# High-Quality Single-Shot Capture of Facial Geometry

Thabo Beeler[1,2] Bernd Bickel[1,2] Paul Beardsley[2] Bob Sumner[2] Markus Gross[1,2]
[1]ETH Zurich [2]Disney Research Zurich


**Figure 1:** *Left: Face model captured using a seven camera studio setup; Center: capture systems; Right: Face model captured using consumer binocular-stereo camera. Facial geometry in all figures is best viewed in the electronic version.*

## Abstract

This paper describes a passive stereo system for capturing the 3D geometry of a face in a single-shot under standard light sources. The system is low-cost and easy to deploy. Results are sub-millimeter accurate and commensurate with those from state-of-the-art systems based on active lighting, and the models meet the quality requirements of a demanding domain like the movie industry. Recovered models are shown for captures from both high-end cameras in a studio setting and from a consumer binocular-stereo camera, demonstrating scalability across a spectrum of camera deployments, and showing the potential for 3D face modeling to move beyond the professional arena and into the emerging consumer market in stereoscopic photography.

Our primary technical contribution is a modification of standard stereo refinement methods to capture pore-scale geometry, using a qualitative approach that produces visually realistic results. The second technical contribution is a calibration method suited to face capture systems. The systemic contribution includes multiple demonstrations of system robustness and quality. These include capture in a studio setup, capture off a consumer binocular-stereo camera, scanning of faces of varying gender and ethnicity and age, capture of highly-transient facial expression, and scanning a physical mask to provide ground-truth validation.

**CR Categories:** I.3.2 [Computer Graphics]: Graphics Systems—Stand-alone systems; I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Color, shading, shadowing, and texture I.4.1 [Image Processing and Computer Vision]: Digitization and Image Capture

## 1 Introduction

### 1.1 Motivation

Capturing a high-quality model of a human face is of interest in multiple domains - for the movie and games industries, in medicine, to provide more natural user-interfaces, and for archival purposes. A key application is the synthesis of a desired sequence of speech and facial expression under arbitrary lighting. This problem has motivated striking results in rendering [Donner and Jensen 2006; Donner et al. 2008], in performance-driven facial animation [Hyneman et al. 2005; Alexander et al. 2009], and in physics-based animation [Sifakis et al. 2005]. However, reproducing realistic human faces is still a challenge for computer graphics because humans are sensitive to facial appearance and quickly sense any anomalies in 3D geometry or dynamics.

This paper is concerned with the capture of 3D geometry of the face. The current method of choice for this task is an active system based on laser, structured light or gradient-based illumination. Active light brings robustness because it effectively augments an object surface with known information. On the other hand, it requires special-purpose hardware and often employs time-multiplexing. Polarization-based methods further constrain deployment to a single camera at a fixed viewpoint. Contrast this with passive stereo vision, which has the potential to be an extremely versatile modality for constructing 3D models - it captures in a single shot, readily adapts to different arrangements and numbers of cameras with no constraint on camera position, seamlessly integrates 3D data captured over multiple distances and at different scales in a scene, captures texture that is intrinsically registered with the recovered 3D data, and uses commodity hardware. However, in the past, the reliability and accuracy of passive stereo have fallen short of what is available from active systems, and it has not been used for capturing high-quality face models.

This paper presents a passive stereo vision system that computes the 3D geometry of the face with reliability and accuracy on a par with a laser scanner or a structured light system. We introduce an image-based embossing technique to capture mesoscopic facial geometry[1], so that the quality of synthesized faces from our system

[1]We use the term mesoscopic for geometry at the scale of pores and fine

equals that achieved with gradient-based illumination. In practical terms, we equal the performance of active systems while attaining the advantages of passive stereo already listed, particularly capture in a single shot under standard light sources, low-cost, and ease-of-deployment. To demonstrate the robustness of the system, we show results for faces of varying gender, ethnicity and age. To demonstrate versatility, we show face models captured off a studio setup and off a consumer binocular-stereo camera, with the latter result suggesting that 3D face scanning is poised to move beyond the professional arena and become a practical application on the desktop.

### 1.2 Related Work

The best current techniques for capturing geometry of a human face are active. The domain grew from original work on stereo capture of a face augmented with skin markings [Parke 1974]. A survey of the area is given in [Pighin and Lewis 2005]. More recent work begins with a hybrid system that combines a recovered depth map and recovered surface normals to generate a model [Nehab et al. 2005]. This technique has been utilized for faces in [Weyrich et al. 2006], and in [Ma et al. 2007] which presented a system for acquiring high-quality surface normals using polarized gradient-based illumination to generate high-resolution 3D reconstructions. Further work includes a hybrid system of structured light and stereo [Weise et al. 2007], and the application of the technique to facial expression transfer in [Weise et al. 2009]. Recent work on single shot photometric stereo is described in [Hernandez et al. 2008]. A hybrid system of active light and augmented skin markings is the basis for the current state-of-the-art example of creating a photoreal human face in [Alexander et al. 2009].

Turning to non-active techniques, Section 2 of this paper describes stereo matching and generation of a 3D mesh. We build on established techniques described in the survey paper [Seitz et al. 2006], and also take inspiration from [Furukawa and Ponce 2007]. Commercial solutions [DI3D 2009] as well as concurrent work by [Bradley et al. 2010] are also applying MVS to the domain of face scanning. The main difference to our system lies in the refinement formulation described in Section 3. Our starting point is the established approach of refining recovered 3D data based on a data-driven photo-consistency term and a surface-smoothing term, which has been a subject of research ranging from [Scharstein and Szeliski 1996] to [Woodford et al. 2008]. Our work differs in the use of a second-order anisotropic formulation of the smoothing term, and we argue that it is particularly suited to faces.

We present an extension to traditional stereo refinement in our method for modeling mesoscopic geometry of the face in Section 3. The data-driven and smoothing terms are augmented with a third term that uses image texture to drive the qualitative recovery of mesoscopic geometry, and we thereby capture fine variations of the geometry that are irrecoverable using stereo disparity alone. Although unrelated to our method, shape recipes are instructive on the relationship between image data and shape data [Torralba and Freeman 2003]. Other works concerned with fine-scale detail are the mesh optimization in [Hiep et al. 2009], the modeling in [Golovinskiy et al. 2006], and extraction of mesostructure from specularity in [Chen et al. 2006]. Our method is similar in spirit to [Glencross et al. 2008] which describes qualitative recovery of 3D information for bas-relief surfaces but the technical approaches are different, and Glencross's system being self-contained while we are proposing an extension within the existing framework of stereo refinement.

Turning to the area of camera calibration, there is a body of theory available in a standard text such as [Hartley and Zisserman 2000]. Our calibration contribution is practical, not theoretical, and described in Section 2.1.

wrinkles, and macroscopic for overall 3D shape of the face.

