning techniques.

A depth sensor provides real-time dynamic scene scanning, in which each pixel contains intensity and range information for a scene point. We aim to automatically recover a sequence of body configurations and surfaces to represent the pose and shape of the subject for every frame of the depth images from a *single* depth sensor. We first introduce the processing pipeline of our approach in Section 4.1. Section 4.2 describes the estimation of the rough body configuration from the previous frame by non-rigid point registration techniques. We refine the body configuration with an estimated full 3D surface model in Section 4.3. Section 4.4 presents the technique of temporal filtering to remove the jittering artifacts. The experimental results and analytic analysis are shown in section 4.5, and the conclusion is made in section 4.6.

While the idea of using a motion database to facilitate motion capture or human modeling has been applied in previous research [117], adapting it to the single-view setup brings new challenges. In particular, the view-dependency of the input depth map. For a given body configuration or a surface mesh model, there may be infinitely many depth maps that the input needs to be compared with. Direct matching of a depth map with a full-body surface mesh model would fail as the motion database may have many models in different poses and inputs from a single depth sensor have at least 50% of the data points missing. We solve this problem through our two-step registration process. The first temporal registration step generates a *view independent* body configuration so that we can find the most similar configuration in the motion database efficiently. Next we perform model-to-input registration via a rendered (e.g., *view-dependent*) depth map that corresponds to the input's perspective. In addition, we perform non-rigid surface fitting to deal with body-size and small body configuration differences between the template and the input. Therefore, instead of building a motion database covering all of the motions, viewpoints, and body sizes, we only need samples to cover the motion; this dramatically reduces the size of the motion database. Our *view-* and *body-size- independent* fitting formulation is our most important technical contribution to human motion modeling.

4.1 Algorithm Overview

Data Capture and Setup: We have data sets from two different sources: depth images and a pre-captured database of sample motions. The depth image stream is the input to our algorithm. To create the sample database, we use a commercial optical motion capture system to capture human motion. The skeleton of the motion has $N_v = 19$ joints, and the skeleton is used to drive a generic human mesh model M. More specifically, we align the skeleton model with the human mesh model in the standard T-pose and use the method in [118] to automatically compute the weight $\rho_{i,k}$ of each vertex i in the mesh model. The mesh surface model can then be animated by the skeleton using linear blending skinning techniques. For the purpose of data registration, we also segment the mesh model in to $N_s = 13$ parts. In summary, the motion database contains different body poses, each pose has a full 3D surface mesh model and an underlying skeleton (e.g., body configuration), as shown in Fig. 4.1. Each pose surface model is denoted as $\mathbf{Y}^{(i)}$, its k^{th} segment denoted as $\mathbf{Y}^{(i,k)}$. The corresponding body configuration is denoted as $\mathbf{V}_Y^{(i)}$, which has a set of joint positions $\{\upsilon_{Y,1}^{(i)}, \ldots, \upsilon_{Y,N_v}^{(i)}\}$.

Problem Definition: We are given a set of body pose models $\{(\mathbf{Y}^{(1)}, \mathbf{V}_Y^{(1)}), \ldots, (\mathbf{Y}^{(N)}, \mathbf{V}_Y^{(N)})\}$ in the motion database. Note that all the full models $\{\mathbf{Y}^{(i)}\}$ have the same set of points and triangles. We are also given a sequence of depth maps: $\mathbf{I}^{(i)}$, which can be turned into an un-structured set of 3D metric points. $\mathbf{I}^{(i)}$ has been

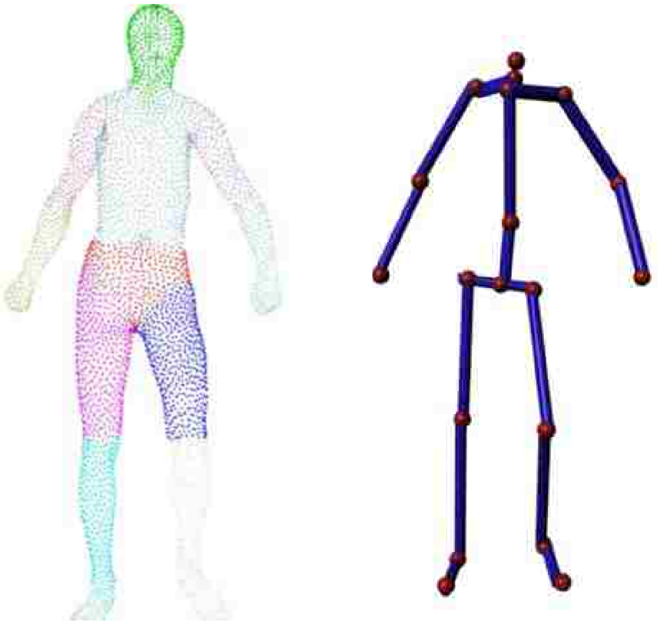


Figure 4.1: We segment the surface into 13 parts, and a different color shows a different part. Our articulated skeleton model has 19 joints (left arm has 3, right arm has 3, head has 1, left leg has 3, right leg has 3, and torso has 6)

segmented from the background and contains only human motions. Given the depth information, segmentation is much easier than that from color. Note that the human motion in the input may be different in terms of body size and movement, compared to ones in the database. We use simple background subtraction for this task. The goal of our algorithm is to recover the body configuration $\mathbf{V}_I^{(i)}$.

We assume that each input depth stream begins with a known pose, e.g., the standard T-pose. An outline of the processing pipeline is given in Figure 4.2. For each frame, we estimate the rough body configuration from the previous frame by non-rigid point registration techniques as described in Section 4.2 (Figure 4.2a-b). The estimated body configuration, which may be incomplete due to occlusions, is used to find in the motion database a full 3D surface model with its corresponding

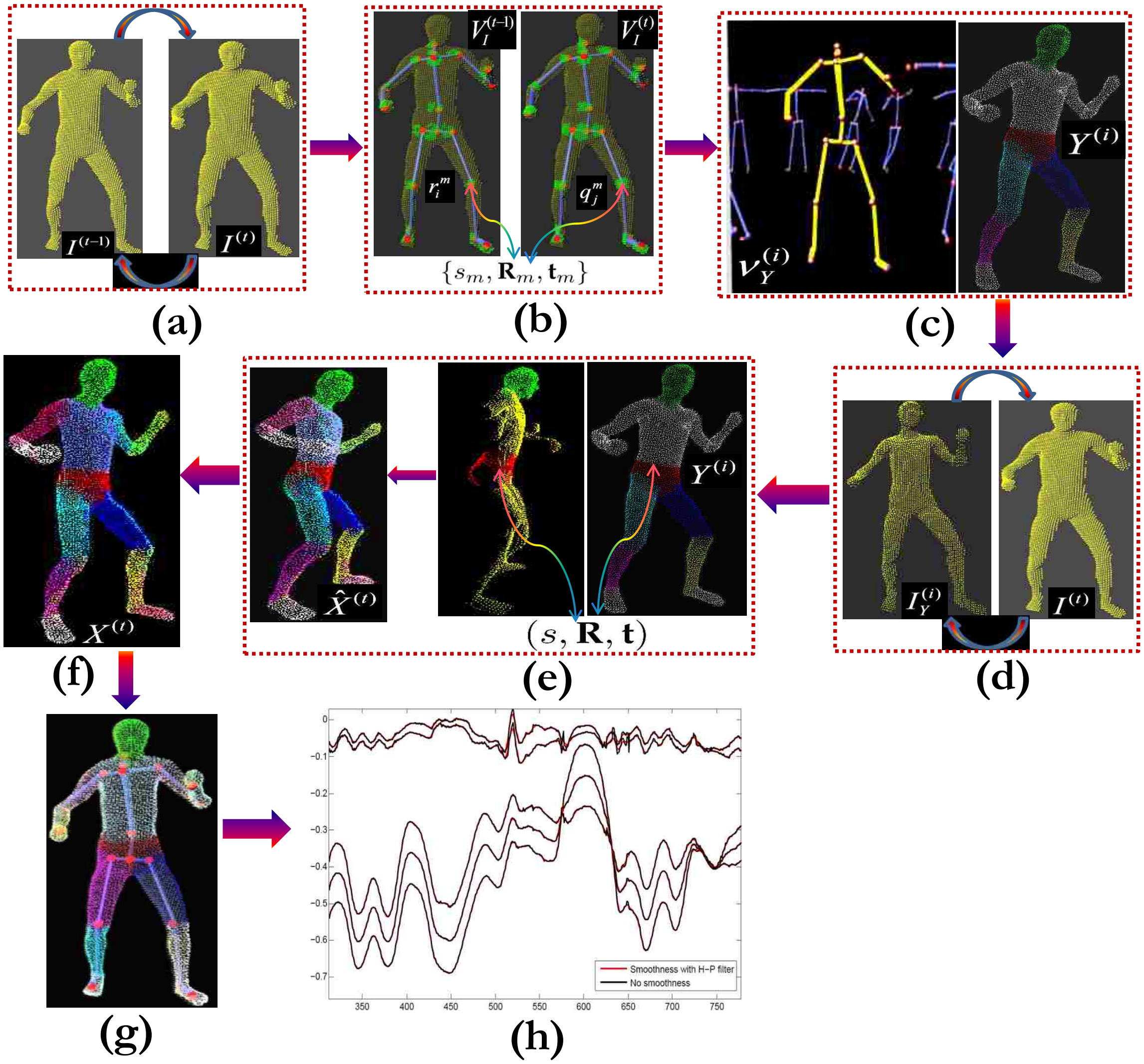


Figure 4.2: The pipeline of skeletal motion capture from a single depth sensor. (a) computation of one-to-one point correspondences between the current frame $\mathbf{I}^{(t)}$ and the previous frame $\mathbf{I}^{(t-1)}$ by non-rigid registration; (b) estimation of the m^{th} joint position $v_{I,m}^{(t)}$ (red) in $\mathbf{I}^{(t)}$, based on the assumption that the corresponding nearest neighbor points (green),$\{r_i^m\}$ and $\{q_j^m\}$, have the linear transformation in neighboring frames; (c) search of the template surface model ($\mathbf{Y}^{(i)}$) and body configuration ($\mathbf{V}_Y^{(i)}$) most similar to the estimated body configuration $\mathbf{V}_I^{(t)}$ in the database; (d) computation of one-to-one point correspondences between the current frame $\mathbf{I}^{(t)}$ and the synthetic depth map $\mathbf{I}_Y^{(i)}$ generated by $\mathbf{Y}^{(i)}$; (e) estimation of the full surface model $\widehat{\mathbf{X}}^{(t)}$ with computation of piece-wise transformations and local occlusion handling; (f) computation of the refined surface model $\mathbf{X}^{(t)}$ by the Laplacian deformation framework; (g) estimation of the refined body configuration by a set of predefined control points. (h) filtering of jittering artifacts with an extended H-P filter.

template body configuration that is most similar to the input (Figure 4.2c). Following this, a full 3D surface model is generated with estimation of the local transforms and local occlusion handling by taking advantage of the template body configuration and the template full surface model (Figure 4.2d-e, Section 4.3). Because this step only captures piece-wise deformations, the non-rigid surface is refined at a later step(Figure 4.2f, Section 4.3). The estimated refined body configuration (Figure 4.2g) serves as input for the next frame. Minor registration error and temporal inconsistency may cause the sequence of recovered surface shapes and body configurations to exhibit jittering artifacts. We therefore apply an extended Hodrick-Prescott filter to reduce the jittering artifact as described in Section 4.4 (Figure 4.2h). Because we estimate the body configuration in the coordinate of the input depth map, any global motion can be easily recovered.

