Noname manuscript No.
(will be inserted by the editor)

Self-calibrated, multi-spectral photometric stereo for 3d face capture.

George Vogiatzis · Carlos Hernández

Received: date / Accepted: date

Abstract This paper addresses the problem of obtaining 3d detailed reconstructions of human faces in real-time and with inexpensive hardware. We present an algorithm based on a monocular multi- spectral photometric- stereo setup. This system is known to capture high-detailed deforming 3d surfaces at high frame rates and without having to use any expensive hardware or synchronized light stage. However, the main challenge of such a setup is the calibration stage, which depends on the lights setup and how they interact with the specific material being captured, in this case, human faces. For this purpose we develop a self-calibration technique where the person being captured is asked to perform a rigid motion in front of the camera, maintaining a neutral expression. Rigidity constrains are then used to compute the head's motion with a structure-from-motion algorithm. Once the motion is obtained, a multi-view stereo algorithm reconstructs a coarse 3d model of the face. This coarse model is then used to estimate the lighting parameters with a stratified approach: In the first step we use a RANSAC search to identify purely diffuse points on the face and to simultaneously estimate this diffuse reflectance model. In the second step we apply non-linear optimization to fit a non-Lambertian reflectance model to the outliers of the previous step. The calibration procedure is validated with synthetic and real data.

Keywords Photometric stereo · Multi-spectral · Faces · motion capture · calibration

G. Vogiatzis
Aston University, Birmingham, B4 7ET, UK
Tel.: +44-121-2043452 E-mail: g.vogiatzis@aston.ac.uk

C. Hernández
Google, Seattle, WA 98103, US
E-mail: carloshernandez@google.com

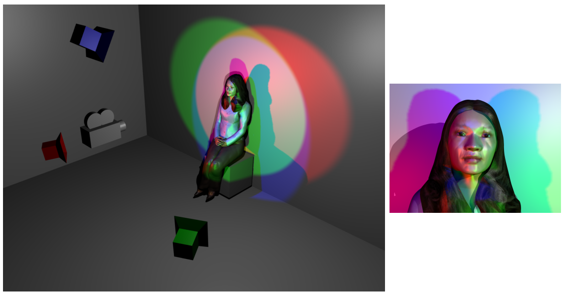

Fig. 1 Acquisition setup. The subject stands in front of three lights of different frequencies (red, green and blue) and a video camera (left). The frame captured is shown on the right. There is a $1-1$ mapping between the RGB triplet measured by the camera sensor and the surface orientation of the scene at each pixel. This leads to a system that can obtain high-detail 2.5d reconstructions of the subject at each frame and hence can be used for facial expression acquisition.

1 Introduction

The 3d capture of human faces is an important task in the fields of computer vision and computer graphics. Recent progress in hardware capabilities make the demand of such technology even greater than before, with applications ranging from medical care to human behavior or computer games. Even though much progress has been made in the recent years in deformable surface capture, faces are specially difficult to capture because humans are very well trained in face recognition and are thus very sensitive to reconstruction errors. Recent progress in facial capture has produced very high quality reconstructions to the point of being able to leap the "uncanny valley" and produce photo-realistic animations that may fool a person into thinking that the avatar is real [14]. However, these types of results can only be achieved with very expensive hardware and thousands of man-hours of interactive editing. In this paper we propose an inexpensive system based on a special case of photometric-stereo

[22] that uses multi-spectral lighting [11,23] (see Fig. 1) and that is able to capture high-detailed 3d faces in real-time. Even though the results show a low frequency shape deformation that is intrinsic to photometric stereo techniques, the algorithm is able to reconstruct very fine details such as skin porosity and wrinkles. Since the method is based on multi-spectral photometric-stereo, the system does not require any time-multiplexing hardware. However it does require a calibration for the material being captured. This means that, in practice, the system has to be calibrated for every different face to be captured. In this work we present a self-calibration algorithm that allows for automatic calibration of the setup and greatly simplifies the whole acquisition pipeline.

1.1 Related work

This paper addresses the problem of deforming shape reconstruction from images and is therefore related to a vast body of computer vision and computer graphics research. However, since faces are quite a specific type of deformable surface, we focus on facial capture systems.

For static faces, range scanner [1] or light stage setups [16] are the state-of-the-art methods to capture both accurate geometry and detailed texture. As for capturing dynamic faces, several facial performance capture systems exist using markers [3], structured light [24,25], stereo [2,8], photometric stereo [11,23] or a combination of several techniques [17]. In terms of accuracy and detail, only the methods with photometric stereo capabilities are able to capture the fine details of the face. Structured light methods such as [24,25] produce very good low frequency shape, but the need of time-multiplexing the patterns creates characteristic artifacts in the shape that need a strong post-processing stage, loosing much of the detail [21]. Stereo methods only work well whenever the face has sufficient texture [8]. In this case, the low frequency of the shape is also very accurate, but due to the nature of the cue being used, fine detail is very difficult to recover. This is in contrast to pure photometric stereo techniques, where the high frequency of the shape is easily recovered, but the low frequency is very noisy, leading to large scale deformations in the shape. Photometric stereo methods come in two variants: multi-spectral and time-multiplexing. Time-multiplexing techniques such as [17,15] need to cope with misalignment artifacts due to the fact that frames taken under different illuminations are also taken at different times. This creates creasing artifacts due to the scene motion between frames. In [15] optic flow is used to align successive frames. Also, since the effective framerate is divided by the number of lights, more expensive hardware is needed in order to obtain real-time capture frame-rates. On the other hand, multi-spectral techniques such as [11,23] (shown in Fig. 1) do not need any time-multiplexing mechanism and only require a video camera and three lights. These methods however cannot cope with different materials in the scene and need to specifically calibrate every time the material changes. In the case of human skin, the variation in skin color among several people requires individual calibration per person.

In [11] the authors propose a simple scheme for calibrating objects that can be flattened and placed on a planar board. The system detects a pattern on the board, from which it can estimate its orientation relative to the camera. By measuring the RGB response corresponding to each orientation of the material they directly estimate the linear mapping. Naturally this method cannot be applied on human faces.

In [13] a two-step process is proposed. Firstly a mirror is used to independently estimate the three light directions. The next step involves capturing three sequences of the object moving in front of the camera. In each sequence, only one of the three lights is switched on at a time and from the pixel intensities measured on the face, the light direction and RGB response of that light can be estimated. Even though this process can be applied on human faces and is very fast, it assumes that the face is Lambertian and fully monochromatic (i.e. all points have the same chromaticity value and potentially different intensity values).

The basis of this work was presented in [10]. In that paper we proposed a self-calibration method, where, before capturing a face, a short calibration sequence is obtained in order to re-calibrate the system specifically for that subject's facial skin. The method is based on using a multi-view stereo algorithm to obtain a low resolution 3d model of a face. This model is then used as a template to photometrically calibrate the rig for that particular subject. The method automatically discovers a subset of points on the face with the same chromaticity and same intensity value, and hence removes the monochromatic assumption of [13]. However, this is achieved at the expense of discarding useful calibration data, namely points on the face with the same chromaticity but differing intensity values. The method typically uses about 2-4% of points on the template.

The calibration method proposed in this paper uses the same low-resolution template as [10]. However this template is used with a robust 2d homography estimation scheme that allows us to automatically discover and use points on the face with the same chromaticity and possibly different intensity values. For the same sequences and identical threshold parameters as [10] this new technique uses around 30-40% of points on the template. Furthermore, we are also able to extend the Lambertian model assumptions by fitting a simple Phong reflectance model as a nonlinear optimization step, initialized by our robust homography solution. This allows us to correct reconstruction artefacts arising from specularities. Figure 2 shows some 3d reconstructions of a video sequence successfully calibrated using the proposed technique.

