Wavelet Model-based Stereo for Fast, Robust Face Reconstruction

Alan Brunton, Jochen Lang, Eric Dubois
School of Information Technology and Engineering
University of Ottawa
Ottawa, Canada
abrunton@site.uottawa.ca

Chang Shu
Institute of Information Technology
National Research Council
Ottawa, Canada
Chang.Shu@nrc-cnrc.gc.ca

***Abstract*—When reconstructing a specific type or class of object using stereo, we can leverage prior knowledge of the shape of that type of object. A popular class of object to reconstruct is the human face. In this paper we learn a statistical wavelet prior of the shape of the human face and use it to constrain stereo reconstruction within a Bayesian framework. We initialize our algorithm with a, typically noisy, point cloud from a standard stereo algorithm, and search our parameter space for the shape that best fits the point cloud. Due to the wavelet basis, our shape parameters can be optimized independently, thus simplifying and accelerating the search. We follow this by optimizing for a secondary prior and observation: smoothing and photoconsistency. Our method is fast, and is robust to noise and outliers. Additionally, we obtain a shape in an parameterized and corresponded shape space, making it ready for further processing such as tracking, recognition or statistical analysis.**

***Keywords*-model-based stereo, wavelet prior, Bayesian framework, graphics processing unit (GPU)**

I. Introduction

Stereo reconstruction of class-specific objects, eg. human faces, may benefit from prior knowledge of the shape of the objects to be reconstructed. The prior models, learned from statistical analysis of 3D shapes, constrain the reconstruction, alleviating some of the most difficult problems in stereo such as occlusion and specularity. In practice, it is necessary to build the prior models in a compact representation so that they can be used efficiently to infer 3D shapes from images. Principal component analysis (PCA) has been widely used for this purpose. For example, Amberg et al. [1] learn a PCA model from 3D scans and vary the model to best fit the input images. However, PCA is a global model, which means every parameter affects the whole shape. For stereo, it is desirable to have a prior model that only influences the shape variations locally. We propose a statistical wavelet model for representing the shape variations and use it in a Bayesian framework to reconstruct human faces from two or more images.

Statistical shape priors have been used within single-view reconstruction techniques, eg. Blanz and Vetter [2]. Using multiple views gives us an observation that expresses the likelihood of a shape based on 3D geometric constraints, i.e. perspective correspondence between the views, as well as surface appearance.

Stereo reconstruction can be broken into binocular stereo and multi-view stereo. Although we only use two images, our approach is more along the lines of multi-view approaches and could straightforwardly be extended to more views. Scharstein and Szeliski [3] provided a survey of the binocular case, and Seitz et al. [4] surveyed and classified the multi-view techniques.

Amberg et al. [1] used model-based stereo to reconstruct human faces. They used a 3D morphable model (3DMM) [2] based on PCA of a training set of faces. In such a framework, shapes are modeled as the mean shape plus a linear combination of eigenvectors which represent eigen shape variation modes, computed from the training set. The mean shape was computed as the mean points' coordinates. PCA is a transform with global support; each shape parameter depends on each vertex position, and vice versa. This means that all parameters have to be optimized together, typically in a complex gradient descent fashion, which may get stuck in a local minimum. In contrast, we use a wavelet basis that has localized support in both space and frequency. This makes the transform computationally efficient, both in terms of time and space, and allows us to use a simple global optimization algorithm. The wavelet basis also affords us greater variability; because the model parameters are independent, we can generate shapes not found in the linear subspace spanned by the training set.

Recently several methods for high-quality stereo capture of human faces have been proposed [5], [6], [7]. While we do not reconstruct with the same precision as some of those methods, we do not require special face-paint, or near-optimal lighting conditions or high-accuracy calibration. Our approach is computationally efficient and can be orders of magnitude faster than these methods. Our approach can incorporate any range information, i.e. our method is not restricted to stereo data. Additionally, we obtain a face mesh that is in a common, corresponded shape space, making our approach ideal for tracking and recognition.

When optimizing a model to best fit a set of observations, it is desirable that the representation be compact, meaning that most of the energy resides in a few coefficients. Thus only those relatively few parameters need to be manipulated

to fit the model. PCA, Fourier and spherical harmonic representations are compact, but the basis functions are not localized in space, hence the coefficients in one scale have to be optimized together, making the optimization problem complex. By using a wavelet basis we are able to solve the optimization in a simple divide-and-conquer approach. Li et al. [8], performed segmentation of 3D Neuroradiological data using a statistical wavelet prior similar to the one we use here. In both cases the model is a generalized B-spline subdivision wavelet [9].

Sun et al. [10] used strong shape priors for stereo. They used an object-specific Markov random field (MRF) to integrate shape priors seamlessly with a weak prior on surface smoothness for articulated and deformable objects. Romeiro and Zickler [11] coupled a 3DMM with an occlusion map defined on the model shape to reconstruct faces in the presence of occluding objects. While we do not handle occluding objects, we can extend our method to incorporate additional objects, and account for the occlusions with little effort. Tonko and Nagel [12] used model-based stereo of non-polyhedral objects for automatic disassembly tasks. Zhao et al. [13] performed stereo reconstruction by first fitting an approximate global parametric and then refining the model using local correspondence processes. Koterba et al. [14] studied the relationship between multi-view Active Appearance Model (AAM) fitting and camera calibration.

