Structure from Motion and Photometric Stereo for Dense 3D Shape Recovery

Reza Sabzevari, Alessio Del Bue, and Vittorio Murino

Istituto Italiano di Tecnologia
Via Morego 30, 16163 Genova, Italy
{reza.sabzevari,alessio.delbue,vittorio.murino}@iit.it
http://www.iit.it

Abstract. In this paper we present a dense 3D reconstruction pipeline from monocular video sequences using jointly Photometric Stereo (PS) and Structure from Motion (SfM) approaches. The input videos are completely uncalibrated both from the multi-view geometry and photometric stereo aspects. In particular we make use of the 3D metric information computed with SfM from a set of 2D landmarks in order to solve for the bas-relief ambiguity which is intrinsic from dense PS surface estimation. The algorithm is evaluated over the CMU Multi-Pie database which contains the images of 337 subjects viewed under different lighting conditions and showing various facial expressions.

Keywords: Structure from Motion, Photometric Stereo, Dense 3D Reconstruction.

1 Introduction

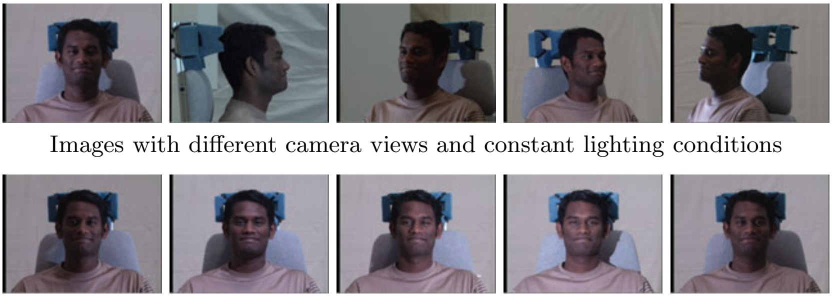
The 3D inference of the objects shape is of paramount interest in many fields of engineering and life science. However, the inference of the depth from a set of images is one of the most challenging inverse problem and various features has been used in order to extract information of 3D surfaces from images. The field of multi-view geometry [8] uses the information of a set of known 2D correspondences in order to estimate the localization of a set of 3D points lying over the surface. Shape from defocus [6] instead uses the blurring effect of images obtained from varying the distance of the lens with respect to the camera sensor. Shape from texture [4] infers the 3D surface bending given the variation of the texture belonging to an object. Differently, shape from shading [13] and photometric stereo [17] compute dense surfaces by analyzing the variations of a pixel subject to different illumination sources. In our work we will focus exclusively on the multi view geometry and photometric aspects of such problems and present a joint algorithm which obtains reliable 3D reconstruction from an uncalibrated monocular sequence of images.

Regarding the multi-view geometry aspect, the sparse 3D surface estimation from a single video requires the extraction of a set of 2D image points for each frame. These selected points are then uniquely matched for the whole video

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part I, LNCS 6978, pp. 660–669, 2011.

sequence thus creating image trajectories. The collection of such 2D trajectories define the motion of the image shape and it is subject to the metric properties of the 3D object and the camera position projecting the 3D points onto the image plane. The localization of the 2D image points is fatally sparse since good features to track and matches are restricted to a few particular regions in the image [14]. However, given a rigid object, very accurate sparse representation of the world can be obtained, even in the presence of interrupted 2D trajectories [12].

Differently, from a photometric perspective, Photometric Stereo (PS) computes dense 3D localization directly from image intensity variations in a single video sequence. Each surface point imaged by a camera reflects a light source with respect to its orientation and surface photometric properties. Thus, if enough views of the same surface point are given at different lighting positions, we may succeed to infer the 3D surface and its photometric properties (i.e. the albedo). However, this approach, in the case of a video sequence, requires the dense matching of each pixel from frame to frame or completely stationary shape with controlled lighting conditions. The latter case is the standard scenario and recent techniques [2] may compute lighting parameter and 3D surface without a prior calibration of the lighting setup. However, a well known problem implicit in the most PS reconstructions is the generalized bas-relief ambiguity. In few words, the same image pixels may correspond to different configurations of 3D surfaces and lighting sources. Choosing the right solution depends generally on a priori information of the 3D shape of the object.



