Robust Variational Reconstruction from Multiple Views

Natalia Slesareva[1], Thomas Bühler[1], Kai Uwe Hagenburg[1], Joachim Weickert[1], Andrés Bruhn[1], Zachi Karni[2], and Hans-Peter Seidel[2]

[1] Mathematical Image Analysis Group, Dept. of Mathematics and Computer Science, Saarland University, Building E1.1, 66041 Saarbrücken, Germany
{slesareva,buehler,hagenburg,weickert,bruhn}@mia.uni-saarland.de
[2] Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany
{karni,hpseidel}@mpi-sb.mpg.de

Abstract. Recovering a 3-D scene from multiple 2-D views is indispensable for many computer vision applications ranging from free viewpoint video to face recognition. Ideally the recovered depth map should be dense, piecewise smooth with fine level of details, and the recovery procedure shall be robust with respect to outliers and global illumination changes. We present a novel variational approach that satisfies these needs. Our model incorporates robust penalisation in the data term and anisotropic regularisation in the smoothness term. In order to render the data term robust with respect to global illumination changes, a gradient constancy assumption is applied to logarithmically transformed input data. Focussing on translational camera motion and considering small baseline distances between the different camera positions, we reconstruct a common disparity map that allows to track image points throughout the entire sequence. Experiments on synthetic image data demonstrate the favourable performance of our novel method.

Keywords: computer vision, variational methods, multi-view reconstruction, structure from motion, partial differential equations.

1 Introduction

Structure from motion is a challenging task in modern computer vision: Extraction of depth information from the images of a single moving camera is useful for such tasks as robot navigation, augmented reality [14], [13] or face recognition. In the latter case structure from motion allows to reconstruct a face from a set of images, obtained by a single moving camera. One of the typical scenarios in this context is a camera that moves horizontally with a constant speed and whose optical axis is fixed orthogonal to the path of motion. While such a setting simplifies the computation, it is still difficult to obtain dense reconstructions that are robust under noise and illumination changes and provide sharp object boundaries.

B.K. Ersbøll and K.S. Pedersen (Eds.): SCIA 2007, LNCS 4522, pp. 173–182, 2007.
© Springer-Verlag Berlin Heidelberg 2007

All these demands can be satisfied by variational techniques, that have proved to be very useful in the context of optic flow estimation [2]. The reconstruction problem is formulated in an energy minimisation framework, under the assumption of global smoothness of the solution. Compared to other methods variational techniques offer a number of specific advantages: They allow transparent modeling without hidden assumptions or post-processing steps. Moreover, their continuous formulation enables rotationally invariant modeling in a natural way. The filling-in effect creates dense depth maps with sub-pixel precision by propagating information over the entire image domain. For these reasons we aim here at exploring the performance of variational methods in the context of 3-D reconstruction from multiple views.

Since the fundamental work of Faugeras and Keriven [5] many different methods for multi-view 3-D reconstruction have been proposed. In most cases a calibrated camera setup is assumed and locally constant intensity of objects in the scene is required. In the core of the minimisation procedure there lays either a gradient descent algorithm such as in [5], [15] or a sophisticated strategy of successive refinement of results as applied in [6]. The results are highly accurate, however the reported computational times take up to several hours. Comparing to other methods, that reconstruct 3-D objects from multiple views using variational framework, like for example in [9], [8] our method produces not a complete model, but only one disparity map. On the other hand the simplicity of our approach allows us to study more sophisticated models that help to improve the robustness of the method with respect to noise and varying illumination.

In this paper we focus on a prototypical scenario of a face recognition system that reconstructs the face surface from images taken by a camera that moves linearly with constant speed within an orthoparallel setting. This allows us to exploit a number of ideas that originate from the computation of optic flow fields. It is well-known that in the orthoparallel case the following relation holds:

$$Z = \frac{b \cdot f}{D} \,. \tag{1}$$

Here, Z denotes the depth of a point in the 3-D world, b is the baseline distance between successive camera positions, f specifies the focal length and D is the disparity, i.e. the distance between the projection of Z on two successive image planes. Formulating our problem in terms of disparity estimation, we obtain a scene reconstruction up to a scaling factor that depends on one intrinsic (focal length) and one extrinsic parameter (baseline).