Wavelets and other multi-resolution techniques have previously been used for stereo for over a decade, but most often [15], [16], [17], [18], [19], [20] a wavelet transform is applied to the images and then the resulting wavelet coefficients are matched in a coarse-to-fine manner, which allows for larger displacements between corresponding pixels. Miled et al. [21] used a wavelet domain representation of the disparity map to regularize stereo reconstruction, but their prior is an edge-preserving smoothing prior, as opposed to a statistical shape prior.

Wavelet shape priors, have been used previously for medical image segmentation [8], [22], and for single-view reconstruction [23], [24].

In this paper we make the following contributions: the use of a statistical wavelet shape prior for fast, robust model-based stereo reconstruction of human faces; a sampling-based Bayesian framework that can incorporate arbitrarily many priors and observations; the re-parameterization of human face scans with a subdivision sampling compatible with the wavelet model. Our approach is robust with respect to noisy observations and works under sub-optimal lighting conditions. While we demonstrate our technique for faces, we emphasize that we can generalize it to any shape that is topologically equivalent to a sphere. The reconstructed shape is captured in a common, registered shape space making it immediately ready for other processing, such as tracking and recognition. Additionally, we leverage a GPU-based implementation that accelerates the bottlenecks in our pipeline.

II. Overview

We begin by learning a statistical wavelet model of the human face, and then use it to robustly fit the model to noisy stereo data followed by stereo matching refinement. Given a database of corresponded triangular meshes of laser scans of human faces, for each face we resample it into a subdivision surface, and then decompose the surface using a wavelet basis [9] into independent components that are localized in space and frequency. We then learn statistics on the distributions of the resulting wavelet coefficients over a training set, and then use them as a statistical prior to guide stereo reconstruction within a Bayesian framework.

Formally, we parameterize the shape of the human face by a high-dimensional vector space $\mathcal{S}$, and learn a model for the prior probability $P(\mathbf{s})$ of a shape parameter vector $\mathbf{s} \in \mathcal{S}$. We then model the observational likelihood of a set of input images $\mathcal{I}$ and a point cloud from a general-purpose stereo algorithm $\mathcal{Y} = \{\mathbf{y}_i : \ i = 0, \ldots, N_y - 1\}$, and we solve for the *maximum a posteriori* (MAP) configuration of $\mathbf{s}$ given $\mathcal{I}$ and $\mathcal{Y}$,

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s}} P(\mathbf{s}|\mathcal{I}, \mathcal{Y}) \tag{1}$$

where $\mathcal{I} = \{I_L, I_R\}$ for the purposes of this chapter. By Bayes' Theorem we have the posterior

$$P(\mathbf{s}|\mathcal{I}, \mathcal{Y}) = \frac{P(\mathbf{s})P(\mathcal{Y}|\mathbf{s})P(\mathcal{I}|\mathbf{s}, \mathcal{Y})}{P(\mathcal{Y})P(\mathcal{I}|\mathcal{Y})} \tag{2}$$

which, because $\mathcal{I}$ and $\mathcal{Y}$ are constant, simplifies to

$$P(\mathbf{s}|\mathcal{I}, \mathcal{Y}) = c_{obs} P(\mathbf{s}) P(\mathcal{Y}|\mathbf{s}) P(\mathcal{I}|\mathbf{s}, \mathcal{Y}) \tag{3}$$

where c_{obs} is a constant of proportionality. We further factor the prior into a component based on the statistics of the training set and a smoothing component: $P(\mathbf{s}) = P_{st}(\mathbf{s})P_{sm}(\mathbf{s})$. More details of the model are given in Section III. We compute the MAP configuration through energy minimization as described in Section IV. Our straightforard optimization technique, which is a combination of Monte Carlo sampling or particle filtering methods and iterative partial maximization, is made feasible by the properties of the wavelet basis. We implement our technique using GPU programming and standard image processing tools in a framework that allows the incorporation additional observations and priors as described in Section V, where we also present results on stereo data and low-resolution, relatively noisy laser range scans.

III. The Model

We model the surface we wish to reconstruct as a wavelet decomposition of a subdivision surface. Although we map the face to a plane, we use a second-generation subdivision wavelet scheme [9] that allows our model to be extended to any surface that can be mapped onto the unit sphere.

In the learning phase of our method, we start with a database or training set of triangular meshes of laser-scanned faces. However, these meshes are not subdivision surfaces, which prevents the use of a wavelet model, so we resample them onto a Catmull-Clark subdivision grid by stereographic projection of a template face. Corresponding vertices of all faces are mapped to the same point on the plane, that of the template, to preserve correspondence. We then decompose the surface into its wavelet coefficients. This, in turn, allows us to compute a statistical model of the wavelet coefficients generated from a database of such scans, and use it as a strong statistical prior to guide surface estimation.

Before we can proceed we must rigidly align the shapes in the training set, i.e. put them all in the same coordinate system, so that the variation in the 3D coordinates of the vertices is due only to the change in the shape, and not to any rotation, translation or scaling. Hence, we first align the triangular meshes with each other using generalized Procrustes alignment (GPA) [25], which iteratively aligns each shape in the set to the average shape of the set. After each iteration the average shape is recomputed using the realigned shapes. We then rigidly align the resulting mean face to the template mesh.

