Modelling Dynamic Scenes by Registering Multi-View Image Sequences

Jean-Philippe Pons Renaud Keriven
Odyssée Laboratory, ENPC
Marne-la-Vallée, France
Jean-Philippe.Pons@certis.enpc.fr

Olivier Faugeras
Odyssée Laboratory, INRIA
Sophia-Antipolis, France
Olivier.Faugeras@sophia.inria.fr

Abstract

In this paper, we present a new variational method for multi-view stereovision and non-rigid three-dimensional motion estimation from multiple video sequences. Our method minimizes the prediction error of the shape and motion estimates. Both problems then translate into a generic image registration task. The latter is entrusted to a similarity measure chosen depending on imaging conditions and scene properties. In particular, our method can be made robust to appearance changes due to non-Lambertian materials and illumination changes. It results in a simpler, more flexible, and more efficient implementation than existing deformable surface approaches. The computation time on large datasets does not exceed thirty minutes. Moreover, our method is compliant with a hardware implementation with graphics processor units. Our stereovision algorithm yields very good results on a variety of datasets including specularities and translucency. We have successfully tested our scene flow algorithm on a very challenging multi-view video sequence of a non-rigid scene.

1. Introduction

Recovering the geometry of a scene from several images taken from different viewpoints, namely *stereovision*, is one of the oldest problems in computer vision. More recently, some authors have considered estimating the dense non-rigid three-dimensional motion field of a scene, often called *scene flow* [16], from multiple video sequences. Both problems require to match different images of the same scene. This is a very difficult task because a scene patch generally has different shapes and appearances when seen from different points of view and over time. To overcome this difficulty, most existing stereovision and scene flow algorithms rely on unrealistic simplifying assumptions that disregard either/both shape/appearance changes.

The oldest and most naive assumption about the photometric properties of the scene is brightness constancy. It only applies to strictly Lambertian scenes and requires a precise photometric calibration of the cameras. Yet it is still popular in the stereovision literature. It motivates the multi-view photo-consistency measure used in voxel coloring, space carving [7], and in the deformable mesh method of [2]. Similarly, the variational formulation of [14] relies on square intensity differences. In a later paper [13], the same authors model the intensity deviations from brightness constancy by a multivariate Gaussian. However, this does not remove any of the severe limitations of this simplistic assumption.

As regards scene flow estimation, many methods [18, 1, 8] use the spatio-temporal derivatives of the input images. Due to the underlying brightness constancy assumption and to the local relevance of spatio-temporal derivatives, these differential methods apply mainly to slowly-moving scenes under constant illumination.

For a better robustness to noise and to realistic imaging conditions, similarity measures embedded in stereovision and scene flow algorithms have to aggregate neighborhood information. In return, they are confronted with geometric distortion between the different views and the different time instants. Some stereovision methods disregard this difficulty and use fixed matching windows. The underlying assumption is the fronto parallel hypothesis: the camera retinal planes are identical and the scene is an assembly of planes parallel to them. Some methods go beyond this hypothesis by taking into account the tangent plane to the object [3, 6, 2, 4]. For example, [6] allows to estimate both the shape and the non-Lambertian reflectance by minimizing the rank of a radiance tensor. The latter is computed by sampling image intensities on a tessellation of the tangent plane. In such approaches, the matching score depends not only on the position of the surface but also on its orientation. Unfortunately, this first-order shape approximation results in a very complex minimizing flow involving second-order derivatives of the matching score. The computation of these terms is tricky, time-consuming and unstable, and, to our knowledge, all authors have resigned to ignore them.

In Section 2, we propose a common variational frame-

work for stereovision and scene flow estimation which correctly handles projective distortion without any approximation of shape and motion and which can be made robust to appearance changes. The metric used in our framework is the ability to predict the other input views from one input view and the estimated shape or motion. This is related to the methodology proposed in [15] for evaluating the quality of motion estimation and stereo correspondence algorithms. But in our method, the prediction error is used for the estimation itself rather than for evaluation purposes.

