Variational Stereovision and 3D Scene Flow Estimation with Statistical Similarity Measures

J.-P. Pons[†], R. Keriven[†], O. Faugeras[*] and G. Hermosillo[*]
[*]INRIA, 2004 Route des Lucioles [†]CERMICS, ENPC
Sophia-Antipolis, France Marne-La-Vallée, France
[jppons,faugeras,ghermosi]@sophia.inria.fr, keriven@cermics.enpc.fr

Abstract

We present a common variational framework for dense depth recovery and dense three-dimensional motion field estimation from multiple video sequences, which is robust to camera spectral sensitivity differences and illumination changes. For this purpose, we first show that both problems reduce to a generic image matching problem after backprojecting the input images onto suitable surfaces. We then solve this matching problem in the case of statistical similarity criteria that can handle frequently occurring non-affine image intensities dependencies. Our method leads to an efficient and elegant implementation based on fast recursive filters. We obtain good results on real images.

1. Introduction

The correspondence problem is the core problem of both structure and motion estimation. To solve this highly ambiguous problem, most methods compare image intensities by their difference, relying on very strong assumptions, such as the lambertian assumption for the stereo problem or the brightness constancy assumption for optical flow.

Correlation techniques can cope with affine changes of image intensities. They have been successfully used both for the stereo problem and in optical flow block matching algorithms. However, these techniques often use a fixed neighborhood, whereas a surface patch of the scene may have different shapes in different cameras and over time. In the stereo problem, the underlying hypothesis is that the camera retinal planes are identical and that the scene is made of fronto parallel planes. In some works, this limitation is alleviated by taking into account the tangent plane to the object [5] or by using adaptative windows [9, 17].

In this paper, in order to avoid projective distorsion, we map depth recovery and three-dimensional motion estimation from multiple calibrated video sequences to a generic image matching problem by backprojecting the input images onto suitable surfaces. This way, no shape approximation such as the tangent plane approximation is needed. Our matching window is not a hard window in the input images as in standard correlation techniques, but a smooth Gaussian window that operates along the objects' surfaces inside the backprojected volume images.

Moreover, in order to cope with non-affine intensity dependencies we use statistical similarity criteria which have already proven successful in multimodal image registration [22, 10, 16, 7].

We have designed a theoretical and computational framework for both problems: we minimize an energy functional. Furthermore, we have proved the well-posedness of the minimization process in both cases.

Three-dimensional structure and motion estimation from multiple video sequences has long been limited to rigid or piecewise-rigid scenes [23, 4, 19] or parametric models [11, 24]. The problem of computing a dense non-rigid 3D motion field, namely *scene flow* [20], from multiple video sequences has been addressed only recently.

Some techniques [25, 3, 12] use the spatio-temporal derivatives of the input images. As pointed out in [20], estimating scene flow from these derivatives without regularization is an ill-posed problem. Indeed, the associated normal flow equations only constrain the scene flow vector to lie on a line parallel to the iso-brightness contour on the object. This is nothing but a 3D version of the aperture problem for optical flow [2]. In [3, 12], several samples of the spatio-temporal derivatives are combined in order to over-constrain scene flow, whereas in [25], the aperture problem is solved by combining the normal flow constraint with a Tikhonov smoothness term. However, due to the underlying brightness constancy assumption, and to the local relevance of spatio-temporal derivatives, differential methods apply mainly to slowly-moving lambertian scenes under constant illumination.

In [21], shape and scene flow are estimated simultane-

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

ously using a plane-sweep carving algorithm in a 6D space. However, this approach still relies on a brightness constancy assumption, has a very high computational and memory cost, and is unable to enforce the smoothness of the recovered shape and motion.

Some other techniques [18, 20, 25, 6] rely on previous optical flow computations. However, the latter may be noisy and/or physically inconsistent through cameras. The heuristic spatial smoothness constraints used in most optical flow methods may also alter the recovered scene flow.