A. Subdivision Resampling

The triangular meshes are resampled onto a quadrilateral Catmull-Clark subdivision surface configured as a regular 2D grid as follows. We stereographically project the template mesh onto a plane aligned with the front of the face and passing through its centroid, saving the mapping as 2D coordinates in the plane. Let this plane be at $z = 1$. Stereographic projection maps the entire surface of a sphere to a plane, mapping a 3D point to the point in the plane at $z = 1$ that is passed through by the line-segment connecting the 3D point and the point $[0\ 0\ -1]^T$. Let $\mathbf{p}_n$ be an arbitrary vertex in the template triangular mesh. This vertex is mapped to a point in the plane $\mathbf{x}_n$ by stereographic projection as follows,

$$\mathbf{x}_n = \begin{bmatrix} x_n/(z_n+1) \\ y_n/(z_n+1) \end{bmatrix}$$

where $\mathbf{p}_n = [x_n\ y_n\ z_n]^T$. Let $\mathbf{p}_{i,n}$ be the same vertex in triangular mesh i from the training set; $\mathbf{p}_{i,n}$ is mapped to the same position in the plane $\mathbf{x}_n$ as is the template vertex $\mathbf{p}_n$, for every mesh i in the training set. In this way, corresponding vertices are always mapped to the same position in the plane, and because all meshes have the same connectivity and all their faces are planar, any point in the plane that is covered by applying this mapping to any one of the meshes corresponds to the same point on any of the other meshes under the same mapping. Thus, we can resample the surfaces while maintaining the correspondence between surfaces.

The planar coordinates are used as texture coordinates to resample the triangular mesh using the rasterization capabilities of the GPU. The locations of grid points are chosen by defining an orthographic projection of the texture coordinates using the rectangle bounding the mesh in the plane. The resolution of the grid is determined by taking an arbitrary base resolution and subdividing it the desired number of times.

B. Wavelet Decomposition

Let us denote the surface by $\mathbf{f} : \mathbb{R}^2 \rightarrow \mathbb{R}^3$, and further by $\mathbf{f}(\mathbf{x})$ at grid point $\mathbf{x} = (x, y)$. The wavelet model is then expressed by

$$\mathbf{f}(\mathbf{x}) = \sum_{n \in V(0)} \mathbf{v}_n^0 \phi_n^0(\mathbf{x}) + \sum_{j=0}^{N} \sum_{m \in W(j)} \mathbf{w}_m^j \psi_m^j(\mathbf{x}) \quad (4)$$

where the terms are defined as follows. The set of vertices in the level-j approximation of the surface is denoted by $V(j)$, and $\mathbf{v}_n^j$ denotes a specific vertex, hence $\mathbf{v}_n^0$ denotes a 3D-vector scaling coefficient. The approximation is refined through subdivision by adding the vertices $\mathbf{w}_m^j \in W(j)$, thus $V(j+1) = V(j) \cup W(j)$. The wavelet coefficients $\mathbf{w}_m^j$ are also 3D vectors. The basis function $\phi_n^0(\mathbf{x})$ denotes the lowest-resolution scaling function centered on the vertex indexed by n, and $\psi_m^j(\mathbf{x})$ denotes the level-j wavelet basis function centered on the vertex indexed by m. The corresponding coefficients $\mathbf{v}_n^0$ and $\mathbf{w}_m^j$ form our shape representation which we wish to best fit to an observation by minimizing an energy function. Specifically, we concatenate the coefficients into a shape parameter vector $\mathbf{s}$, with the lowest resolution coefficients coming first. Let us denote the (3D) coefficient indexed by k by $\mathbf{s}^k$.

Because these basis functions have only local support, $\mathbf{f}(\mathbf{x})$ only depends on a few coefficients, and the coefficients can be computed in linear time. Both decomposition and reconstruction are comprised of a series of lifting operations of $O(1)$ complexity at each node at each resolution level. Let N_v denote the number of vertices in the full resolution mesh. Since there are $N_v/4^{N-j}$ nodes in level j, and $N_v + N_v/4 + N_v/16 + \ldots < 2N_v$, the total number of times the lifting operations must be applied is $O(N_v)$ in either the decomposition or reconstruction. The lifting operations effectively predict one coefficient using its neighbors in the grid, subtract the prediction from the true value leaving the residual component that is not correlated to the neighbors according to that prediction model or filter. Because the transform is biorthogonal, we may assume the coefficients are fully decorrelated, i.e. independent, and can be optimized individually.

C. Statistical Prior

We now define the observation and prior components of our model. We model the prior probability of the wavelet coefficients as independent Gaussian distributions, and compute statistics on a database of face shapes. Let $\mathbf{s}_i$

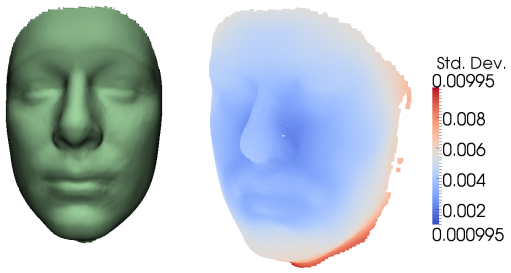

Figure 1. Left: The surface $\bar{\mathbf{f}}$ reconstructed from the mean shape vector $\bar{\mathbf{s}}$. Right: false color visualization of the magnitude of the standard deviation of the model parameters associated with each vertex in the full-resolution grid mesh, $|\sigma^k|$.

denote the shape parameter vector of face i in the database $\mathcal{F} = \{\mathbf{f}_i, \mathbf{s}_i, \mathbf{r}_i\}_1^F$, where F is the number of faces in our database and $\mathbf{r}_i$ will be defined shortly. Our prior model is defined by three shape quantities. The first is simply the mean shape parameter vector

$$\bar{\mathbf{s}}^k = \frac{1}{n}\sum_{i=1}^{F} \mathbf{s}_i^k \tag{5}$$

for $k = 0, \ldots, C-1$, where C is the number of wavelet coefficients and each 3D coefficient vector $\mathbf{s}^k$ in a shape parameter vector $\mathbf{s}$ can be treated independently because of the decorrelating and localizing properties of the wavelet basis functions. The face surface $\bar{\mathbf{f}}$ that is reconstructed by applying (4) to $\bar{\mathbf{s}}$ is shown in Figure 1.