### 1.3 Contributions

This paper makes a systemic contribution and two technical contributions. The systemic contribution is to demonstrate a state-of-the-art passive stereo vision system for face scanning, and to argue that past weaknesses have been overcome to yield a technology that is single-shot, low-cost, easy to deploy, and has impact in two areas. Firstly in the area of professional capture of high-quality face models, we argue that passive stereo is on an equal footing with active systems. Secondly in the emerging area of consumer stereo photography, we show that face scanning can be accomplished using a consumer binocular-stereo camera, indicating that the technology is ready to expand beyond the professional domain. Moving to our technical contributions, the primary contribution is the modeling of mesoscopic geometry in Section 3.3, and the second contribution is the calibration method in Section 2.1. We also describe extensions to generic stereo refinement methods in Sections 3.1 and 3.2 that tailor the processing to faces.

## 2 Face Scanning

This section describes the end-to-end system as shown in Figure 2. Camera calibration is a pre-processing stage and is described in Section 2.1. The run-time system begins with pairwise stereo matching, and uses a pyramidal approach in which results at lower-resolutions guide the matching at higher-resolutions as described in Section 2.2. At each layer of the pyramid, matches are computed at pixel level to give dense matches across the face, and the matches are used to generate a 3D mesh as described in Section 2.3. The mesh is refined using a modification of the traditional approach, in which photo-consistency and smoothing terms are augmented with a novel term that captures fine detail at the pore-level. This is described in Section 3. Low-level details are omitted in places for space reasons, but a full description of the system is available in [Beeler et al. 2010]. An excellent overview and categorization of MVS is found in [Seitz et al. 2006];

### 2.1 Calibration

The theoretical foundation of camera calibration is well established, and our focus has been on the practical matter of achieving a straightforward and reliable calibration for a face-capture system. The method requires a small number of views, typically one to three views, of a sphere augmented with fiducials as shown in Figure 3. Each fiducial is a double circle. The center points of the circles provide the correspondences between cameras, as well as providing a known metric distance $D_F$ that can be used to set scale. The fiducials are not used to provide known 3D coordinates - hence the sphere need not be perfect, and the fiducials can be placed by hand with arbitrary distribution. Fiducials were printed on sticky paper, and the slight distortion in fixing a flat sticker to a spherical surface was not found to be a problem.

The approach has the following advantages. Firstly it is suited to face capture because a calibration sphere that is approximately head-sized and placed at the intended position of the subject is therefore well-placed for the cameras. The sphere need not lie completely within the camera images so there is no fine-tune positioning. Unlike a calibration plane, a sphere has no preferred direction in space, making it appropriate for a setup in which circumpositioned cameras are directed inward towards an object of interest. Unlike an LED-based calibration, the method requires only a small number of views and provides sub-pixel accurate features.

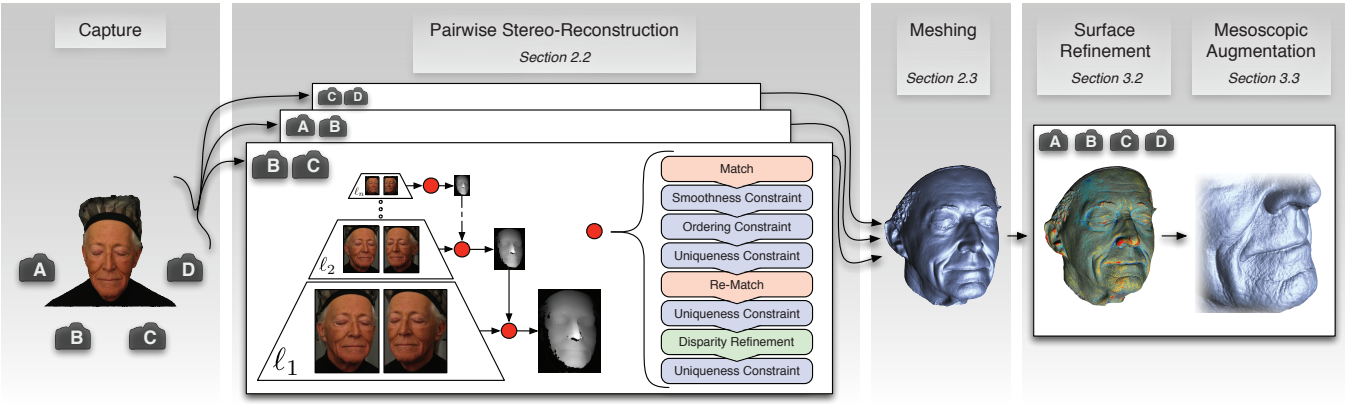


**Figure 2: The proposed system -** *The subject is captured with multiple cameras. This figure shows a four-camera setup, but the system can incorporate an arbitrary number of cameras.*

Finally, the calibration sphere occupies the same workspace as the subject's head in the run-time system. Thus, we ensure that calibration data is collected - and the calibration is therefore well-estimated - in exactly the same workspace as will be used at run-time.

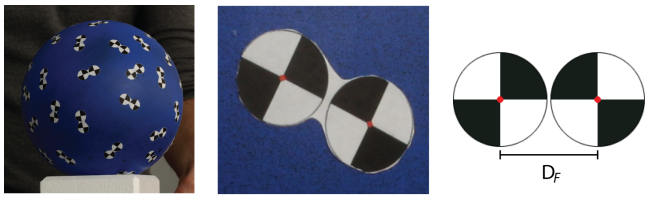


**Figure 3:** *Fiducials are placed randomly on the calibration sphere. A fiducial consists of two checkerboard circles with red dots at their centers. The fiducial defines a known distance $D_F$.*

Calibration is fully automatic. The algorithm is

- Segment out the sphere in each image using cues of background subtraction and known sphere color.
- Fit an ellipse to the silhouette of the segmented sphere. Ignore the area outside the ellipse in the remainder.
- Detect the circles via spots of known color at their centers. Compute the center positions to sub-pixel accuracy using function *findCornerSubPix* in [Intel 2001].
- Set up an approximately Euclidean coordinate frame with the origin at the center of the calibration sphere by using (a) a rough estimate of camera focal length plus (b) the known metric dimension of the sphere.
- Compute the 3D coordinates of the detected circle centers in this coordinate frame, Each fiducial thereby generates one 3D triangular facet with vertices at the sphere center and the fiducial's two circle centers.
- Match the 3D triangular facets between pairs of cameras using RANSAC [Hartley and Zisserman 2000]. Note that a single putative match is sufficient to compute a rotation between two cameras in the approximately-euclidean frame, while the translation has been factored out by the choice of sphere center as the origin. Thus the sample size for the RANSAC sampling is one.
- Use the computed correspondences between pairs of cameras to construct correspondences across all cameras.
- Discard the approximately-euclidean frame, and use the computed correspondences as input to the calibration system at [Svoboda ] to determine camera intrinsics and extrinsics up to unknown scale.
- Set the scale of the computed coordinate frame using the known distance $D_F$.
- Define a 3D 'capture-zone' as the intersection of the camera viewing frustums, to delimit the active region within which 3D processing will be done at run-time.