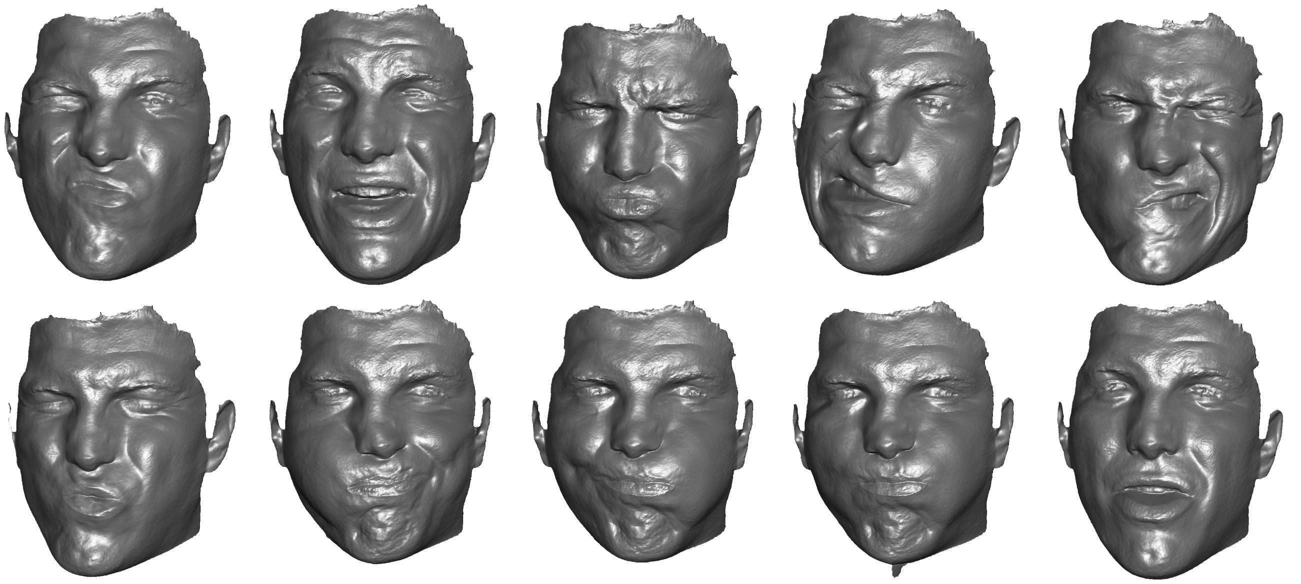

Fig. 2 Acquisition of 3d facial expressions using [11] together with the shadow processing of [13]. The system was calibrated using the self-calibration technique described in this paper.

The rest of the paper will look at the system in more detail.

2 Color photometric stereo

In classic three-source photometric stereo we are given three images of a Lambertian scene, taken from the same viewpoint, and illuminated by three distant light sources. The light sources emit the same light frequency spectrum from three different non-coplanar directions. The aim of the algorithm is to estimate from these three images the surface orientation of the object in each pixel.

Let $c_i(x,y)$ with $i = 1 \dots 3$ denote the pixel intensity of pixel (x,y) in the i-th image. We assume that in the i-th image the surface point is illuminated by a distant light source whose direction is denoted by the vector $\mathbf{l}_i$ and whose spectral distribution is $E_i(\lambda)$. We also assume that the surface point absorbs incoming light of various wavelengths according to the reflectance function $R(x,y,\lambda)$. Finally, let the response of the camera sensor at each wavelength be given by $S(\lambda)$ and $\mathbf{n}(x,y)$ the surface local normal. Then the pixel intensity $c_i(x,y)$ is given by

$$c_i(x,y) = \mathbf{l}_i^\top \mathbf{n}(x,y) \int E(\lambda) R(x,y,\lambda) S(\lambda) d\lambda. \quad (1)$$

The value of this integral is known as the surface *albedo* ρ so that (1) becomes a simple dot product

$$c_i = \mathbf{l}_i^\top \rho \mathbf{n}. \quad (2)$$

If we write $\mathbf{L} = \begin{bmatrix} \mathbf{l}_1 \ \mathbf{l}_2 \ \mathbf{l}_3 \end{bmatrix}^\top$ and $\mathbf{c} = \begin{bmatrix} c_1 \ c_2 \ c_3 \end{bmatrix}^\top$ then the system has exactly one solution for the surface orientation which is given by

$$\mathbf{n} = \frac{\mathbf{L}^{-1}\mathbf{c}}{||\mathbf{L}^{-1}\mathbf{c}||}. \quad (3)$$

Once we compute the normals, the surface can be recovered by integrating the normal field.

The core of the facial capture algorithm is based on the technique of color photometric stereo [19]. The key observation is that in an environment where red, green, and blue light is simultaneously emitted from different directions, a Lambertian surface will reflect each of those colors simultaneously without any mixing of the frequencies. The quantities of red, green and blue light reflected are a linear function of the surface normal direction. A color camera can measure these quantities from a single RGB image. In [11] it was shown how this idea can be used to obtain a reconstruction of a deforming object. Because color photometric stereo is applied on a single image, one can use it on a video sequence without having to multiplex the illumination between frames. In color photometric stereo each of the three camera sensors can be seen as one of the three images of classic photometric stereo. The pixel intensity of pixel (x,y) for the i-th sensor is given by

$$c_i(x,y) = \sum_j \mathbf{l}_j^\top \mathbf{n}(x,y) \int E_j(\lambda) R(x,y,\lambda) S_i(\lambda) d\lambda. \quad (4)$$

Note that now the sensor sensitivity S_i and spectral distribution E_j are different per sensor and per light source respectively. To be able to determine a unique mapping between

RGB values and normal orientation we need to assume a monochromatic surface. We therefore require that

$$R\left(x,y,\lambda\right)=\rho\left(x,y\right)\alpha\left(\lambda\right) \tag{5}$$

where $\rho\left(x,y\right)$ is the monochromatic albedo of the surface point and $\alpha\left(\lambda\right)$ is the characteristic chromaticity of the material. Let

$$v_{ij}=\int E_j\left(\lambda\right)\alpha\left(\lambda\right)S_i\left(\lambda\right)d\lambda \tag{6}$$

be the i^{th}-row and j^{th}-column element of matrix $\mathbf{V}$. Then the vector of the three sensor responses at a pixel is given by

$$\mathbf{c}=\mathbf{V}\cdot\mathbf{L}\rho\mathbf{n}. \tag{7}$$

The j^{th} column vector $\mathbf{v}_j$ of matrix $\mathbf{V}$ provides the response measured by the three sensors when a unit of light from source j is received by the camera. The normal is obtained by

$$\mathbf{n}=\frac{\left(\mathbf{V}\cdot\mathbf{L}\right)^{-1}\mathbf{c}}{\left\|\left(\mathbf{V}\cdot\mathbf{L}\right)^{-1}\mathbf{c}\right\|} \tag{8}$$

In order to completely calibrate the system, we only need to estimate the matrix $\mathbf{V}\cdot\mathbf{L}$ up to an unknown scale as seen from eq. (8). The next section will focus on how to estimate this matrix from a simple calibration procedure while in section 4 we will look at how to estimate a more complex nonlinear mapping that also models specular reflectance.

3 Self-Calibration of color photometric stereo system

When reconstructing 3d faces, the calibration method proposed in [13] could be used. However, although the estimation of the light directions $\mathbf{l}_i$ can be very accurate, the estimation of the color vectors $\mathbf{v}_i$ is much noisier. This is particularly true when computing the relative lengths of the vectors, *i.e.* the relative strengths of each light when interacting with the skin. The main reason for this is that [13] uses all points on the face for calibration, assuming monochromatic reflectance. Since this assumption is not true in general, the accuracy of the calibration suffers. In order to avoid these problems, we propose to use a completely automatic self-calibration process where, starting from a calibration video sequence, a coarse 3d shape of the face is computed, and the lights are estimated in a robust way so that the shape and the calibration matrix explain the video sequence as well as possible.