While we can perform statistical analysis on each $\mathbf{s}^k$ independently of other values of k, we must consider their three components together. Each $\mathbf{s}^k$ is a 3D vector representing either the scale (absolute value) or the detail (relative value) of the shape at a particular frequency and spatial location. However, the coordinate axes in general do not correspond to the main directions of variation in $\mathcal{F}$ of $\mathbf{s}_i^k,\ i = 1, \ldots, F$. Therefore, we perform PCA on each set of coefficient vectors, to obtain 3D vectors $\mathbf{r}_i^k$ that represent position along the directions of greatest variation, and 3×3 matrices U^k that transform these coordinates to our original world coordinate system, as in

$$\mathbf{s}_i^k = \bar{\mathbf{s}}^k + U^k \mathbf{r}_i^k \tag{6}$$

where we write $\mathbf{s}^k = [x_s^k, y_s^k, z_s^k]^T$ and $\mathbf{r}^k = [x_r^k, y_r^k, z_r^k]^T$ to denote the components of these vectors, and $\mathbf{r} = [\mathbf{r}^{0T}, \mathbf{r}^{1T}, \ldots]^T$ to denote the complete vector of statistical shape parameters.

Due to the orthogonality of the wavelet basis functions and the basis of the principal component analysis, we may justify assuming that $\mathbf{r}^k$ is independent from $\mathbf{r}^m$ for $m \neq k$, and that the components x_r^k, y_r^k and z_r^k form zero-mean Gaussian distributions that are independent from each other. From the training set we can learn the standard deviation of each component $\sigma^k = [\sigma_x^k, \sigma_y^k, \sigma_z^k]^T$. The standard deviation across the surface is show in the right side of Figure 1. This allows us to write the prior probability of a surface $\mathbf{f}$ as

$$P(\mathbf{f}) = P(\mathbf{s}) = P(\mathbf{r}) \tag{7}$$

where $\mathbf{f}$ relates to $\mathbf{s}$ by (4), and $\mathbf{s}$ relates to $\mathbf{r}$ by (6) and

$$P(\mathbf{r}) = \prod_k P(\mathbf{r}^k) \tag{8}$$

and

$$P(\mathbf{r}^k) = \left(\frac{1}{\sigma_x^k\sqrt{2\pi}} e^{-\frac{(x_r^k)^2}{2(\sigma_x^k)^2}}\right) \cdot \left(\frac{1}{\sigma_y^k\sqrt{2\pi}} e^{-\frac{(y_r^k)^2}{2(\sigma_y^k)^2}}\right) \cdot \left(\frac{1}{\sigma_z^k\sqrt{2\pi}} e^{-\frac{(z_r^k)^2}{2(\sigma_z^k)^2}}\right) \tag{9}$$

D. Observations

To model the likelihood of observing the point cloud $\mathcal{Y}$ and the images $\mathcal{I}$, given a shape parameter vector $\mathbf{s}$, we reconstruct the surface $\mathbf{f}(\mathbf{x})$ from $\mathbf{s}$ using (4). We assume the point cloud is a noisy approximation of the surface $\mathbf{f}$, with approximately zero-mean Gaussian noise. Hence, we model the probability of observing the point set, given the current mesh, as an exponential distribution on the sum-of-squared distances of the model vertices to their nearest neighbors in the point cloud. In practice, most stereo algorithms have some systemic error in addition to noise, but we alleviate this by using a truncation threshold on the nearest neighbor distance in addition to using a prior.

We further assume the surface is approximately Lambertian, and project the surface into both images and perform stereo matching during a refinement stage. Although human skin can be both quite specular and translucent, we counteract the effects through the statistical prior and through matching techniques. We use robust matching costs to account for outliers due to specularities and we explicitly take self-occlusion into account. Our framework can be extended to include additional occluders. We further use an anisotropic second-order smoothing energy to regularize the refinement. We define our likelihood as an exponential distribution, where because they are deterministically related,

$$P(\mathcal{I}|\mathbf{r}, \mathcal{Y}) = P(\mathcal{I}|\mathbf{s}, \mathcal{Y}) = P(\mathcal{I}|\mathbf{f}, \mathcal{Y}) \tag{10}$$

and

$$P(\mathcal{I}|\mathbf{f}, \mathcal{Y}) \propto \exp(-E_M(\mathbf{f}, \mathcal{Y}, I_L, I_R)) \tag{11}$$

where E_M is a matching cost defined in Section IV. The matching depends on the point cloud in that the nearest neighbor distance is used to determine how far to sample during stereo refinement, when the mesh $\mathbf{f}$ is optimized with respect to 11.

While a smoothness constraint is conceptually a prior, i.e. the prior knowledge that the face is piecewise smooth, because it is applied to the mesh vertices and not to the

wavelet coefficients, we treat it as part of the refinement process, optimized in conjunction with the matching cost.

IV. The Energy Function and Minimization

In this section, we derive an energy or objective function from our probability model, and describe how we minimize that function. Our energy function consists of two parts, again representing observation and prior of our model.