## 2.2 Pairwise Stereo-Reconstruction

In this section we describe the individual steps of our stereo reconstruction. First, the face is segmented out of the images using cues of background subtraction and skin color. Matching is done pairwise between neighboring cameras[2], and at pixel level to establish dense matches across the face. For a given camera-pair, the first step is to rectify the images to obtain row-aligned epipolar geometry. An image pyramid is generated for each rectified image by factor-of-two subsampling using Gaussian convolution. The image resolution at the lowest-resolution layer of the pyramid is chosen to be around $150 \times 150$ pixels, but this is approximate and the criteria is simply that the major facial features are still visible.

Each layer of the pyramid is then processed as follows: First, matches are computed for all pixels as described in Section 2.2.1. Next, we check smoothness, uniqueness and ordering constraints for each pixel (see Section 2.2.2). Pixels that do not fulfill these constraints are re-matched using a limited search area (Section 2.2.1). The limited search area ensures that smoothness and ordering constraints hold on the re-matched pixels. The uniqueness constraint however needs to be enforced once more. The disparity maps are then refined. An in-depth description is deferred to Section 3.1, since it is an instantiation of the refinement formulation introduced in Section 3. Finally, the uniqueness constraint is enforced again.

Matching starts at the lowest-resolution layer of the pyramid. The resulting disparity map provides input to the the next higher layer, where it is used to constrain the search area for matching, and so on

[2]Pairing of cameras in a multi-camera system is done manually, although it would be straightforward to automate if needed.

up to the highest-resolution layer of the pyramid. As demonstrated in Figure 4, this leads to a hierarchical refinement of the reconstruction over the layers of the pyramid.

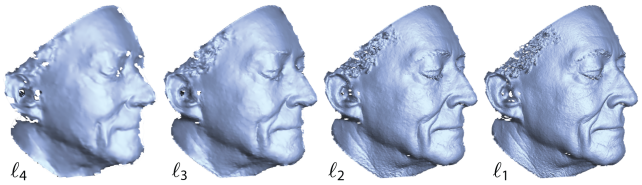


**Figure 4:** *Reconstructions of a stereo-pair at 4 different layers of the pyramid starting from the coarsest layer $\ell_4$ (160px × 160px). The dimensions double at each layer up to the highest resolution layer $\ell_1$ (1280px × 1280px).*

#### 2.2.1 Pixel Matching

Following the taxonomy of [Scharstein and Szeliski 2002], the system employs a winner-take-all block-matching algorithm using normalized cross-correlation (NCC) as matching cost over a square window (3x3). Matching is performed along the epipolar line only. Pixel $p$ in image $\mathcal{I}$ is matched against all pixels in image $\mathcal{J}$ within a given search area and the best match is retained. The disparity at $p$ is computed to sub-pixel accuracy by computing NCC values for $p$ against the matching pixel $q$ and its two neighbors in image $\mathcal{J}$, fitting a polynomial of degree three, and finding the position in image $\mathcal{J}$ where it is at minimum.

Matching is performed twice per layer. The initial matching computes putative matches for all pixels using the disparity estimates of the preceding layer (or the 'capture-zone' if no preceding guesses are present) to constrain the search area. Next, we check for each pixel smoothness, uniqueness and ordering constraints (see Section 2.2.2). Pixels that do not fulfill these constraints are re-matched using the disparity estimates of the neighboring pixels that fulfilled the constraints to limit the search area.

#### 2.2.2 Constraints

The system can make use of constraints that hold for human faces in the given setting. Pixels in image $\mathcal{I}$ are matched against image $\mathcal{J}$, and vice-versa from image $\mathcal{J}$ to image $\mathcal{I}$. Acceptance of a match at pixel $p$ in image $\mathcal{I}$ is subject to three constraints -

- **Smoothness Constraint -** computed disparity at $p$ is consistent with neighbors in a surrounding window. In our implementation this is achieved by enforcing that more than half of all neighbors in a $3 \times 3$ neighborhood differ by a disparity less than one pixel.
- **Uniqueness Constraint -** the matching needs to be bijective: if $p$ in image $\mathcal{I}$ matches to $q$ in image $\mathcal{J}$ then $q$ must also match to $p$. To take different foreshortening into account we tolerate a disparity mismatch of up to one pixel in our implementation.
- **Ordering Constraint -** computed disparity at $p$ does not exceed the disparity of its right-neighbor pixel by more than one pixel.

### 2.3 Meshing

This section uses established techniques. Each camera-pair in Section 2.2 produces one disparity map, which is used to compute a corresponding array of 3D points and a corresponding array of surface normals. Since we estimate a dense disparity map, the normals are computed using finite differences on the points. 3D points and surface normals are collected across all camera pairs[3]. Outliers are removed using a simplified approach of [Merrell et al. 2007]. If two 3D points project onto the same pixel in a given camera view, both with normals facing towards that camera, and without an intermediate point with normal facing away from the camera, then the associated topology is incorrect, and the 3D point with the higher foreshortening angle is rejected. The resulting set of 3D points and normals is input to a Poisson surface reconstruction [Kazhdan et al. 2006]. The output is a triangular mesh, each vertex consisting of a 3D point plus surface normal. This mesh is then refined as described in Section 3.2.

## 3 Refinement

This section describes the refinement method that was utilized in Section 2. The refinement consists of a linear combination of two terms: a photometric consistency term $d_p$ that favors solutions with high NCC and a surface consistency term $d_s$ that favors smooth solutions. These terms are balanced both by a user-specified smoothness parameter $w_s$ and a data-driven parameter $w_p$, which ensures that the photometric term has greatest weight in regions with good feature localization. The refinement is performed both on the disparity map and later on the surface and we will discuss the individual realizations in Sections 3.1 and 3.2, resp. Both refinements are implemented as iterative processes. In practice they were found to preserve the volume and to converge quickly to the desired solution. Figure 5 shows the convergence for the disparity refinement. Since the convergence is close to exponential at the beginning, we terminate the refinement before convergence is reached to strike a balance between quality and computational effort. This is especially valuable for lower-resolution layers of the disparity pyramid, since the next higher layer is going to refine the disparities anyway and we therefore need only to eliminate the gross errors.

