Self-calibrating a real-time monocular 3d facial capture system

Carlos Hernández
Toshiba Research Europe
carlos.hernandezesteban@gmail.com

George Vogiatzis
Toshiba Research Europe
george.vogiatzis@gmail.com

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 robust estimator which allows for detailed real-time 3d capture of faces. The calibration procedure is validated with two real sequences.

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 photorealistic animations that may fool a person into thinking that the avatar is real [13]. 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 [20] that uses multi-spectral lighting [10, 21] 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.

2. 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 [14] 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 [22, 23], stereo [2, 8], photometric stereo [10, 21] or a combination of several techniques [15]. 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 [22, 23] 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 [19]. 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

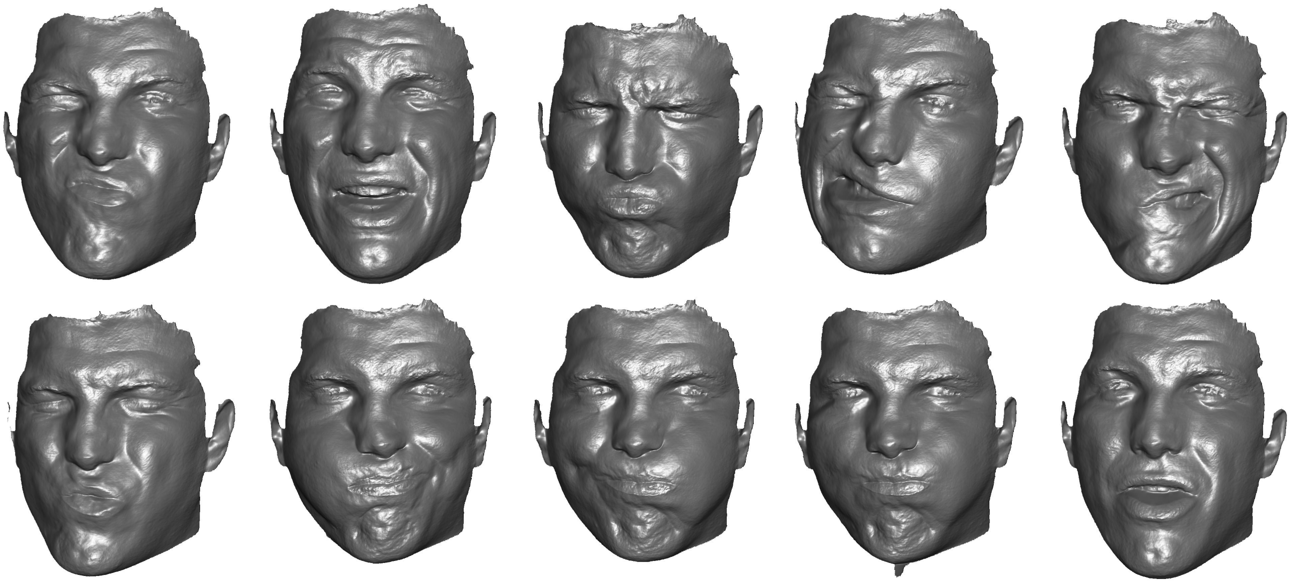

Figure 1. Acquisition of 3d facial expressions using [10] together with the shadow processing of [12]. The system was calibrated using the self-calibration technique described in this paper.

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 [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 wrinkling artifacts due to the scene motion between 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 [10, 21] do not need any time-multiplexing mechanism and only requires a video camera and three light projectors. 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 [10] 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 [12] 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 fully monochromatic.

In this paper we propose a very accurate self-calibration method, where, before capturing a face, a short calibration sequence is obtained in order to re-calibrate the system specifically for that person's facial skin. The method automatically discovers points on the face with the same albedo, and hence removes the assumption of [12]. Figure 1 shows some 3d reconstructions of a video sequence successfully calibrated using the proposed technique.

3. 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.