4.2 Pose Estimation from Temporal Registration (PETR)

We process each input frame sequentially. Generally, there are only small changes of body configuration between two consecutive frames. Therefore, the skeleton of frame t can be estimated from the results of the previous frame $t-1$. We choose to use non-rigid point registration techniques to align the partial models (e.g., the depth maps) $\mathbf{I}^{(t-1)}$ and $\mathbf{I}^{(t)}$ together to find point correspondences, which is used to estimate the skeleton movement. In our system, similar to the face modeling algorithm, we adopt the coherent point drift (CPD) algorithm, which is a robust probabilistic method for both rigid and non-rigid registration of point sets [119].

We apply CPD in both directions: first registering $\mathbf{I}^{(t)}$ to $\mathbf{I}^{(t-1)}$ and then vice versa, as shown in Figure 4.2a. If the matching from both ways is consistent, a one-to-one point correspondence is declared. Hence, after this registration procedure, we have obtained a set of one-to-one correspondences, denoted as $C = \{(r_i, q_j)\}$ where $r \in \mathbf{I}^{(t-1)}$ and $q \in \mathbf{I}^{(t)}$, from which we will estimate the skeleton motion as the second step. Note that since we assume that the skeleton motion in frame $t-1$ is known, we only need to estimate the relative motion between them. For each joint $v_{I,m}^{(t-1)}$ in frame $t-1$, K nearest points in $\{r_i\}$ are selected, denoted as $\{r_i^m\}$. With these points' correspondences $\{q_j^m\}$ in frame t, we can estimate a linear transformation using the Procrustes analysis (PA) [120]:

$$\{s_m, \mathbf{R}_m, \mathbf{t}_m\} = \text{procrustes}(\{r_i^m, q_j^m\}) \tag{4.1}$$

where $\mathbf{R}_m$ is a 3×3 rotation matrix, $\mathbf{t}_m$ is a 3×1 translation vector, and s_m is the scaling parameter. Therefore, the m^{th} joint position $v_{I,m}^{(t)}$ in $\mathbf{I}^{(t)}$ can be estimated as follows (Figure 4.2b):

$$v_{I,m}^{(t)} = s_m \cdot \mathbf{R}_m \cdot v_{I,m}^{(t-1)} + \mathbf{t}_m \tag{4.2}$$

A finale note here is that we also assume that $\mathbf{I}^{(t-1)}$ has been segmented into different parts, $\{r_i^m\}$ must belong to the segmentation group m. Due to occlusions, there could be some joint positions that are lost in frame t. This will be handled in the second stage of the algorithm described in the next section.

4.3 Data-driven Body Configuration Refinement

The error will inevitably accumulate in the above incremental body configuration update scheme. It will get worse when there exists occlusion, which is very common in a single-view setup where at least 50% of the human body is occluded in every frame. In this section we will describe our approach to use the pre-captured motion database to complete and refine the body configuration.

For a given initial body configuration $\mathbf{V}_I^{(t)}$ (note that this configuration can be incomplete), we search in the motion database to find the most similar body configuration. To do this, we again apply PA to estimate the linear transformation between $\mathbf{V}_I^{(t)}$ and every template body configuration in the database (Figure 4.2c), the one with the minimum residue distance after the transformation is the winner, denoted as $\mathbf{V}_Y^{(i)}$. It has a corresponding complete mesh model $\mathbf{Y}^{(i)}$. We render this full model from the perspective of the input depth camera to generate a synthetic depth map $\mathbf{I}_Y^{(i)}$. The perspective of the input camera can be easily estimated by using the input 3D points as a calibration object. We also perform intrinsic calibration beforehand to make the camera pose estimation problem easier. Since $\mathbf{Y}^{(i)}$ is already segmented, the rendered $\mathbf{I}_Y^{(i)}$ is also segmented with color-coding of different body parts. Then we apply the same technique in the previous section to align $\mathbf{I}_Y^{(i)}$ with $\mathbf{I}^{(t)}$ (Figure 4.2d). After this alignment, $\mathbf{I}^{(t)}$ is segmented into different parts.

Once the correspondence between the input and the full model ($\mathbf{I}_Y^{(i)}$) is established we can compute the $(s, \mathbf{R}, \mathbf{t})$ between each body part and their matching points (Figure 4.2e). Furthermore, at this time we can use the complete surface model to make a best-effort guess about the occluded part in the input image. More specifically, for a segment $\mathbf{Y}^{(i,k)}$ that has no corresponding points in the input, we will use its most immediate visible ancestor node's $(s, \mathbf{R}, \mathbf{t})$ to transform segment $\mathbf{Y}^{(i,k)}$ into the coordinate of the input frame. We denote the piece-wise transformed model as $\widehat{\mathbf{X}}^{(t)}$, e.g., the complete surface model that corresponds to $\mathbf{I}^{(t)}$.

Laplacian Surface Refinement: $\widehat{\mathbf{X}}^{(t)}$ is reconstructed based on articulated motion without deformation, e.g., each segment is independently estimated via rigid transformation. While it is commonly used in motion capture, the positions of all surface vertices need to be refined to fit the dense point cloud $\mathbf{I}^{(t)}$ as illustrated in Figure 4.2f. More specifically, we compute Laplacian coordinates Δ of $\mathbf{Y}^{(i)}$ as $\Delta = \mathbf{L}\mathbf{Y}^{(i)}$, where $\mathbf{L}$ is the cotangent Laplacian matrix. We refine the surface from the Laplacian coordinates and vertices position constraints by solving the following linear least squares system:

$$\begin{aligned} \mathbf{X}^{(t)} &= \arg\min_{\mathbf{X}}\{w_L\|\mathbf{L}\mathbf{X}-\Delta\|_2^2 + w_C\textstyle\sum_k\|\mathbf{X}_i^{(k)} - q_j^{(k)}\|_2^2 \\ &\qquad\qquad + w_T\|\mathbf{X}-\widehat{\mathbf{X}}^{(t)}\|_2^2\} \end{aligned} \tag{4.3}$$

where $\mathbf{X}_i^{(k)}$ is the i^{th} vertex in $\mathbf{X}$, $q_j^{(k)}$ is the j^{th} vertex from $\mathbf{I}^{(t)}$, and k is the index of match vertices between $\widehat{\mathbf{X}}^{(t)}$ and $\mathbf{I}^{(t)}$. The first item uses the weight w_L to determine the smoothness of the resulting surface. $w_L = 0$ is the smoothest, while $w_L = 1$ preserves the full details of the original surface. The second term ensures that a set of position constraints are satisfied with the weight w_C. The third term is a regularization term for solving the ill-posed problem, because over 50% of the vertices are missing in our input. The weight w_T determines how close the resulting surface $\mathbf{X}^{(t)}$ is to the previous estimated surface $\widehat{\mathbf{X}}^{(t)}$.

This refinement step, which is based on the original Laplacian deformation framework [121], has been similarly used in several recent papers that aim to capture both motion and appearance [122, 6, 123]. Compared with these works, the difference is that we only use a single camera (depth only without any color information). Therefore we have to add the last regularization term based on the template surface model. Figure 4.3 shows the effectiveness of this term. The reconstructed model is similar to the template model. Our goal is to focus only on the body configuration, therefore we trade off the need for an array of cameras for the lack of appearance modeling.

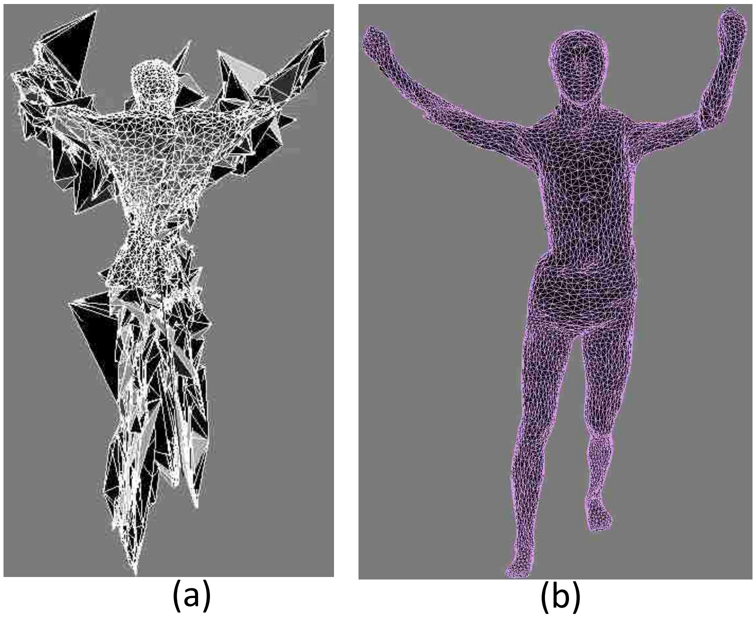


Figure 4.3: The example demonstrates the effectiveness of the regularization term. (a) The reconstructed surface using our method without the regularization term, which has obvious distortions; (b) The reconstructed surface of our method with the regularization term, which looks much better than (a).

Body Configuration Refinement: At this point, we have a full surface model $\mathbf{X}^{(t)}$ that has the same topology as the template model $\mathbf{Y}^{(i)}$ and whose visible part closely conforms to the input depth image $\mathbf{I}^{(t)}$. Given $\mathbf{X}^{(t)}$, $\mathbf{Y}^{(i)}$, and $\mathbf{V}_Y^{(i)}$, we can refine the body configuration $\mathbf{V}_I^{(t)}$ using Procrustes Analysis (PA) as in Eq. 4.1 and 4.2 accordingly (Figure 4.2g). But instead of using the nearest points in $\mathbf{X}^{(t)}$, we use a set of predefined control points in $\mathbf{X}^{(t)}$'s each segment to update $\upsilon_{I,m}^{(t)}$. In this way we can select surface points that are less likely to have non-rigid deformation between the skin and the bone, for example, the front side of the knees and the out-side of the elbows. The updated body configuration, as well as the segmented input depth map, are used as input for the next frame.

From the above discussion we can see that the first temporal registration step only provides a rough body configuration estimation to facilitate the nearest neighbor search in the motion database. So as long as the first frame in the input sequence is from a known pose, our algorithm can be bootstrapped.

In summary this refinement step improves the motion capture results in two ways. First, it avoids the pose drift problem since the pose is updated from a template body configuration $\mathbf{V}_Y^{(i)}$, rather than the previous frame. Secondly, it provides a means to fill in severely or even completely occluded parts. In addition, since we adopted non-rigid point registration techniques followed by a surface optimization step, our method is robust to personal size and height variations. Various captured motion from a single subject is sufficient to create the motion database.