The calibration step is based on the fact that, even if faces are difficult to reconstruct using a passive method such as multi-view stereo [20], some algorithms can provide a sufficiently accurate reconstruction so that a robust light estimation algorithm such as [12] obtains a good estimate of the light configuration. For this purpose, a calibration sequence is recorded were the person being captured performs a rigid head motion, such as the one shown in Fig. 3. Since the expression of the face does not change during the sequence, rigidity can be used to perform standard structure-from-motion [26] in order to obtain both the camera motion (which is equivalent to the rigid head motion) and a sparse-set of 3d points (see Fig. 4). The next two sections describe in more detail the two steps involved in the calibration process: reconstruction of a coarse 3d face model and illumination estimation.

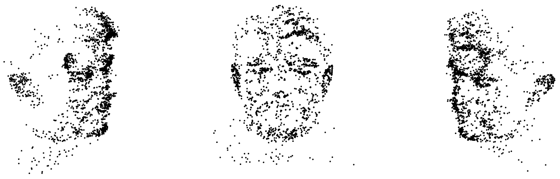

Fig. 4 Sparse set of 3d points after using a structure-from-motion algorithm on the sequence of Fig. 3. From left to right, the 3d points are shown from three different viewpoints roughly at -45 degrees, 0 degrees, and 45 degrees.

3.1 Estimating an approximate 3d face

Once the head motion is available, we can compute a dense model with a multi-view stereo algorithm. It is worth noting that the camera calibration may be inaccurate with a reprojection error of several pixels. This is due to the fact that faces have relatively few interesting points that can be well localized and tracked throughout long sequences with a small reprojection error (mainly the corner of the eyes and the mouth). Nevertheless, the calibration does not have to be very accurate as we only need a coarse shape estimate.

Figure 5 top shows the 3d reconstruction obtained with [9]. Note that the shape does not contain much detail and only the low frequencies of the shape are correct. However, as shown in the following section, this coarse shape is sufficient to estimate the lighting using [12] as only 8 dof of the matrix $\mathbf{V}\cdot\mathbf{L}$ have to be computed.

3.2 Robust estimation of light sources from a coarse shape

The estimation of the calibration matrix $\mathbf{V}\cdot\mathbf{L}$ was inspired by the photometric calibration scheme described in [12]. In that work, an initial coarse 3d shape is obtained from silhouettes, while in our case the initial shape is obtained from a multi-view stereo algorithm. We now describe the light estimation algorithm in our particular framework.

Calibrating our color photometric stereo setup involves estimating the mapping from surface orientation $\mathbf{n}$ to the RGB triplet $\mathbf{c}$ measured in the camera sensor. To perform

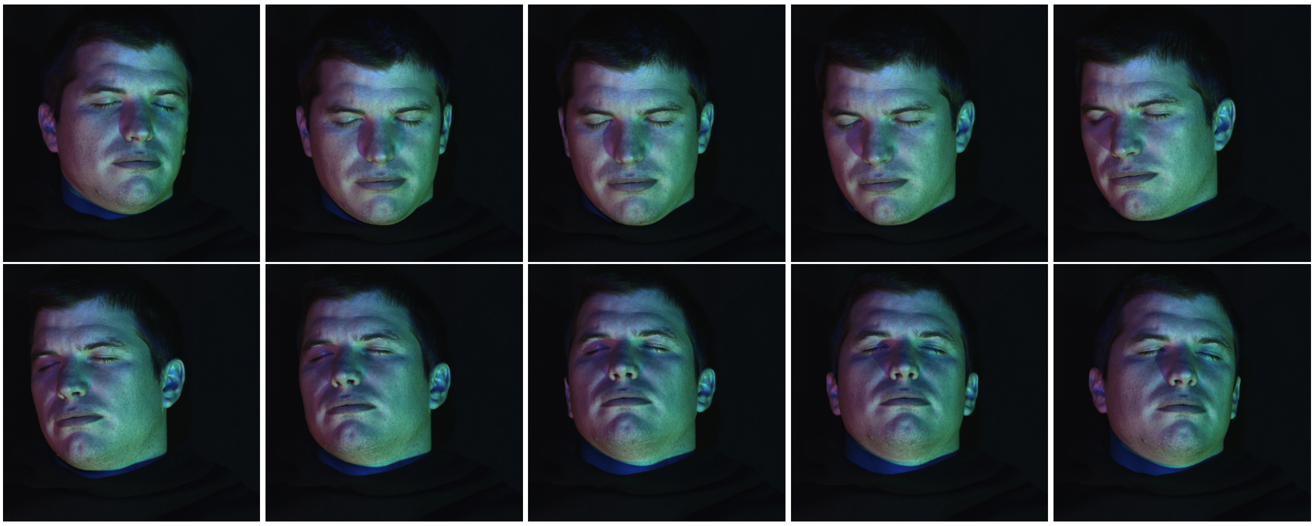

Fig. 3 Face calibration sequence under a three-source color photometric setup.

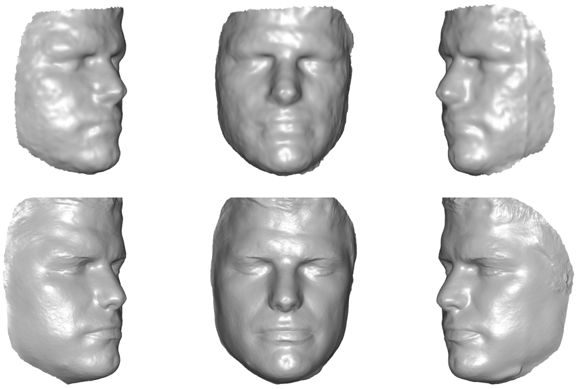

Fig. 5 Top: Coarse shape obtained with the multi-view stereo algorithm [9] on the sequence of Fig. 3. Bottom: refined shape after successful light estimation and photometric stereo evolution using the scheme of [12].

this estimation we need a set of rgb/normal pairs $(\mathbf{c}, \mathbf{n})$ in order to fit some type of parametric representation of this mapping. This set of correspondences is readily provided by our approximate 3d face model. The assumption here is that despite its inaccuracies this model will contain enough correct correspondences to accurately fix the parameters of the mapping. The only question that remains is how to robustly fit the mapping to the correct points only while disregarding possibly inaccurate points. One of the simplest and most effective robust model fitting techniques is RANSAC [6] which works by randomly samling minimal subsets of the set of correspondences. For each such subset the corresponding mapping is estimated and then all other correspondences are tested to see if they conform with it. At the end the algorithm returns the mapping with the largest number of correspondences in agreement. RANSAC based estimation is widely used for structure-from-motion problems where it has been known to find the right model in datasets with a large percentage of outlier correspondences.

In our problem, if the coarse shape contains enough correct points or inliers, then repeatedly sampling a subset of M random pairs $(\mathbf{c}, \mathbf{n})$ on the shape will give a high probability that at least one of those subsets consists of M inliers. At the same time, one can expect that the outliers do not generate a consensus in favor of any particular illumination model while the inliers do so in favor of the correct model. This observation motivated [12] to use a robust RANSAC scheme [6] to separate inliers from outliers and estimate the light matrix. The scheme can be summarized as follows:

1. Pick M random points on the coarse 3d model and, from their RGB intensities and normals, estimate mapping hypothesis.
2. Every point on the surface $\mathbf{x_m}$ will now vote for this hypothesis *if* its predicted image intensities are within a given threshold τ of the observed image intensities $\mathbf{c}_m$ where τ allows for quantization errors, image noise, etc. We use the L2 norm to calculate The distance between the predicted and observed rgb triplets.
3. Repeat 1 and 2 a set number of times always keeping the mapping hypothesis with the largest number of votes.

In practice, since we have a calibrated video sequence and not just a single frame, the algorithm uses all the frames in order to vote for a light hypothesis. This heavily increases the amount of data available.