Our formulation is completely decoupled from the nature of the image similarity measure used to assess the quality of the prediction. It can be the normalized cross correlation, some statistical measures such as the mutual information [17], or any other application-specific measure. Through this choice, we can make the estimation robust to camera spectral sensitivity differences, non-Lambertian materials and illumination changes. In Section 3, we detail two similarity measures that can be used in our framework.

Our method processes entire images from which projective distortion has been removed, thereby avoiding the complex machinery usually needed to match windows of different shapes. Moreover, its minimizing flow is much simpler than in [3, 6]. This results in elegant and efficient algorithms. In Section 4, we describe our implementation and we present our experimental results

2. Minimizing the Prediction Error

Our method consists in maximizing, with respect to shape and motion, the similarity between each input view and the predicted images coming from the other views. We adequately warp the input images to compute the predicted images, which simultaneously removes projective distortion. Numerically, this can be done at a low computational cost using texture-mapping graphics hardware (*cf* Section 4). For example, in the case of stereovision, we back-project the image taken by one camera on the surface estimate, then we project it to the other cameras to predict the appearance of the other views. The closer the shape estimate is to the actual geometry, the more similar the predicted images will be to the corresponding input images, modulo noise, calibration errors, appearance changes and semi-occluded areas. This is the core principle of our approach.

Interestingly, this can be formulated as a generic image registration task. The latter is entrusted to a measure of image similarity, chosen depending on imaging conditions and scene properties. This measure is basically a function mapping two images to a scalar value. The more similar the two images are, the lower the value of the measure is. We incorporate this measure and a regularization term in an energy functional. Here we focus primarily on the design of the matching term and we propose a basic regularization term. To minimize our energy functionals, we use a gradient descent, embedded in a multi-resolution coarse-to-fine strategy to decrease the probability of getting trapped in irrelevant local minima.

2.1. Stereovision

In the following, let a surface $S \subset \mathbb{R}^3$ model the shape of the scene. We note $I_i : \Omega_i \subset \mathbb{R}^2 \rightarrow \mathbb{R}^d$ the image captured by camera i. The perspective projection performed by the latter is denoted by $\Pi_i : \mathbb{R}^3 \rightarrow \mathbb{R}^2$. Our method takes into account the visibility of the surface points. In the sequel, we will refer to S_i as the part of S visible in image i. The reprojection from camera i onto the surface is denoted by $\Pi_{i,S}^{-1} : \Pi_i(S) \rightarrow S_i$. With this notation in hand, the reprojection of image j in camera i via the surface writes $I_j \circ \Pi_j \circ \Pi_{i,S}^{-1} : \Pi_i(S_j) \rightarrow \mathbb{R}^d$. We note M a generic measure of similarity between two images.

The matching term $\mathcal{M}$ is the sum of the dissimilarity between each input view and the predicted images coming from all the other cameras. Thus, for each ordered pair of cameras (i, j), we compute the similarity between I_i and the reprojection of I_j in camera i via S, on the domain where both are defined, i.e. $\Omega_i \cap \Pi_i(S_j)$, in other words after discarding semi-occluded regions:

$$\mathcal{M}(S) = \sum_i \sum_{j \neq i} \mathcal{M}_{ij}(S) \,, \tag{1}$$

$$\mathcal{M}_{ij}(S) = M|_{\Omega_i \cap \Pi_i(S_j)} \left(I_i \,,\; I_j \circ \Pi_j \circ \Pi_{i,S}^{-1} \right) \,. \tag{2}$$

Unlike several existing methods [3, 6, 2, 4], we do not follow a minimal surface approach, i.e. our energy functional is not an integral on the surface estimate. The minimal surface approach mixes data fidelity and regularization, which makes it difficult to tune the regularizing behavior, as discussed in [12]. In contrast, our energy functional is the sum of a matching term computed in the images and of a user-defined regularization term.