4.4 Temporal Filtering

Since we choose to independently update the body configuration for each frame there is some visible jitter in the recovered motion. We apply Hodrick-Prescott (H-P) trend filtering [124] to remove these artifacts (Figure 4.2h). The traditional H-P filter only deals with scalar data. We extend it to the vector case. A vertex $x_i^{(t)}$ in $\mathbf{X}^{(t)}$ is considered as a time series from different t. Its movement consists of a slowly varying trend component $a_i^{(t)}$ and a more rapidly varying random component

$b_i^{(t)}$. That is, $x_i^{(t)} = a_i^{(t)} + b_i^{(t)}, t = 1, 2, \dots T$. The goal of trend filtering is to isolate $a_i^{(t)}$, or equivalently, $b_i^{(t)}$ from $x_i^{(t)}$. H-P filtering estimates $a_i^{(t)}$ by minimizing the following weighted sum objective function with two competing objectives: $a_i^{(t)}$ needs to be smooth, and $b_i^{(t)}$ should be small,

$$\sum_{t=1}^{T} \|a_i^{(t)} - x_i^{(t)}\|^2 + \lambda \sum_{t=2}^{T-1} \|x_i^{(t-1)} - 2x_i^{(t)} + x_i^{(t+1)}\|^2 \tag{4.4}$$

where $\lambda \geq 0$ is the penalty parameter that controls the trade-off between the smoothness of $x_i^{(t)}$ and the size of the residual $a_i^{(t)} - x_i^{(t)}$. The loss function of Eq. 4.4 can be written in the following matrix form,

$$\|\mathbf{a}^{(t)} - \mathbf{x}^{(t)}\|_F^2 + \lambda \|D\mathbf{a}^{(t)}\|_F^2 \tag{4.5}$$

where $\mathbf{x}^{(t)} = (x_1^{(t)}, \dots, x_T^{(t)})^T \in \mathbf{R}^{T\times 3}$, $\|u\|_F$ is the Frobenius norm, $\mathbf{y}^{(t)} = (a_1^{(t)}, \dots, a_T^{(t)})^T \in \mathbf{R}^{T\times 3}$, and $D \in \mathbf{R}^{(T-2)\times T}$ is the second-order difference matrix,

$$D = \begin{bmatrix} 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & -2 & 1 & \\ & & & & 1 & -2 & 1 \end{bmatrix}$$

By taking derivatives of the objective function in Eq. 4.5 with respect to $\mathbf{y}^{(t)}$ (see Appendix A), we can get the solution,

$$\mathbf{a}^{(t)} = (\mathbf{I} + \lambda D^T D)^{-1} \mathbf{x}^{(t)} \tag{4.6}$$

As we will demonstrate in our results, the visual quality of the captured motion is greatly improved after H-P trend filtering compared to the typical low-pass filtering that uses a sliding window.

4.5 Experiments and Results

We implement our algorithms in MATLAB (www.mathworks.com) and built a motion database that contains several thousand different 3D poses, covering motion activities of walking, bending, picking, stretching, turning etc, all captured from a single subject. We test our method on 11 depth image sequences with five subjects. The results show that our approach can capture skeletal motions from sequences of varying complexity, ranging from simple stretching, walking, to full-body rotations. The results can be seen in the accompanying video.

Qualitative Evaluation

The results in the accompanying video ($www.vis.uky.edu/ \sim xwang/resultsthesis.mov$). demonstrates that our method can correctly estimate motion configurations from a wide spectrum of scenes. Here, we show some examples of motion sequences from synthetic data, ranging from simple motions, including walking (Figure 4.4a), stretching (Figure 4.4b) and kicking (Figure 4.4c), to more complex motions, such as swinging (Figure 4.5a) and body rotation (Figure 4.5b). In addition to these basic motions, the video also indicates that our approach can recover some challenging activities and details of motion models. For example, our approach can deal with body occlusion (Figure 4.5c) and dancing (Figure 4.6a). Figure 4.6b shows our approach is invariant to body size. Note that, we use real data from a TOF sensor for Figure 4.6b.

Our approach not only estimates body configurations from the depth images with the partial occlusion and total occlusion in body parts (Figure 4.5c), but also is capable of handling severe deformations to different poses and even shape, as shown in the red box in Figure 4.7, in which only a profile view of the subject is captured.

Quantitative Evaluation

We evaluate the performance of our approach by comparing it to commercial motion capture systems. We also discuss how our approach effectively corrects the drifting problem in reconstruction of skeletal motions, and deals with the jittering artifacts in the reconstruction results. Furthermore, the robustness of the reconstruction algorithm is also tested in terms of different levels of noise.

Comparison with optical motion capture systems. Our results are compared with the ground truth obtained from an 8-camera VICON motion capture system. It should be noted that the markers using in the VICON system contaminate the depth maps, therefore, we have to use synthetically generated depth maps for this evaluation. We report the mean absolute errors (MAE) [4] over all joint locations of all actions between the ground truth and estimated results,

$$\varepsilon_{avg} = \frac{1}{m}\sum_{i=1}^{m}\|\mathbf{m}_i - \hat{\mathbf{m}}_i\|_2$$

The MAE over all the joint positions in our experiments is about 2 centimeters. Table 4.1 shows the MAE position error for different sequences. To facilitate the comparison with other approaches, we show the mean and standard deviation of the relative MAE error norms in various sequences. It should be noted that the average error from the most recent method [4] that also uses a single depth sensor is much higher than ours, 100mm vs. 20mm.

Table 4.1: Average MAE, mean, variance, minimum, and maximum error in centimeters (cm) over all the joint positions for different sequences.

Case	Mean	Minimum	Maximum	Variance
kicking	1.26	0.60	2.00	0.1061
rotation	1.57	0.98	2.23	8.61e-2
stretching	1.36	0.85	2.17	7.16e-2
walking	1.52	1.00	2.22	9.21e-2
swinging	1.43	0.81	2.15	7.39e-2

Figure 4.8 shows a quantitative error analysis of the stretching sequence (frames 0 to 269) for the left elbow and right elbow. The black star shows the joint positions of the ground truth from motion capture. Red circles represent the results of our approach. It can be seen that the relative error between the black curve and the red curve is very small in both figures.

Drifting correction. Figure 4.9 shows the position error curves of using our method (red) and only its first step (Section 4.2), estimating pose with temporal registration only (PETR) (blue) respectively, from a fighting sequence of Actor II. The errors are computed against the ground truth. We only show four joints in Figure 4.9. It can be seen that the reconstruction error quickly becomes so large that the results from PETR are useless. In contrast, our full approach can successfully recover the body configurations of an entire sequence and considerably improve the reconstruction results, which is revealed from the small differences between our results (red) and the ground truth. The example demonstrates how our method corrects the drifting problem by applying refinement with the full surface model, which resolves ambiguities with additional template model priors.

Smoothness to skeletal motions. The reconstruction results of body configuration often have temporal jittering artifacts. The experimental results of the joint in the left leg without the temporal filtering technique are illustrated in Figure 4.10. The ground truth is represented by the black line. The cyan line indicates the performance of our method with temporal filtering, whereas, the blue line shows the performance of our approach without temporal filtering. This demonstrates the importance of the temporal filtering techniques, which significantly improves the smoothness of motion changes, and successfully removes the outliers. Here, we set $\lambda = 10$ to control the smoothness. Moreover, we compare our method with the low-pass filter (red). The red curve is smoother than the blue curve, but it still has rapid changes at some points as in the blue curve. This shows that the low-pass filter can reduce the temporal jittering artifacts a little bit, but it cannot remove the artifact completely. The accompanying video further indicates the effectiveness of temporal filtering. This result demonstrates that the estimation is stable, and the overall quality of the skeleton motion is consistently high.

Dependency to databases. We investigate how the performance of our algorithm is in terms of various sizes of the database. In this experiment, we use the synthetic database with 7980 frames, including walking, swinging, and occlusion. The data is sampled in the different rates to simulate different sizes of the database. In Figure 4.11, the reconstruction errors are reported with different levels of sampling, such as 1, 2, 4, 8, 12, 16, 20, 24, and 28, corresponding to the original sampling rate of 120fps, 60fps, 30fps, 15fps, 10fps, 5fps, and 4.3fps. It demonstrates that as the sampling rate decreases (the size of database is decreasing), our algorithms get progressively less accurate, but it still has high accuracy under reasonable sampling rates of the database. On the other hand, for the complicated movements as the occlusion sequence, denser samples are required to obtain good reconstruction. Overall our approach is quite insensitive to sampling rate.

Testing in a public database. We also evaluate our approach using the public database [4], which consists of 28 real-world motion sequences with varying complexity ranging from simple hand lifting to challenging motions, like a tennis swing. Half of the motion sequences have 100 frames, and the others contain 400 frames. All of them are recorded at 25fps with the resolution 177 × 144 from a TOF sensor. Furthermore, the data contains the ground truth of 3D markers, which are captured by a commercial active marker motion capture. Table 4.2 shows a quantitative error analysis of our algorithm in this database. It can be seen that our approach has a high and stable performance in different sequences. The reconstruction errors of our algorithm are obviously smaller than the errors of the state-of-the-art techniques [4], while it is worse than ones from synthetic data as in Table 4.1. This is due to the significant noise in depth images, although we use an extended Locally Optimal Projection (LOP) algorithm [125] to reduce the noise level. The example of noisy depths and discussion of this algorithm are presented in [125].

Table 4.2: The reconstruction errors [mm] of our approach using a public database [4]

Sequences	0	1	3	5	6	7	8	9	14	16	27
Errors	36.9	37.2	37.7	39.1	39.2	36.6	37.4	37.3	35.9	39.2	40.1

Robustness to noisy constraints. The test is performed to investigate the performance of our algorithm in terms of various levels of input noise. In this experiment, we select 100 frames of depths from a sequence of body rotation. Zero-mean Gaussian noise of different standard deviations are added to the depths to simulate the noise presented in depths. The reconstruction errors are reported in Figure 4.12 with different levels of noise. It can be seen that as the noise increases, our algorithm gets progressively less accurate, but still works well under a reasonable level of noise.

Speed. The average timings for each individual step of our algorithm, which is implemented in MATLAB, are shown in Table 4.3. These timings are measured on an Intel Core 2 CPU Desktop with 2.66 GHz. We believe a C or GPU-based implementation can significantly improve the performance.

Table 4.3: Average running times of individual steps per frame

Step	Time
Skeleton estimation (Sec. 4.2)	$\sim$ 75s
Database search (Sec. 4.3)	$\sim$ 3s
Piece-wise model estimation (Sec. 4.3)	$\sim$ 46s
Surface refinement (Sec. 4.3)	$\sim$ 4s

We briefly exemplify the usability of our algorithm for one application in the following.

Multi-person Motion Capture. Finally we demonstrate our method's occlusion handling capability by extending it to capture multiple persons. This is demonstrated with a fighting sequence of 927 frames. Note that none of the template models have multiple persons in it. For the majority of frames (95%), we apply simple segmentation to cut out each person using connected component analysis with the depth information. The other frames, in which the two persons touch each other, can be segmented manually or automatically with the color images. The segmented subject has more occlusions than the single-person case. As demonstrated in the accompanying video, our approach generates satisfactory results under these challenging frames. Some examples are shown in Figure 4.13. To our best knowledge, this is the first time multi-person motion capture is reported with a single sensor. The overlaps between the two persons, even after correct segmentation, can cause erroneous silhouette description, which will lead to failure in many silhouette-based motion capture approaches.

4.6 Conclusion

In this work, we have developed a novel approach to estimate body configurations and surface deformations from a TOF depth sensor. We study the limitations of related work in motion recovery and address the problem using registration, and surface fitting techniques. Our approach also generates surface deformations with full correspondences and correct topology. The key insight is to use nonrigid registration to globally estimate the skeleton, followed by locally handling the occlusion and generate the full surface model by taking advantage of template priors. Moreover, our approach refines the full surface model by fitting it to the input depth with Laplacian coordinates setting. The refined surface model in turn corrects the occasional skeletal motion mistakes. We conduct experiments to evaluate our approach by comparing the reconstructed results to the ground truth. The experimental results demonstrated that our method is robust to a wide variety of activities, and generalizes well with more complex scenes in which there exist partial or extreme occlusion, appearance and clothing variations, or multiple persons present.