Depending on how the mapping between normals and rgb pixel intensities is formulated and how the inlier set is defined, one can have two variants of this RANSAC algorithm. These two variants involve minimal subsets of three and four correspondences respectively. The advantage of the first variant (3-point algorithm) is that due to the smaller minimal subset it can potentially be faster to converge to the

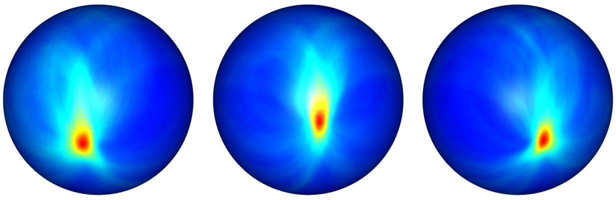

Fig. 6 Inliers for 3-point algorithm as a function of direction of the three rows of $\mathbf{V} \cdot \mathbf{L}$. This image refers to the sequence of Fig. 3 using the coarse shape of Fig. 5 top and $\tau = 4$. The image intensities are quantized in the range from 0 to 255.

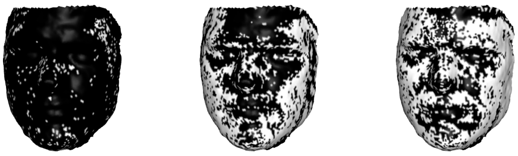

Fig. 8 Distribution of inliers (in white) for $\tau = 2$. From left to right: 3pt algorithm, 4pt algorithm, Phong model. Note that the greasy forehead is classified as an outlier (black) in both 3pt and 4pt algorithms. After optimization of the Phong model parameters, the forehead points fall within the inlier threshold.

right solution. On the other hand the second variant (4-point algorithm) typically admits a much wider set of correspondences as inliers. Both variants are discussed below.

3.2.1 Three point algorithm

If we are given three points $\mathbf{x_a}, \mathbf{x_b}, \mathbf{x_c}$ with an unknown but *equal* albedo ρ, their (non co-planar) normals $\mathbf{n_a}, \mathbf{n_b}, \mathbf{n_c}$, and the corresponding collected RGB intensities $\mathbf{c}_a, \mathbf{c}_b, \mathbf{c}_c$, we can uniquely determine the matrix $\rho\mathbf{V} \cdot \mathbf{L}$ that describes the mapping from normals to rgb triplets as follows:

$$\rho\mathbf{V} \cdot \mathbf{L} = [\mathbf{n_a}\, \mathbf{n_b}\, \mathbf{n_c}]^{-1} [\mathbf{c_a}\, \mathbf{c_b}\, \mathbf{c_c}] \,. \tag{9}$$

It is worth noting that, even though we are estimating the simplest illumination model, *i.e.* the 3×3 matrix $\mathbf{V} \cdot \mathbf{L}$, the algorithm could easily be extended to estimate a first order spherical harmonic illumination [4], *i.e.* a 3×4 matrix modeling three distant light sources plus ambient light. The RANSAC algorithm would be exactly the same, except that now it would need to pick a minimum of four points instead of three to build an illumination hypothesis. However, in all the experiments ambient light was negligible, so this extension was not necessary.

We show in Fig. 6 the number of inliers per light direction, *i.e.* per row of $\mathbf{V} \cdot \mathbf{L}$ optimized for the best scale. The space that RANSAC explores in this example is well behaved, with a clearly defined global optimum.

We show in the top row of Fig. 7 the impact of the threshold τ on the number of inliers (in white). We can distinguish how the mouth and the eyes are never selected as inliers for two different reasons. While the mouth is an outlier because of its different monochromatic intensity (different shade of red than rest of the face), the eyes are outliers because they moved during the rigid motion capture, so the reconstruction in that region is not correct. However we can further observe that even in the rest of the face the algorithm only picks a small percentage of points as inliers.

The biggest drawback of the 3-point algorithm is that the set of inliers that validate a particular hypothesis is a set of points with equal monochromatic intensity. As one can imagine, in datasets with significant variability in intensity, the number of points that have any particular intensity will be small. This means that in such cases the normal to rgb mapping estimation will be less robust as it is based on smaller datasets. However if the dataset is largely of constant monochromatic intensity then this approach may be faster to converge than the 4-point variant described below.

3.2.2 Four point algorithm

The key to describing the four point algorithm is noticing that the normal-to-rgb mapping of (7) is a $\mathbb{R}^3 \to \mathbb{R}^3$ map where the scale of one of the two vectors is unknown. It also happens to be the case however that this scale is actually not important for photometric stereo since we are really interested in the unit normal. If instead of the Euclidean 3d spaces of $(c_1, c_2, c_3)^T$ and $\rho\,(n_1, n_2, n_3)^T$ we consider the projective 2d spaces of $(c_1/c_3, c_2/c_3)^T$ and $(n_1/n_3, n_2/n_3)^T$ then the mapping of (7) is just $\mathbb{P}^2 \to \mathbb{P}^2$. This type of map is also known as a 2d homography [26] and is a very common image coordinate transformation induced by acquiring two images of a plane or when the camera motion between the two frames is a pure rotation. In our case the two spaces are not image coordinates. The first space is loosely equivalent to the hue and saturation color coordinates of the rgb triplet $(c_1, c_2, c_3)^T$ while the second space is the coordinates of the local surface gradient vector $\nabla z = \left(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}\right)^T$. Our homography is described by the matrix $\mathbf{V} \cdot \mathbf{L}$. We know that a 2d homography is completely determined if we have four correspondences between the two spaces. This leads naturally to a RANSAC algorithm that uses a minimal set of four correspondences. In fact this algorithm is identical to the well known homography estimation algorithm that is used in SfM systems [26].

The benefit of this approach is that the inlier set can be virtually all points on the coarse 3d face model that have correct geometry (correct position and surface orientation) and satisfy the monochromatic assumption. In particular, as opposed to the 3-point algorithm we are allowed to have inlier sets that have varying monochromatic intensity (i.e. brighter or darker points) as long they have the same chromaticity (hue and saturation). In sequences with significant variation in intensity this will lead to significantly larger inliers sets and improved robustness compared to the 3-point algorithm.

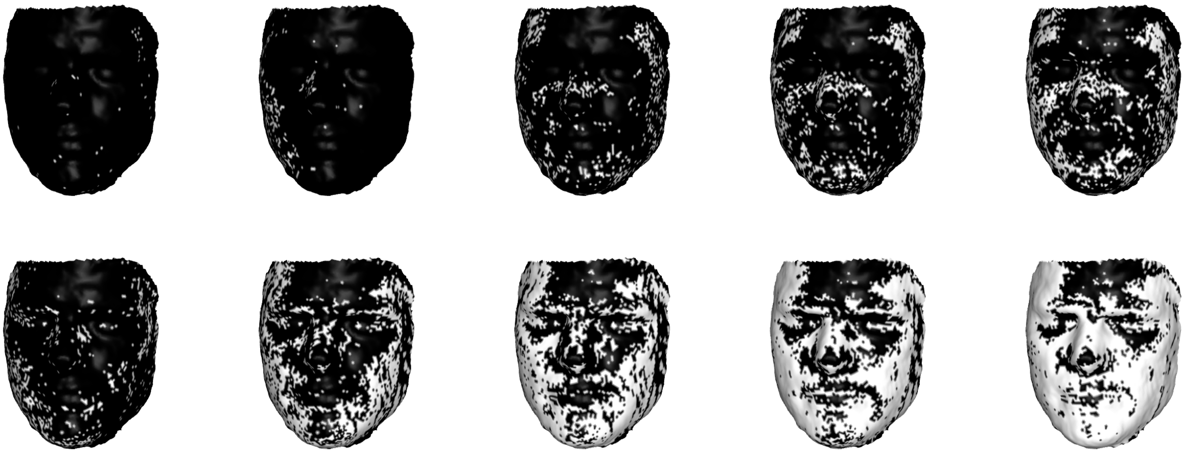

Fig. 7 Distribution of inliers (in white) as a function of the threshold τ. From left to right, $\tau = 1$, $\tau = 2$, $\tau = 3$, $\tau = 4$, $\tau = 5$. The image intensities are quantized in the range from 0 to 255. The first row shows the inliers for the 3pt algorithm while the second row is the same for the 4pt algorithm. This experiment used the calibration sequence of figure 3