Our method for scene flow estimation neither needs previous optical flow computations nor makes use of ambiguous spatio-temporal image derivatives. It proceeds by directly evolving a 3D vector field so as to fit to the intensity changes in all cameras. It is robust to illumination changes through the use of statistical similarity criteria. Moreover, it can recover large displacements thanks to a multi-resolution coarse-to-fine strategy.

In the sequel, we focus on the case of two cameras not to overload notations. Our framework extends easily to the N-camera case simply by summing the statistical criteria over all pairs of cameras. The rest of this paper is organized as follows. Section 2 defines the statistical similiarity criteria to be used in subsequent sections. In Section 3, we present our stereovision method. Section 4 describes our novel method for scene flow estimation.

2. Statistical intensity similarity criteria

We consider two similarity criteria which assume different relations between the image intensities. The cross correlation (CC) is a measure of the affine dependency. The mutual information (MI) [22, 10] measures how the intensity distributions of two images fail to be independent. Our two criteria can also take two different forms. A global form computed for the entire image, and a local form computed on corresponding regions. The latter can cope with non-stationary joint probability distributions of the intensities.

The global criteria $\mathbf{CC}^g$ and $\mathbf{MI}^g$ are computed from the global joint probability density function, estimated by the Parzen window method [13], with a Gaussian window with variance $\beta > 0$ as the Parzen kernel. For two images I_1 and I_2 defined over a bounded domain Ω of $\mathbb{R}^2$, the joint probability density function is given by

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

$$\text{or more concisely: } P(\mathbf{i}) = \frac{1}{|\Omega|} \int_\Omega G_\beta \left(\mathbf{I}(\mathbf{x}) - \mathbf{i}\right) \, d\mathbf{x} \,.$$

The local criteria use a space-dependant Gaussian-weighted version of this estimator

$$P_{\mathbf{x}_0}(\mathbf{i}) = \frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \int_\Omega G_\beta(\mathbf{I}(\mathbf{x}) - \mathbf{i}) \, G_\gamma(\mathbf{x} - \mathbf{x}_0) \, d\mathbf{x} \,,$$

where $\mathcal{G}_\gamma(\mathbf{x}_0) = \int_\Omega G_\gamma(\mathbf{x} - \mathbf{x}_0) \, d\mathbf{x}$. The variance $\gamma > 0$ controls the neighborhood size. New local similarity measures $\mathbf{CC}^l$ and $\mathbf{MI}^l$ are obtained by integration over Ω of the local estimations of the cross correlation and the mutual information.

3. Variational stereovision with statistical measures

Our approach proceeds by deforming a surface so as to match the backprojected images of the different cameras. The matching criterion is one of the statistical measures defined in Section 2. The surface deformation is driven by the minimization of an energy functional through a gradient descent method.

3.1. Notations

We model the objects of the scene as the graph, in a cartesian setting, of an unknown function f defined over a bounded domain Ω of $\mathbb{R}^2$ with smooth boundary $\partial\Omega$, and belonging to a dense linear subspace F_1 of the Hilbert space $H_1 = L^2(\Omega, \mathbb{R})$.

We note I_k the image captured by camera k, and $\mathcal{I}_k$ the backprojection of I_k onto the entire 3D space. That is, $\mathcal{I}_k(x, y, z)$ is the intensity of the pixel obtained by projecting the 3D point (x, y, z) onto image k. Thus, the camera geometry is encapsulated in function $\mathcal{I}_k$. The gradient of $\mathcal{I}_k$ can be readily obtained from the gradient of image I_k and the coefficients of the perspective projection matrix of camera k. We note $\boldsymbol{\mathcal{I}}$ the pair $(\mathcal{I}_1, \mathcal{I}_2)$. We denote by $\mathbf{S}$ the parameterized surface $(x, y) \mapsto (x, y, f(x, y))$, so that the backprojection of image k onto the surface is given by $\mathcal{I}_k \circ \mathbf{S}$. We note $\mathbf{M}$ the 3D point $(x, y, f(x, y))$.