We now compute the variation of the matching term with respect to an infinitesimal vector displacement δS of the surface. Figure 1 displays the camera setup and our notations. Using the chain rule, we get

$$\left.\frac{\partial \mathcal{M}_{ij}(S + \epsilon\, \delta S)}{\partial \epsilon}\right|_{\epsilon=0} = \int_{\Omega_i \cap \Pi_i(S_j)} \underbrace{\partial_2 M(\mathbf{x}_i)}_{1\times d} \underbrace{D I_j(\mathbf{x}_j)}_{d\times 2} \underbrace{D\Pi_j(\mathbf{x})}_{2\times 3} \underbrace{\left.\frac{\partial \Pi_{i,S+\epsilon\,\delta S}^{-1}(\mathbf{x}_i)}{\partial \epsilon}\right|_{\epsilon=0}}_{3\times 1} d\mathbf{x}_i \,,$$

where $\mathbf{x}_i$ is the position in image i and $D.$ denotes the Jacobian matrix of a function.

When the surface moves, the predicted image changes. Hence the variation of the matching term involves the derivative of the similarity measure with respect to its second argument, denoted by $\partial_2 M$. Its meaning is detailed in

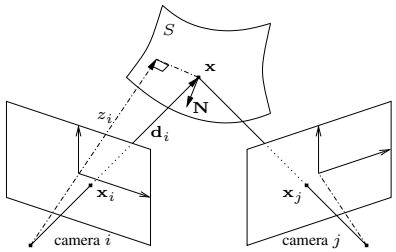

Figure 1. The camera setup and our notations.

Section 3. We then use a relation between the motion of the surface and the displacement of the reprojected surface point $\mathbf{x} = \Pi_{i,S}^{-1}(\mathbf{x}_i)$:

$$\left.\frac{\partial \Pi_{i,S+\epsilon\,\delta S}^{-1}(\mathbf{x}_i)}{\partial \epsilon}\right|_{\epsilon=0} = \frac{\mathbf{N}^T \delta S(\mathbf{x})}{\mathbf{N}^T \mathbf{d}_i}\ \mathbf{d}_i\ ,$$

where $\mathbf{d}_i$ is the vector joining the center of camera i and $\mathbf{x}$, and $\mathbf{N}$ is the outward surface normal at this point. Finally, we rewrite the integral in the image as an integral on the surface by the change of variable $d\mathbf{x}_i = -\mathbf{N}^T \mathbf{d}_i\, d\mathbf{x}/z_i^3$, where z_i is the depth of $\mathbf{x}$ in camera i, and we obtain that the gradient of the matching term is

$$\nabla \mathcal{M}_{ij}(S)(\mathbf{x}) = -\delta_{S_i \cap S_j}(\mathbf{x}) \left[\partial_2 M(\mathbf{x}_i) DI_j(\mathbf{x}_j) D\Pi_j(\mathbf{x}) \frac{\mathbf{d}_i}{z_i^3}\right] \mathbf{N}\ , \tag{3}$$

where $\delta_{\cdot}$ is the Kronecker symbol. As expected, the gradient cancels in the regions not visible from both cameras. Note that the term between square brackets is a scalar function.

The regularization term is typically the area of the surface, and the associated minimizing flow is a mean curvature motion. The evolution of the surface is then driven by

$$\frac{\partial S}{\partial t} = \left[-\lambda\, H + \sum_i \sum_{j \neq i} \delta_{S_i \cap S_j}\, \partial_2 M\, DI_j\, D\Pi_j \frac{\mathbf{d}_i}{z_i^3}\right] \mathbf{N}\ , \tag{4}$$

where H denotes the mean curvature of S, and λ is a positive weighting factor.

2.2. Scene flow

Two types of methods prevail in the scene flow literature. In the first family of methods [16, 18], scene flow is constructed from previously computed optical flows in all the input images. However, the latter may be noisy and/or physically inconsistent through cameras. The heuristic spatial smoothness constraints applied to optical flow may also alter the recovered scene flow. The second family of methods [18, 1, 8] relies on spatio-temporal image derivatives. These differential methods apply mainly to slowly-moving Lambertian scenes under constant illumination.

Our method does not fall into any of these two categories. It directly evolves a 3D vector field to register the input images captured at different times. It can recover large displacements thanks to a multi-resolution strategy and can be made robust to illumination changes through the design of the similarity measure.