Let $c_i(x, y)$ with $i = 1 \ldots 3$ 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 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 [17]. 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 [10] 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 $R(x,y,\lambda) = \rho(x,y) \, \alpha(\lambda)$. Where $\rho(x,y)$ is the monochromatic albedo of the surface point and $\alpha(\lambda)$ is the characteristic chromaticity of the material. Let

$$v_{ij} = \int E_j(\lambda) \, \alpha(\lambda) \, S_i(\lambda) \, d\lambda$$

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

$$\mathbf{c} = \mathbf{V} \cdot \mathbf{L} \rho \mathbf{n}. \quad (5)$$

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.

In order to completely calibrate the system, only the knowledge of the product matrix $\mathbf{V} \cdot \mathbf{L}$ is required.

4. Calibration

When reconstructing 3d faces, the calibration method proposed in [12] 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 [12] 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.

In the following we describe the three steps involved in the calibration process: camera calibration, shape reconstruction and light matrix estimation.

4.1. Sequence capture and camera calibration

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 [18], some algorithms can provide a sufficiently accurate reconstruction so that a robust light estimation algorithm such as [11] 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. 2. Since the expression of the face does not change during the sequence, rigidity can be used to perform standard structure-from-motion [24] in order to obtain both the camera motion and a sparse-set of 3d points (see Fig. 3).

4.2. Coarse shape estimation

Once camera calibration 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.

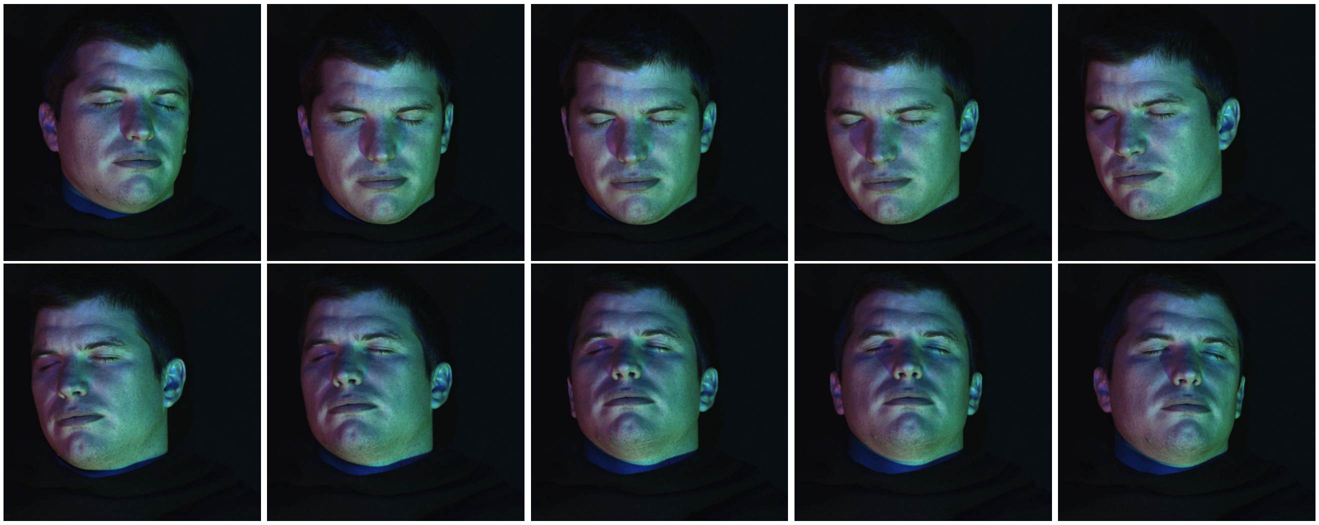

Figure 2. Face calibration sequence under a three-source color photometric setup.