3.2. Variational formulation

We define the stereo problem as the minimization of a cost functional $\mathcal{E}_1(f) = \mathcal{M}_1(f) + \mathcal{R}_1(f)$, where $\mathcal{M}_1(f)$ measures the statistical dissimilarity of the backprojected images $\mathcal{I}_1 \circ \mathbf{S}$ and $\mathcal{I}_2 \circ \mathbf{S}$, while $\mathcal{R}_1(f)$ defines regularizing constraints on f.

Note that, thanks to the backprojecting step, our method matches against intensities along the objects' surfaces in constrast with standard correlation techniques which rely on an approximation of the objects' shapes. Moreover, our local similarity measures operate on smooth Gaussian windows in the backprojected volume images, in contrast with the popular rigid rectangular windows in the input images.

A classical choice for the regularization term $\mathcal{R}_1$ is

$$\mathcal{R}_1(f) = \alpha \int_\Omega \phi \left(\nabla f(\mathbf{x})\right) \, d\mathbf{x}$$

for some function $\phi : \mathbb{R}^2 \mapsto \mathbb{R}_+$, and some weighting parameter α. The choice of ϕ enforces different smoothness assumptions on f. The classical *Tikhonov* regularization often used in ill-posed problems, correponds to $\phi(.) = \frac{1}{2}|.|^2$. Several other ϕ functions have been designed in order to preserve depth discontinuities [15]. In our implementation, we have considered the Perona-Malik, the Rudin and the Aubert functions.

We seek a minimum $\hat{f} = \arg\min_{f \in F_1} \mathcal{E}_1(f)$. One can show that a necessary condition of optimality is the so-called Euler equation $\nabla_{H_1} \mathcal{E}_1(f) = \mathbf{0}$. Rather than trying to solve directly the Euler equation, which is impossible in most cases, we follow a gradient descent strategy starting from a guess f_0. That is, we solve the following evolution problem for f, which becomes a function $[0, +\infty[\rightarrow H_1$:

$$\left\{ \begin{array}{ll} f(0) & = f_0 \in H_1 \\ \frac{df}{d\tau} & = -\nabla_{H_1} \mathcal{E}_1(f(\tau)) \quad , \quad \tau > 0. \end{array} \right. \tag{1}$$

We call a global classical solution of equation (1) a function $f \in C^0([0, +\infty[, H_1) \cap C^1(]0, +\infty[, H_1) \cap C^0(]0, +\infty[, F_1)$ which satisfies equation (1).

The explicit computation of the gradient of the statistical dissimilarity term $\mathcal{M}_1$ was carried out using the same pattern as in [8, 7]. The gradient in H_1 of $\mathcal{M}_1$ for the two global criteria is given by

$$\nabla \mathcal{M}_1(f)(\mathbf{x}) = \frac{1}{|\Omega|} \sum_{k=1,2} \left(G_\beta \star \partial_k E_f\right)(\boldsymbol{\mathcal{I}}(\mathbf{M})) \frac{\partial \mathcal{I}_k}{\partial z}(\mathbf{M}) \,, \tag{2}$$

where $\star$ indicates a convolution with respect to the two intensities and E_f depends on the criterion:

$$\begin{aligned} E_f^{\mathbf{CC}^g}(\mathbf{i}) &= \frac{-1}{v_1\, v_2} \left[2 v_{1,2}(i_1 i_2 - i_1 \mu_2 - i_2 \mu_1) \right. \\ &\quad \left. - \mathbf{CC}^g \left(i_2 v_1 (i_2 - 2\mu_2) + i_1 v_2 (i_1 - 2\mu_1)\right)\right] \\ E_f^{\mathbf{MI}^g}(\mathbf{i}) &= -\left(1 + \log \frac{P_f(\mathbf{i})}{P_f(i_1)\, P_f(i_2)}\right) \,, \end{aligned} \tag{3}$$

where P_f is the global joint probability distribution, $\mu_k,\ v_k,\ k = 1, 2$ are the averages and the variances and $v_{1,2}$ is the covariance of the backprojected images $\mathcal{I}_1 \circ \mathbf{S}$ and $\mathcal{I}_2 \circ \mathbf{S}$.