Images with different camera views and constant lighting conditions

Images with same camera view and varying lighting conditions

Fig. 1. An example of the set of images in the Multi-Pie database for subject 42

In this work, we design a 3D dense reconstruction pipeline that joins SfM and PS techniques in order to obtain reliable 3D reconstructions from image sequences. The generalized bas-relief ambiguity is reduced by obtaining a reliable localization of a set of sparse 2D points extracted from the images and used to obtain a metric 3D reconstruction of the shape. In this case, we can define a photo-geometric relation between the 3D sparse model and the dense 3D surface obtained with SfM and PS respectively. By solving for the transformation, we obtain the correct alignment of the two reconstructions up to an overall scale.

There are similar works in the literature that attempt to include SfM constraints into a dense reconstruction problem. Lim et al. [11] tried to recover correct depths from multiple images of a moving object illuminated by time varying lighting. They used multiple views of the object to generate a coarse planar surface based on the recovered 3D points and then they used PS in an iterative process to recover dense surface and align it into the recovered 3D point. Zhang et al. [18] use an iterative algorithm to solve a sub-constrained optical flow formulation. They use SfM to compute the camera motion and initialize the lighting on sparse features. Then, they iteratively recover the shape and lighting in a coarse-to-fine manner using an image pyramid. All such methods have known features and drawbacks. Other works such as [9] uses specific setups with colored lights or [1] active patterns using projectors in order to constraints the photometric ambiguities. Since our proposed method uses solely standard image sequences, we deal with a less constrained case than the two previously mentioned approaches.

The testing framework used to verify the effectiveness of our 3D reconstruction algorithm is the CMU Multi-Pie face database [7]. This database contains more than $750,000$ images of 337 subjects recorded in up to four sessions over the span of five months. Subjects were imaged under 15 view points and 19 illumination conditions while displaying a range of facial expressions. Figure 1 shows samples of the database for different views and varying lighting conditions. No information is provided by this database to recover the 3D position of any point on the subjects. Notice that in such scenario, we deliberately choose not to use any a priori information about the calibration for both the cameras and lighting conditions of the experiments. In such way, we pose ourself in the most challenging scenario and with the largest modeling freedom. Many applications could be considered for such scenario, e.g. model-based face recognition, face morphing or creating 3D face databases using inexpensive off-the-shelf facilities instead of expensive 3D laser scanners.

The paper is structured as it follows. The next section is dedicated to the formulation of problem and it is described how the metric upgrade affects the results. Then, in Section 3, results obtained by applying the proposed approach on the MultiPIE database are discussed. And finally, in section 4 some brief remarks conclude this paper.

2 Sparse and Dense 3D Reconstruction

We first formulate the SfM and PS problems in the mathematical context of bilinear matrix factorization. In general, either the 2D image trajectories used by SfM and the image pixel variations in time can be both described by bilinear matrix models. For the case of SfM, the bilinear model contains the 3D shape coordinates and the camera projection matrices. Similarly, the PS case results in two factors that contain the object surface normals with the albedo and the lighting directions.

2.1 Structure from Motion

Structure from Motion algorithms simultaneously reconstruct the 3D position and camera matrices using a set of 2D points extracted from an image sequence. The inter-image relations are linked by the fact that a unique shape is projected into the images by a moving camera. Thus the 2D image trajectories created by this mapping can be used to estimate the 3D position of a shape if a sufficient baseline is given. A set of popular approaches compute simultaneously the 3D structure and camera motion via a factorization approach using solely the collection of such 2D trajectories. In more detail, the 3D structure and the camera projection matrices can be expressed as a bilinear matrix model.