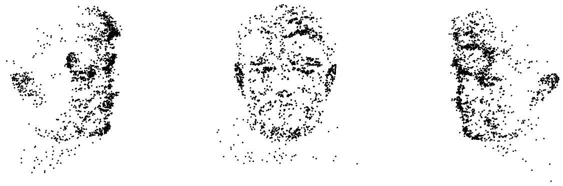

Figure 3. Sparse set of 3d points after using a structure-from-motion algorithm on the sequence of Fig. 2. From left to right, the 3d points are shown from three different viewpoints roughly at -45 degrees, 0 degrees, and 45 degrees.

Figure 4 top shows the 3d reconstruction obtained with [9]. Note that the shape does not contain much de tail 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 [11] as only 8 dof of the matrix $\mathbf{V} \cdot \mathbf{L}$ have to be computed.

4.3. Robust estimation of light sources from a coarse shape

The estimation of the calibration matrix $\mathbf{V} \cdot \mathbf{L}$ is based on the algorithm described in [11]. In that work the initial coarse shape is a 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.

Similar to the photometric stereo algorithm, the core of the calibration step is based on equation (5). In order to use this equation to perform photometric stereo, the given inputs are the collected intensities $\mathbf{c}$ and the light matrix $\mathbf{V} \cdot \mathbf{L}$ while the unknowns are the surface normals $\mathbf{n}$. For calibration purposes, the inputs are pairs of image RGB intensities and surface normals while the unknowns are the light matrix $\mathbf{V} \cdot \mathbf{L}$.

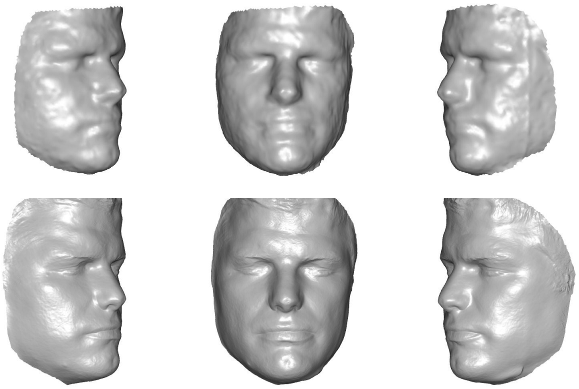

Figure 4. Top: Coarse shape obtained with the multi-view stereo algorithm [9] on the sequence of Fig. 2. Bottom: refined shape after successful light estimation and photometric stereo evolution using the scheme of [11].

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 $\rho\mathbf{V} \cdot \mathbf{L}$ as

$$\rho\mathbf{V} \cdot \mathbf{L} = \left[\mathbf{n_a}\,\mathbf{n_b}\,\mathbf{n_c}\right]^{-1}\left[\mathbf{c_a}\,\mathbf{c_b}\,\mathbf{c_c}\right]. \tag{6}$$

For multiple images, these same three points can provide the light matrix in each image up to a global unknown scale factor. The problem is then how to obtain three such points.

The answer is that, if the coarse shape contains enough correct points or inliers, then repeatedly sampling a triplet of random points on the shape will give a high probability that at least one of those triplets contains three inliers. At the same time, one can expect that the outliers do not generate a consensus in favor of any particular illumination model

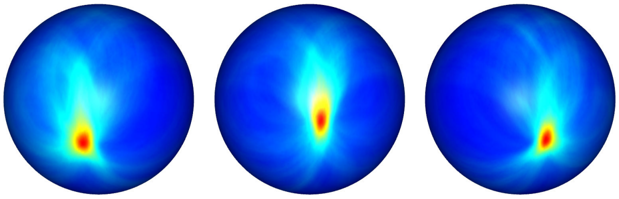

Figure 5. Estimated light on the sequence of Fig. 2 using the coarse shape of Fig. 4 top and $\tau = 4$. The image intensities are quantized in the range from 0 to 255.

while the inliers do so in favor of the correct model. This observation motivated [11] 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 three random points on the surface and, from their RGB intensities and normals, estimate an illumination hypothesis for $\rho\mathbf{VL}$.