The gradient in H_1 of $\mathcal{M}_1$ for the local criteria is given by

$$\nabla \mathcal{M}_1(f)(\mathbf{x}) = \sum_{k=1,2} \left(G_\gamma \star \left(G_\beta \star \frac{1}{\mathcal{G}_\gamma} \partial_k E_f\right)\right)(\boldsymbol{\mathcal{I}}(\mathbf{M}), \mathbf{x}) \frac{\partial \mathcal{I}_k}{\partial z}(\mathbf{M}) \,, \tag{4}$$

where the first convolution acts on the two spatial variables while the second is still on the intensity variables. $E_f^{\mathbf{CC}^l}$ and $E_f^{\mathbf{MI}^l}$ are space-dependent versions of equation (3).

The gradient in H_1 of $\mathcal{R}_1$ is given by

$$\nabla \mathcal{R}_1(f) = -\alpha \ \mathrm{div}\ (\nabla \phi \circ \nabla f) \ . \tag{5}$$

Self occlusions can be taken into account by restricting the integration domain of the similarity criteria to the portion of the surface visible from both cameras. If we carry out the derivation of these modified criteria under the widely accepted technical assumption that the visibility remains constant, we get the same expressions as equations (2) and (4), except that the gradients are now supported by the visible domain. This approach is not included in our current implementation because it is of limited practical interest in the case of depth maps.

The following theorem, detailed in [14], addresses the well-posedness of the minimization process:

Theorem 1 *If the following assumptions are satisfied:*

- ϕ *is a positive definite quadratic form,*
- $\forall k$, $\mathcal{I}_k$ *and* $\nabla \mathcal{I}_k$ *are bounded and Lipschitz continuous,*

then equation (1) has a unique global classical solution.

Consider the *Tikhonov* case, and suppose image itensities are bounded. Let us enforce $\mathcal{I}_k(x, y, z) = 0\,, \forall |z| > C$, which simply states that we do not consider arbitrarily high depth values. Let us substitute to $\mathcal{I}_k$ its convolution with a 3D Gaussian kernel of variance $\sigma > 0$: $\mathcal{I}_k^\sigma = G_\sigma \star \mathcal{I}_k$. Then Theorem 1 applies. Moreover, the Gaussian smoothing stage is compatible with a multi-resolution strategy. Indeed, the energy functional $\mathcal{E}_1$ may be nonconvex due to its data term, so that the gradient descent may be trapped in a local minimizer. As a consequence, its asymptotic state depends on the initial guess f_0.

In order to avoid convergence to physically irrelevant local minima, we adopt a multi-resolution coarse-to-fine strategy as in [1]. The flow equation (1) is applied to a set of smoothed and subsampled volume images $\mathcal{I}_k^\sigma$. In our experiments, the initial guess for the coarsest resolution is a plane with a suitable constant depth z_0.

3.3. Numerical experiments

The gradients of the statistical criteria described in Section 2 can be implemented efficiently thanks to fast recursive filtering . The computation time is of a few seconds per frame for medium resolution reconstructions.

Figure 1 shows the stereo pair used in our first series of experiments and compares the reconstructed surfaces obtained for different statistical criteria. The poor results of

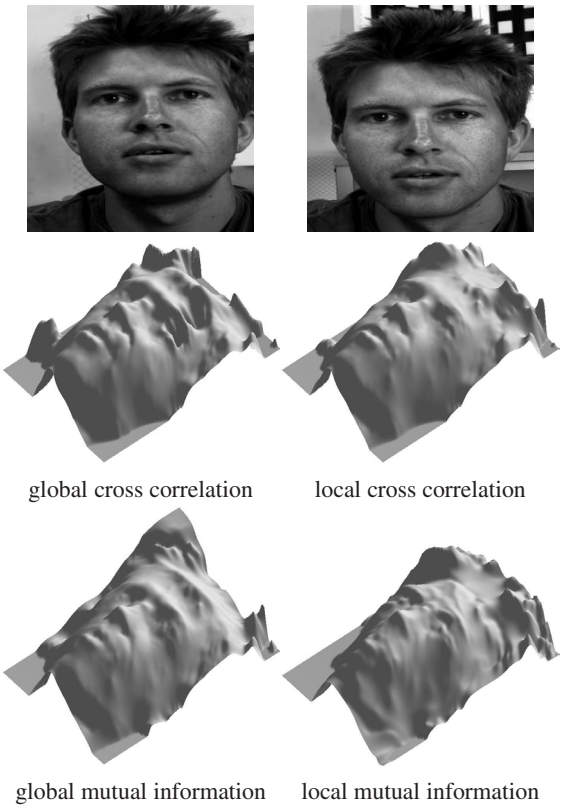

Figure 1. Stereo pair (top) and reconstructed surface for different statistical criteria.