The main limitation of our approach is that it cannot reproduce the surface accurately and reconstruct small details of the subject, although the skeleton motion is globally captured. This is due to noise suppression operations in depth and the inherited smoothness from Laplacian operator. In addition, the low resolution of a TOF sensor can lead to the sparse correspondences between the template and the input. Therefore, the appearance of the output surface deformations are similar to the appearance of the template. Looking into the future, we will improve the surface accuracy and details with higher resolution of depth sensors and benefit from a large volume data as the work in [117].

A Solution to H-P Filter in Vector Case

From the definition of Frobenius norm, $\|A\|_F^2 = tr(A^T A)$, and $tr(A+B) = tr(A) + tr(B)$, we can get,

$$\begin{aligned}\|\mathbf{a}^{(t)} - \mathbf{x}^{(t)}\|_F^2 &= tr((\mathbf{a}^{(t)})^T \mathbf{a}^{(t)}) - tr((\mathbf{a}^{(t)})^T \mathbf{x}^{(t)}) \\ &\quad - tr((\mathbf{x}^{(t)})^T \mathbf{a}^{(t)}) - tr((\mathbf{x}^{(t)})^T \mathbf{x}^{(t)})\end{aligned}$$

$$\|D\mathbf{a}^{(t)}\|_F^2 = tr((\mathbf{a}^{(t)})^T D^T D \mathbf{a}^{(t)})$$

We take derivatives with respect to $\mathbf{a}^{(t)}$ with the facts, $\frac{\partial}{\partial X} tr(X^T A) = A$, $\frac{\partial}{\partial X} tr(A X^T) = A$, and $\frac{\partial}{\partial X} tr(X^T B X) = BX + B^T X$, and we find,

$$\frac{\partial}{\partial \mathbf{a}^{(t)}} \|\mathbf{a}^{(t)} - \mathbf{x}^{(t)}\|_F^2 = -\mathbf{x}^{(t)} - \mathbf{x}^{(t)} + 2\mathbf{a}^{(t)}$$

$$\frac{\partial}{\partial \mathbf{a}^{(t)}} \|D\mathbf{a}^{(t)}\|_F^2 = 2D^T D \mathbf{a}^{(t)}$$

and

$$\frac{\partial}{\partial \mathbf{a}^{(t)}}\{\|\mathbf{a}^{(t)} - \mathbf{x}^{(t)}\|_F^2 + \lambda\|D\mathbf{a}^{(t)}\|_F^2\} =$$

$$- 2\mathbf{x}^{(t)} + 2\mathbf{a}^{(t)} + 2\lambda D^T D\mathbf{a}^{(t)}$$

we ignore the constant $tr((\mathbf{x}^{(t)})^T\mathbf{x}^{(t)})$, and by setting it to zero, it follows,

$$(\mathbf{I} + \lambda D^T D)\mathbf{a}^{(t)} \quad = \quad \mathbf{x}^{(t)}$$

therefore,

$$\mathbf{a}^{(t)} \quad = \quad (\mathbf{I} + \lambda D^T D)^{-1}\mathbf{x}^{(t)}$$

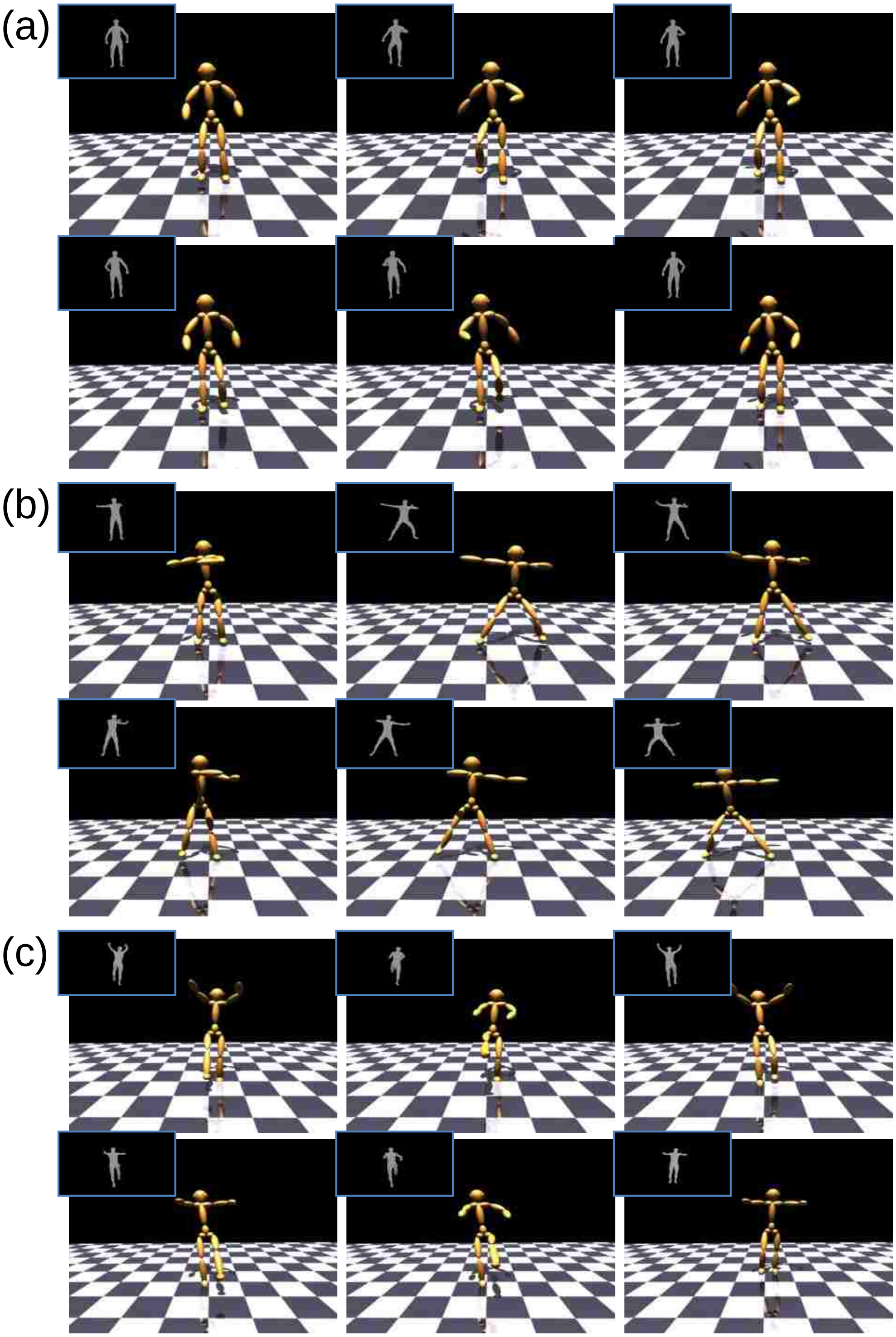


Figure 4.4: Some examples of walking(a), stretching (b), and kicking (c) sequences with our approach. The insets show the synthetic input depth maps; The recovered body configurations are used to drive the model.

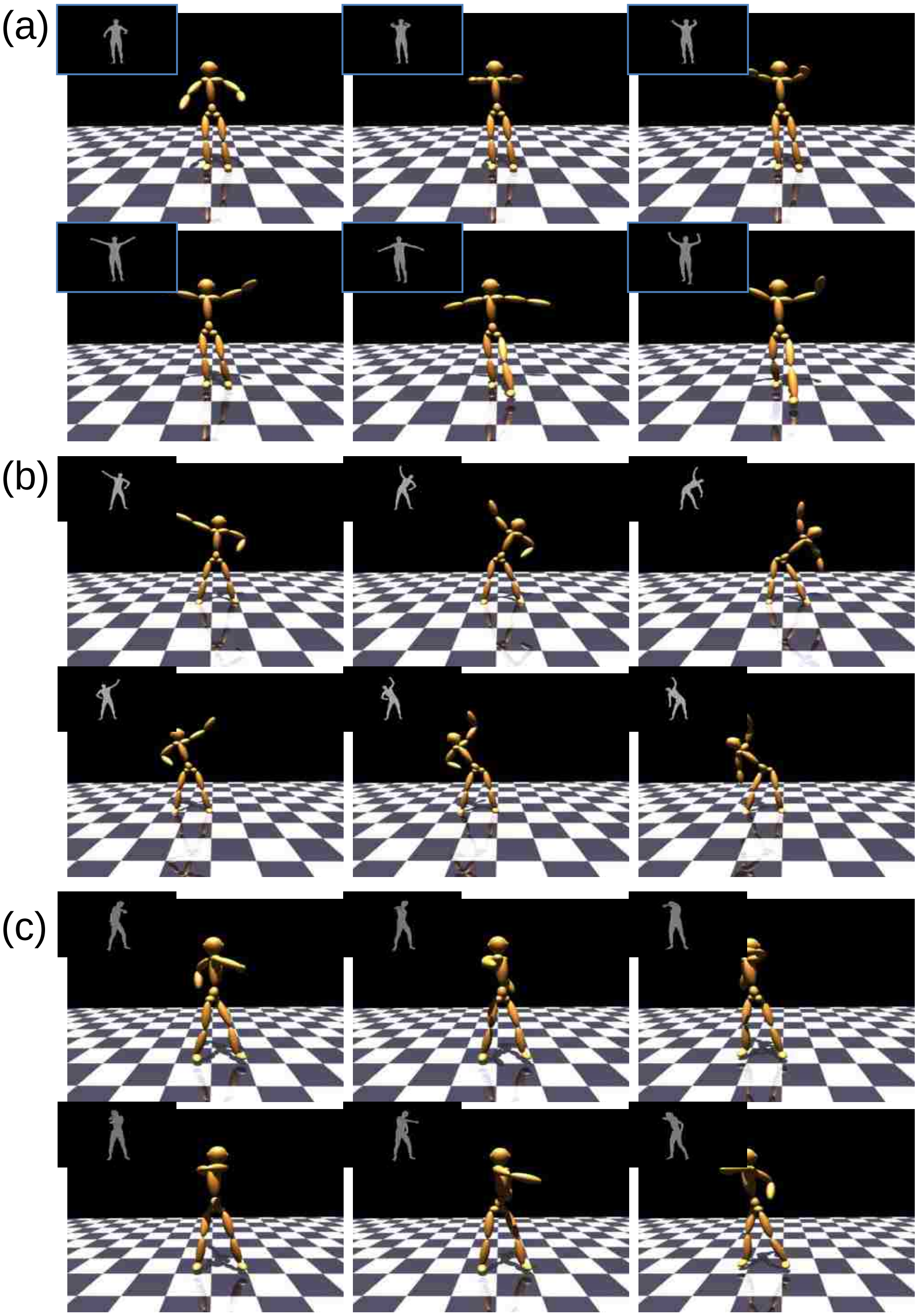
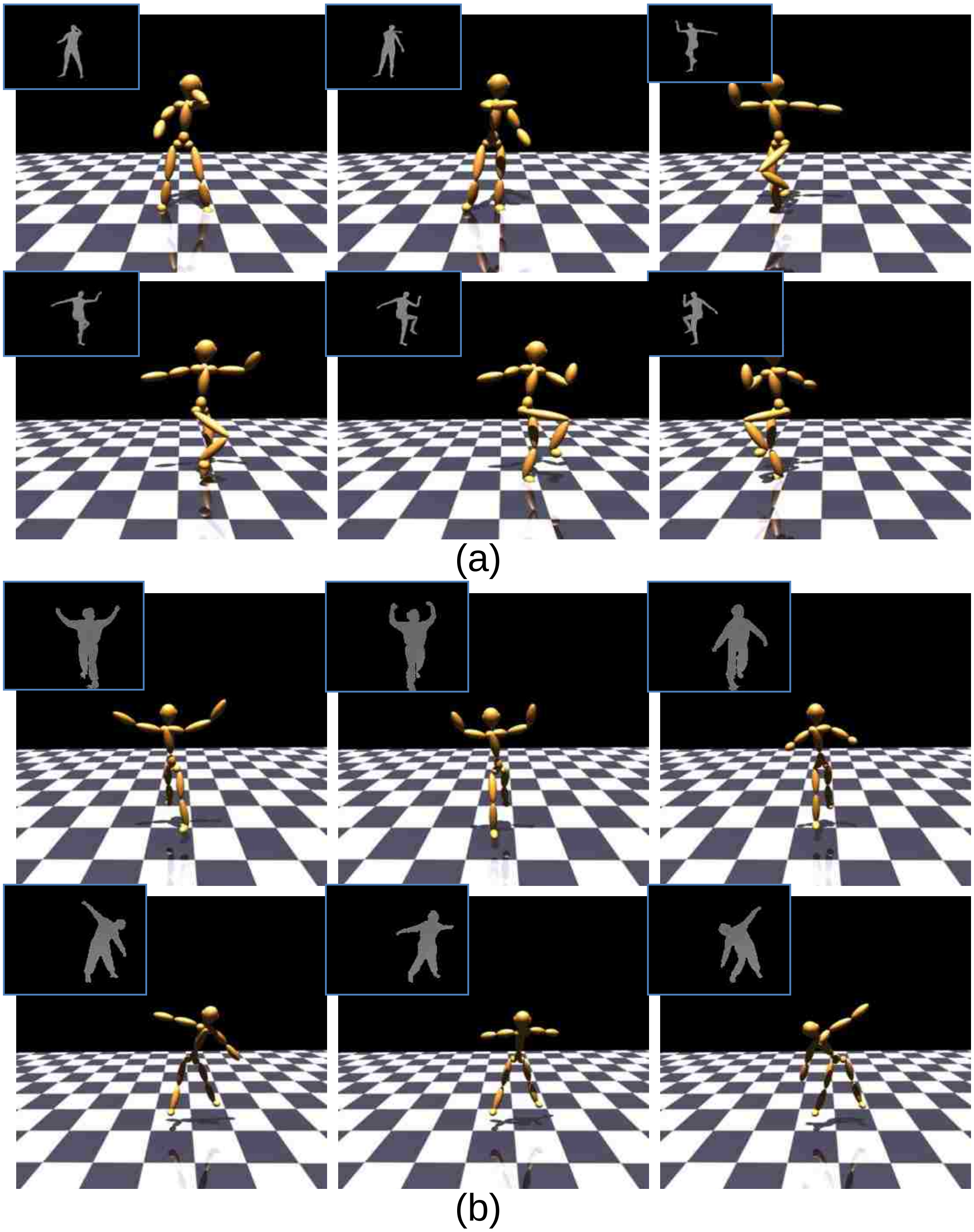