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 $\mathbf{c}_m$, *i.e.*
$$|\rho\mathbf{V} \cdot \mathbf{L} \cdot \mathbf{n}_m - \mathbf{c}_m| < \tau, \qquad (7)$$
where τ allows for quantization errors, image noise, etc.

3. Repeat 1 and 2 a set number of times always keeping the illumination 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, making the scheme extremely robust.

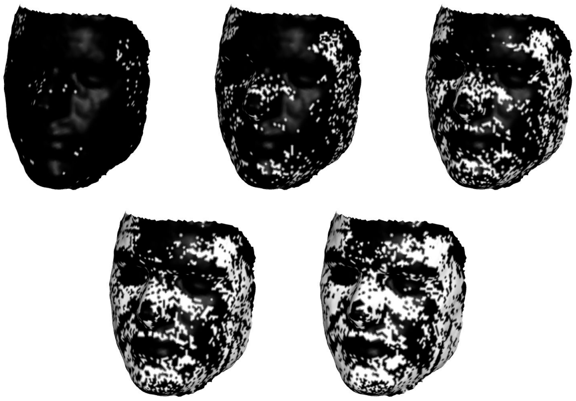

Figure 6. Distribution of inliers (in white) as a function of the threshold τ. From left to right, $\tau = 2$, $\tau = 4$, $\tau = 6$, $\tau = 8$, $\tau = 10$. The image intensities are quantized in the range from 0 to 255.

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. 5 the number of inliers per light direction, *i.e.* per row of $\mathbf{V} \cdot \mathbf{L}$ optimized for the best scale. We can appreciate how the space that RANSAC explores is very well behaved, with a clearly defined global optimum. We show in Fig. 6 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 albedo (more 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.

After estimating the light matrix, we can optionally refine the initial coarse geometry with the photometric cue by evolving the surface using a scheme such as [16] or [11]. We show in Fig. 4 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.

5. Experimental results

We have run the same algorithm on a second sequence shown in Fig. 7. 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. 9 top. The sparse set of 3d points (shown in Fig. 8) 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 4.3, giving the light estimates shown in Fig. 10. 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. 11) and we can also refine the coarse shape in order to obtain a high resolution static face capture (see Fig. 9 bottom).

Once the calibration step is completed, we can reconstruct video footage of that same person under the same setup using [10](see Fig. 12). Note that, wherever the constant chromaticity assumption is not verified, *e.g.* on the lips of the face, the normal estimation suffers from a bas