global cross correlation suggest that no global affine dependency exists for the considered stereo pair. Local cross correlation and global mutual information yield good results, whereas local mutual information performs worse. Indeed, the amount of information used for matching images is smaller in this case. Hence, local mutual information should be reserved to extreme imaging conditions in which all other criteria fail. Figure 2 shows a high resolution reconstruction obtained with local cross correlation and some views with backprojected texture.

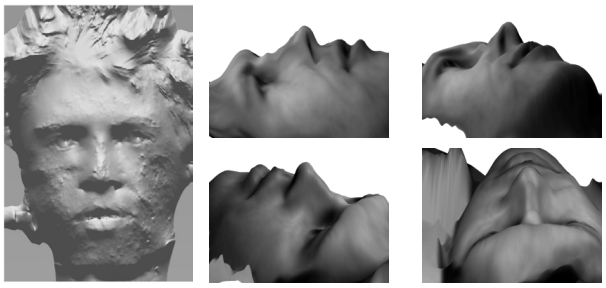

Figure 2. Some detailed views of the reconstructed surface.

Figure 3 illustrates the robustness of our method to camera spectral sensitivity differences. We have transformed the intensities of the initial stereo pair to simulate different sensors. While local cross correlation fails in this case,

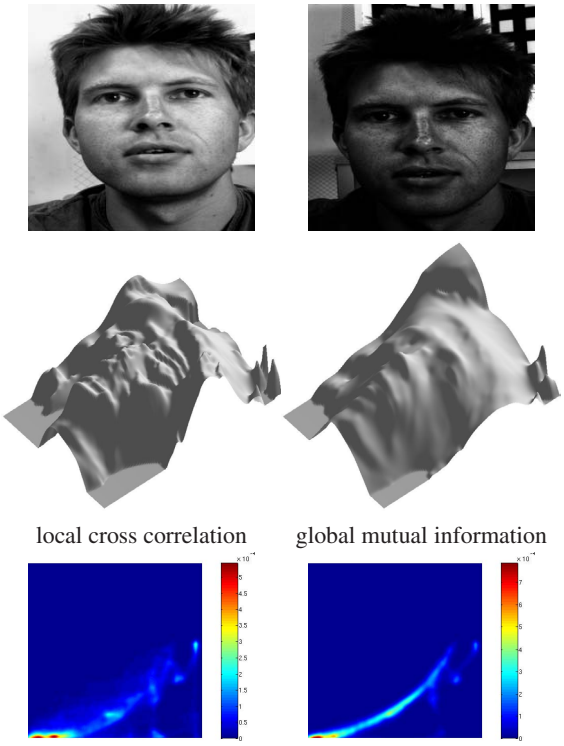

Figure 3. Modified stereo pair (top). Reconstructed surface with local cross correlation (middle left) and with global mutual information (middle right). Joint probability distribution of backprojected images onto the initial (bottom left) and the final (bottom right) surface with global mutual information.

global mutual information yields good results. Figure 3 also represents the evolution of the joint probability distribution of backprojected images in this case: we can see that the mutual information criterion tends to cluster the joint probability distribution, and that the non-affine dependency we had imposed could be recovered correctly.

4. Variational scene flow estimation with statistical measures

We evolve a 3D vector field defined on an estimation of the object surface so as to match the backprojected image at time instant t and the backprojected image onto the predicted surface at time instant $t+1$ in all cameras. This way, we enforce the agreement between the estimated scene flow

and the 2D displacements in all cameras, without an explicit use of previous optical flow computations or of ambiguous spatio-temporal image derivatives. As in Section 3, the vector field evolution is driven by the minimization of an energy functional through a gradient descent method.