In more detail, by defining the non-homogeneous coordinate of a point j in frame i as the vector $\mathbf{w}_{ij} = (u_{ij}\ v_{ij})^T$, we may write the measurement matrix $\mathtt{W}$ that gathers the coordinates of all the points in all the views as:

$$\mathtt{W} = \begin{bmatrix} \mathbf{w}_{11} & \dots & \mathbf{w}_{1p} \\ \vdots & \ddots & \vdots \\ \mathbf{w}_{g1} & \dots & \mathbf{w}_{gp} \end{bmatrix} = \begin{bmatrix} \mathtt{W}_1 \\ \vdots \\ \mathtt{W}_g \end{bmatrix} \tag{1}$$

where g is the number of frames and p the number of points. In general, the rank of $\mathtt{W}$ is constrained to be $\text{rank}\{\mathtt{W}\} \leq r$ where $r \ll \min\{2g, p\}$

In the case of a rigid object viewed by an orthographic camera, if we assume the measurements in $\mathtt{W}$ are registered to the image centroid, the camera motion matrices $\mathtt{R}_i$ and the 3D points $\mathbf{S}_j$ can be expressed as:

$$\mathtt{R}_i = \begin{bmatrix} r_{i1} & r_{i2} & r_{i3} \\ r_{i4} & r_{i5} & r_{i6} \end{bmatrix} \quad \text{and} \quad \mathbf{S}_j = \begin{bmatrix} X_j & Y_j & Z_j \end{bmatrix}^T \tag{2}$$

where $\mathtt{R}_i$ is a 2×3 matrix that contains the first two rows of a rotation matrix (i.e. $\mathtt{R}_i\mathtt{R}_i^T = \mathtt{I}_{2\times 2}$) and $\mathbf{S}_j$ is a 3-vector containing the metric coordinates of the 3D point. Thus a 2D point j in a frame i is given by $\mathbf{w}_{ij} = \mathtt{R}_i\, \mathbf{S}_j$. We can collect all the image measurements and their respective bilinear components $\mathtt{R}_i$ and $\mathbf{S}_j$ in a global matrix as in Eq. (1). Thus we can formulate the factorization model of the image trajectories as

$$\mathtt{W} = \mathtt{R}_{2g\times 3}\ \ \mathtt{S}_{3\times p} \tag{3}$$

where the bilinear components $\mathtt{R}$ and $\mathtt{S}$ are defined as:

$$\mathtt{R} = \begin{bmatrix} \mathtt{R}_1 \\ \vdots \\ \mathtt{R}_g \end{bmatrix} \quad \text{and} \quad \mathtt{S}_{sfm} = \begin{bmatrix} \mathbf{S}_1 & \cdots & \mathbf{S}_p \end{bmatrix}. \tag{4}$$

Expressing the camera projections and 3D points in such matrix form makes evident the rank constraint of $\mathtt{W}$.

Given the rank relation: $\text{rank}(\mathtt{W}) \leq \min\{\text{rank}(\mathtt{R}), \text{rank}(\mathtt{S}_{sfm})\}$, we have that the rank of the measurement is at most equal to three. This constraints is used to obtain a closed form solution for the 3D position of the points and the camera matrices as presented in the seminal paper of Tomasi and Kanade [15]. In the case of different imaging conditions remember that the simplistic assumption of an orthographic camera model has been extended to more complex affine cameras [10] or either projective ones [16].

2.2 Photometric Stereo

The principle at the base of PS is that an object illuminated by a light source will reflect light with respect to the surface orientation, light direction and intrinsic photometric properties of the shape. Thus, we can use a collection of the data representing the lighting variations of the pixels in order to infer the photometric properties of the shape. Notice that in this case we treat the object as being static and the light source moving – the aim here is to find a dense 3D reconstruction (i.e. for each pixel position in the image) of the object shape.