### 3.1 Disparity Map Refinement

Sub-pixel disparity values are updated in every iteration as a linear combination of $d_p$ and $d_s$, where $d_p$ is an adjustment in the direction of improved photometric-consistency, and $d_s$ is an adjustment in the direction of improved surface-consistency. Individual steps are -

**Compute $d_p$ -** Given current pixel $p$ in image $\mathcal{I}$ and its match $q$ in image $\mathcal{J}$, compute the $\overline{\text{NCC}}$ of $p$ with $q-1$, $q$, $q+1$ where the offsets indicate the left- and right-neighbors of $q$. We use $\overline{\text{NCC}} = (1 - \text{NCC})/2$, which resembles an error function ranging from 0 (no error) to 1 (complete dissimilarity). The respective $\overline{\text{NCC}}$s are labeled $\xi_{-1}, \xi_0, \xi_{+1}$ and $d_p$ is calculated as

$$d_p = \begin{cases} p - q - 0.5 & \xi_{-1} < \xi_0, \xi_{+1} \\ p - q + 0.5\frac{(\xi_{-1} - \xi_{+1})}{\xi_{-1} + \xi_{+1} - 2\xi_0} & \xi_0 < \xi_{-1}, \xi_{+1} \\ p - q + 0.5 & \xi_{+1} < \xi_{-1}, \xi_0 \end{cases}$$

**Compute $d_s$ -** The formulation of surface-consistency has been designed for human faces, where disparity varies smoothly with just a few (extreme) depth discontinuities. These discontinuities suggest the use of anisotropic kernels [Robert and Deriche 1996], which adapt to the local gradient to avoid smoothening across boundaries. For human faces however, regions of high gradient are mostly due to different foreshortening of the camera pairs and smoothening should not be attenuated within these regions. Following [Woodford et al. 2008] we employ second-order properties, but use them

[3]There is no special processing if multiple camera-pairs cover the same part of the face giving rise to overlapping 3D data.

within an anisotropic formulation over a two dimensional domain. The equation is discretized as

$$d_s = \frac{w_x(d_{x-1,y} + d_{x+1,y}) + w_y(d_{x,y-1} + d_{x,y-1})}{2(w_x + w_y)} \quad (1)$$

where $w_x = \exp(-(|d_{x-1,y} - d_{x,y}| - |d_{x+1,y} - d_{x,y}|)^2)$. These weights render the harmonic equation anisotropic, reducing smoothing across depth discontinuities.

**Compute $d'$ -** The equation is $d' = (w_p d_p + w_s d_s) / (w_p + w_s)$, where $w_s$ is a user-specified smoothness parameter and $w_p$ is

$$w_p = \begin{cases} \xi_{-1} - \xi_0 & \xi_{-1} < \xi_0, \xi_{+1} \\ 0.5(\xi_{-1} + \xi_{+1} - 2\xi_0) & \xi_0 < \xi_{-1}, \xi_{+1} \\ \xi_{+1} - \xi_0 & \xi_{+1} < \xi_{-1}, \xi_0 \end{cases}$$

Thus the photometric term has greatest weight in textured areas of the image where the image data is most informative about feature localization.

The refinement is terminated after a predefined number of iterations (40 for the lower-resolution layers and 180 for highest layer). See Figure 5 for justification.

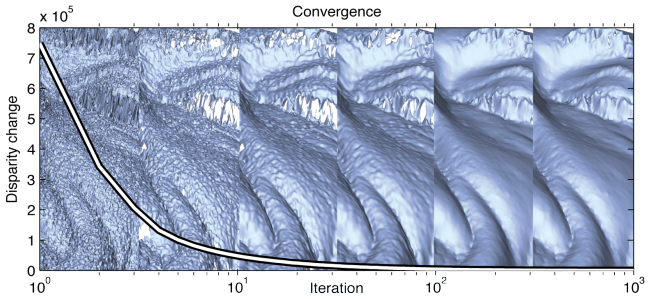


**Figure 5:** *Convergence of the refinement over the first 1000 iterations at layer $\ell_1$ using $w_s = 0.005$. The images in the back show samples of the surface at iterations 0,1,5,10,100 and 1000, resp. The initial convergence is close to exponential (note the log scale for the iteration axis) and as can be seen from the samples the quality does not change noticeably between iterations 100 and 1000. We thus stop the refinement at iteration 180.*

### 3.2 Surface Refinement

The surface refinement differs from the disparity map refinement in that we need to refine in continuous 3-space. To keep computation tractable we restrict the refinement to along the normal direction $\mathbf{n}$ at $\mathbf{X}$ and define a refinement resolution $\delta$ (usually 0.1 mm). The normals are not changed during the refinement. This again results in a discrete one-dimensional refinement and we proceed analogously to Section 3.1 by iterating over all points and computing updates for $\mathbf{X}$ as a linear combination of $\mathbf{X}_p$ and $\mathbf{X}_s$, where $\mathbf{X}_p$ is an adjustment in the direction of improved photometric-consistency, and $\mathbf{X}_s$ is an adjustment in the direction of improved surface-consistency. Individual steps are -

**Compute $\mathbf{X}_p$ -** Generate the points $\mathbf{X}_{-\delta} = \mathbf{X} - \delta\mathbf{n}$ and $\mathbf{X}_{+\delta} = \mathbf{X} + \delta\mathbf{n}$. Define as reference view the visible camera with the least foreshortened view of $\mathbf{X}$. Measure a photo-consistency error for a point by taking the $\overline{\text{NCC}}$ between a 3x3 patch centered at the projection in the reference image and the corresponding patches in all other images where the patch is visible. Compute $\delta_p$ analogously to $d_p$ given error values $\xi_{-\delta}, \xi_0$ and $\xi_{+\delta}$ for $\mathbf{X}_{-\delta}$, $\mathbf{X}_0$ and $\mathbf{X}_{+\delta}$, resp.

**Compute $\mathbf{X}_s$ -** The surface-consistency estimate $\mathbf{X}_s$ is computed using mean-curvature flow [Meyer et al. 2003].

**Compute $\mathbf{X}'$ -** Compute $\mathbf{X}' = (w_p\mathbf{X}_p + w_s\mathbf{X}_s)/(w_p + w_s)$ where $w_p$ and $w_s$ are the same as in Section 3.1.

### 3.3 Mesoscopic Augmentation

The refinement in Section 3.2 results in surface geometry that is smooth across skin pores and fine wrinkles, because the disparity change across such a feature is too small to detect[4]. The result is flatness and lack of realism in synthesized views of the face. On the other hand, visual inspection shows the obvious presence of pores and fine wrinkles in the images. This is due to the fact that light reflected by a diffuse surface is related to the integral of the incoming light. In small concavities, such as pores, part of the incoming light is blocked and the point thus appears darker. This fact has been exploited by various authors (e.g. [Glencross et al. 2008]) to infer local geometry variation. In this section we propose a method to embed this observation into our surface refinement framework. It is qualitative, and the geometry that is recovered is not metrically correct. However, augmenting the macroscopic geometry with fine-scale features does produce a significant improvement in the perceived quality of reconstructed face geometry.