Since the camera moves slowly with a constant speed, we obtain a series of consecutive disparity maps that are identical. Hence, it is sufficient to compute a single joint disparity map.

Our paper is organised as follows. The next section describes our variational model and its underlying assumptions in detail. Its PDE formulation is given by the Euler-Lagrange equation sketched in Section 3. Experiments in Section 4 illustrate the performance of our approach. The paper is concluded by a summary in Section 5.

2 Variational Framework

We assume a single camera that acquires images while moving slowly with a constant speed along the x-axis. Thus, approximately the same displacement field (disparity map) $\lambda(x, y)$ occurs between each pair of subsequent frames and can be recovered as minimiser of a single energy functional:

$$E(\lambda) = E_D(\lambda) + \alpha E_S(\lambda)\,, \tag{2}$$

where $E_D(\lambda)$ is a data term, $E_S(\lambda)$ is a smoothness term, and the regularisation parameter $\alpha > 0$ determines the desired amount of smoothness.

Let $f^i(\mathbf{x})$ denote the grey value of frame i at location $\mathbf{x} = (x, y)$. In order to render our method robust against noise, we first convolve with a Gaussian K_σ of standard deviation $\sigma > 0$. By applying a logarithmic transform to the result, the multiplicative effects of global illumination changes are transformed into additive perturbations. This leads to the images $g^i(\mathbf{x})$ for $i = 1,...,N$, which serve as input data for our variational approach.

For the data term $E_D(\lambda)$ we choose a gradient constancy assumption between corresponding structures within consecutive frames g^i and g^{i+1}:

$$\nabla g^{i+1}(x + \lambda, y) = \nabla g^i(x, y)\,. \tag{3}$$

It ignores any additive perturbations on $g^i(\mathbf{x})$ caused by global illumination changes between consecutive frames $f^i(\mathbf{x})$. Penalising deviations from this constancy assumption between all consecutive frame pairs in a statistically robust way [7] can be achieved by use of the data term

$$E_D(\lambda) = \int_\Omega \frac{1}{N} \sum_{i=1}^{N-1} \Psi\left(|\nabla g^{i+1}(x + \lambda, y) - \nabla g^i(x, y)|^2\right) d\mathbf{x}\,, \tag{4}$$

where $\Omega \subset \mathbb{R}^2$ denotes our rectangular image domain, and $\Psi(s^2) := \sqrt{s^2 + \epsilon^2}$ is a L^1 penaliser with a small regularising constant $\epsilon > 0$ ensuring differentiability.

Since the baseline distance between consecutive frames is supposed to be small for our application, we can simplify our data term by the Taylor linearisations

$$\begin{aligned} \partial_x g^{i+1}(x + \lambda, y) &\approx \partial_x g^{i+1}(x, y) + \partial_{xx} g^{i+1}(x, y)\,\lambda\,, \\ \partial_y g^{i+1}(x + \lambda, y) &\approx \partial_y g^{i+1}(x, y) + \partial_{xy} g^{i+1}(x, y)\,\lambda\,. \end{aligned}$$

Introducing the matrices

$$J^i = \begin{pmatrix} (g^{i+1}_{xx})^2 + (g^{i+1}_{xy})^2 & (g^{i+1}_x - g^i_x) g^{i+1}_{xx} + (g^{i+1}_y - g^i_y) g^{i+1}_{xy} \\ (g^{i+1}_x - g^i_x) g^{i+1}_{xx} + (g^{i+1}_y - g^i_y) g^{i+1}_{xy} & (g^{i+1}_x - g^i_x)^2 + (g^{i+1}_y - g^i_y)^2 \end{pmatrix}$$

and the vector $\mathbf{w} := (\lambda(\mathbf{x}), 1)^\top$ allows to reformulate the data term in a compact way as a sum of robustified quadratic forms:

$$E_D(\lambda) = \int_\Omega \frac{1}{N} \sum_{i=0}^{N-1} \Psi\left(\mathbf{w}^\top J^i \mathbf{w}\right) d\mathbf{x}\,. \tag{5}$$