The chosen photometric model is based on a spherical harmonics representation of lighting variations [2] and it allows to frame PS as a factorization problem with normality constraints on one of the bilinear factors. Given a set of images of a Lambertian object with varying illumination, it is possible to extract the dense normals to the surface of the object $\mathbf{n}$, the albedos ρ and the lighting directions $\mathbf{l}$. For a 1^{st} order spherical harmonics approximation, the brightness at image pixel j at frame i can be modeled as:

$$Y_{ij} = \mathbf{l}_i^\top \ \rho_j [1 \ \mathbf{n}_j^\top]^\top = \mathbf{l}_i \mathbf{s}_j$$

where $\mathbf{l}_i \in \mathbb{R}^4$, $\rho_j \in \mathbb{R}$, $\mathbf{z}_j \in \mathbb{R}^3$ with $\mathbf{n}_j^\top \mathbf{n}_j = 1$. A compact matrix form can be obtained for each pixel y_{ij} as:

$$\mathtt{Y} = \begin{bmatrix} y_{11} & \cdots & y_{1t} \\ \vdots & \ddots & \vdots \\ y_{f1} & \cdots & y_{ft} \end{bmatrix} = \begin{bmatrix} \mathbf{l}_1^\top \\ \vdots \\ \mathbf{l}_f^\top \end{bmatrix} \begin{bmatrix} \rho_1 \begin{bmatrix} 1 \\ \mathbf{n}_1 \end{bmatrix} \cdots \rho_t \begin{bmatrix} 1 \\ \mathbf{n}_t \end{bmatrix} \end{bmatrix} = \mathtt{L}\,\mathtt{N} \tag{5}$$

where a single image i is represented by the vector $\mathbf{y}_i = \begin{bmatrix} y_{i1} \cdots y_{it} \end{bmatrix}^T$. The $f \times 4$ matrix $\mathtt{L}$ contains the collection of the lighting directions while the $4 \times t$ matrix $\mathtt{N}$ the values for the normals and the albedos. Thus the 1^{st} order spherical harmonics model enforces a rank four constraint over the image brightness of the scene. Similarly to the SfM case, it is possible to factorize the pixel values in $\mathtt{Y}$ to obtain a closed form solution that complies with the normal constraints (i.e. $\mathbf{n}_j^\top \mathbf{n}_j = 1$) as presented in [2].

Notice that we solve for the surface normals associated to each pixel. Normals integration is then required to recover the final 3D surface from the surface normals. Thus, after applying the overall PS algorithm we obtain a matrix $\mathtt{S}_{ps}$ of size $3 \times t$ containing the 3D coordinates of the surface. However this final step give a solution which is up to an unknown Generalized Bas Relief (GBR) transformation [3]. Figure 2 shows qualitatively the difference between a correct solution and a metric 3D surface. In order to find an unique solution, we use the SfM 3D metric shape to resolve the GBR ambiguity. How to estimate the correct transformation which respects the shape metric using SfM represents the main novel issue and the core of the work, which will be described in the next session.

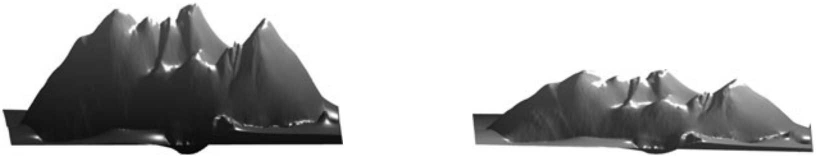


Fig. 2. Left image shows the surface before the upgrade. The right image shows the surface after the metric upgrade for subject 42.