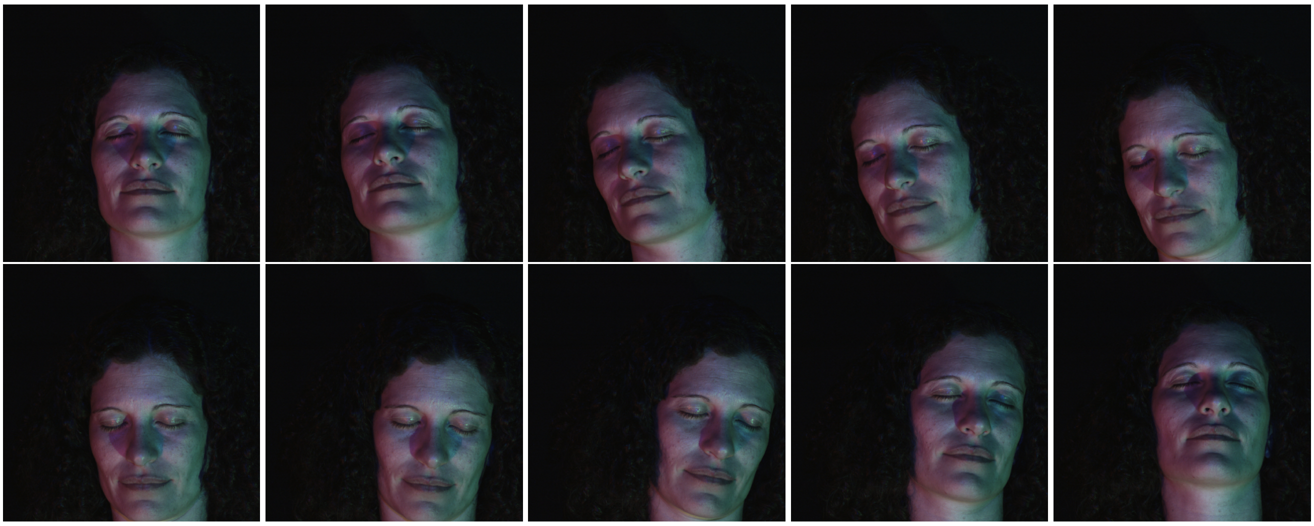

Fig. 9 Face calibration sequence under a three-source color photometric setup.

A potential drawback to using a RANSAC scheme that requires four samples instead of three in the minimal set is the fact that such a scheme might require more iterations in order to identify the correct solution. At the same time the 4-point scheme has a larger inlier set which decreases the number of iterations required. In fact it is straightforward to establish the necessary and sufficient condition under which the 4-point scheme requires less RANSAC iterations (on average) to find the solution. Let π_3 and π_4 be the percentages of inliers under the 3-point and 4-point schemes respectively. In general we can expect that $\pi_4 > \pi_3$. One can show that if and only if $\pi_4 > \pi_3^{3/4}$ then the 4-point scheme requires less iterations than the 3-point scheme to reach the same level of confidence that the true solution will be found.

In Fig. 10 we show the effect of the threshold parameter τ on the inliers picked by the 3 and 4 point algorithms. We have run both schemes on the calibration data of sequence 3 and plot the number of inliers for the same threshold value. We notice that the 4 point algorithm designates a much higher percentage of points as inliers than the 3 point algorithm.

This is also shown in the bottom of Fig. 7 where we show these inliers on the coarse face model. Note that now parts of the lips have been designated as inliers. However the eyes are still never picked due to their deformation as explained above. Finally the greasy forehead and tip of the nose are never picked because they exhibit non-Lambertian, specular effects. The following section outlines how this Lambertian reflectance assumption can be removed.

To get a better understanding of the numerical properties of the algorithm we also conducted an experiment with synthetically generated data of a textured shiny sphere (shown

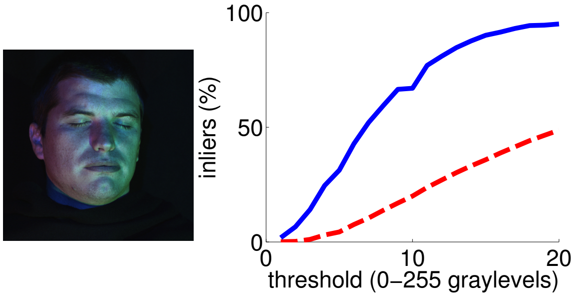

Fig. 10 This graph plots for several different threshold values, the percentage of inliers used by the 3 point and 4 point algorithms (red and blue curves respectively) when applied to the sequence of figure 3. The 4 point variant is able to treat many more points on the template as inliers.

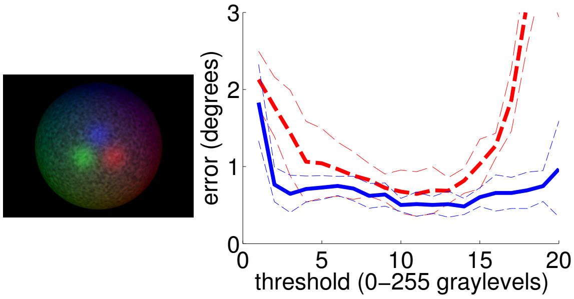

Fig. 11 Sensitivity analysis of 3pt and 4pt algorithms. We used both RANSAC variants to estimate a light matrix in the image on the left. Different amounts of noise were added to the image and the accuracy of the estimate (measured in degrees between light directions) is shown. Each experiment was repeated 20 times. The blue curves are the mean and std. deviations of the accuracy of the 4pt variant while the red curve is the 3pt variant. The 4 point RANSAC scheme seems to be much more tolerant to noise levels. This is due to the fact that it uses a much wider pool of inliers than its 3pt counterpart.

in the left of Fig. 11). In this case accurate ground truth of the scene illumination is known so we can perform some sensitivity analysis and comparison of the 3 point and 4 point variants. To that end we corrupted the synthesized image of the sphere with noise and measured the accuracy of the obtained calibrations under both schemes. Figure 11 plots the error in the calibration for both schemes as we vary the threshold τ. The error was measured as the mean angle between rows of the estimated and ground truth matrix $\mathbf{V}\cdot\mathbf{L}$. All experiments were repeated 20 times and we also show 2-standard-deviation intervals for the error values. Our findings confirm that the 4 point algorithm is able to offer better estimates across all threshold values. The 'smile' shape of these plots is due to the fact that if the threshold is too small, the set of inliers is too small and the estimation is noisy with high error. On the other hand if the threshold is too big then some outliers are entering the estimation which again increases the error. The 4 point algorithm also appears to be much flatter which means that it is more stable with regards to choosing a threshold value τ.

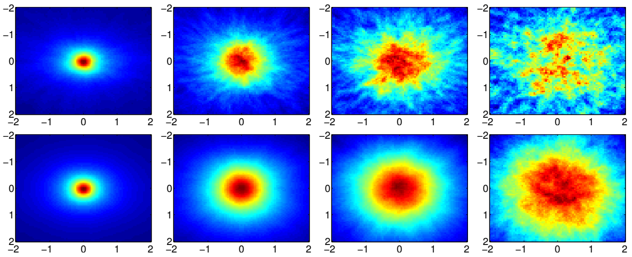

Fig. 12 Effect of noise in search space. For the synthetic data used in figure 11 we render the inliers corresponding to 3pt algorithm (1st row) and 4pt algorithm (2nd row) as a function of a displacement of the x and y coordinates of the last row of $\mathbf{V}\cdot\mathbf{L}$. The middle point in each image corresponds to zero displacement from the ground truth value of $\mathbf{V}\cdot\mathbf{L}$. For all experiments we used a threshold value of $\tau = 10$. From left to right are different noise levels (std. dev. of noise is 1,2,3 and 4 graylevels respectively). Note that the shape of the search space becomes much less smooth in the 3pt case, in the presence of some image noise.