The role of the smoothness term $E_S(\lambda)$ in our energy functional is to penalise deviations from smoothness in the unknown disparity field $\lambda(\mathbf{x})$. Instead of a standard quadratic smoothness term (based on the L^2 norm), we use the anisotropic image-driven regulariser of Nagel and Enkelmann [12]:

$$E_S(\lambda) = \int_\Omega \nabla\lambda^\top D(\nabla g)\nabla\lambda \, d\mathbf{x} \,. \tag{6}$$

Here, $D(\nabla g)$ is a normalised and regularised projection matrix orthogonal to ∇g. It is given by

$$D(\nabla g) = \frac{1}{|\nabla g|^2 + 2\nu^2} \begin{pmatrix} g_y^2 + \nu^2 & -g_x g_y \\ -g_x g_y & g_x^2 + \nu^2 \end{pmatrix}$$

with some small regularisation parameter ν.

Now we can write down the complete energy functional by combining the data term (5) and the smoothness term (6):

$$E(\lambda) = \int_\Omega \Big(\frac{1}{N} \sum_{i=0}^{N-1} \Psi\left(\mathbf{w}^\top J^i \mathbf{w}\right) + \alpha \nabla\lambda^\top D \nabla\lambda \Big) \, d\mathbf{x} \,. \tag{7}$$

3 Euler-Lagrange Equation

From the calculus of variations [4] we know that a necessary condition for a function $\lambda(x, y)$ to be a minimiser of the energy functional (7) is given by the Euler-Lagrange equation

$$\sum_{i=0}^{N-1} \frac{1}{N} \Psi'\Big(\mathbf{w}^\top J^i \mathbf{w}\Big)(J^i_{11}\lambda + J^i_{12}) - \operatorname{div}(D(\nabla g)\nabla\lambda) = 0$$

with reflecting boundary conditions.

This nonlinear partial differential equation can be solved with the help of two nested fixed point iterations: The outer loop fixes nonlinearities with previously computed values of λ, while the inner loop solves the resulting linear problem with the well-known successive overrelaxation (SOR) method [16].

4 Experiments

We evaluate the performance of the algorithm with the help of two synthetic sequences created in OpenGL: The first one illustrates a female head, as shown in Figure 1, while the second one represents a more challenging task – reconstruction of a tree illustrated in Figure 2. The performance of the method was tested on original sequences and versions with varying illumination as well as variants with noise. Moreover, the results for the original sequences were compared to the publicly available two-frame graph cuts method of Kolmogorov and Zabih [10].

In our experiments both sequences contain up to 8 images with small displacements of up to one pixel between successive camera positions. The ground truth maps were obtained by rescaling and transforming the original OpenGL Z-buffers into disparity maps. Consequently, for comparison with a ground truth we compute the *average absolute disparity error (AADE)*

$$\mathbf{AADE} = \frac{1}{M} \sum_{i=1}^{M} |d_i^{\mathbf{truth}} - d_i^{\mathbf{estimate}}| \,,$$

where M denotes the number of pixels.

There are just two model parameters that require adjustment: A smoothness parameter α and a standard deviation of a Gaussian σ for the preprocessing step. Other numerical parameters were kept fixed and constant for all experiments. The computation in all cases was stopped when the normalised L^1 norm of the updates at a certain iteration k became sufficiently small:

$$\frac{\sum_i |\lambda_i^k - \lambda_i^{k-1}|}{\sum_i \lambda_i^k} < \eta$$

In our experiments values for η vary between 10^{-6} to 10^{-8}. The average time, required for the evaluation of our experiments on an Intel Pentium 4 CPU with 3.2 GHz is in the order of 10 to 40 minutes (for 8 images degraded with Gaussian noise). More sophisticated solvers such as multigrid methods, however, may allow even for runtimes of less than a second [3].