Let now S^t model the shape of the scene and I_i^t be the image captured by camera i at time t. Let $\mathbf{v}^t : S^t \to \mathbb{R}^3$ be a 3D vector field representing the motion of the scene between t and $t+1$. The matching term $\mathcal{F}$ is the sum over all cameras of the dissimilarity between the images taken at time t and the corresponding images at $t+1$ warped back in time using the scene flow.

$$\mathcal{F}(\mathbf{v}^t) = \sum_i \mathcal{F}_i(\mathbf{v}^t)\ , \tag{5}$$

$$\mathcal{F}_i(\mathbf{v}^t) = M\left(I_i^t\ ,\ I_i^{t+1} \circ \Pi_i \circ (\Pi_{i,S^t}^{-1} + \mathbf{v}^t)\right)\ . \tag{6}$$

Its gradient writes

$$\nabla^T \mathcal{F}_i(\mathbf{v}^t) = -\delta_{S_i^t} \frac{\mathbf{N}^T \mathbf{d}_i}{z_i^3}\, \partial_2 M\, DI_i^{t+1}\, D\Pi_i\ . \tag{7}$$

In this case, the regularization term is typically the harmonic energy of the flow over the surface, and the corresponding minimizing flow is an intrinsic heat equation based on the Laplace-Beltrami operator.

3. Some Similarity Measures

In this section, we present two similarity measures that can be used in our framework: cross correlation and mutual information [17]. Cross correlation assumes a local affine dependency between the intensities of the two images, whereas mutual information can cope with general statistical dependencies. We have picked these two measures among a broader family of statistical criteria proposed in [5] for multimodal image registration. In the following, we consider two scalar images $I_1,\ I_2 : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$. The measures below can be extended to vector (e.g. color) images by summing over the different components.

The minimizing flows given in Section 2 involve the derivative of the similarity measure with respect to the second image, denoted by $\partial_2 M$. The meaning of this derivative is the following: given two images $I_1,\ I_2 : \Omega \to \mathbb{R}^d$, we note $\partial_2 M(I_1, I_2)$ the function mapping Ω to the row vectors of $\mathbb{R}^d$, verifying for any image variation δI:

$$\left.\frac{\partial M(I_1, I_2 + \epsilon\, \delta I)}{\partial \epsilon}\right|_{\epsilon=0} = \int_\Omega \partial_2 M(I_1, I_2)(\mathbf{x})\, \delta I(\mathbf{x})\, d\mathbf{x}\ . \tag{8}$$

Cross correlation is still the most popular stereovision matching measure. Most methods settle for fixed rectangular correlation windows. In this case, the choice of the window size is a difficult trade-off between match reliability and oversmoothing of depth discontinuities due to projective distortion. In our method, we match distortion-free images, so the size of the matching window is not related

to a shape approximation. The matter here is in how big a neighborhood the assumption of affine dependency is valid. Typically, non-Lambertian scenes require to reduce the size of the correlation window, making the estimation less robust to noise and outliers. In our implementation, instead of hard windows, we use smooth Gaussian windows. They make the continuous formulation of our problem more elegant and they can be implemented efficiently with fast recursive filtering. Due to space limitations, we invite the reader to refer to our technical report [10] for the full expressions of M and $\partial_2 M$ in this case.

Mutual information is based on the joint probability distribution of the two images, estimated by the Parzen window method with a Gaussian of standard deviation β:

$$P(i_1, i_2) = \frac{1}{|\Omega|} \int_\Omega G_\beta \left(I_1(\mathbf{x}) - i_1 \, , \, I_2(\mathbf{x}) - i_2 \right) d\mathbf{x} \, . \quad (9)$$

We note P_1, P_2 the marginals. Our measure is the opposite of the mutual information of the two images:

$$M^{MI}(I_1, I_2) = - \int_{\mathbb{R}^2} P(i_1, i_2) \log \frac{P(i_1, i_2)}{P_1(i_1) P_2(i_2)} \, di_1 \, di_2 \, . \quad (10)$$

Its derivative with respect to the second image writes

$$\begin{aligned} \partial_2 M^{MI}(I_1, I_2)(\mathbf{x}) &= \zeta(I_1(\mathbf{x}), I_2(\mathbf{x})) \, , \\ \zeta(i_1, i_2) &= \frac{1}{|\Omega|} \, G_\beta \star \left(\frac{\partial_2 P}{P} - \frac{P_2'}{P_2} \right) (i_1, i_2) \, . \end{aligned} \quad (11)$$

4. Experimental Results

We have implemented our method in the level set framework [9], motivated by its numerical stability and its ability to handle topological changes automatically. However, our method is not specific to a particular surface model: an implementation with meshes would be straightforward.