Finally in Fig. 12 we used different levels of noise to contaminate the synthetic sphere image and we are showing the number of inliers around the ground truth value for the 3 and 4 point algorithms. The x and y coordinates of the images shown correspond to a displacement of the x and y coordinates of the third row of $\mathbf{V}\cdot\mathbf{L}$, so the $(0,0)$ corresponds to the actual ground truth calibration. We observe that the shape of the cost function explored by RANSAC becomes progressively less smooth for the 3 point algorithm (Fig. 12 top row). The 4 point variant however has a much more regular shape with a well defined maximum even at high levels of noise (Fig. 12 bottom row).

4 Fitting a specular reflectance model

The RANSAC scheme described in the previous section, in both its 3-point and 4-point variants is based on Lambertian reflectance assumptions. More specifically it is assumed that there is a set of points on the coarse 3d face model, whose reflectance is purely diffuse with no specular component. This assumption is justified in cases where (a) the face is trully perfectly diffuse (e.g. through the use of special make-up) or (b) where the face is so shiny that the specularity is very localized. In this latter case only a small subset of the points will have been contaminated and one can still calibrate the system using the rest of the 3d shape. However there are cases where the specular reflectance extends to a large range of viewing angles, producing a big specularity on the images. Such an example can be seen in Fig. 8 (middle) where we show the inliers of the 4-point algorithm on the face template. Notice that there were no inliers registered in the forehead because due to skin grease that area had a significant specular reflectance component. In an video sequence of that face we can expect to have a significant number of pixels corrupted by specularities.

Clearly, if we use the mapping estimated by RANSAC on these pixels, the surface orientation we obtain would be incorrect. When we then try to integrate these orientations to obtain 3d shape we will observe characteristic 'bulging' artifacts. Such an example is shown in the first two images of Figure 13 where the forehead seems to be slightly protruding. To remedy this problem we would need to fit a more complex reflectance model that includes specular reflectance. In this work we experimented with fitting a simple Phong model. Even though this is a very simplistic model that has well known limitations, it was adequate for the purposes of photometric stereo reconstruction. The Phong reflectance model for a scene with three color lights and three pixel sensors is given by

$$\mathbf{c} = \mathbf{c}^{diff} + \mathbf{c}^{spec} \tag{10}$$

where the diffuse component is as previously

$$\mathbf{c}^{diff} = \rho \mathbf{V} \cdot \mathbf{L} \mathbf{n} \tag{11}$$

while the specular component is given by

$$c_i^{spec}(x,y) = \sum_j \left[\mathbf{l}_j^T \left(2\mathbf{n}\mathbf{n}^T - I \right) \mathbf{v} \right]^\alpha \times \int E_j(\lambda) R_{spec}(x,y,\lambda) S_i(\lambda) d\lambda$$

In this equation the sensor sensitivity S and light source spectral distribution E are the same as previously. However in general the material will have a different reflectance function for the specular component given by R_{spec}. Parameter α is known as the specular *hardness* and it controls the size of the specular lobe. A large value corresponds to a narrow specular lobe while a small value makes it wide. The vector $\mathbf{v}$ is the viewpoint direction while vector $\left(2\mathbf{n}\mathbf{n}^T - I\right)\mathbf{v}$ is the specular direction.

To make the problem tractable, as previously, we will assume monochromaticity. For the same reasons, this time we will also require the specular monochromatic albedo of the material to be constant. If the specular albedo was allowed to vary on the surface, the mapping between normal and rgb triplet would no longer be invertible. This is because there would be four unknowns per pixel (two for direction and two for the albedos) and only three constraints from the three sensors). The assumption appears to be validated in practice. In vector form, the specular reflection component can be written as

$$\mathbf{c}^{spec} = \mathbf{W} \left[\mathbf{L} \left(2\mathbf{n}\mathbf{n}^T - I \right) \mathbf{v} \right]^\alpha \tag{12}$$

where $\mathbf{W}$ is a 3×3 matrix and α is a scalar parameter and both of which are constant for all pixels.

Having obtained an estimate of the Lambertian reflectance model $\mathbf{V} \cdot \mathbf{L}$ as well as an inlier/outlier classification for each point using RANSAC (see previous section) we now fit the specular model described above. We need to estimate matrices $\mathbf{W}$, and $\mathbf{L}$ as well as α. The matrix $\mathbf{V} \cdot \mathbf{L}$ is assumed to be given by our previous RANSAC estimation so the value of $\mathbf{V}$ is automatic given $\mathbf{L}$. Also, we need to obtain the diffuse albedo $\rho()$ for each point on the face template model. Our approach is to minimise the L2 norm of the differences between observed rgb triplets and the ones synthesised through the model:

$$\min_{\alpha, L, W, \rho} \sum_{(x,y)} \left\| \mathbf{c} - \rho \mathbf{V} \cdot \mathbf{L}\mathbf{n} + \mathbf{W} \left[\mathbf{L} \left(2\mathbf{n}\mathbf{n}^T - I \right) \mathbf{v} \right]^\alpha \right\|^2 . \tag{13}$$

Since we already have identified points of the template that adhere to the Lambertian reflectance, our cost function need only be optimized for the outlier points. It is worth pointing out that given a particular choice of light direction matrix $\mathbf{L}$ and hardness parameter α the rest of the unknown variables can be obtained through solving a linear least squares problem. This makes the optimization process considerably more efficient. We apply a simple nonlinear optimisation scheme (in our experiments we used Matlab's `lsqnonlin` function). We initialize $\mathbf{L}$ to the normalised columns of $\mathbf{V} \cdot \mathbf{L}$ while α is typically initialised to 1.

Figure 13 shows a 3d face model obtained by a Phong model whose parameters were estimated from the template of 5. To invert the Phong reflectance model per pixel of the input image we minimise the difference between the synthesized and observed rgb triplets for each pixel. The minimization is performed with respect to each pixel's surface orientation and diffuse albedo:

$$\min_{\mathbf{n}, \rho} \left\| \mathbf{c} - \rho \mathbf{V} \cdot \mathbf{L}\mathbf{n} + \mathbf{W} \left[\mathbf{L} \left(2\mathbf{n}\mathbf{n}^T - I \right) \mathbf{v} \right]^\alpha \right\|^2 \tag{14}$$

where the reflectance parameters $\mathbf{L}$, $\mathbf{W}$, α are those obtained from the previous calibration step. Once again, for any given surface orientation $\mathbf{n}$ the diffuse albedo is trivially computed via least squares so the search can be limited to $\mathbf{n}$. To optimize the cost with respect to $\mathbf{n}$ we sample 64 locations on the unit sphere and we use the location with the minimum cost to initialize a gradient descent optimization. In our experiments this simple strategy never failed to converge to the global optimum.

After estimating the parameters of the reflectance model and inverting the model to get surface orientations for each input image, we can optionally refine the initial coarse geometry with the photometric cue by evolving the surface using a scheme such as [18] or [12]. We show in Fig. 5 bottom how, by merging the multi-view stereo cue and the photometric stereo cue, the low frequency shape of the multi-view stereo solution is kept, while the high frequency shape of the photometric stereo cue is "added" creating a very detailed and realistic static reconstruction of the face.