4.1 Face Sequence

In this experiment we were using 8 images of a head scene, created in 3DS Max 8.0 and imported to OpenGL. The performance of the algorithm was tested on the original sequence, its degraded version, contaminated with Gaussian noise of $\sigma = 25$ and a sequence made from the same scene but under conditions of varying illumination (see Figure 1). In all experiments we observed that the main details of the head have been reconstructed in a realistic way: One can recognise that the reconstructed object represents a human face with clearly shaped nose, lips and eye slots. All experiments in this subsection have been carried out with two slightly different error measures: Once, the overall AADE of the computed disparity map was used; the other time, the AADE measurement was restricted to the face region by using a mask shown in the Figure 1. Table 1 shows optimal parameters and error measures for the first setting, while Table 2 presents the results for the second setting. Further on we observe that the results are fairly robust under noise and varying illumination: All essential features of the face remain recognisable and in accordance with the ground truth map.

Additionally we have investigated the influence of the number of images on the reconstruction quality. The clear difference in error measurements confirms our expectations: A larger number of images produces more stable results, since the amount of correspondences and, therefore, the reliability of the result increases.

Table 1. Results for the *Head scene*. $AADE_f$ = Average absolute disparity error computed for the whole disparity map. Disparity values for these experiments vary in the interval (0.1, 1). The parameters α and σ have been optimised.

	2 Frames	4 Frames	6 Frames	8 Frames	Noise, 8 Fr.	Illum., 8 Fr.
α	0.5	0.12	0.06	0.05	4.9	0.6
σ	2.5	2.7	2.8	2.9	5.7	1.7
$AADE_f$	0.0357	0.0298	0.0284	0.0286	0.0819	0.0387