Our first data cost is the sum-of-squared distances of the model vertices $\mathbf{f}$ to their nearest neighbors in the point cloud $\mathcal{Y}$, obtained as an initial estimate from a conventional stereo algorithm. On initialization, we find the nearest neighbor of each vertex in $\mathbf{f}$. The energy is then

$$E_{NN} = \sum_{\mathbf{x}} \min\left(\|\mathbf{f}(\mathbf{x}) - \mathbf{y}_{\mathbf{x}}\|, \tau_{NN}\right) \tag{12}$$

where $\mathbf{y}_{\mathbf{x}} \in \mathcal{Y}$ is the nearest neighbor of $\mathbf{f}(\mathbf{x})$, and τ_{NN} is a truncation constant to mitigate the effect of outliers, noise and incomplete data (i.e. holes in the point cloud where the initial stereo estimate could not reconstruct the surface).

Our matching energy E_M is defined over the surface map and the input images,

$$E_M(\mathbf{f}, \mathcal{Y}, \mathcal{I}) = \sum_i \sum_{j \neq i} \sum_{\mathbf{x}} w_M(\mathbf{x}) D(I_i, I_j, \mathbf{f}(\mathbf{x})) \tag{13}$$

where $i, j \in [0, |\mathcal{I}|-1]$ denote the each pair of reference and matching images in the set, D is a point-wise dissimilarity measure, and $w_M(\mathbf{x})$ is a per-vertex weight. The dissimilarity D is defined as

$$D(I_i, I_j, \mathbf{p}) = 1 - NCC(I_i(\mathbf{q}), I_j(H_{ij}\mathbf{q})) \tag{14}$$

where $NCC(\cdot,\cdot)$ is the normalized cross-correlation (averaged over the red, green and blue channels) of two image patches containing an equal number of samples, $I_i(\mathbf{q})$ denotes the image patch around point $\mathbf{q}$, which is the projection of $\mathbf{p}$ into image I_i, and H_{ij} is the homography mapping a point in the image plane of I_i to the plane tangent to the surface at point $\mathbf{p}$ with normal $\mathbf{n}$ to the image plane of I_j. This is given by

$$H_{ij} = K_j \left(R_{ij} - \frac{\mathbf{t}_{ij}\mathbf{n}^T}{d_i} \right) K_i^{-1} \tag{15}$$

where K_i and K_j are the intrinsic calibration matrices of I_i and I_j, the matrix $[R_{ij}|\mathbf{t}_{ij}]$ transforms coordinates relative to I_i to coordinates relative to I_j, and d_i is the depth of $\mathbf{p}$ with respect to I_i. Thus, H_{ij} projectively maps points in the image patch surround $\mathbf{q}$ in image I_i, to the tangent plane to the surface at point $\mathbf{p}$, to the correspondging point in I_j.

The per-vertex weight is designed to reflect the reliability of the dissimilarity or photo-consistency function D. As described further below, the refinement stage iteratively takes the current mesh $\mathbf{f}$, and for each vertex $\mathbf{f}(\mathbf{x})$ with normal $\mathbf{n}(\mathbf{x})$, samples the smoothing and photoconsistency functions at points above and below the mesh vertex along the normal direction. Let these sample points be denoted by $\mathbf{p}_m$, and thus we have matching cost samples $D(I_i, I_j, \mathbf{p}_m)$, where m is an index. Let $\hat{D}(\mathbf{x}) = \min_m D(I_i, I_j, \mathbf{p}_m)$ denote the minimum dissimilarity or matching cost for a given vertex. Then the matching weight is given by

$$w_M(\mathbf{x}) = \sum_m \left(D(I_i, I_j, \mathbf{p}_m) - \hat{D}(\mathbf{x}) \right) \tag{16}$$

which is greatest when there is one low miminum of the matching cost, and the rest of the samples are high. Hence, the weight is highest when the matching cost gives the most distinctive information about the surface. We further use depth-buffering to determine if a point is visible in both the reference and matching images. If it is not, $w_M(\mathbf{x})$ is set to zero.

For a smoothing energy term we use the distance of a vertex to an average of the neighboring vertices. We first compute the smoothed vertex position

$$\bar{\mathbf{f}}(\mathbf{x}) = \frac{w_x(\mathbf{f}(\mathbf{x}_l) + \mathbf{f}(\mathbf{x}_r)) + w_y(\mathbf{f}(\mathbf{x}_u) + \mathbf{f}(\mathbf{x}_d))}{2(w_x + w_y)} \tag{17}$$

where $\mathbf{x}_l = (x-1, y)$, $\mathbf{x}_r = (x+1, y)$, $\mathbf{x}_u$ and $\mathbf{x}_d$ are defined similarly. The horizontal weight is defined as

$$w_x = \exp\left(-\left(\|\mathbf{f}(\mathbf{x}_l) - \mathbf{f}(\mathbf{x})\| - \|\mathbf{f}(\mathbf{x}_r) - \mathbf{f}(\mathbf{x})\| \right)^2 \right)$$

and w_y is defined similarly. This smoothing is a variation of the one used by Beeler et al. [6] for disparity map refinement. Because we have a quadrilateral mesh, we can apply it to the vertex coordinates. The smoothing energy is then

$$E_{sm}(\mathbf{f}) = \sum_{\mathbf{x}} \left\| \mathbf{f}(\mathbf{x}) - \bar{\mathbf{f}}(\mathbf{x}) \right\| \tag{18}$$

reflecting how each vertex in $\mathbf{f}$ deviates from the anisotropic average of its neighbors (17).