The predicted images can be computed very efficiently thanks to graphics card hardware-accelerated rasterizing capabilities. In our implementation, we determine the visibility of surface points in all cameras using OpenGL depth buffering, we compute the reprojection of an image to another camera via the surface using projective texture mapping, and we discard semi-occluded areas using shadow-mapping [11]. The bottleneck in our current implementation is the computation of the similarity measure. Since it only involves homogeneous operations on entire images, we could probably resort to a graphics processor unit based implementation with fragment shaders.

4.1. Stereovision

Table 1 describes the stereovision datasets used in our experiments. All are real images except "Buddha". "Cactus" and "Gargoyle" are courtesy of Pr. K. Kutulakos (University of Toronto). "Buddha" and "Bust" are publicly available from the OpenLF software (LFM project, Intel).

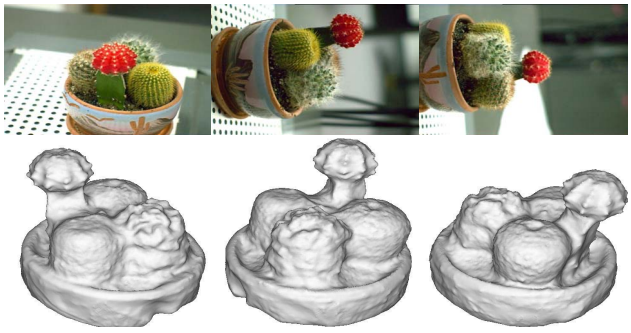

Figure 2. Some images from the "Cactus" dataset and our results.

Figure 3. Some images from the "Gargoyle" dataset and our results.

We have used either cross correlation (CC) or mutual information (MI). Both perform well on these complex scenes. "Buddha" and "Bust" are probably the more challenging datasets: "Buddha" is a synthetic scene simulating a translucent material and "Bust" includes strong specularities. However, cross correlation with a small matching window (variance of 4 pixels) yields very good results.

Using all possible camera pairs is not necessary since, when two cameras are far apart, no or little part of the scene is visible in both views. Consequently, in practice, we only pick pairs of neighboring cameras. In all our experiments, the initial surface is a coarse bounding box of the scene. We show our results in Figures 2, 3, 4 and 5. For each dataset, we display some of the input images, the ground truth when available, then our results.

The overall shape of the objects is successfully recovered, and a lot of details are captured: the stings of "Cactus", the ears and the pedestal of "Gargoyle", the nose and the collar of "Buddha", the ears and the mustache of "Bust". A few defects are of course visible. Some of them can be explained. The hole around the stick of "Gargoyle" is not fully recovered. This may be due to the limited number of images (16): some parts of the concavity are visible only in one camera. The depression in the forehead of "Bust" is related to a very strong specularity: intensity is almost

Name	#Images	Image size	#Image pairs	Measure	Level set size	Time (sec.)
Cactus	30	768×484	60	CC	128^3	1670
Gargoyle	16	719×485	32	MI	128^3	905
Buddha	25	500×500	50	CC	128^3	530
Bust	24	300×600	48	CC	$128 \times 128 \times 256$	1831

Table 1. Description of the stereovision datasets used in our experiments.

Figure 4. Some images from the "Buddha" dataset, ground truth and our results.