Figure 4.5: Some examples of swinging (a), rotation (b), and occlusion (c) sequences with our approach. The insets show the synthetic input depth maps; The recovered body configurations are used to drive the model.



(a)

(b)

Figure 4.6: Some examples of a kicking sequence (a) with our approach. The insets show the synthetic input depth maps; Some examples of a child sequences with our approach. The insets show the real input depth maps. The recovered body configurations are used to drive the model.

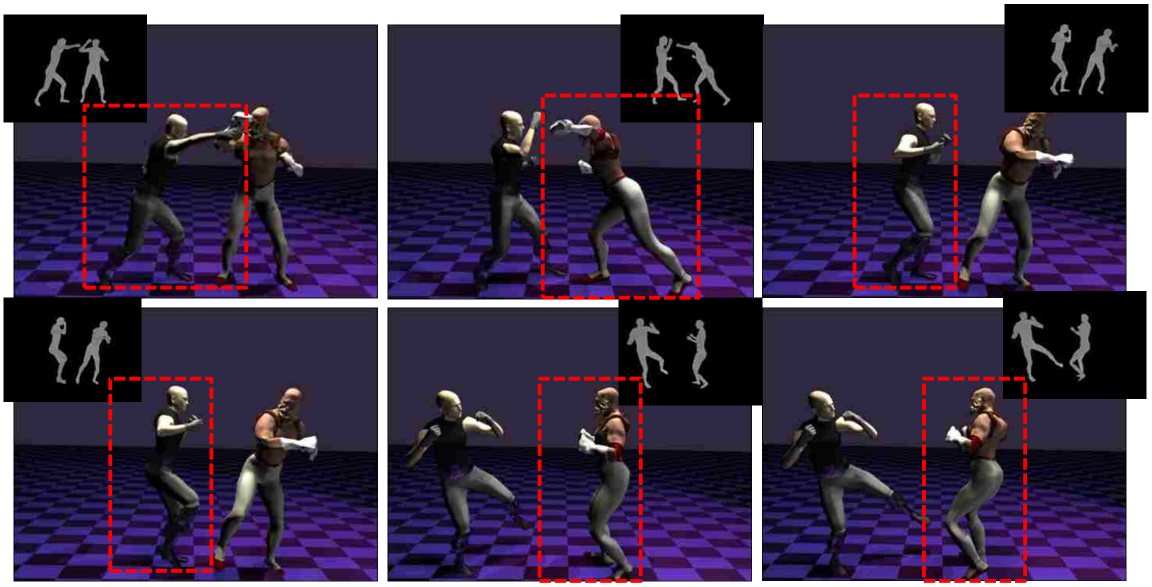


Figure 4.7: Some examples of extreme occlusion to our approach. The insets show the input depth maps; The recovered body configurations are used to drive the model.

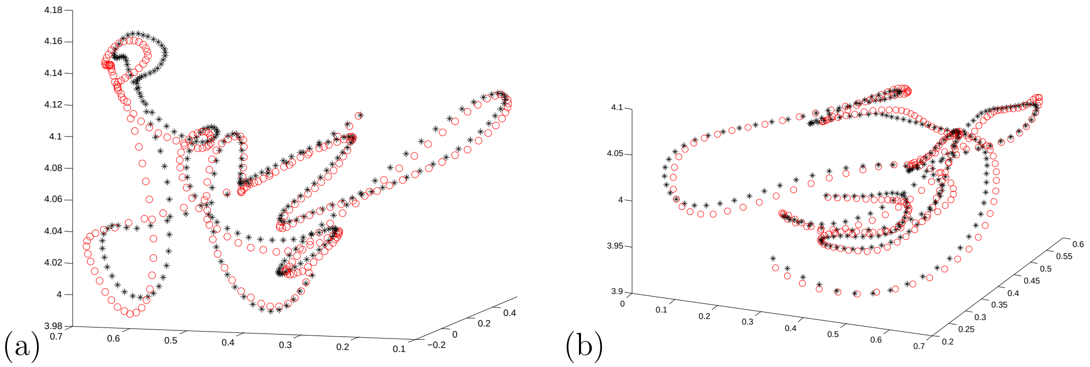


Figure 4.8: Quantitative error analysis in joint positions of the left elbow (a) and right elbow (b) for a stretching sequences (frames 0 to 269). *Black(*)*: Ground truth as obtained by a 8-camera motion capture system. *Red(o)*: the recovered body configuration sequence.

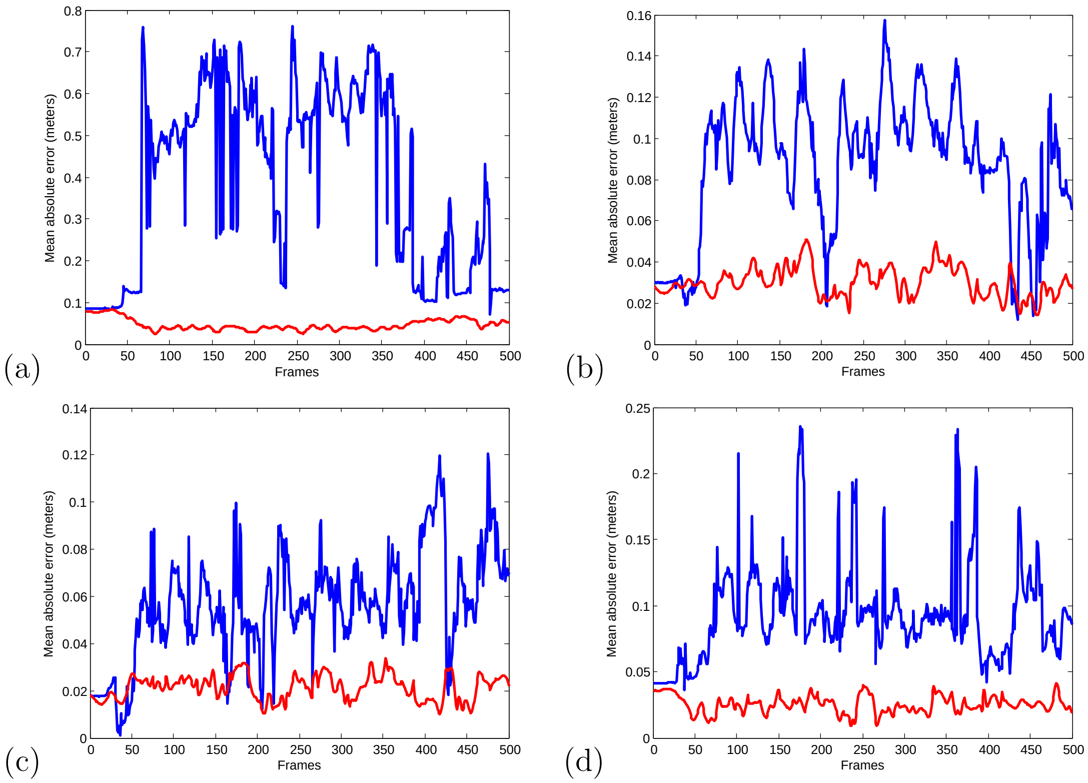


Figure 4.9: Error curves for PETR-only (blue) and the complete method (red), referring to the position absolute differences of the left forearm (a), right thigh (b), right leg (c), and left thigh (d). The Y axis is in meters, note that the Y scale of each subgraph is different.