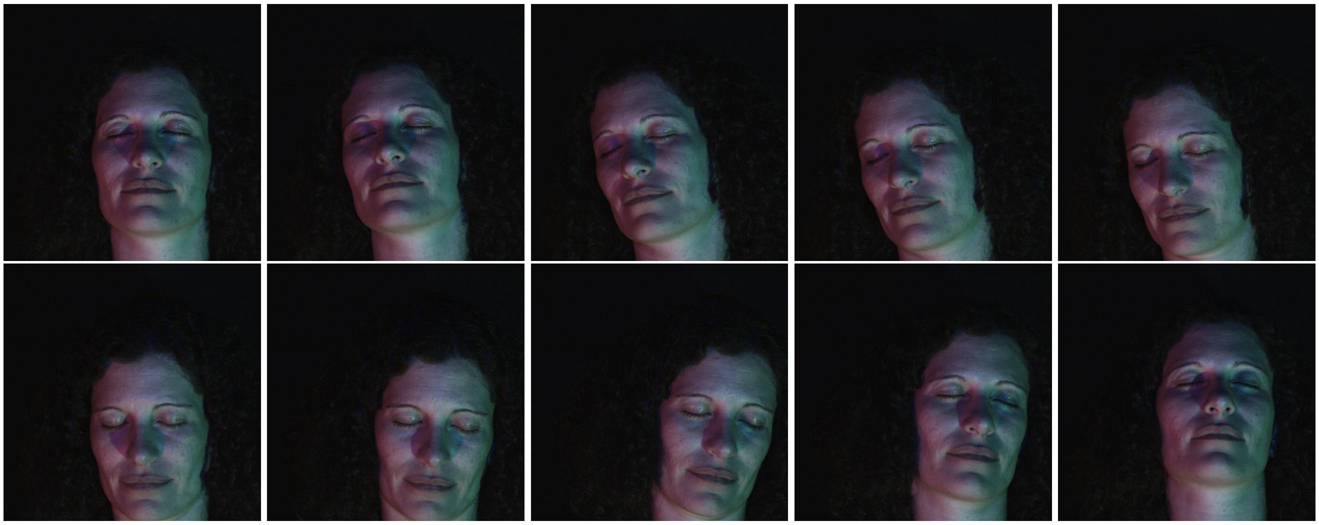

Figure 7. Face calibration sequence under a three-source color photometric setup.

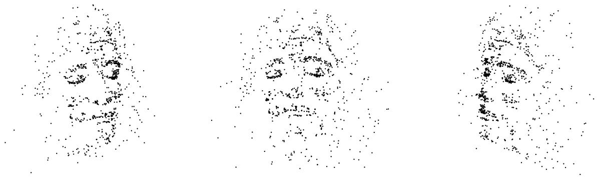

Figure 8. Sparse set of 3d points after using a structure-from-motion algorithm on the sequence of Fig. 7. From left to right, the 3d points are shown from three different viewpoints roughly at -45 degrees, 0 degrees, and 45 degrees.

relieve 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 be 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 [10], 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).

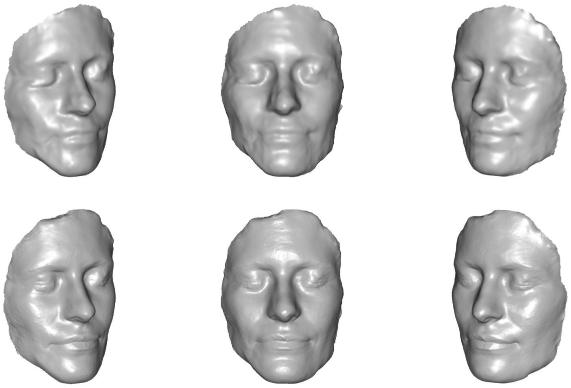

Figure 9. Top: Coarse shape obtained with the multi-view stereo algorithm [9] on the sequence of Fig. 7. Bottom: refined shape after successful light estimation and photometric stereo evolution using the scheme of [11].

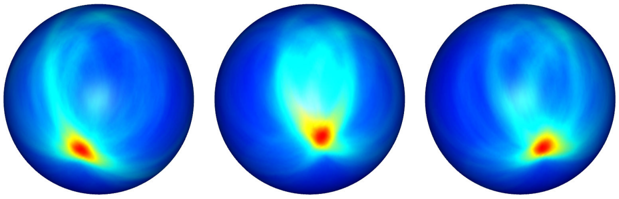

Figure 10. Estimated light on the sequence of Fig. 2 using the coarse shape of Fig. 4 top and $\tau = 4$. The image intensities are quantized in the range from 0 to 255.

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

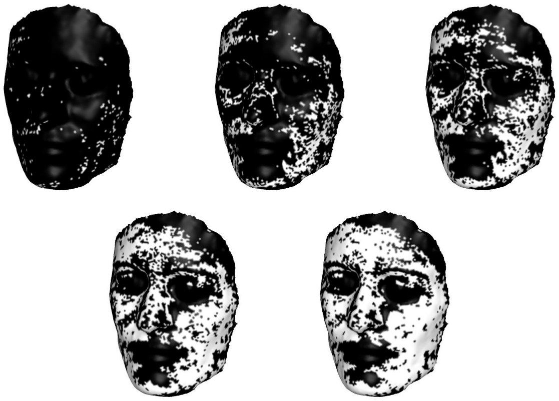

Figure 11. Distribution of inliers (in white) as a function of the threshold τ. From left to right, $\tau = 2$, $\tau = 4$, $\tau = 6$, $\tau = 8$, $\tau = 10$. The image intensities are quantized in the range from 0 to 255.