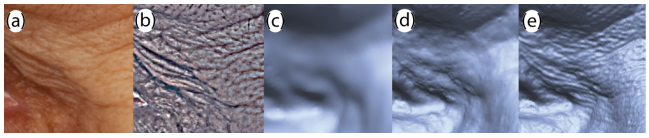


**Figure 6:** *This figure demonstrates the effect of the mesoscopic-consistency term. The captured image (a) is filtered to extract the mesoscopic detail (b). In (c), the Poisson-reconstructed surface is shown. The refinement described in Section 3.2 already enhances the coarse geometry (d), but only the mesoscopic formulation is capable of reproducing the fine-scale details (e).*

For the mesoscopic augmentation we are only interested in features that are too small to be recovered by the stereo algorithm. We therefore first compute high-pass filtered values $\mu$ for all points $\mathbf{X}$ using the projection of a Gaussian $\mathcal{N}$

$$\mu(\mathbf{X}) = \frac{\sum_{c\in\mathcal{V}} \alpha_c \left(\mathcal{I}_c(\mathbf{X}) - [\mathcal{N}_{\mathbf{\Sigma}_c} \otimes \mathcal{I}_c](\mathbf{X})\right)}{\sum_{c\in\mathcal{V}} \alpha_c}, \quad (2)$$

where $\mathcal{V}$ denotes the set of visible cameras, $\mathbf{\Sigma}_c$ the covariance matrix of the projection of the Gaussian $\mathcal{N}$ into camera $c$, and the weighting term $\alpha_c$ is the cosine of the foreshortening angle observed at camera $c$. The variance of the Gaussian $\mathcal{N}$ is chosen such that high spatial frequencies are attenuated. It can either be defined directly on the surface using the known maximum size of the features or in dependence of the matching window $m$ as described in [Beeler et al. 2010].

The next steps are based on the assumption that variation in mesoscopic intensity is linked to variation of the geometry. For human skin we found that this is mostly the case. Spatially bigger skin features tend to be smooth and are thus filtered out as shown in Figure 7. The idea is thus to adapt the local high-frequency geometry of the mesh to the mesoscopic field $\mu(\mathbf{X})$. The geometry should locally form a concavity whenever $\mu(\mathbf{X})$ decreases and a convexity

[4] This is a function of image resolution, not a limitation of the algorithm.

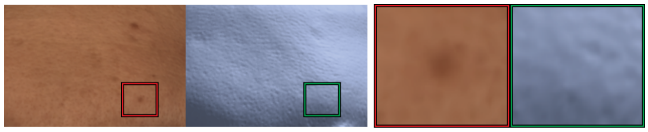

**Figure 7:** *This figure shows part of a forehead on the left and a zoomed-in patch on the right. Note that the mesoscopic augmentation adds high-frequency details such as pores, while coarser features, such as the spot showed on the right, usually do not influence the geometry, since they usually do not contain very high spatial frequencies.*

when it increases. The new position estimate $\mathbf{X}_\mu = \mathbf{X} + \delta_\mu \mathbf{n}$ is thus computed using the correctional factor

$$\delta_\mu = \eta \frac{\sum_{i \in \mathcal{R}} w_i \left(\mu(\mathbf{X}) - \mu(\mathbf{X}_i)\right) \left(1 - \frac{|<\mathbf{X}-\mathbf{X}_i, \mathbf{n}>|}{\|\mathbf{X}-\mathbf{X}_i\|}\right)}{\sum_{i \in \mathcal{R}} w_i}, \quad (3)$$

where the sum is taken over the one-ring neighborhood $\mathcal{R}$ of $\mathbf{X}$. The weights $w_i$ are computed as $w_i = \exp(-\|\mathbf{X} - \mathbf{X}_i\|)$ and $\eta$ is a user-specified parameter (the embossing strength). This equation has the property that the correctional factor $\delta_\mu$ is attenuated when there is little or no high-frequency content in the image and it is even further attenuated whenever the geometric gradient is large. This reduces the impact of the mesoscopic term on high-frequency features that can be reconstructed by the MVS, such as hair. On the other hand, the correction factor reaches its maximum when the mesoscopic gradient is large and the geometric gradient small - augmenting flat surfaces with high-frequency detail.

The update of the 3D point now uses all three adjusted points $\mathbf{X}_p$, $\mathbf{X}_s$ and $\mathbf{X}_\mu$ to compute $\mathbf{X}' = (w_p\mathbf{X}_p + w_s\mathbf{X}_s + w_\mu\mathbf{X}_\mu)/(w_p + w_s + w_\mu)$. The weights $w_p$ and $w_s$ are the same as in Section 3.2 and $w_\mu$ is defined as

$$w_\mu = \rho \frac{3\xi_0}{\delta(\xi_{-\delta} + \xi_0 + \xi_{+\delta})}, \quad (4)$$

with $\rho$ being a user specified term that controls the influence of the mesoscopic term. Figure 6 shows an example of how the mesoscopic term enriches the results. To emphasize the simplicity of the refinement we provide pseudocode in Algorithm 1. The algorithms for the refinements described in Sections 3.1 and 3.2 are very similar and of less complexity, since the mesoscopic term is not present. The function computes the position update $\mathbf{X}'$ for $\mathbf{X}$ using the normal $\mathbf{n}$. The parameters and their typical values are: resolution $\delta = 0.05$ in mm, surface smoothness $w_s = 0.03$, mesoscopic weight $\rho = 0.07$ and embossing strength $\eta = 0.2$.