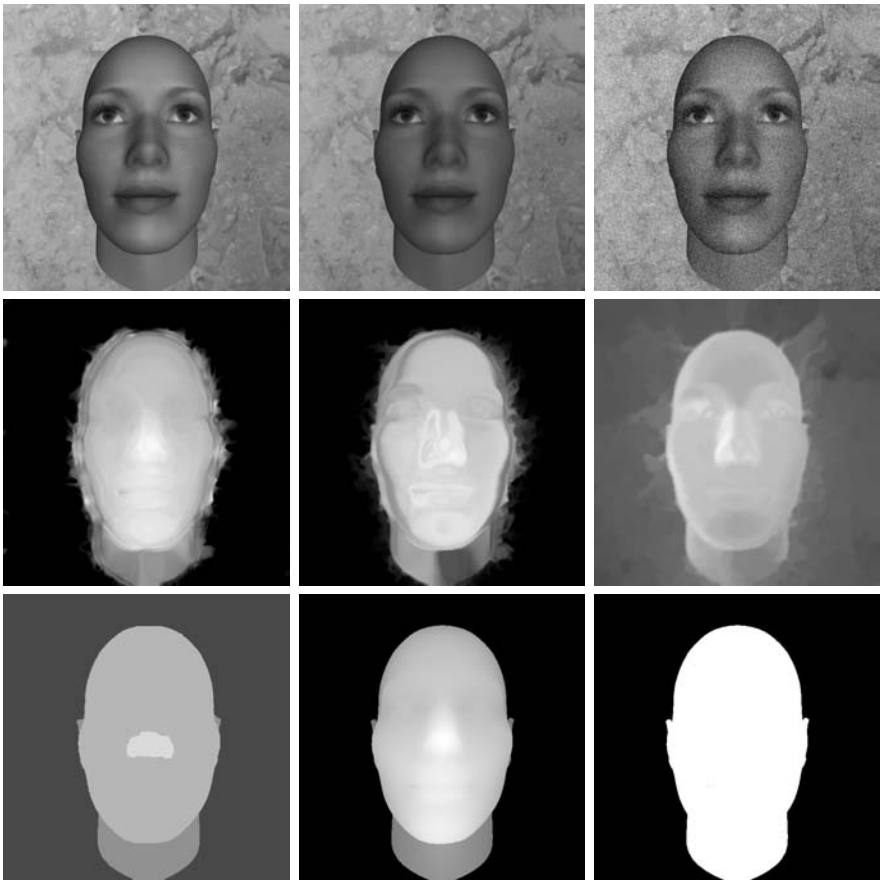

Fig. 1. *Head sequence*, top to bottom, left to right. *First row:* Original frame 1, frame 7 of a sequence with varying illumination, frame 1 of the sequence degraded with Gaussian noise of $\sigma = 25$. *Second row:* Typical results of reconstruction for 8 images of the original sequence, the sequence with illumination changes, and the noisy sequence. *Third row:* Graph cuts result (8 disparity levels), ground truth and mask.

Table 2. Results for the *Head scene*. $AADE_m$ = Average absolute disparity error computed for the face only. Disparity values for these experiments vary in the interval (0.1, 1). The parameters α and σ have been optimised.

	2 Frames	4 Frames	6 Frames	8 Frames	Noise, 8 Fr.	Illum., 8 Fr.
α	0.4	0.27	0.21	0.14	4.0	31
σ	4.1	3.9	4.0	4.1	10.1	2.2
$AADE_m$	0.0244	0.0204	0.0193	0.0192	0.0569	0.0371

Finally, let us compare our results to the one obtained by using the graph cuts method of Kolmogorov and Zabih [10]. Since this method relies on large displacements, we computed the disparity map between the first and the eighth image of the noise free sequence and divided the obtained result by 7 (number of images minus one). The corresponding disparity map which is presented in Figure 1 illustrates a very precise reconstruction of the silhouette of the head with clear distinction of the ears and the neck. However, the main features of the face were completely lost. Evidently, the algorithm is not able to reconstruct these features, because this would require to estimate the displacements at the corresponding locations with sub-pixel precision. But even for relatively large displacements it is well-known that reconstructions of graph cuts methods for such smoothly varying surfaces suffer from similar stair-casing effects [11], this time, however, due to the strong non-convexity of typical regularisers. Our observations are confirmed by the higher AADE for the graph cuts method for both the face region and the whole sequence which is given by $AADE_f = 0.0766$ and $AADE_m = 0.030$, respectively.

4.2 Tree Sequence

In this experiment we reconstruct an object of a very complex structure with fine level of details. Additional difficulty for the algorithm represents a homogeneous region, that corresponds to the sky above the landscape. As before, we make our task even more challenging by degrading the original sequence with Gaussian noise of $\sigma = 25$ and also varying the illumination in the scene.

For the original sequence we observe a very detailed reconstruction: Separate branches of the tree were estimated in accordance to the model, the overall silhouette of the tree was preserved quite well, even the difference in depth between neighbouring leaves appears to be very close to the ground truth map. The homogeneous region, corresponding to the sky was also estimated satisfactory: Since hardly any information is available in the sky region that allows for a direct estimation of the motion, our method propagates this information via the smoothness term. Again, the reconstruction process shows robustness with respect to noise and varying illumination: Both disparity maps show high similarity to the ground truth map with slightly higher values of AADE. In this experiment the difference between AADE values for the original sequence and those with noise and illumination change is not so large as in the previous experiment for the head sequence. This can be explained with the complexity of