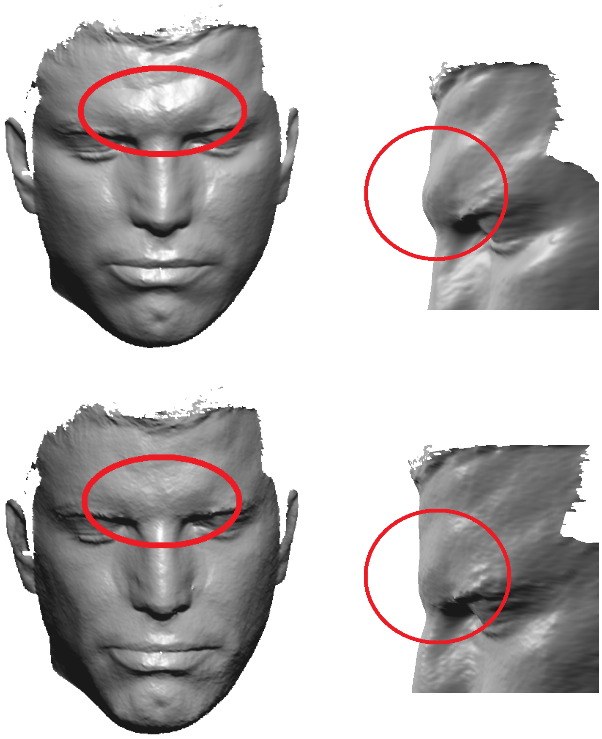

Fig. 13 Color photometric stereo using the Phong model. We used the calibration sequence of figure 3 to fit the parameters of a Phong model. This was then used to reconstruct a single frame of the sequence. The first two images show what happens if we just use the Lambertian mapping estimated in section 3.2 while the second two images show the reconstruction under the Phong model described in section 4. Notice the characteristic 'bulging' artefact that appears when specularities are treated with the Lambertian model. These artefacts are eliminated when the Phong model is used.

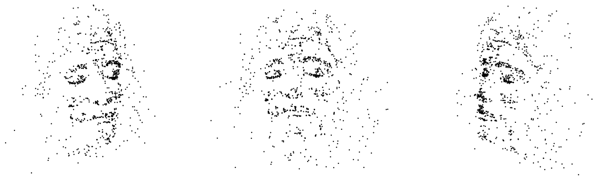

Fig. 14 Sparse set of 3d points after using a structure-from-motion algorithm on the sequence of Fig. 9. From left to right, the 3d points are shown from three different viewpoints roughly at -45 degrees, 0 degrees, and 45 degrees.

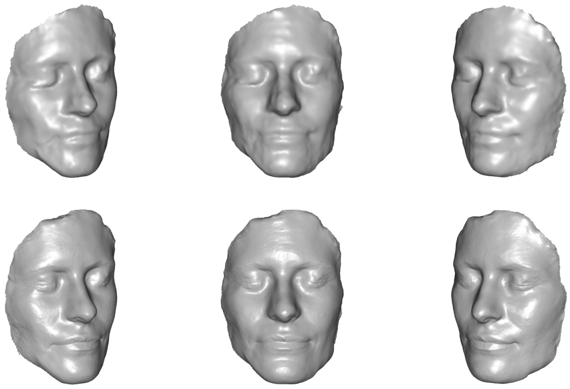

Fig. 15 Top: Coarse shape obtained with the multi-view stereo algorithm [9] on the sequence of Fig. 9. Bottom: refined shape after successful light estimation and photometric stereo evolution using the scheme of [12].

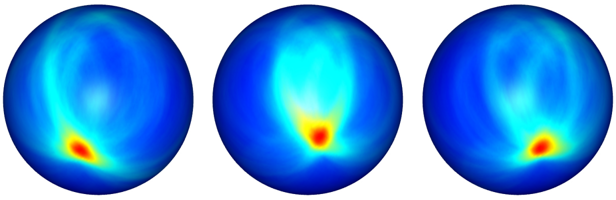

Fig. 16 Inliers for 3-point algorithm as a function of direction of the three rows of $\mathbf{V} \cdot \mathbf{L}$. This image refers to the sequence of Fig. 9 using the coarse shape of Fig. 15 top and $\tau = 4$. The image intensities are quantized in the range from 0 to 255.

5 Experimental results

We have run the same algorithm on a second sequence shown in Fig. 9. After structure-from-motion, the camera motion and the video sequence are fed into the multi-view stereo algorithm in order to produce a coarse shape of the face shown in Fig. 15 top. The sparse set of 3d points (shown in Fig. 14) is only used to define a rough bounding box in order to speed-up the multi-view stereo algorithm. Once the coarse shape is computed, we can run the light calibration step described in Section 3.2, giving the light estimates shown in Fig. 16. Again, in order to have an idea of how good the estimate is, we can visualize the distribution of inliers w.r.t the RANSAC threshold τ (see Fig. 17) and we can also refine the coarse shape in order to obtain a high resolution static face capture (see Fig. 15 bottom).

Once the calibration step is completed, we can reconstruct video footage of that same person under the same setup using [11](see Fig. 18). Note that, wherever the constant chromaticity assumption is not verified, *e.g.* in the eyes or inside the mouths, the normal estimation suffers from a bas-relief ambiguity deformation [5]. However the impact of such ambiguity in the final shape depends on the size of the region. If the region is small compared to the rest of the image, as it is the case with the lips, the low frequency of the shape will not be very distorted since it is computed as an integration process of the entire image. As for the high frequency, it will bump the surface in a realistic way even if, over all, the normals are distorted.

As an improvement to [11], we use a real-time implementation of the algorithm. Since the reconstruction algorithm itself is just a per-pixel 3×3 matrix-vector multiplication followed by a Poisson integration step[7], this can be achieved real-time at 60 Hz by using an FFT-based integration implemented on a gpu (with the CUDA libraries).

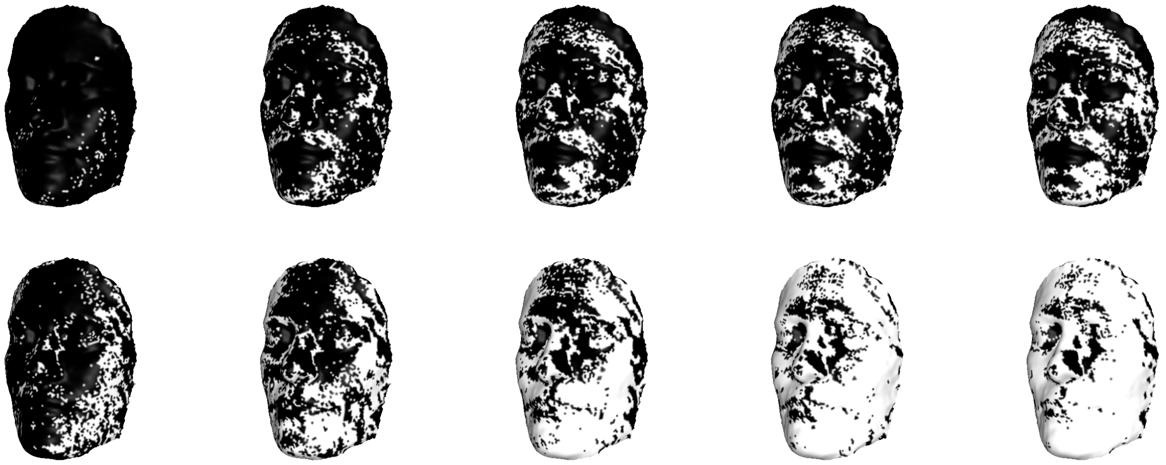

Fig. 17 Distribution of inliers (in white) as a function of the threshold τ. From left to right, $\tau = 1$, $\tau = 2$, $\tau = 3$, $\tau = 4$, $\tau = 5$. The image intensities are quantized in the range from 0 to 255. The first row shows the inliers for the 3pt algorithm while the second row is the same for the 4pt algorithm. This experiment used the calibration sequence of figure 9

6 Conclusion

We have presented a self-calibration method for monocular 3d face capture using a color photometric stereo framework. The method is based on a preliminary video capture of the person where a rigid motion is performed with a neutral facial expression. This enables us to use a structure-from-motion algorithm followed by a multi-view stereo algorithm in order to reconstruct a coarse 3d shape of the static face. The same calibration video can then be used together with the shape in order to robustly estimate the color response of the face under the photometric stereo setup. Once the system is calibrated, reconstruction of 3d faces can be achieved in a live real-time manner.

The main weakness of the proposed reconstruction framework is the low frequency noise in the 3d shape, which is characteristic of photometric stereo algorithms. A promising research direction is to combine this technique with other cues such as MVS [8] that can constrain the low-frequency of the shape.