The prior or regularization term in our energy function is taken directly as the negative logarithm of the prior $P(\mathbf{r})$. That is,

$$E_{st}(\mathbf{r}) = \sum_k \left(\frac{(x_r^k)^2}{2(\sigma_x^k)^2} + \frac{(y_r^k)^2}{2(\sigma_y^k)^2} + \frac{(z_r^k)^2}{2(\sigma_z^k)^2} \right) . \tag{19}$$

We combine these to get our energy or objective function,

$$E(\mathbf{s}) = w_{st}E_{st}(\mathbf{r}) + w_{NN}E_{NN}(\mathbf{f}) + E_M(\mathbf{f}) + w_{sm}E_{sm}(\mathbf{f}) \tag{20}$$

where $\mathbf{f}$ and $\mathbf{r}$ are related to $\mathbf{s}$ as before, and w_{st}, w_{NN} and w_{sm} are user-controlled parameters.

To minimize (20) we break the optimization into two parts. We first use sampling methods based on the learned distributions of our model parameters to minimize E_{stat} and E_{NN}. As noted by Li et al. [8], we can serialize the parameter sampling because the orthogonality of the wavelet basis and the principal components allows us to assume independence between parameters. This leads to a complexity of $O(PS)$ for P parameters and S samples

instead of $O(P^S)$ without the independence assumption. We examine two sampling strategies: uniform sampling and stochastic sampling. In the first case, we sample each parameter (x_r^k,y_r^k or z_r^k) uniformly within three standard deviations (eg. $\pm 3\sigma_x^k$). In the second case, values of, for example, x_r^k are chosen at random from the distribution $\mathcal{N}(0, \sigma_x^k)$.

After each sample value x_r^k, the wavelet coefficient $\mathbf{s}_k$ is reconstructed using the PCA eigenvectors U^k and the mean shape coefficient $\bar{\mathbf{s}}^k$, then the face surface $\mathbf{f}$ is reconstructed using (4).

One of the main drawbacks of statistical shape priors, especially for face reconstruction, is that they produce overly regularized results that do not deviate sufficiently from the mean shape. This is particularly a concern for faces, where much of the identifying detail is contained in finer scales. Hence, after optimizing the first five levels of coefficients using the sampling method described above, we follow with an iterative mesh refinement stage that minimizes E_M and E_{smooth} together. Following refinement we transform the surface back into the model parameters. We formulate this two-stage optimization as iterative partial maximization, where we optimize the model in terms of one part of the energy function and then in terms of the other. The first part is the combined energy of the statistical prior and the nearest neighbor distance. The second part is the smoothing and matching energies.

As mentioned above, refinement proceeds by sampling along the normal directions for each vertex in the mesh. Thus for mesh vertex $\mathbf{f}(\mathbf{x})$ with normal $\mathbf{n}(\mathbf{x})$, we have sample points $\mathbf{p}_m = \mathbf{f}(\mathbf{x}) + m\delta_{\mathbf{x}}\mathbf{n}(\mathbf{x})$ for $m = -N_r, \ldots, N_r$, where $\delta_{\mathbf{x}}$ is a user-controlled step size parameter. The sample that minimizes the combination of smoothing and matching energies is taken as the new vertex position. That is,

$$\mathbf{f}(\mathbf{x}) \leftarrow \arg\min_m E_{ref}(\mathbf{p}_m)$$

where E_{ref} is the per-vertex combined smoothing and matching energy,

$$E_{ref}(\mathbf{p}_m) = w_{sm} \left\| \mathbf{p}_m - \bar{\mathbf{f}}(\mathbf{x}) \right\| + w_M(\mathbf{x}) D(I_i, I_j, \mathbf{p}_m) \tag{21}$$

favoring smooth surfaces where matching information is missing or unreliable. Refinement is performed by iteratively sampling in this way for each reference-matching image pair in succession.

V. Experiments

This section documents our experimental validation of our approach, including the implementation, and the results obtained.

A. Implementation

Our implementation uses CUDA, OpenGL, OpenCV and CLAPACK. Our images were captured using a Canon Rebel EOS XTi 400D, a 10 Mpixel digital SLR camera, and downsampled by a factor of two. For calibration we used publicly available structure-from-motion software [26]. For initial stereo estimates we used PMVS [27] and OpenCV's graph cut stereo [28].

We perform registration manually, selecting landmarks first on the template face model, then selecting the same points in the same order in the initial stereo/range data. From this an initial estimate of the similarity transform between the model space and the initial data is computed using linear least-squares. While this is a major limitation of our method in its current instantiation, this step could be replaced by automatic mutli-view face detection and localization in the input images, for example by a method such as that of Koterba et al. [14]. Such an automated registration method would likely be equally or more accurate as manually selecting landmarks in noisy point-clouds. This might also remove the need for the initial stereo estimate, as the model could be fitted directly to the disparity space images (DSI) of the input images.

For our wavelet we started with a base mesh of 2×3 and subdivide eight times to get the full-resolution grid mesh of 129×257. We perform the resampling of the training set using OpenGL and GLSL. Each vertex in the template mesh is stereographically projected onto a plane aligned with the front of the face. Then, each corresponding vertex in every other mesh is mapped to the same position in the plane, thus preserving correspondence. The initial stereographic projection is performed on the CPU, but the resampling of the meshes in the training set is performed on the GPU. This makes the learning part of the approach very fast. To learn from a training set of 100 faces takes only a few minutes, plus the time to first perform the GPA to align the faces.