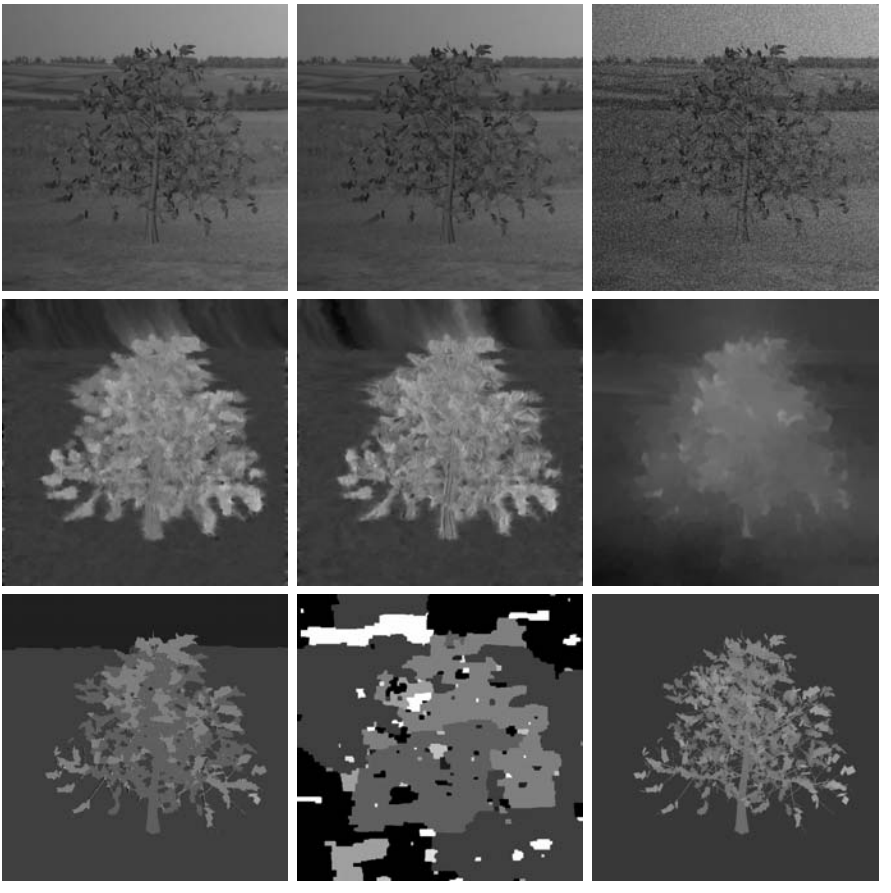

Fig. 2. *Tree sequence*, top to bottom, left to right. *First row:* Original Frame 1, frame 7 with varying illumination, frame 1 degraded with Gaussian noise of $\sigma = 25$. *Second row:* Typical results of reconstruction from 8 images of original sequence, with illumination changes and noise. *Third row:* Graph cuts results for noise free and noisy image sequence (8 disparity levels), ground truth.

the reconstructed object which leads to larger errors already in the undisturbed sequence.

The result of the graph cuts method for the noise free sequence between the first and eighth image (see Fig. 2) shows quite accurate reconstruction of the scene. Separate branches and overall shape of the tree were reconstructed very well and in accordance with the ground truth map. However, once again small variations of the disparity values cannot be estimated appropriately (the different disparity layers within the tree are very well visible). The corresponding AADE of $AADE = 0.0793$ for the graph cuts method is nevertheless close to ours. This is due to the accurate spatial reconstruction of the shape of the tree. For the noisy image sequence, however, the graph cuts method gives very poor results.

Table 3. Results for the *Tree sequence*. AADE = Average absolute disparity error. Disparity values for these experiments vary in the interval (0.1, 1).

	2 Frames	4 Frames	6 Frames	8 Frames	Noise, 8 Fr.	Illum., 8 Fr.
α	0.7	0.2	0.1	0.03	27.0	0.12
σ	1.8	2.0	2.3	2.66	2.7	2.18
AADE	0.0718	0.0644	0.0622	0.0616	0.0635	0.0650

Although we applied the same presmoothing strategy as for our stereo method, the disparity map contains many artifacts and the overall shape of the tree is very hard to recognise.