**Algorithm 1** $-\mathbf{X}'$ = refinePointMesoscopic($\mathbf{X}, \mathbf{n}, \delta, w_s, \rho, \eta$)
*- NCC($\mathbf{X}$,$\mathbf{n}$) computes the normalized cross correlation of a surface patch at $\mathbf{X}$ with normal $\mathbf{n}$ by projecting it into all visible images*
*- $\bar{\kappa}$ denotes the mean curvature*

$\xi_{-\delta} = (1 - \text{NCC}(\mathbf{X} - \delta\mathbf{n}, \mathbf{n}))/2$
$\xi_0 = (1 - \text{NCC}(\mathbf{X}, \mathbf{n}))/2$
$\xi_{+\delta} = (1 - \text{NCC}(\mathbf{X} + \delta\mathbf{n}, \mathbf{n}))/2$
**if** $\xi_{-\delta} < \xi_{+\delta}, \xi_0$ **then**
  $\delta_p = -0.5\delta$
  $w_p = (\xi_{-\delta} - \xi_0)/\delta$
**else if** $\xi_{+\delta} < \xi_{-\delta}, \xi_0$ **then**
  $\delta_p = 0.5\delta$
  $w_p = (\xi_{+\delta} - \xi_0)/\delta$
**else**
  $\delta_p = 0.5(\xi_{-\delta} - \xi_{+\delta})/(\xi_{-\delta} + \xi_{+\delta} - 2\xi_0)\delta$
  $w_p = 0.5(\xi_{-\delta} + \xi_{+\delta} - 2\xi_0)/\delta$
**end if**
$\delta_s = -\bar{\kappa}\mathbf{n}$
$\delta_\mu = \eta \frac{\sum_{i \in \mathcal{R}} \exp(-\|\mathbf{X}-\mathbf{X}_i\|)(\mu(\mathbf{X})-\mu(\mathbf{X}_i))(1-|\langle \mathbf{X}-\mathbf{X}_i, \mathbf{n}\rangle|/\|\mathbf{X}-\mathbf{X}_i\|)}{\sum_{i \in \mathcal{R}} \exp(-\|\mathbf{X}-\mathbf{X}_i\|)}$
$w_\mu = 3\rho\xi_0/\delta(\xi_{-\delta} + \xi_0 + \xi_{+\delta})$
$\mathbf{X}' = \mathbf{X} + (w_p\delta_p + w_s\delta_s + w_\mu\delta_\mu)/(w_p + w_s + w_\mu)\mathbf{n}$

# 4 Results

## 4.1 Capture Process

Results were obtained using two capture systems - a studio setup and a consumer binocular-stereo camera (see Figure 1). The studio setup consists of seven cameras arranged around the subject. Neighboring camera pairs subtend an angle of about 20° at the head and the outermost cameras subtend an angle of about 110°. There are two Canon 500D cameras on each side, and three Canon 5D cameras that are arranged in a triangle and dedicated to the frontal view of the face. The cameras were synchronizable to 0.1 seconds, which is sufficient for static subjects, but not for capture of transient facial expression. We handle this by working in a darkened room, sending a signal to all cameras to open their apertures for two seconds and triggering the external flash with a one second delay. The cameras in the studio setup were manual-focus, and they were refocused and the calibration repeated for each new subject. The consumer stereo camera that we used is the Fuji Real 3D W1 shown in Figure 1. It has a stereo baseline of 77mm, and probably marks the appearance of a new market in consumer stereo photography. The auto-focus of the Fuji could not be disabled. This meant that the calibration parameters must have changed in the interval between capturing the calibration sphere and capturing the face, but this did not cause any obvious degradation of the results[5]. For both systems, images were down-sampled once due to the Bayer color-filter pattern before input to the software.

The compute time, from image input to output of a 3D model, takes around 20 minutes[6]. This is for software that has had no optimization or parallelization yet, and we believe that we can reduce compute time to a few minutes and possibly further. A related matter of practical usefulness is that the stereo matching is pyramidal and it is straightforward to quickly generate models at the lower-resolution layers for preview and checking. Model generation takes a few seconds at the lowest-resolution (150x150 pixel) layer for example.

## 4.2 Quantitative Evaluation

This section contains results for a physical mask of known ground-truth. The mask was created by taking a plaster-cast of a face, scanning with laser, and printing on an Object Connex 500 3D printer. Figure 9 shows the mask which is half a face, not a full face, due to an unwanted limitation at the time of our experiments. Error is measured as perpendicular distance between the registered ground-truth model and recovered model. The errors are listed in Table 1 and their distribution is shown in Figure 9. For comparison, the phys-

[5]In fact, this was a beneficial side-effect of the calibration method that the surface of the calibration sphere coincides with the subsequent position of the surface of the face, so auto-focus does not much change camera parameters.

[6]Compute times with the seven-camera studio setup and the Fuji binocular camera were similar, The reason is that the Fuji images are noisier and the refinements in Section 3 took longer, counteracting the effect of fewer cameras.

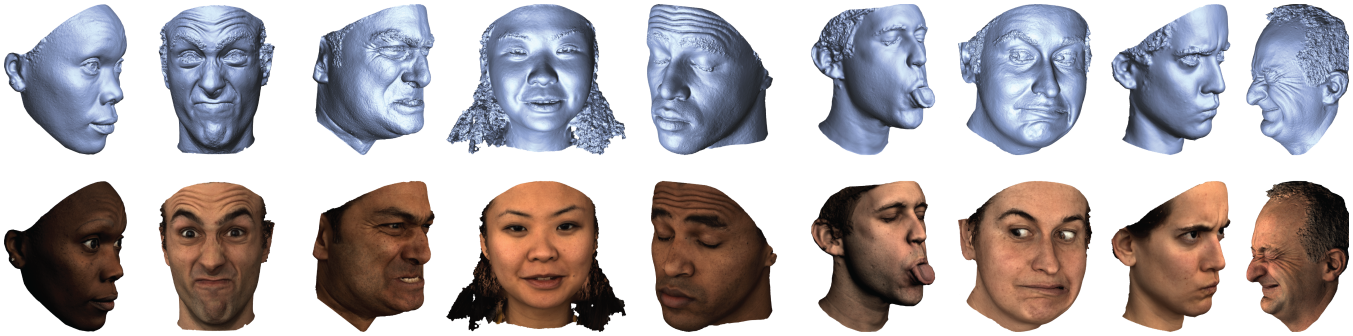

**Figure 8:** *Recovered models and synthesized views, for viewpoints different from the original camera images, across subjects of varying appearance. Our focus has been on skin, and it may be that the hair and specular components - like the eyes, teeth and tongue - benefit from custom algorithms. But the 3D reconstruction is reasonable across all these parts.*

ical resolution of the 3D printer is 0.042 mm. The error statistics include regions like the nostrils (scale $\sim$ 5 mm), where the algorithm did not reconstruct because the nostril interior is invisible in the images. Thus, the errors are an over-estimate in the sense that they include this source, but removal would have invalidated the objectivity of the result. Figure 10 provides a visual comparison of the ground-truth and recovered models. The details in the recovered model are slightly less defined but recovery of mesoscopic geometry is substantially correct.

| | Average [mm] | Median [mm] | Angular [$^\circ$] |
|---|---|---|---|
| PMVS | $0.132 \pm 0.19$ | 0.096 | $6.989 \pm 7.97$ |
| W/o Meso | $0.092 \pm 0.13$ | 0.070 | $8.690 \pm 7.69$ |
| With Meso | $\mathbf{0.088 \pm 0.12}$ | **0.067** | $\mathbf{5.675 \pm 6.03}$ |