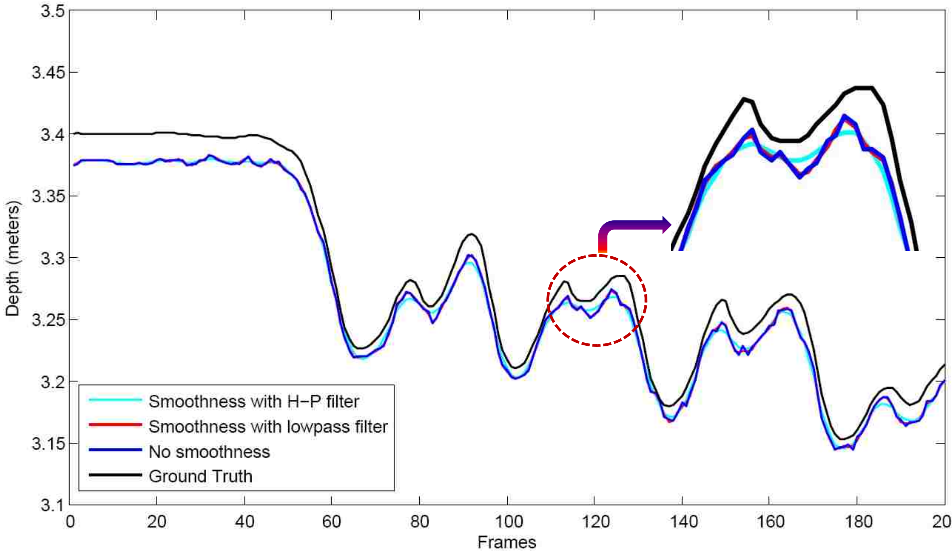


Figure 4.10: Smoothness results in depth values with H-P filter (red), with low-pass filter (green), and without smoothness. Ground-truth depth values are presented in solid black lines.

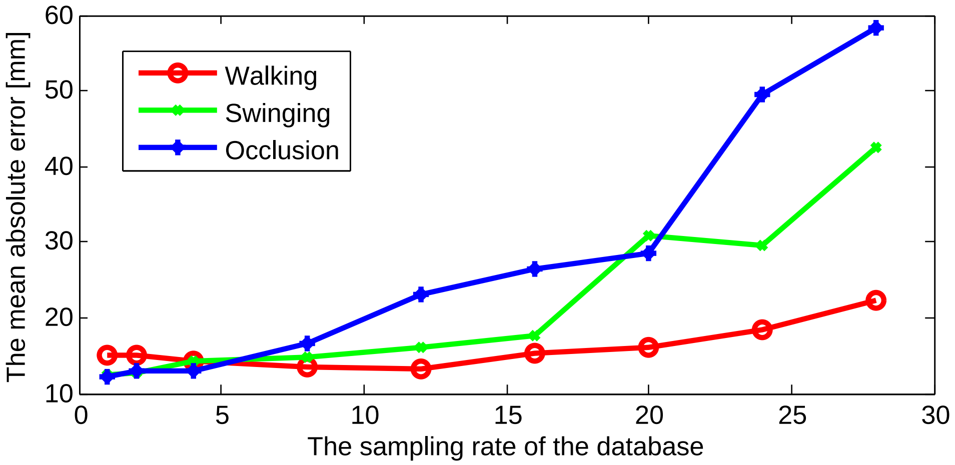


Figure 4.11: The reconstruction errors vs. the sampling rate in the database of the walking sequence.

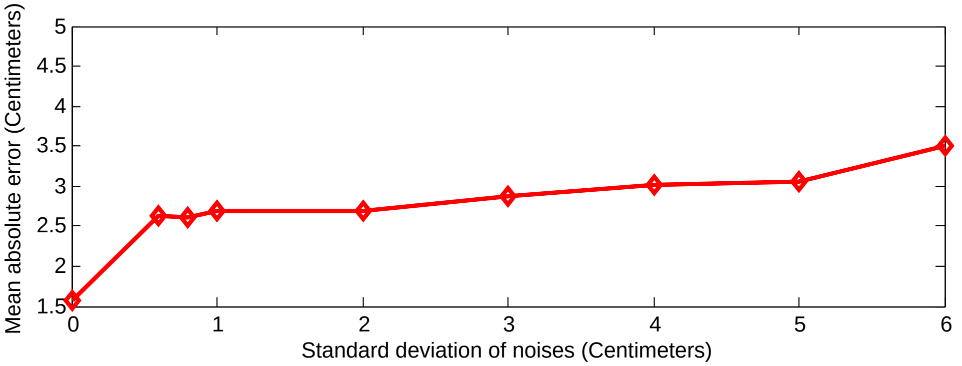


Figure 4.12: Motion reconstruction errors vs. the input noise with different levels.

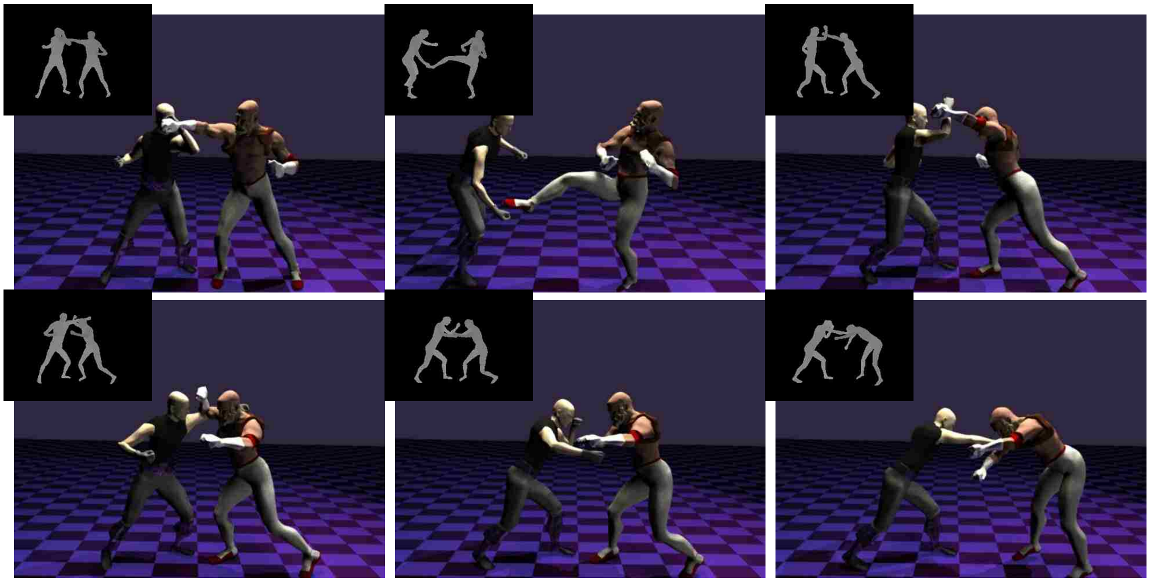


Figure 4.13: Some examples of fighting sequences reconstructed with our approach. The insets show the input depth maps; The recovered body configurations are used to drive the model.

Chapter 5 Conclusion and Future Work

In this thesis, we have developed new solutions to the single-view-reconstruction problem, concerning the modeling of both human face and motion. It has been demonstrated that prior models can be used to solve the under-constrained problem in SVR under very challenging conditions. This is because the prior knowledge restricts the space of possible solutions.

Based on this observation, we study the approaches in formulating the prior knowledge required to solve the ambiguities. We have developed an approach to learn the local models with non-linear dimensionality reduction in the problem of human face modeling, instead of directly relying on knowledge about the precise material properties of the target surfaces. The learned local model captures the underlying image and shape dynamics of both the image and that of the 3D shape. Specifically, the local image models for each patch of facial images and the local surface models for each patch of the 3D shape are learned using a non-linear dimensionality reduction technique. The correspondences between these local models are then learned by a manifold alignment method. By combining the local shapes, the global shape of a face can be reconstructed directly using a single least-square system of equations. Experimental results of real and synthetic data show that our approach can yield accurate shape recovery from out-of-training samples with a variety of pose and illumination variations.

Following the discussion of our method to reconstruct face models , we introduce a new model-based approach to deal with the ambiguities in the problem of human motion modeling. Previous learning-based approaches are not applicable to this SVR problem due to many challenges from the internal and external variations in single view video-based marker-less motion capture. We use the depth cue from a single depth sensor to reduce some challenges, i.e., appearance, illumination, and background; depth cues also provide us metric 3D information. However, using a single sensor, instead of a camera array, results in view-dependent and incomplete measurement of object motion. We have developed a novel two-stage template fitting algorithm that is invariant to subject size and view-point variations, and robust to occlusions. Starting from a known pose, our algorithm first estimates a body configuration through temporal registration, which is used to search the template motion database for a best match. The best match body configuration as well as the corresponding surface mesh model are deformed to fit the input depth map, filling in occluded parts from the input and compensating differences in pose, body-size, and height between the input image and the template. Our approach does not require any markers, user-interaction, or appearance-based tracking. Experiments demonstrate that our approach achieves an average motion tracking accuracy of 20mm, and is capable of dealing with severe occlusions even with depth images containing more than one person.

Future Work Even though the approaches, we have developed have advantages over the state-of-the-art in SVR, many further improvements are possible. One of the major limitations of our approach to human face modeling is a lack of details in the results of reconstruction due to the smoothness constraints in GP-LVM, as well as the unavailability of enough 3-D facial shapes. One solution requiring further investigation is the combination of our pipeline with a depth sensor. The sensor provides 3-D metric information, which may be enforced in our least-square recovery system.

In the work of human motion modeling we restrict ourselves to model the motions that are similar to those in the training database. The simplest approach is to use Inverse Kinematics as an online step. We can generate complete and partial surface models similar to the input, based on the estimated skeleton. Thus, we can remove the dependence on the training database and model the entire range of human motion.

In our current method, we treat each joint separately and assume that it is independent of the other joints. The recovery results may be improved by enforcing length constraints. Moreover, one problem with non-rigid registration concerns computation, which grows quickly with the number of 3-D points. We plan to investigate efficient methods so that dense point clouds can be used.

A main weakness of our approach is that pose estimation from temporal registration is very time consuming. A possible solution is to use sparse coding techniques, from which we can reconstruct the skeleton from a single depth image. Thus, we eliminate or reduce the dependency of temporal registration and do not have the drifting problem.

In recent years sparse coding has drawn considerable interest in signal processing. The assumption is that signals, such as images, can be decomposed into sparse linear combinations of atoms; these atoms being contained in a dictionary that consists of over-complete sets of vectors. Generally, the sparse representation (SR) of a signal is computed by optimizing an objective function with two items. One item measures the sparsity of the signal and the other measures the reconstruction error. Sparse representation is relatively robust against distortions and missing data in the depth images, and provides us a mechanism to deal with significant occlusions. We can formulate motion modeling from a single depth image as a machine-learning problem. There could be two stages: learning and estimation. In the learning stage, we can first capture a database of depth images containing a human subject in motion. From the training database, we can learn the dictionaries for depth images and motion models, respectively. These dictionaries capture the general frequent patterns and local structures in all training images and motions. Using the dictionary, a sparse representation of each depth image or motion model can be then computed via ℓ_1-regularized least squares techniques. In the estimation stage, for a given depth image, we can estimate its SR using the learned depth dictionary. The resulting SR is then approximated as a linear combination of its neighbors in the SR space. The weight coefficients can then be transferred to the motion model's SR space to recover the full motion.

Bibliography

[1] M. Reiter, R. Donner, G. Langs, and H. Bischof, "3d and infrared face reconstruction from rgb data using canonical correlation analysis," in *Proc. of Intl. Con. on Pattern Recognition*, 2006, pp. 425–428.

[2] M. Castelan and E. R. Hancock, "A simple coupled statistical model for 3d face shape recovery," in *Proc. of Intl. Con. on Pattern Recognition*, 2006, pp. 231–234.

[3] Z. Lei, Q. Bai, R. He, and S. Li, "Face shape recovery from a single image using cca mapping between tensor spaces," *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1–7, June 2008.

[4] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real time motion capture using a single time-of-flight camera," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2010.