References

1. Cyberware, inc. http://cyberware.com 2
2. Dimensional imaging. http://www.di3d.com 2
3. Mova. http://www.mova.com 2
4. Basri, R., Jacobs, D., Kemelmacher, I.: Photometric stereo with general, unknown lighting. Int. J. Comput. Vision **72**(3), 239–257 (2007). DOI http://dx.doi.org/10.1007/s11263-006-8815-7 6
5. Belhumeur, P.N., Kriegman, D.J., Yuille, A.L.: The bas-relief ambiguity. Int. J. Comput. Vision **35**(1), 33–44 (1999) 10
6. Fischler, M., Bolles, R.: Random sample consensus: A paradigm for model-fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981) 5
7. Frankot, R.T., Chellappa, R.: A method for enforcing integrability in shape from shading algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **10**(4), 439–451 (1988). DOI http://dx.doi.org/10.1109/34.3909 10
8. Furukawa, Y., Ponce, J.: Dense 3d motion capture for human faces. In: IEEE Conference on Computer Vision and Pattern Recognition (2009) 2, 11
9. Hernández, C., Schmitt, F.: Silhouette and stereo fusion for 3d object modeling. Computer Vision and Image Understanding **96**(3), 367–392 (2004) 4, 5, 10
10. Hernández, C., Vogiatzis, G.: Self-calibrating a real-time monocular 3d facial capture system. In: Proceedings International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT), 2010 (2010) 2
11. Hernández, C., Vogiatzis, G., Brostow, G., Stenger, B., Cipolla, R.: Non-rigid photometric stereo with colored lights. In: IEEE International Conference on Computer Vision (2007) 2, 3, 10, 12
12. Hernández, C., Vogiatzis, G., Cipolla, R.: Multi-view photometric stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(3), 548–554 (2008) 4, 5, 9, 10
13. Hernández, C., Vogiatzis Cipolla, R.: Shadows in three-source photometric stereo. In: IEEE European Conference on Computer Vision (2008) 2, 3, 4
14. Image Metrics, U.I.f.C.T.: Emily project. SIGGRAPH 2008 Demo session (2002) 1
15. Kim, H., Wilburn, B., Ben-Ezra, M.: Photometric stereo for dynamic surface orientations. In: Proc. European Conf. on Computer Vision (2010) 2
16. Ma, W., Hawkins, T., Peers, P., Chabert, C., Weiss, M., Debevec, P.: Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In: Eurographics Symposium on Rendering, pp. 183–194 (2007) 2
17. Ma, W., Jones, A., Chiang, J., Hawkins, T., Frederiksen, S., Peers, P., Vukovic, M., Ouhyoung, M., Debevec, P.: Facial performance synthesis using deformation-driven polynomial displacement maps. ACM Transactions on Graphics **27**(5) (2008) 2
18. Nehab, D., Rusinkiewicz, S., Davis, J., Ramamoorthi, R.: Efficiently combining positions and normals for precise 3d geometry. In: Proc. of the ACM SIGGRAPH, pp. 536–543 (2005) 9
19. Petrov, A.: Light, color and shape. Cognitive Processes and their Simulation (in Russian) pp. 350–358 (1987) 3

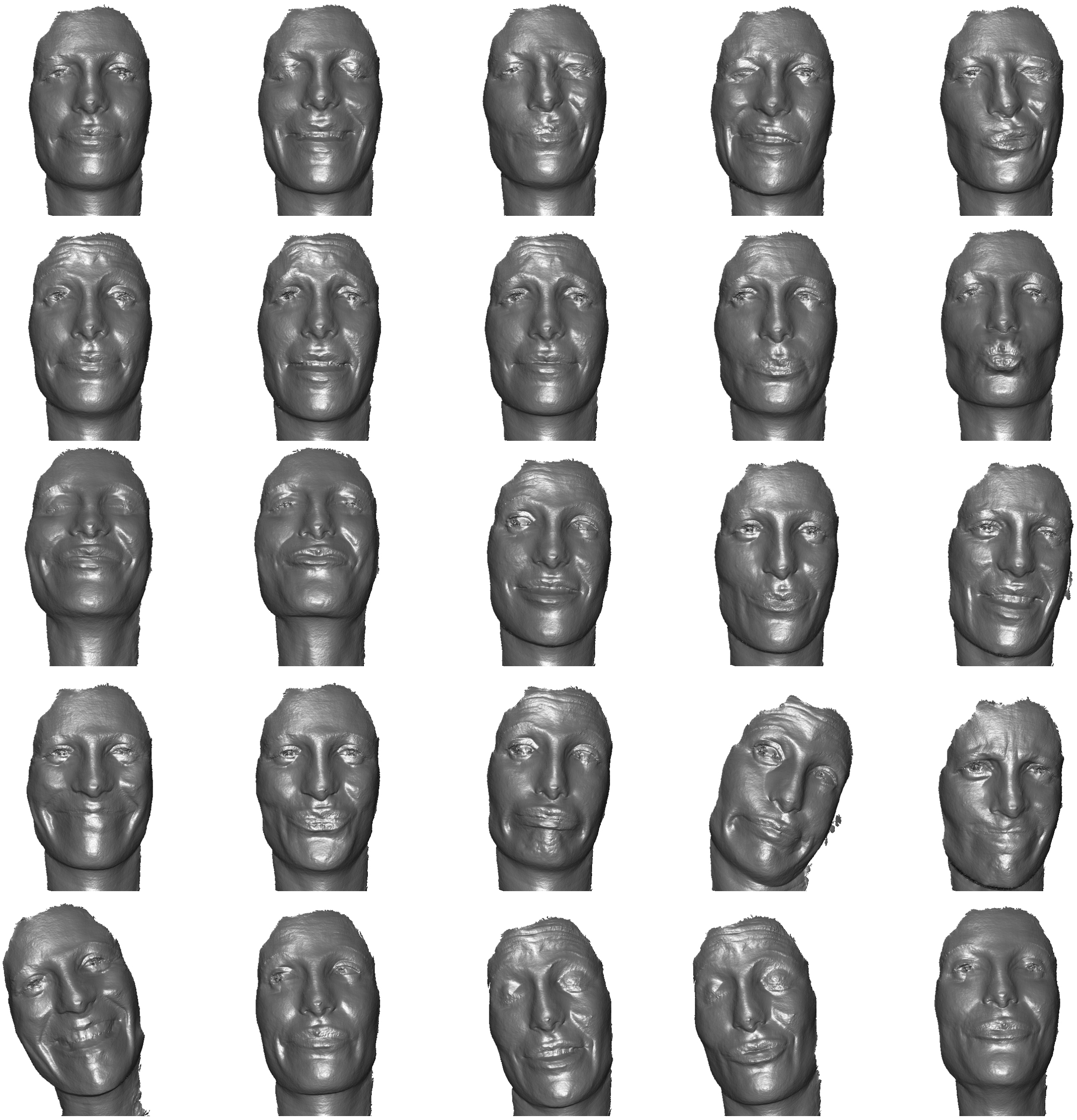

Fig. 18 Acquisition of 3d facial expressions using [11]. The system was calibrated using the self-calibration technique described in this paper.

20. Seitz, S., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol. 1, pp. 519–528 (2006) 4
21. Weise, T., Leibe, B., Gool, L.V.: Fast 3d scanning with automatic motion compensation. In: IEEE Conference on Computer Vision and Pattern Recognition (2007) 2
22. Woodham, R.: Photometric method for determining surface orientation from multiple images. Optical Engineering **19**(1), 139–144 (1980) 2
23. Woodham, R.J.: Gradient and curvature from the photometric-stereo method, including local confidence estimation. J. Opt. Soc. Am. A **11**(11), 3050–3068 (1994) 2
24. Zhang, L., Snavely, N., Curless, B., Seitz, S.M.: Spacetime faces: high resolution capture for modeling and animation. In: SIGGRAPH '04, pp. 548–558 (2004) 2
25. Zhang, S., Huang, P.S.: High-resolution, real-time three-dimensional shape measurement. Optical Engineering **45**(12) (2006) 2
26. Zisserman, A., Hartley, R.: Multiple View Geometry. Springer-Verlag (2000) 4, 6