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. 1

[2] Dimensional imaging. http://www.di3d.com. 1

[3] Mova. http://www.mova.com. 1

[4] R. Basri, D. Jacobs, and I. Kemelmacher. Photometric stereo with general, unknown lighting. *Int. J. Comput. Vision*, 72(3):239–257, 2007. 5

[5] P. N. Belhumeur, D. J. Kriegman, and A. L. Yuille. The bas-relief ambiguity. *Int. J. Comput. Vision*, 35(1):33–44, 1999. 6

[6] M. Fischler and R. Bolles. 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] R. T. Frankot and R. Chellappa. A method for enforcing integrability in shape from shading algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 10(4):439–451, 1988. 6

[8] Y. Furukawa and J. Ponce. Dense 3d motion capture for human faces. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 1, 7

[9] C. Hernández and F. Schmitt. Silhouette and stereo fusion for 3d object modeling. *Computer Vision and Image Understanding*, 96(3):367–392, December 2004. 4, 6

[10] C. Hernández, G. Vogiatzis, G. Brostow, B. Stenger, and R. Cipolla. Non-rigid photometric stereo with colored lights. In *IEEE International Conference on Computer Vision*, 2007. 1, 2, 3, 5, 6, 8

[11] C. Hernández, G. Vogiatzis, and R. Cipolla. Multiview photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):548–554, 2008. 3, 4, 5, 6

[12] C. Hernández and R. Vogiatzis, Cipolla. Shadows in three-source photometric stereo. In *IEEE European Conference on Computer Vision*, 2008. 2, 3

[13] U. I. f. C. T. Image Metrics. Emily project. SIGGRAPH 2008 Demo session, 2002. 1

[14] W. Ma, T. Hawkins, P. Peers, C. Chabert, M. Weiss, and P. Debevec. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Eurographics Symposium on Rendering*, pages 183–194, 2007. 1

[15] W. Ma, A. Jones, J. Chiang, T. Hawkins, S. Frederiksen, P. Peers, M. Vukovic, M. Ouhyoung, and P. Debevec. Facial performance synthesis using deformation-driven polynomial displacement maps. *ACM Transactions on Graphics*, 27(5), 2008. 1, 2

[16] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi. Efficiently combining positions and normals for precise 3d geometry. In *Proc. of the ACM SIGGRAPH*, pages 536–543, 2005. 5

[17] A. Petrov. Light, color and shape. *Cognitive Processes and their Simulation (in Russian)*, pages 350–358, 1987. 3

[18] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multiview stereo reconstruction algorithms. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 519–528, 2006. 3

[19] T. Weise, B. Leibe, and L. V. Gool. Fast 3d scanning with automatic motion compensation. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2007. 1

[20] R. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):139–144, 1980. 1

[21] R. J. Woodham. Gradient and curvature from the photometric-stereo method, including local confidence estimation. *J. Opt. Soc. Am. A*, 11(11):3050–3068, 1994. 1, 2

[22] L. Zhang, N. Snavely, B. Curless, and S. M. Seitz. Spacetime faces: high resolution capture for modeling and animation. In *SIGGRAPH '04*, pages 548–558, 2004. 1