[5] R. Hietmeyer, "Biometric identification promises fast and secure processing of airline passengers," *The Intl. Civil Aviation Organization Journal*, vol. 55, no. 9, pp. 10–11, 2000.

[6] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun, "Performance capture from sparse multi-view video," *ACM Trans. on Graphics*, vol. 27, no. 3, 2008.

[7] L. Gu and T. Kanade, "3d alignment of face in a single image," *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, pp. 1305–1312, June 2006.

[8] M. Salzmannn, R. Urtasun, and P. Fua, "Local deformation models for monocular 3d shape recovery," *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, June 2008.

[9] M. Salzmann, R. Urtasun, and P. Fua, "Local deformation models for monocular 3d shape recovery," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, June 2008, pp. 1 –8.

[10] M. J. Brooks and B. K. P. Horn, *Shape from shading*, B. K. P. Horn, Ed. Cambridge, MA, USA: MIT Press, 1989.

[11] A. P. Witkin, "Recovering surface shape and orientation from texture," *Artificial Intelligence*, vol. 17, pp. 17–45, 1981.

[12] M. Salzmann, R. Hartley, and P. Fua, "Convex optimization for deformable surface 3-d tracking," in *Proc. of Intl. Conf. on Computer Vision*, October 2007, pp. 1–8.

[13] M. Salzmann, V. Lepetit, and P. Fua, "Deformable surface tracking ambiguities," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

[14] M. Salzmann, J. Pilet, S. Ilic, and P. Fua, “Surface deformation models for nonrigid 3d shape recovery,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1481–1487, August 2007.

[15] X. Llado, A. D. Bue, and L. Agapito, “Non-rigid 3d factorization for projective reconstruction,” in *British Machine Vision Conf.*, September 2005.

[16] L. Torresani, A. Hertzmann, and C. Bregler, “Learning non-rigid 3d shape from 2d motion,” in *Advances in Neural Information Processing Systems*, 2003.

[17] J. Xiao and T. Kanade, “Uncalibrated perspective reconstruction of deformable structures,” in *Proc. of Intl. Conf. on Computer Vision*, vol. 2, October 2005, pp. 1075–1082.

[18] V. Blanz and T. Vetter, “A morphable model for the synthesis of 3d faces,” in *Proc. of ACM SIGGRAPH*, 1999, pp. 187–194.

[19] I. Matthews and S. Baker, “Active appearance models revisited,” *Int. J. of Computer Vision*, vol. 60, pp. 135–164, 2003.

[20] M. Salzmann, S. Ilic, and P. Fua, “Physically valid shape parameterization for monocular 3-d deformable surface tracking,” in *British Machine Vision Conf.*, September 2005.

[21] D. Metaxas and D. Terzopoulos, “Constrained deformable superquadrics and nonrigid motion tracking,” in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 1991, pp. 337–343.

[22] L. Cohen and I. Cohen, "Deformable models for 3-d medical images using finite elements and balloons," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, June 1992, pp. 592–598.

[23] T. McInerney and D. Terzopoulos, "A finite element model for 3d shape reconstruction and nonrigid motion tracking," in *Proc. of Intl. Conf. on Computer Vision*, 1993, pp. 518–523.

[24] A. P. Pentland, "Automatic extraction of deformable part models," *Int. J. of Computer Vision*, vol. 4, no. 2, pp. 107–126, 1990.

[25] H. Delingette, M. Hebert, and K. Ikeuchi, "Deformable surfaces: A free-form shape representation," in *SPIE Geometric Methods in Computer Vision*, vol. 1570, 1991, pp. 21–30.

[26] K. S. Bhat, C. D. Twigg, J. K. Hodgins, P. K. Khosla, Z. Popović, and S. M. Seitz, "Estimating cloth simulation parameters from video," in *2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, July 2003, pp. 37–51.

[27] L. Tsap, D. Goldof, and S. Sarkar, "Nonrigid motion analysis based on dynamic refinement of finite element models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 5, pp. 526–543, May 2000.

[28] M. Brand, "Morphable 3d models from video," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, June 2001, pp. 456–463.

[29] L. Torresani, A. Hertzmann, and C. Bregler, "Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 878–892, 2008.

[30] R. Hartley and R. Vidal, "Perspective nonrigid shape and motion recovery," in *Proc. of Europ. Conf. on Computer Vision*, 2008, pp. 276–289.

[31] T. F. Cootes, G. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[32] M. Salzmann, J. Pilet, S. Ilic, and P. Fua, "Surface deformation models for nonrigid 3d shape recovery," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1481–1487, 2007.

[33] W. Y. Zhao and R. Chellappa, "Symmetric shape-from-shading using self-ratio image," *Int. J. of Computer Vision*, vol. 45, no. 1, pp. 55–75, 2001.

[34] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning, with application to clustering with side-information," in *Advances in Neural Information Processing Systems*, 2002, pp. 505–512.

[35] T. Hastie and R. Tibshirani, "Discriminant adaptive nearest neighbor classification," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, no. 6, pp. 607–616, 1996.

[36] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, “Learning a mahalanobis metric from equivalence constraints,” *Journal of Machine Learning Research*, vol. 6, pp. 937–965, 2005.

[37] J. Goldberger, S. T. Roweis, G. E. Hinton, and R. Salakhutdinov, “Neighbourhood components analysis,” in *Advances in Neural Information Processing Systems*, 2004.

[38] K. Weinberger, J. Blitzer, and L. Saul, “Distance metric learning for large margin nearest neighbor classification,” in *Advances in Neural Information Processing Systems*, 2006, pp. 1473–1480.

[39] S. C. H. Hoi, W. Liu, M. R. Lyu, and W. ying Ma, “Learning distance metrics with contextual constraints for image retrieval,” in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2006, pp. 2072–2078.

[40] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of Eugenics*, vol. 7, no. 7, pp. 179–188, 1936.

[41] M. Sugiyama, “Local fisher discriminant analysis for supervised dimensionality reduction,” in *Proc. of Intl. Conf. on Machine learning*, 2006.

[42] S.-J. Kim, K. Koh, M. Lustig, and S. Boyd, “An efficient method for compressed sensing,” *IEEE Intl. Conf. on Image Processing*, vol. 3, pp. 117–120, October 2007.

[43] A. Globerson and S. Roweis, “Metric learning by collapsing classes,” in *Advances in Neural Information Processing Systems*, vol. 18, 2006, pp. 451–458.

[44] L. Yang, R. Jin, R. Sukthankar, and Y. Liu, "An efficient algorithm for local distance metric learning," in *Proc. of National conf. on Artificial intelligence*, 2006, pp. 543–548.

[45] L. Yang, R. Jin, and R. Sukthankar, "Bayesian active distance metric learning," in *Proc. of Annual Conference on Uncertainty in Artificial Intelligence*, 2007, pp. 442–449.

[46] M. Schultz and T. Joachims, "Learning a distance metric from relative comparisons," in *Advances in Neural Information Processing Systems*, 2004.

[47] H. C. Hongch and D. yan Yeung, "Locally linear metric adaptation for semi-supervised clustering," in *Proc. of Intl. Conf. on Machine learning*, 2004, pp. 153–160.

[48] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[49] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[50] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[51] N. Hu, W. Huang, and S. Ranganath, “Head pose estimation by non-linear embedding and mapping,” in *IEEE Intl. Conf. on Image Processing*, vol. 2, 2005, pp. 342–345.

[52] B. Raytchev, I. Yoda, and K. Sakaue, “Head pose estimation by nonlinear manifold learning,” in *Proc. of Intl. Con. on Pattern Recognition*, 2004, pp. 462–466.

[53] X. He and P. Niyogi, “Locality preserving projections,” in *Advances in Neural Information Processing Systems*, 2003.

[54] L. Chen, L. Zhang, Y. Hu, M. Li, and H. Zhang, “Head pose estimation using fisher manifold learning,” in *Proc. of the IEEE Intl. Workshop on Analysis and Modeling of Faces and Gestures*, 2003, pp. 203–207.

[55] Y. Fu and T. S. Huang, “Graph embedded analysis for head pose estimation,” in *Proc. of Intl. Conf. on Automatic Face and Gesture Recognition*, 2006, pp. 3–8.

[56] V. Balasubramanian, J. Ye, and S. Panchanathan, “Biased manifold embedding: A framework for person-independent head pose estimation,” in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2007, pp. 1–7.

[57] D. D. Ridder, O. Kouropteva, O. Okun, M. Pietikainen, and R. Duin, “Supervised locally linear embedding,” in *Intl. Conf. on Artificial Neural Networks and Neural Information Processing*, vol. 2714, 2003, pp. 333–341.

[58] C.-G. Li and J. Guo, “Supervised isomap with explicit mapping,” in *IEEE Intl. Conf. on Innovative Computing, Information and Control*, 2006.

[59] M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios, and N. Koudas, “Non-linear dimensionality reduction techniques for classification and visualization,” in *Intl. Conf. on Knowledge Discovery and Data Mining*, 2002, pp. 645–651.

[60] X. Geng, D.-C. Zhan, and Z.-H. Zhou, “Non-linear dimensionality reduction techniques for classification and visualization,” *IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 35, no. 6, pp. 1098–1107, 2005.

[61] Q. Zhao, D. Zhang, and H. Lu, “Supervised lle in ica space for facial expression recognition,” in *Intl. Conf. on Neural Networks and Brain*, vol. 3, 2005, pp. 1970–1975.

[62] V. N. Balasubramanian, S. Krishna, and S. Panchanathan, “Person-independent head pose estimation using biased manifold embedding,” *EURASIP Journal on Advances in Signal Processing*, 2007.

[63] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, “Shape from shading: A survey,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 8, pp. 690–706, 1999.

[64] I. Shimshoni, Y. Moses, and M. Lindenbaum, “Shape reconstruction of 3d bi-laterally symmetric surfaces,” *Int. J. of Computer Vision*, vol. 39, pp. 97–110, 2000.

[65] E. Prados and O. Faugeras, “Shape from shading: a well-posed problem?” *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 870–877, June 2005.

[66] E. Prados, F. Camilli, and O. Faugeras, “A unifying and rigorous shape from shading method adapted to realistic data and applications,” *J. Math. Imaging Vis.*, vol. 25, no. 3, pp. 307–328, 2006.

[67] I. Kemelmacher and R. Basri, “Molding face shapes by example,” *Proc. of Europ. Conf. on Computer Vision*, pp. 277–288, 2006.

[68] T. Sim and T. Kanade, “Combining models and exemplars for face recognition: An illuminating example,” in *Proc. of the CVPR 2001, Workshop on Models on Models versus Exemplars in Computer Vision*, December 2001.

[69] J. J. Atick, P. A. Griffin, and A. N. Redlich, “Statistical approach to shape from shading: Reconstruction of 3d face surfaces from single 2d images,” *Neural Computation*, vol. 8, pp. 1321–1340, 1996.

[70] W. A. P. Smith and E. R. Hancock, “Recovering facial shape and albedo using a statistical model of surface normal direction,” in *Proc. of Intl. Conf. on Computer Vision*, 2005, pp. 588–595.

[71] R. Dovgard and R. Basri, “Statistical symmetric shape from shading for 3d structure recovery of faces,” in *Proc. of Europ. Conf. on Computer Vision*, 2004, pp. 108–116.

[72] V. Blanz and T. Vetter, “A morphable model for the synthesis of 3d faces,” in *Proc. of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194.