The (inverse) wavelet transform (4) must be performed for every sample value of every sampled coefficient. In our current framework, we optimize the first five levels of coefficients ($P = 561 \times 3 = 1683$ model parameters) using independent parameter sampling, with $S = 50$ samples per parameter. This means the surface must be reconstructed from the model parameters 84150 times, hence the speed of the transform is crucial to the speed of the overall apprach. In practice, some parameters have very small variation in the database (eg. $\sigma_x^k < 10^{-12}$) and we do not sample those parameters, so the total number of inverse wavelet transforms that must be performed is 60750 in our current setup. Each transform takes 0.202ms using our CUDA-based GPU implementation, for a total of 12.272s spent reconstructing the surface. This is compared to slightly over 1 ms per transform using a highly optimized CPU implementation. (Note that with the dimensions we are using, the entire wavelet data fits in the CPU cache, making this virtually optimal CPU performance.) The GPU wavelet transform splits the computation into blocks that overlap by two vertices/coefficients on all sides, reads the coefficients

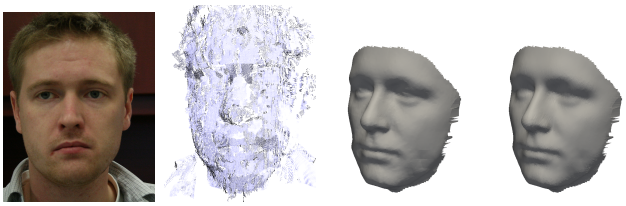

Figure 2. Left to right: input image, initial point cloud, reconstruction before refinement (level 4), after refinement.

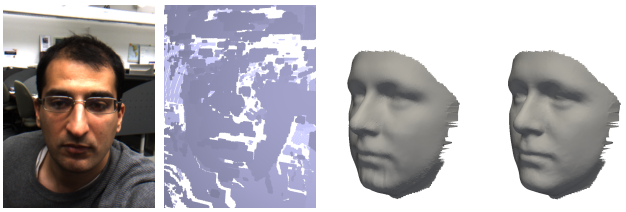

Figure 3. Left to right: input image, initial point cloud, reconstruction, mean face for comparison. Note that the face is reconstructed in spite of the fact that the subject is wearing glasses which causes severe problems for the initial stereo estimate.

Figure 4. Fitting the model to a laser scan. Left to right: point cloud, reconstruction before refinement, after refinement, original mesh.

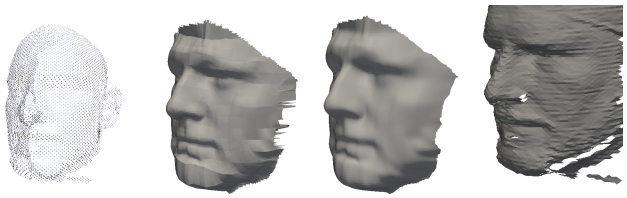

Figure 5. Fitting the model to a laser scan. Left to right: point cloud, reconstruction before refinement, after refinement, original mesh.

into shared memory, and performs the lifting operations in shared memory. If the number of blocks required is less than the number of multi-processors on the GPU then the transform can be performed in-place, writing to the same global memory it reads from. One evaluation of E_{NN} takes 0.671ms, for a total of 41.550s over the entire algorithm. In total, the parameter sampling takes just over 67s.

The refinement stage takes 0.794s for 200 iterations (per reference-matching image pair), three samples per vertex per iteration, and two images. It also breaks the computation into overlapping blocks. For each vertex in each block, we use a reference thread and a matching thread, which share the computation. One thread computes the normal, while the other computes the anisotropic average, both using the neighboring vertices in shared memory. The reference thread then samples the window in the reference image, while the matching thread samples the window in the matching image, each thread storing the samples in shared memory. (We use 3×3 windows for NCC.) The remainder of the NCC computation is divided between the two threads, and the resulting matching cost for each sample saved for computation of the per-vertex matching weight.

B. Results

Figure 2 shows the results of our reconstruction algorithm applied to a stereo pair, with an initial point cloud from a general stereo algorithm [27]. Despite the noise in the original point cloud, the reconstruction after parameter sampling (second from right) captures the shape of the face with some artifacts due to the independence of the shape parameters, while the post-refinement reconstruction smooths the artifacts while preserving shape detail. Note how the reconstruction captures the fact that the mouth is slightly lower on the left side than on the right side, as in the input image. Figure 3 shows another result, this time with the initial point cloud from graph cuts [28]. It is again quite noisy, and it also exhibits fronto-paralle bias. Note that the subject is wearing glasses, and predictably the initial point contains only outliers around the eyes. Nonetheless our method constructs a plausible surface for the entire face. The mean face is shown next to the reconstruction for comparison. Note how the nose is elongated and the cheek bones are more prominent in the reconstruction as in the input image.

Figure 4 and 5 show the results of reconstruction by fitting the prior model to laser-scan data. Since these point clouds are more reliable than stereo data, we increase the weight w_{NN}. Note how the shape of the nose in Figure 4 is captured accurately without the noise that is present in the original mesh (far right). The reconstruction captures the shape of the nose and cheek bones in both cases while smoothing the surface and increasing the resolution. The noise or artifacts in the reconstructions are along the outside of the face where the prior is less reliable. Since there are no images to go with these point clouds, the refinement consists only of smoothing.