As before, we have experimented with smaller data sets of 2, 4 and 6 consequent images. Resulting error measures, presented in the Table 3 show consequent improvement as the number of images in the sequence grows.

5 Summary and Outlook

We have proposed a variational technique for a specific task of 3-D reconstruction for multiple views with small baseline distances. The method has been tailored towards applicability under more challenging conditions by incorporating various concepts that allow to handle data sets with varying illumination and noise. We have evaluated the performance of the approach with two sets of synthetic data with good results. The phenomena which are not taken into account so far are occlusions and specular reflections. This is a part of our ongoing work, whereby for the handling of specular reflections ideas from [9] and [1] are expected to be useful. An extension of our algorithm to arbitrary camera ego-motion is another topic of current research.

Acknowledgements

The authors thank Christian Morbach for providing his code for importing the 3ds files into OpenGL. Natalia Slesareva also gratefully acknowledges funding by the International Max-Planck Research School.

References

1. Birkbeck, N., Cobzas, D., Sturm, P., Jägersand, M.: Variational shape and reflectance estimation under changing light and viewpoints. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 536–549. Springer, Heidelberg (2006)
2. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optic flow estimation based on a theory for warping. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3024, pp. 25–36. Springer, Heidelberg (2004)

3. Bruhn, A., Weickert, J., Kohlberger, T., Schnörr, C.: A multigrid platform for real-time motion computation with discontinuity-preserving variational methods. International Journal of Computer Vision 70(3), 257–277 (2006)
4. Elsgolc, L.E.: Calculus of Variations. Pergamon, Oxford (1961)
5. Faugeras, O., Keriven, R.: Variational principles, surface evolution, PDE's, level set methods, and the stereo problem. IEEE Transactions on Image Processing 7(3), 336–344 (1998)
6. Furukawa, Y., Ponce, J.: Carved visual hulls for image-based modeling. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 564–577. Springer, Heidelberg (2006)
7. Huber, P.J.: Robust Statistics. Wiley, New York (1981)
8. Pons, R.K.J.-P., Faugeras, O.: Modelling dynamic scenes by registering multi-view image sequences. In: International Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 822–827 (2005)
9. Jin, H., Yezzi, A.J., Soatto, S.: Variational multiframe stereo in the presence of specular reflections. In: Proceedings of the 1st International Symposium on 3D Data Processing Visualization and Transmission, June 2002, pp. 626–631 (2002)
10. Kolmogorov, V., Zabih, R.: Computing visual correspondences with occlusions using graph cut methods. In: Proc. Eighth International Conference on Computer Vision, Vancouver, Canada, July 2001, vol. 2, pp. 588–594. IEEE Computer Society Press, Los Alamitos (2001)
11. Li, G., Zucker, S.W.: Differential geometric consistency extends stereo to curved surfaces. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 44–57. Springer, Heidelberg (2006)
12. Nagel, H.-H., Enkelmann, W.: An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. IEEE Transactions on Pattern Analysis and Machine Intelligence 8, 565–593 (1986)
13. Pressigout, M., Marchand, E.: Model-free augmented reality by virtual visual servoing. In: IAPR Int. Conf. on Pattern Recognition, ICPR'04, Cambridge, UK, August 2004, vol. 2, pp. 887–891 (2004)
14. Stricker, D.: Tracking with reference images: A real-time and markerless tracking solution for out-door augmented reality applications. In: Virtual Reality, Archaeology, and Cultural Heritage International Symposium (VAST01), Glyfada, Greece, November 2001 (2001)
15. Yezzi, A.J., Soatto, S.: Structure from motion for scenes without features. In: Proc. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 2003, vol. 1, pp. 525–532. IEEE Computer Society Press, Los Alamitos (2003)
16. Young, D.M.: Iterative Solution of Large Linear Systems. Dover, New York (2003)