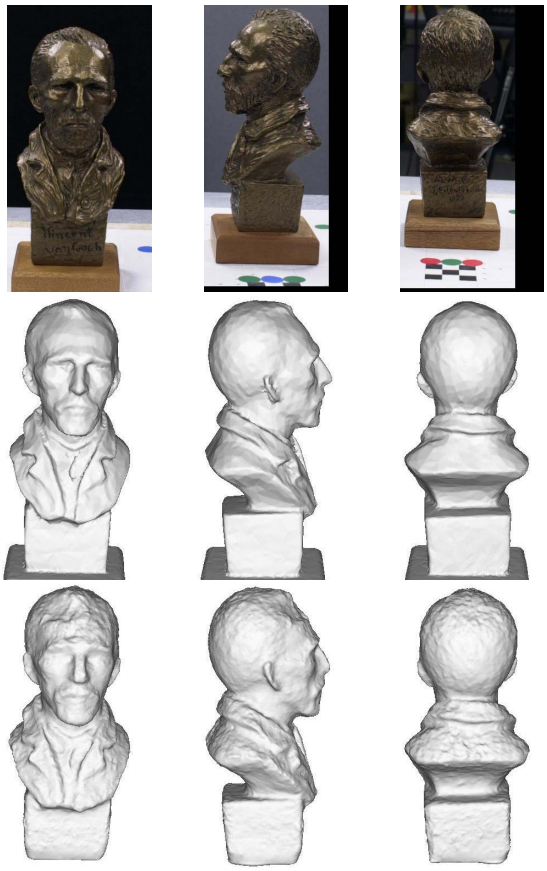

Figure 5. Some images from the "Bust" dataset, pseudo ground truth and our results.

saturated in some images.

Finally, compared with the results of the non-Lambertian stereovision method of [6] on the same datasets, our reconstructions are significantly more detailed and above all our computation time is considerably smaller. It does not exceed thirty minutes on a 2 GHz Pentium IV PC under Linux.

4.2. Stereovision + scene flow

We have tested our scene flow algorithm on a challenging multi-view video sequence of a non-rigid scene. The "Yiannis" sequence is taken from a collection of datasets that were made available to the community by P. Baker and J. Neumann (University of Maryland) for benchmark purposes. This sequence shows a character talking while rotating his head. It was captured by 22 cameras at 54 fps plus 8 high-resolution cameras at 6 fps. Here we focus on the 30 synchronized sequences at the lower frame rate to demonstrate that our method can handle large displacements.

We have applied successively our stereovision and scene flow algorithms: once we know the shape S^t, we compute the 3D motion $\mathbf{v}^t$ with our scene flow algorithm. Since $S^t + \mathbf{v}^t$ is a very good estimate of S^{t+1}, we use it as the initial condition in our stereovision algorithm and we perform a handful of iterations to refine it. This is mush faster than restarting the optimization from scratch.

Figure 6 displays the first four frames of one of the input sequence and our estimation of shape and 3D motion at corresponding times. We successfully recover the opening and closing of the mouth, followed by the rotation of the head while the mouth opens again. Moreover, we capture displacements of more than twenty pixels. Our results can be used to generate time-interpolated 3D sequences of the scene. See the *Odyssée Lab* web page for more results.

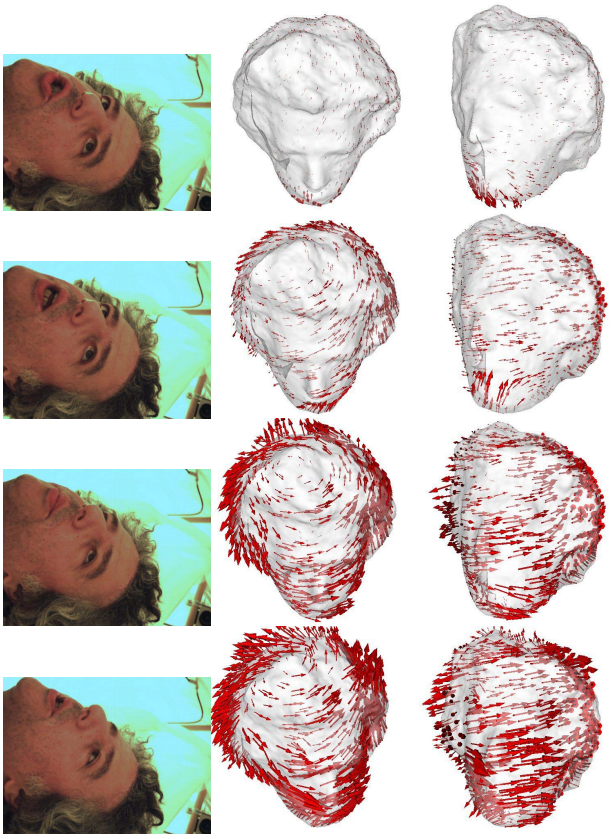

Figure 6. First images of one sequence of the "Yiannis" dataset and our results.