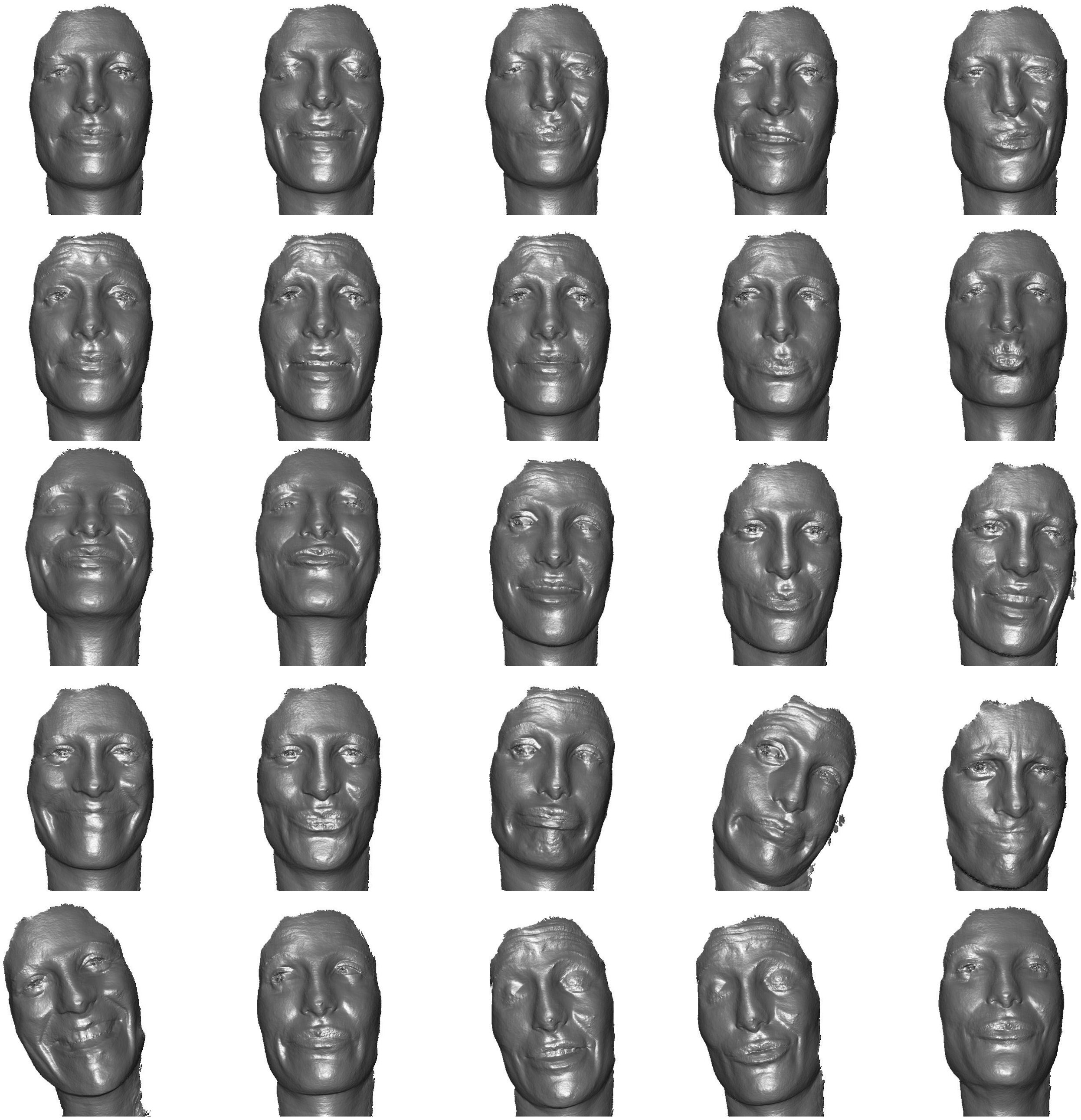

Figure 12. Acquisition of 3d facial expressions using [10]. The system was calibrated using the self-calibration technique described in this paper.

[23] S. Zhang and P. S. Huang. High-resolution, real-time three-dimensional shape measurement. *Optical Engineering*, 45(12), 2006. 1

[24] A. Zisserman and R. Hartley. *Multiple View Geometry*. Springer-Verlag, 2000. 3