2.3 Photometric Stereo Metric Upgrade

The photometric stereo step estimates at each image pixel position the 3D surface $\mathtt{S}_{ps}$. Notice however that a GBR transformation $\mathtt{H}$ such as [3]:

$$\mathtt{H} = \begin{bmatrix} 1\,0\,0 \\ 0\,1\,0 \\ u\,v\,\lambda \end{bmatrix}, \tag{6}$$

can be multiplied to the recovered shape giving $\tilde{\mathtt{S}} = \mathtt{HS}_{ps}$. The shape $\tilde{\mathtt{S}}$ is still a valid solution to the PS problem. Thus, we need to fix the GBR transformation that reflects the correct depth of the surface. If a set $\mathcal{O}$ of metric 3D coordinates in $\mathtt{S}_{ps}$ is available, we might be able to estimate the GBR parameters that define the correct metric surface. Such correspondences can be obtained through the mentioned SfM algorithm in Section 2.1. First, we extract a set of 2D points from a multi view image sequence such as the one showed in Figure 3. These points will form the matrix $\mathtt{W}$ as in Eq. (1). However notice that not all the points are visible in each view thus the matrix $\mathtt{W}$ will have missing entries. This leads to a factorization problem with missing data which can be solved with general purpose optimizers such as the BALM [5]. After this step, we have a set of sparse 3D metric coordinates which can be used to solve for the GBR ambiguity.

A further problem should be definitely solved. First, the SfM 3D shape $\mathtt{S}_{sfm}$ is up to an unknown rotation and in general it is not aligned with respect to the 3D surface estimated with PS (i.e. $\mathtt{S}_{ps}$). This can be solved with the assumption that one of the views in the SfM sequence corresponds to the view of the sequence used by the PS algorithm. This is always true for the Multi-Pie database and, more in general, this is a strict assumption of our method. If we have such

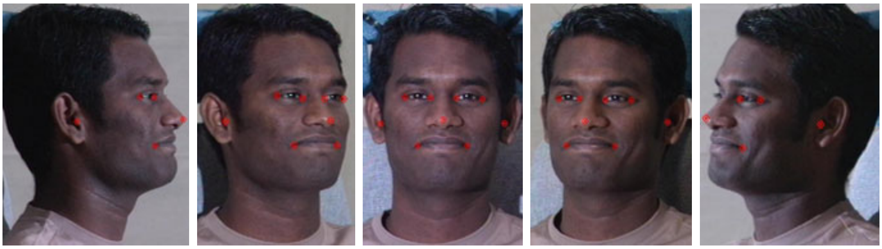


Fig. 3. 2D points on the image (Subject 42, Session 1, Recording 2)

image in common, the correspondence between the image point used by PS and SfM is also given. We call $\bar{\mathtt{S}}_{ps}$ as the $3 \times p$ matrix containing the corresponding points between PS and SfM sequences. Thus, we can define the following *photo-geometric transformation* $\mathtt{A}$ such as:

$$\mathtt{S}_{sfm} = \mathtt{H}\ \mathtt{R}_{rel}\ \bar{\mathtt{S}}_{ps} = \mathtt{A}\ \bar{\mathtt{S}}_{ps} \tag{7}$$

where $\mathtt{R}_{rel}$ is a 3×3 rotation matrix that aligns the PS and SfM 3D points. The solution can be found by computing the matrix $\mathtt{A}$ with standard Least Squares that simultaneously aligns and solves for the GBR shape ambiguity.

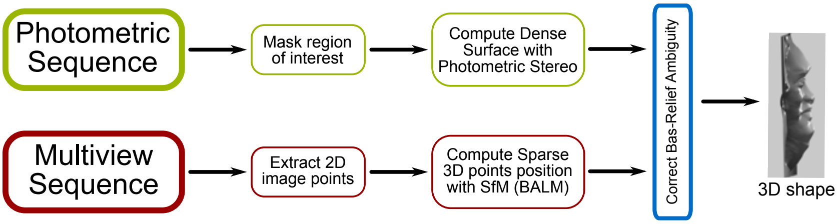

Photometric Sequence
Mask region of interest
Compute Dense Surface with Photometric Stereo
Multiview Sequence
Extract 2D image points
Compute Sparse 3D points position with SfM (BALM)
Correct Bas-Relief Ambiguity
3D shape