**Table 1:** *Absolute error values for the ground-truth experiment described in Section 4.2. The angular error measures the angular difference of the normals. The refinement with mesoscopic term is superior in all cases. We included a comparison to PMVS [Furukawa and Ponce 2007] for completeness, however we want to point out that their method is a general-purpose MVS, while ours is tailored to face-reconstruction.*

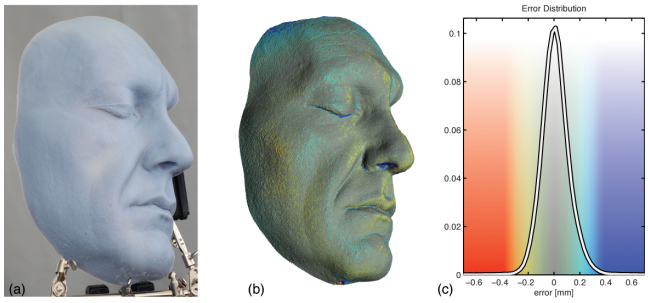


**Figure 9:** *(a) Physical mask of known ground-truth; (b) Recovered model color-coded by error; (c) Distribution of the signed absolute error between the ground-truth and the registered recovered model.*

Measured errors are not directly applicable to real faces because the surface reflectance of the face mask is different from human skin, with reduced specularity for example, but the results are informative to first-order. These experiments have also suggested an interesting possibility for future work - latest-generation 3D printers have sophisticated material handling, and it might be possible to print a mask whose reflectance properties are a better approximation to skin.

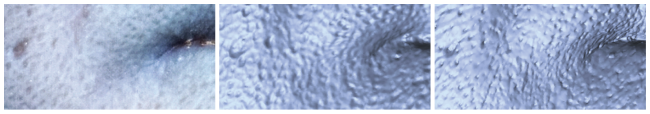

**Figure 10:** *Left: image of the physical mask; Center: rendering of the recovered model; Right: rendering of the physical mask.*

### 4.3 Qualitative Evaluation

Figure 8 shows results for a variety of subjects of varying gender, ethnicity, age, and facial expression. Figure 11 demonstrates high-fidelity reconstruction for a subject with geometric variation in the skin at a range of scales. Figure 12 shows both the subtle deformations of mesoscopic detail in distorted areas as well as their consistency in regions that do not undergo deformation. Figure 13 shows results for a subject with dark-colored skin.

**Figure 11:** *Recovered model for a face with geometric variation in the skin at a range of scales.*

Figure 14 shows models recovered for highly-transient facial expression. The subject slapped his own cheek causing a fast-moving

**Figure 14:** *Top: Images of a subject slapping himself and causing a shock-wave in the face. Bottom: the respective reconstructions.*

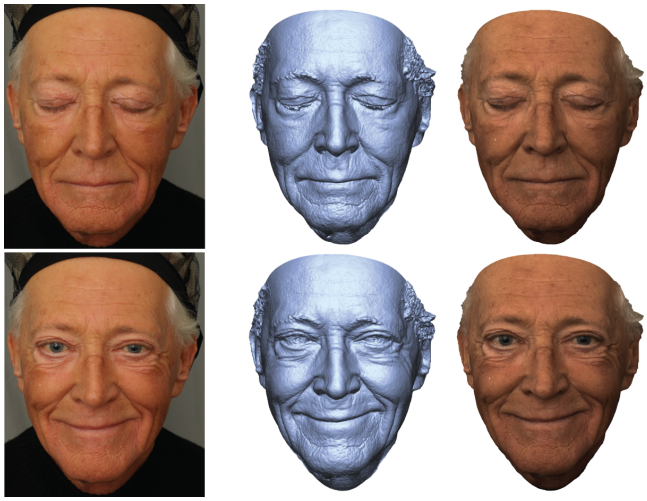

**Figure 12:** *Recovered models of a subject for two different expressions.*

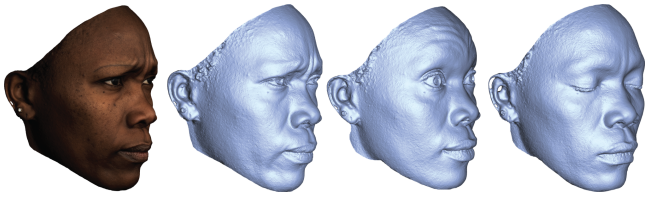

**Figure 13:** *Recovered model for a face with dark-colored skin.*

shock-wave across the face. As discussed in Section 4.1, our current studio setup is not capable of continuous capture, and the figure is showing results for multiple captures at different times, not a single shock-wave. The results illustrate the advantage of single-shot capture - a time-multiplexed system would require specialized high-speed hardware and high light-levels for this case.

Figure 15 shows results for capture from the Fuji camera. Image capture with the Fuji under normal ambient light yielded very noisy images, most likely due to the relatively small 1/2.3 inch sensor size. Doing the capture with a bright diffuse light source solved this problem and yielded the required image quality. The face model has less coverage than with the studio setup, because this is a small baseline stereo camera taking a frontal view.

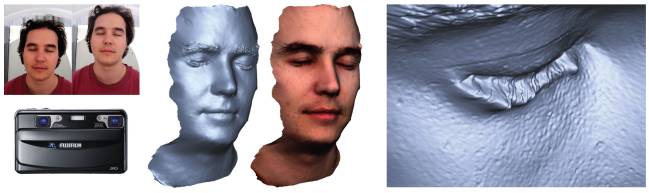

**Figure 15:** *Left: Images from the Fuji binocular-stereo camera. Center: the recovered model. Right: Close-up of a region around the eye.*

## 5 Discussion

**Robustness:** We used two contrasting capture systems, the first a studio setup with seven prosumer SLR cameras plus indirect lighting, and the second a consumer binocular-stereo camera. This illustrates system behavior on a spectrum ranging from careful capture of high-quality images to point-and-shoot capture of lower-quality images with lens distortion. It further illustrates system behavior on the spectrum of varying camera configuration, ranging from cameras all around the front hemisphere of the head to binocular stereo with a small baseline. Both capture methods yield good quality face models, providing evidence that our calibration method and run-time system are robust to changing camera configuration and changing image characteristics. We have built face models for around twenty different subjects at this stage, with multiple captures for some of the subjects. Our system works on all captures, without the need to tune the software for individual cases.

**Current Limitations:** Specularity on the face is a problem when doing capture under direct lighting, occurring for example when the tip of the nose reflects a bright light source. Specular areas typically distort the mesh. Ways to deal with this include preventing it from happening in the first place by using indirect lighting or cross-polarization, or post-processing to explicitly detect the affected area and create a plausible reconstruction.