VI. Discussion

We have presented a method for fast and robust model-based stereo using a statistical wavelet shape prior. We have demonstrated this approach for human faces using both stereo and laser-scan data. The most interesting direction for future work is to introduce a temporal term into the model and use this framework for stereo tracking. Such an approach would distinguish between tracking the changes in shape from tracking the changes in position relative to the cameras. This could be used to help avoid drift that occurs in optical

flow-based tracking. We currently register the initial stereo data to the shape template with user selected landmarks, but to automate this process is a natural avenue for future work.

ACKNOWLEDGMENT

This work was supported in part by the National Research Council of Canada (NRC) and in part by the Natural Sciences and Engineering Research Council of Canada (NSERC). The authors would like to thank Houman Rastgar for his participation as a test subject.

REFERENCES

[1] B. Amberg, A. Blake, A. Fitzgibbon, S. Romdhani, and T. Vetter, "Reconstructing high quality face-surfaces using model based stereo," in *Proc. ICCV 2007*, 2007.

[2] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of SIGGRAPH '99*, 1999, pp. 187–194.

[3] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal Computer Vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[4] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'2006)*, 2006, pp. 519–526.

[5] D. Bradley, W. Heidrich, T. Popa, and A. Sheffer, "High resolution passive facial performance capture," in *ACM Transactions on Graphics (SIGGRAPH)*, 2010.

[6] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross, "High-quality single-shot capture of facial geometry," *ACM Trans. on Graphics (Proc. SIGGRAPH)*, vol. 29, no. 3, 2010.

[7] Y. Furukawa and J. Ponce, "Dense 3d motion capture for human faces," in *IEEE Internation Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[8] Y. Li, T.-S. Tan, I. Volkau, and W. L. Nowinski, "Model-guided segmentation of 3d neuroradiological image using statistical surface wavelet model," in *Proc. CVPR 2007*, 2007.

[9] M. Bertram, M. A. Duchaineau, B. Hamann, and K. I. Joy, "Generalized b-spline subdivision-surface wavelets for geometry compression," *IEEE Trans. Visualization and Computer Graphics*, vol. 10, no. 3, pp. 326–338.

[10] Y. Sun, P. Kohli, M. Bray, and P. Torr, "Using strong shape priors for stereo," in *Proceedings of the 5th Indian conference on computer vision, graphics and image processing (ICVGIP 2006)*, 2006, pp. 882–893.

[11] F. Romeiro and T. Zickler, *Model-based Stereo with Occlusions*, 2007.

[12] M. Tonko and H.-H. Nagel, "Model-based stereo-tracking of non-polyhedral objects for automatic disassembly experiments," *International Journal of Computer Vision*, vol. 37, pp. 99–118, 2000.

[13] W. Zhao, N. Nandhakumar, and P. Smith, "Model-based interpretation of stereo imagery of textured surfaces," *Machine Vision and Applications*, vol. 10, pp. 201–213, 1997.

[14] S. Koterba, S. Baker, I. Matthews, C. Hu, J. Xiao, J. Cohn, and T. Kanade, "Multi-view aam fitting and camera calibration," in *Proceedings of the Tenth International Conference on Computer Vision (ICCV'05)*, 2005.

[15] J. Magarey and A. Dick, "Multiresolution stereo image matching using complex wavelets," in *Proceedings of the 14th International Conference on Pattern Recognition (ICPR'98)*, vol. 1, 1998.

[16] J. Li, H. Zhao, X. Zhou, and C. Shi, "Robust stereo image matching using a two-dimensional monogenic wavelet transform," *Optics Letters*, vol. 34, pp. 3514–3516, 2009.

[17] C. Liu, W. Pei, S. Niyokindi, J. Song, and L. Wang, "Micro stereo matching based on wavelet transform and projective invariance," *Measurement Science and Technology*, vol. 17, 2006.

[18] M. Shim, "Wavelet-based stereo vision," *Lecture Notes in Computer Science Biologically Motivated Computer Vision*, vol. 1811/2000, pp. 335–385, 2000.

[19] G. Caspary and Y. Zeevi, "Wavelet-based multiresolution stereo vision," in *Proceedings of the 16th International Conference on Pattern Recognition (ICPR'02)*, vol. 3, 2002.

[20] A. Bhatti and S. Nahavandi, *Stereo Correspondence Estimation based on Wavelets and Multiwavelets Analysis*, 2008.

[21] W. Miled, J. Pesquet, and M. Parent, "Wavelet-constrained regularization for disparity map estimation," in *Proceedings of the European Signal and Image Processing Conference*, 2006.

[22] S. Essafi and G. Langs, "Hierarchical 3d diffusion wavelet shape priors," in *IEEE 12th International Conference on Computer Vision (ICCV)*, 2009, pp. 1717–1724.

[23] Y. Chen and R. Cipolla, "Learning shape priors for single view reconstruction," in *International Conference on 3D Data Imaging and Modeling (3DIM)*, 2009, pp. 1425–1432.

[24] M. Levine and Y. Yu, "State-of-the-art of 3d facial reconstruction methods for face recognition based on a single 2d training image per person," *Pattern Recognition Letters*, vol. 30, pp. 908–913, 2009.

[25] I. Dryden and K. Mardia, *Statistical Shape Analysis*. Wiley, 2002.

[26] N. Snavely, S. Seitz, and R. Szeliski, "Photo tourism: Exploring image collections in 3d," in *Proceedings of SIGGRAPH*, 2006.

[27] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," in *Proc. CVPR 2007*, 2007.

[28] V. Kolmogorov, "Graph based algorithms for scene reconstruction from two or more views," Ph.D. dissertation, Cornell University, 2003.