Fig. 4. Scheme of the 3D dense reconstruction algorithm pipeline

In summary, Figure 4 depicts the algorithm pipeline for 3D dense reconstruction. Our approach uses two different sequences: one contains 20 single view images with different illuminations to be used for the PS reconstruction, and the other sequence contains 15 images showing multiple views of the same subject used to extract 2D image points for the BALM algorithm. First, we preprocess the photometric sequence in order to select only the part of the image where the skin is present. A treshholding technique on Hue channel of the sequence with 20 images is used and refined with morphological operators to remove the background and clothing. A dense 3D surface is obtained applying the photometric stereo method on the masked images. On the other hand, in the SfM phase, some corresponding points are marked in 15 images, as it can be seen in Figure 3. As some of the points may be invisible in some views the resulted matrix will have some missing entries. To compute the 3D position of marker points and dealing with such missing data as well, the BALM method is used as the SfM engine, which results in a sparse reconstruction. Finally, the resulted reconstructions, dense and sparse one, are merged to reduce the bas-relief ambiguity effect. At this step, the sparse points are projected on to the image plane and their corresponding points on the surface are extracted. Having these two sets of points resulted by PS and SfM methods, we can solve Eq. (7) for $\mathtt{A}$. As soon as we find the *photo-geometric transformation* $\mathtt{A}$ which relates these two point clouds, we can apply it on the dense surface of PS and correct all the points of such surface.

3 Results on Multi-Pie Database

This section shows our results for two sample subjects of the MultiPIE database (42 and 46). Figure 5 presents the dense surface computed with photometric stereo (top left) and the surface with the attached texture (top right). As it is apparent in this figure, the elevation of surface points from the image plane does not comply with the metric condition. Figure 6 shows the dense surface after bas-relief correction.

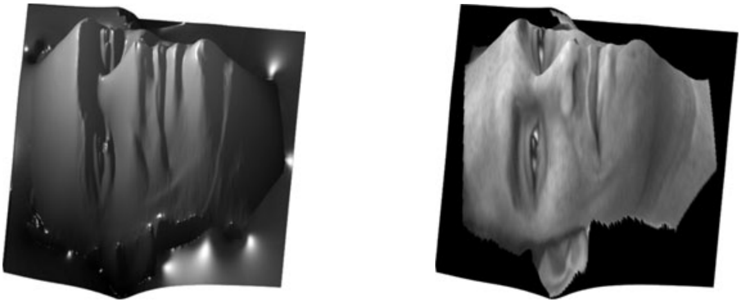


Fig. 5. Dense Surface and texture from Photometric Stereo reconstruction for subject 42

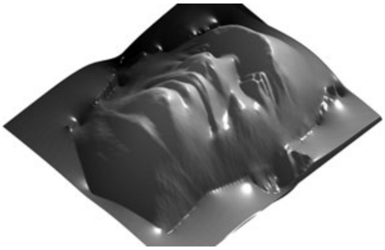


Fig. 6. Surface after bas-relief correction for subject 42

Figures 7 and 8 illustrate the results of proposed approach for another subject in the database. Computed dense surface and the surface with the attached texture are presented in Fig. 7. The dense surface after bas-relief correction is presented in Fig. 8.