## 6 Conclusions and Future Work

The best current methods to obtain high-quality face models use active light, and they offer reliability and accuracy. For example, laser is noted for its ability to produce point measurements of sub-millimeter accuracy, while gradient-based illumination has the ability to accentuate detail and enhance recovery of fine-scale 3D geometry. However active methods impose constraints such as the

need for special-purpose hardware, for subjects to be still, or for projected light that is intrusive due to high-brightness or strobing.

In contrast, passive stereo vision uses single-shot capture under standard light sources. And commodity cameras now routinely have the image resolution to reveal individual skin pores, so that faces provide the kind of dense evenly-distributed texture that is perfect for stereo matching and 3D reconstruction. This paper has demonstrated the capabilities of a state-of-the-art passive stereo system for face scanning. It competes with active systems in reliability and quality for high-end applications, but it is low-cost, and versatile enough to work off a consumer stereo camera. We demonstrated an augmented type of stereo refinement to qualitatively recover pore-scale geometry and yield improved visual realism in synthesized faces. Our current system is in snapshot mode, but leads naturally on to future work on image sequences. In conclusion, we believe that this work demonstrates that passive stereo has matured into a robust technology for capturing models of the face, and that its advantages will support new types of deployment.

## Acknowledgements

We would like to thank Nick Apostoloff and Nori Kanazawa from Image Movers Digital for their valuable input and xyzrgb for providing the ground-truth model. We also wish to thank Peter Kaufmann for his framework and all of our subjects for letting us capture and publish their faces, especially Manuel Lang.

## References

ALEXANDER, O., ROGERS, M., LAMBETH, W., CHIANG, M., AND DEBEVEC, P. 2009. The Digital Emily Project: Photoreal facial modeling and animation. *ACM Trans. Graph.*.

BEELER, T., BICKEL, B., BEARDSLEY, P., SUMNER, B., AND GROSS, M. 2010. High-quality single shot capture of facial geometry: Implementation details. Tech. Rep. 671, ETH Zurich.

BRADLEY, D., HEIDRICH, W., POPA, T., AND SHEFFER, A. 2010. High resolution passive facial performance capture. *ACM Trans. Graph.*.

CHEN, T., GOESELE, M., AND SEIDEL, H. 2006. Mesostructure from specularity. *CVPR*.

DI3D. 2009. Dimensional imaging. *http://www.di3d.com*.

DONNER, C., AND JENSEN, H. 2006. A spectral bssrdf for shading human skin. *Eurographics Symposium on Rendering*.

DONNER, C., WEYRICH, T., D'EON, E., RAMAMOORTHI, R., AND RUSINKIEWICZ, S. 2008. A layered, heterogeneous reflectance model for acquiring and rendering human skin. *ACM Trans. Graph.*.

FURUKAWA, Y., AND PONCE, J. 2007. Accurate, dense, and robust multi-view stereopsis. *CVPR*.

GLENCROSS, M., WARD, G., MELENDEZ, F., JAY, C., LIU, J., AND HUBBOLD, R. 2008. A perceptually validated model for surface depth hallucination. *ACM Trans. Graph.*.

GOLOVINSKIY, A., MATUSIK, W., PFISTER, H., RUSINKIEWICZ, S., AND FUNKHOUSER, T. 2006. A statistical model for synthesis of detailed facial geometry. *ACM Trans. Graph.*.

HARTLEY, R., AND ZISSERMAN, A. 2000. *Multiple View Geometry*, second ed. Cambridge University Press.

HERNANDEZ, C., VOGIATZIS, G., AND CIPOLLA, R. 2008. Shadows in three-source photometric stereo. *ECCV*.

HIEP, V., KERIVEN, R., LABATUT, P., AND PONS, J. 2009. Towards high-resolution large-scale multi-view stereo. *CVPR*.

HYNEMAN, W., ITOKAZU, H., WILLIAMS, L., AND ZHAO, X. 2005. Human face project. *SIGGRAPH 2005 Courses*.

INTEL. 2001. Opencv reference manual. *http://developer.intel.com*.

KAZHDAN, M., BOLITHO, M., AND HOPPE, H. 2006. Poisson surface reconstruction. In *SGP*.

MA, W., HAWKINS, T., PEERS, P., CHABERT, C., WEISS, M., AND DEBEVEC, P. 2007. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. *Rendering Techniques*.

MERRELL, P., AKBARZADEH, A., WANG, L., MORDOHAI, P., FRAHM, J., YANG, R., NISTER, D., AND POLLEFEYS, M. 2007. Real-time visibility-based fusion of depth maps. *ICCV*.

MEYER, M., DESBRUN, M., SCHRÖDER, P., AND BARR, A. H. 2003. Discrete differential-geometry operators for triangulated 2-manifolds. In *Visualization and Mathematics III*.

NEHAB, D., RUSINKIEWICZ, S., DAVIS, J., AND RAMAMOORTHI, R. 2005. Efficiently combining positions and normals for precise 3d geometry. *ACM Trans. Graph.*.

PARKE, F. 1974. A parametric model for human faces. *PhD Thesis, University of Utah*.

PIGHIN, F., AND LEWIS, J. 2005. Digital face cloning. *ACM Trans. Graph.*.

ROBERT, L., AND DERICHE, R. 1996. Dense depth map reconstruction: A minimization and regularization approach which preserves discontinuities. In *ECCV*.

SCHARSTEIN, D., AND SZELISKI, R. 1996. Stereo matching with non-linear diffusion. *CVPR*.

SCHARSTEIN, D., AND SZELISKI, R. 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*.

SEITZ, S., CURLESS, B., DIEBEL, J., SCHARSTEIN, D., AND SZELISKI, R. 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms. *CVPR*.

SIFAKIS, E., NEVEROV, I., AND FEDKIW, R. 2005. Automatic determination of facial muscle activations from sparse motion capture marker data. *ACM Trans. Graph.*.

SVOBODA, T. Multi camera self-calibration. *http://cmp.felk.cvut.cz/ svoboda/SelfCal/index.html*.

TORRALBA, A., AND FREEMAN, W. 2003. Properties and applications of shape recipes. *CVPR*.

WEISE, T., LEIBE, B., AND GOOL, L. V. 2007. Fast 3D scanning with automatic motion compensation. *CVPR*.

WEISE, T., LI, H., GOOL, L., AND PAULY, M. 2009. Face/off: live facial puppetry. *SCA*.

WEYRICH, T., MATUSIK, W., PFISTER, H., BICKEL, B., DONNER, C., TU, C., MCANDLESS, J., LEE, J., NGAN, A., JENSEN, H., AND GROSS, M. 2006. Analysis of human faces using a measurement-based skin reflectance model. *ACM Trans. Graph.*.

WOODFORD, O., TORR, P., REID, I., AND FITZGIBBON, A. 2008. Global stereo reconstruction under second order smoothness priors. *CVPR*.