[73] S. Romdhani and T. Vetter, “Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior,” in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2005, pp. 986–993.

[74] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, “From few to many: Generative models for recognition under variable pose and illumination,” in *Proc. of Intl. Conf. on Automatic Face and Gesture Recognition*, 2000, pp. 277–284.

[75] T. B. Moeslund, A. Hilton, and V. Krüger, “A survey of advances in vision-based human motion capture and analysis,” *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 90–126, 2006.

[76] R. Poppe, “Vision-based human motion analysis: An overview,” *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 4–18, 2007.

[77] A. Agarwal and B. Triggs, “Recovering 3d human pose from monocular images,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 1, pp. 44–58, 2006.

[78] H. Ning, W. Xu, Y. Gong, and T. Huang, “Discriminative learning of visual words for 3d human pose estimation,” in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[79] C. Sminchisescu, A. Kanaujia, and D. Metaxas, "bm^3e : Discriminative density propagation for visual tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 11, pp. 2030–2044, 2007.

[80] K. Grauman, G. Shakhnarovich, and T. Darrell, "Inferring 3d structure with a statistical image-based shape model," in *Proc. of Intl. Conf. on Computer Vision*, vol. 1, 2003, pp. 641–647.

[81] A. Elgammal and C.-S. Lee, "Inferring 3d body pose from silhouettes using activity manifold learning," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, 2004, pp. 681–688.

[82] G. Shakhnarovich, P. Viola, and T. Darrell, "Fast pose estimation with parameter-sensitive hashing," in *Proc. of Intl. Conf. on Computer Vision*, vol. 2, 2003, pp. 750–757.

[83] A. Fathi and G. Mori, "Human pose estimation using motion exemplars," in *Proc. of Intl. Conf. on Computer Vision*, 2007, pp. 1–8.

[84] L. Ren, G. Shakhnarovich, J. K. Hodgins, H. Pfister, and P. Viola, "Learning silhouette features for control of human motion," *ACM Trans. on Graphics*, vol. 24, no. 4, pp. 1303–1331, 2005.

[85] J. Deutscher, A. Blake, and I. Reid, "Articulated body motion capture by annealed particle filtering," in *Proc. of Intl. Conf. on Computer Vision*, vol. 2, 2000, pp. 126–133.

[86] C. Sminchisescu, A. Kanaujia, Z. Li, and D. N. Metaxas, "Conditional visual tracking in kernel space," in *Advances in Neural Information Processing Systems*, 2005, p. 11.

[87] C. Sminchisescu and A. Jepson, "Generative modeling for continuous non-linearly embedded visual inference," in *Proc. of Intl. Conf. on Machine learning*, 2004, p. 96.

[88] R. Urtasun, D. J. Fleet, A. Hertzmann, and P. Fua, "Priors for people tracking from small training sets," in *Proc. of Intl. Conf. on Computer Vision*, vol. 1, 2005, pp. 403–410.

[89] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models," in *Advances in Neural Information Processing Systems*, 2006, pp. 1441–1448.

[90] S. Hou, A. Galata, F. Caillette, N. Thacker, and P. Bromiley, "Real-time body tracking using a gaussian process latent variable model," in *Proc. of Intl. Conf. on Computer Vision*, 2007, pp. 1–8.

[91] C. Sminchisescu, A. Kanaujia, and D. Metaxas, "Learning joint top-down and bottom-up processes for 3d visual inference," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 1743–1752.

[92] D. Grest, V. Krger, and R. Koch, "Single view motion tracking by depth and silhouette information," in *Scandinavian Conference on Image Analysis*, 2007.

[93] Y. Pekelny and C. Gotsman, “Articulated object reconstruction and markerless motion capture from depth video,” *Computer Graphics Forum*, vol. 27, no. 2, pp. 399–408, 2008.

[94] H. Wang, L. Dong, J. O’Daniel, R. Mohan, A. S. Garden, K. K. Ang, D. A. Kuban, M. Bonnen, J. Y. Chang, and R. Cheung, “Validation of an accelerated ’demons’ algorithm for deformable image registration in radiation therapy,” *Physics in Medicine and Biology*, vol. 50, pp. 2887–2905, 2005.

[95] T. Riklin-Raviv and A. Shashua, “The quotient image: Class based recognition and synthesis under varying illumination conditions,” in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 1999, p. 2566.

[96] A. Myronenko, X. Song, and M. A. . Carreira-Perpin.an, “Non-rigid point set registration: Coherent point drift,” in *Advances in Neural Information Processing Systems*, 2006.

[97] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[98] A. Myronenko and X. B. Song, “On the closed-form solution of the rotation matrix arising in computer vision problems,” *CoRR*, 2009.

[99] R. W. Sumner and J. Popović, “Deformation transfer for triangle meshes,” in *Proc. of ACM SIGGRAPH*, 2004, pp. 399–405.

[100] N. D. Lawrence, "Gaussian process latent variable models for visualization of high dimensional data," in *Advances in Neural Information Processing Systems*, 2004.

[101] D. J. C. MacKay, "Introduction to gaussian processes," in *Neural Networks and Machine Learning*, ser. NATO ASI Series. Kluwer Academic Press, 1998, pp. 133–166.

[102] N. D. Lawrence, "Learning for larger datasets with the gaussian process latent variable model," in *AISTATS*, 2007.

[103] C. Wang and S. Mahadevan, "Manifold alignment using procrustes analysis," in *Proc. of Intl. Conf. on Machine learning*, 2008, pp. 1120–1127.

[104] W. Zhao, R. Chellappa, A. Rosenfeld, and P. Phillips, "Face recognition: A literature survey," 2003.

[105] Y. Adini, Y. Moses, and S. Ullman, "Face recognition: The problem of compensating for changes in illumination direction," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 721–732, 1997.

[106] Y. Hu, L. Chen, Y. Zhou, and H. Zhang, "Estimating face pose by facial asymmetry and geometry," in *Proc. of Intl. Conf. on Automatic Face and Gesture Recognition*, 2004, pp. 651–656.

[107] S. O. Ba and J.-M. Odobez, "A probabilistic framework for joint head tracking and pose estimation," in *Proc. of Intl. Con. on Pattern Recognition*, 2004, pp. 264–267.

[108] S. Z. Li, X. Lu, X. Hou, X. Peng, and Q. Cheng, "Learning multiview face subspaces and facial pose estimation using independent component analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 14, no. 6, pp. 705–712, 2005.

[109] S. McKenna and S. Gong, "Real-time face pose estimation," *Real-Time Imaging, Special Issue on Visual Monitoring and Inspection*, vol. 4, no. 5, pp. 333–347, 1998.

[110] M. Bichsel and A. Pentland, "Automatic interpretation of human head movements," in *International Joint Conference on Artificial Intelligence, Workshop on Looking At People*, 1993.

[111] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, pp. 269–271, 1959.

[112] M.-H. Yang, "Extended isomap for classification," *Proc. of Intl. Con. on Pattern Recognition*, vol. 3, p. 30615, 2002.

[113] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Advances in Neural Information Processing Systems*, 2004.

[114] D. F. Specht, "A general regression neural network," *Neural Networks, IEEE Transactions on*, vol. 2, no. 6, pp. 568–576, November 1991.

[115] M. Tipping, "The relevance vector machine," in *Advances in Neural Information Processing Systems*, 2000.

[116] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression (pie) database of human faces," Robotics Institute, Carnegie Mellon University, Tech. Rep., January 2001.

[117] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "Scape: shape completion and animation of people," *ACM Trans. on Graphics*, vol. 24, no. 3, pp. 408–416, 2005.

[118] I. Baran and J. Popović, "Automatic rigging and animation of 3d characters," *ACM Trans. on Graphics*, vol. 26, no. 3, p. 72, 2007.

[119] A. Myronenko and X. Song, "Point set registration: Coherent point drift," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 99, 2010.

[120] D. G. Kendall, "A survey of the statistical theory of shape," *Statistical Science*, vol. 4, no. 2, pp. 87–99, 1989.

[121] M. Botsch, "On linear variational surface deformation methods," *IEEE Trans. on Visualization on Computer Graphics*, vol. 14, pp. 213–230, 2007.

[122] D. Vlasic, I. Baran, W. Matusik, and J. Popović, "Articulated mesh animation from multi-view silhouettes," *ACM Trans. on Graphics*, vol. 27, no. 3, pp. 1–9, 2008.

[123] J. Gall, C. Stoll, E. de Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel, "Motion capture using joint skeleton tracking and surface estimation," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2009, pp. 1746–1753.

[124] R. J. Hodrick and E. C. Prescott, “Postwar u.s. business cycles: An empirical investigation,” *Journal of Money, Credit and Banking*, vol. 29, no. 1, pp. 1–16, February 1997.

[125] M. Ye, X. Wang, and R. Yang, “Accurate and robust skeletal motion capture from a single depth sensor,” Department of Computer Science, University of Kentucky, Tech. Rep. TR-514-10, 2010.

VITA

Xianwang Wang

- Birth Date: 13rd Oct 1973
- Birth Place: China

Education

- **2005-2010**, PhD Candidate in Computer Science, University of Kentucky. Research Area: Computer vision, Machine learning, and Image processing
- **1998-2001**, Master of Engineering, Computer Science, Sichuan University, China. Research Area: Pattern Recognition and Artificial Intelligence.
- **1994-1998**, Bachelor of Engineering, Computer Science, Sichuan University, China.

Pending Patents

1. **Xianwang Wang** and Jing Xiao, "Context Constrained Novel View Interpolation," U.S. Provisional Patent Application (Serial No.: 61/262,015).

2. **Xianwang Wang**, Mao Ye, and Ruigang Yang, "Markerless Human Motion Capture from a Single Depth Sensor via Sparse Coding," Provisional Patent Application by University of Kentucky (INV09/1676).

Publications

1. **Xianwang Wang** and R. Yang. "Learning 3D Shape from a Single Facial Image via Non-linear Manifold Embedding and Alignment," IEEE conference on Computer Vision and Pattern Recognition (CVPR), 2010.

2. **Xianwang Wang**, Q. Zhang, Q. Han, R. Yang, M. Carswell, B. Seales, and E. Sutton. "Endoscopic Video Texture Mapping on Pre-built 3D Anatomical Objects without Camera Tracking," Journal of IEEE Transactions on Medical Imaging (TMI), 2010.

3. **Xianwang Wang**, X. Huang, J. Gao, and R. Yang. "Illumination and Person-Insensitive Head Pose Estimation Using Distance Metric Learning," European Conference on Computer Vision (ECCV), 2008.

4. **Xianwang Wang**, Q. Zhang, R. Yang, B. Seales, and M. Carswell,. "Feature-based Texture Mapping from Video Sequence," Symposium on Interactive 3D Graphics and Games (i3D), 2008.

5. X. Huang, **Xianwang Wang**, J. Gao, and R. Yang. "Estimating Pose and Illumination Direction for Frontal Face Synthesis," IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Biometrics Workshop, 2008.

6. Y. Wang, K. Liu, Q. Hao, **Xianwang Wang**, D. Lau and L. Hassebrook.. "Hybrid Structured Light Illumination and Stereo Vision 3-D Reconstruction," Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010(under 2nd revision).

7. **Xianwang Wang**, M. Ye, R. Yang, and L. Ren, "Accurate and Robust Skeletal Motion Capture from a Single Depth Sensor," submitted to IEEE Transactions on Visualization and Computer Graphics (TVCG), 2010.