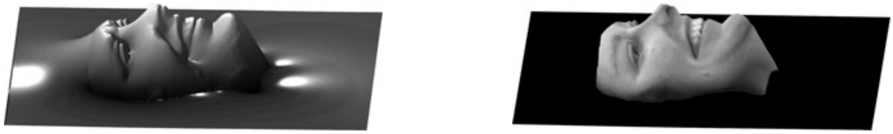


Fig. 7. Dense Surface and texture from Photometric Stereo reconstruction for subject 46

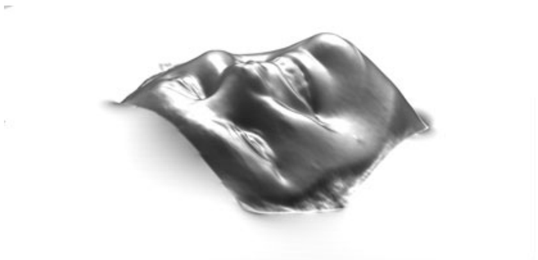


Fig. 8. Surface after bas-relief correction for subject 46

4 Conclusions

In this paper, we have presented a 3D reconstruction pipeline to obtain dense 3D metric surfaces using both Photometric Stereo and Structure from Motion techniques. The method has been tested using the Multi-Pie database in an uncalibrated scenario. The 3D reconstructions are satisfactory, however we plan to use more complex photometric models in order to grasp finer details of the objects that may strongly diverge from the Lambertian surface assumption (e.g. glasses, hairs). Another point for future investigations is to couple more deeply both SfM and PS techniques with the aim to achieve a simultaneous estimation of both photometric and 3D structure components. Such future work will be tested on ground truth data in order to be able to compare our reconstructed surfaces with real ones.

References

1. Aliaga, D.G., Xu, Y.: A self-calibrating method for photogeometric acquisition of 3d objects. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(4), 747–754 (2010)
2. Basri, R., Jacobs, D., Kemelmacher, I.: Photometric stereo with general, unknown lighting. International Journal of Computer Vision 72(3), 239–257 (2007)

3. Belhumeur, P.N., Kriegman, D.J., Yuille, A.L.: The bas-relief ambiguity. International Journal of Computer Vision 35(1), 33–44 (1999)
4. Blostein, D., Ahuja, N.: Shape from texture: Integrating texture-element extraction and surface estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 11(12), 1233–1251 (2002)
5. Del Bue, A., Xavier, J., Agapito, L., Paladini, M.: Bilinear factorization via augmented lagrange multipliers. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 283–296. Springer, Heidelberg (2010)
6. Favaro, P., Soatto, S.: A geometric approach to shape from defocus. IEEE Transactions on Pattern Analysis and Machine Intelligence, 406–417 (2005)
7. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and Vision Computing 28(5), 807–813 (2010)
8. Hartley, R., Zisserman, A.: Multiple view geometry, vol. 642. Cambridge University Press, Cambridge (2000)
9. 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)
10. Kanatani, K., Sugaya, Y., Ackermann, H.: Uncalibrated factorization using a variable symmetric affine camera. IEICE Transactions on Information and Systems 90(5), 851 (2007)
11. Lim, J., Ho, J., Yang, M., Kriegman, D.: Passive photometric stereo from motion. In: Proceedings of the Tenth IEEE International Conference on Computer Vision, vol. 2, pp. 1635–1642. IEEE Computer Society, Los Alamitos (2005)
12. Marques, M., Costeira, J.: Estimating 3d shape from degenerate sequences with missing data. Computer Vision and Image Understanding 113(2), 261–272 (2009)
13. Prados, E., Faugeras, O.: Shape from shading. In: Handbook of Mathematical Models in Computer Vision, pp. 375–388 (2006)
14. Shi, J., Tomasi, C.: Good features to track. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 593–600 (1994)
15. Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: A factorization approach. International Journal of Computer Vision 9(2) (1992)
16. Triggs, B.: Factorization methods for projective structure and motion. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, pp. 845–851 (1996)
17. Woodham, R.: Photometric method for determining surface orientation from multiple images. Optical Engineering 19(1), 139–144 (1980)
18. Zhang, L., Curless, B., Hertzmann, A., Seitz, S.M.: Shape and motion under varying illumination: Unifying structure from motion, photometric stereo, and multi-view stereo. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, vol. 2, p. 618. IEEE Computer Society, Los Alamitos (2003)