5. Conclusion

We have presented a novel method for multi-view stereovision and scene flow estimation which minimizes the prediction error. Our method correctly handles projective distortion without any approximation of shape and motion, and can be made robust to appearance changes. To achieve this, we adequately warp the input views and we register the resulting distortion-free images with a user-defined similarity measure. We have implemented our stereovision method in the level set framework and we have obtained results comparing favorably with state-of-the-art methods, even on complex non-Lambertian real-world images including specularities and translucency. Using our algorithm for motion estimation, we have successfully recovered the 3D motion of a non-rigid scene.

References

[1] R. Carceroni and K. Kutulakos. Multi-view scene capture by surfel sampling: From video streams to non-rigid 3D motion, shape and reflectance. *The International Journal of Computer Vision*, 49(2–3):175–214, 2002.

[2] Y. Duan, L. Yang, H. Qin, and D. Samaras. Shape reconstruction from 3D and 2D data using PDE-based deformable surfaces. In *European Conference on Computer Vision*, volume 3, pages 238–251, 2004.

[3] O. Faugeras and R. Keriven. Variational principles, surface evolution, PDE's, level set methods and the stereo problem. *IEEE Transactions on Image Processing*, 7(3):336–344, 1998.

[4] B. Goldlücke and M. Magnor. Space-time isosurface evolution for temporally coherent 3D reconstruction. In *International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 350–355, 2004.

[5] G. Hermosillo, C. Chefd'hotel, and O. Faugeras. Variational methods for multimodal image matching. *The International Journal of Computer Vision*, 50(3):329–343, Nov. 2002.

[6] H. Jin, S. Soatto, and A. Yezzi. Multi-view stereo beyond Lambert. In *International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 171–178, 2003.

[7] K. Kutulakos and S. Seitz. A theory of shape by space carving. *The International Journal of Computer Vision*, 38(3):199–218, 2000.

[8] J. Neumann and Y. Aloimonos. Spatio-temporal stereo using multi-resolution subdivision surfaces. *The International Journal of Computer Vision*, 47:181–193, 2002.

[9] S. Osher and J. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton–Jacobi formulations. *Journal of Computational Physics*, 79(1):12–49, 1988.

[10] J.-P. Pons, R. Keriven, and O. Faugeras. Modelling dynamic scenes by registrating multi-view image sequences. Technical Report 5321, INRIA, 2004.

[11] M. Segal, C. Korobkin, R. van Widenfelt, J. Foran, and P. Haeberli. Fast shadows and lighting effects using texture mapping. *Computer Graphics*, 26(2):249–252, 1992.

[12] S. Soatto, A. Yezzi, and H. Jin. Tales of shape and radiance in multi-view stereo. In *International Conference on Computer Vision*, volume 2, pages 974–981, 2003.

[13] C. Strecha, R. Fransens, and L. Van Gool. Wide-baseline stereo from multiple views: a probabilistic account. In *International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 552–559, 2004.

[14] C. Strecha, T. Tuytelaars, and L. Van Gool. Dense matching of multiple wide-baseline views. In *International Conference on Computer Vision*, volume 2, pages 1194–1201, 2003.

[15] R. Szeliski. Prediction error as a quality metric for motion and stereo. In *International Conference on Computer Vision*, volume 2, pages 781–788, 1999.

[16] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):475–480, 2005.

[17] P. Viola and W. M. Wells III. Alignement by maximization of mutual information. *The International Journal of Computer Vision*, 24(2):137–154, 1997.

[18] Y. Zhang and C. Kambhamettu. On 3D scene flow and structure estimation. In *International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 778–785, 2001.