4.1. Notations

We denote by $\mathbf{S}^t$ the estimated surface at time instant t, modelled as the graph of a function $f^t \in F_1$. We model the 3D scene flow between t and $t+1$ as an unknown function $\mathbf{v}^t$ belonging to a dense linear subspace F_2 of the Hilbert space $H_2 = L^2\left(\Omega, \mathbb{R}^3\right)$. The predicted surface at time instant $t+1$ is given by $\mathbf{S}^t + \mathbf{v}^t$. We do not constrain the deformed surface $\mathbf{S}^t + \mathbf{v}^t$ to agree with $\mathbf{S}^{t+1}$ in order to make scene flow estimation robust to surface estimation errors.

In this section, we note I_k^t the image captured by camera k at time instant t, and we define $\mathcal{I}_k^t$ as in paragraph 3.1. We note $\mathbf{M}$ the 3D point $(x, y, f^t(x,y))$ and $\mathbf{V}$ the vector $\mathbf{v}^t(x,y)$.

4.2. Variational formulation

We consider the minimization of a cost functional $\mathcal{E}_2(\mathbf{v}^t) = \mathcal{M}_2(\mathbf{v}^t) + \mathcal{R}_2(\mathbf{v}^t)$, where $\mathcal{M}_2(\mathbf{v}^t)$ measures the global or local statistical dissimilarity between the backprojected images $\mathcal{I}_k^t \circ \mathbf{S}^t$ at time instant t and the backprojected images $\mathcal{I}_k^{t+1} \circ (\mathbf{S}^t + \mathbf{v}^t)$ onto the predicted surface at time instant $t+1$, while $\mathcal{R}_2$ defines regularizing constraints and smoothness assumptions on $\mathbf{v}^t$. In our experiments, we consider a *Tikhonov* regularization, but any other physics-based or application-specific smoothness term $\mathcal{R}_2$ could be considered.

The gradient in H_2 of $\mathcal{M}_2$ for global criteria is given by

$$\nabla \mathcal{M}_2(\mathbf{v}^t)(\mathbf{x}) = \frac{1}{|\Omega|} \sum_k \left(G_\beta \star \partial_2 E_{\mathbf{v}}^k\right) \left(\mathcal{I}_k^t(\mathbf{M}) \, , \, \mathcal{I}_k^{t+1}(\mathbf{M}+\mathbf{V})\right) \nabla \mathcal{I}_k^{t+1}(\mathbf{M}+\mathbf{V}), \quad (6)$$

where $E_{\mathbf{v}}^k$ is defined from images $\mathcal{I}_k^t \circ \mathbf{S}^t$ and $\mathcal{I}_k^{t+1} \circ (\mathbf{S}^t + \mathbf{v}^t)$ as in equation (3).

The gradient in H_2 of $\mathcal{M}_2$ for local criteria is given by

$$\nabla \mathcal{M}_2(\mathbf{v}^t)(\mathbf{x}) = \sum_k \left(G_\gamma \star \left(G_\beta \star \frac{1}{\mathcal{G}_\gamma} \partial_2 E_{\mathbf{v}}^k\right)\right) \left(\mathcal{I}_k^t(\mathbf{M}) \, , \, \mathcal{I}_k^{t+1}(\mathbf{M}+\mathbf{V}) \, , \, \mathbf{x}\right) \nabla \mathcal{I}_k^{t+1}(\mathbf{M}+\mathbf{V}) \, . \quad (7)$$

Under the same assumptions as those of Section 3, we have proved (see [14]) the well-posedness of the minimization process.

4.3. Numerical experiments

Figure 4 shows some frames of the input stereo sequence and the computed scene flow between the first two frames. We clearly see that the overall movement of the head and the closing of the mouth are recovered but it is somewhat difficult to evaluate the details of the flow in this figure. In order to show the precision of the computed scene flow, we have synthetized a motion-compensated 3D sequence from the initial surface and texture, and from the successive scene flows. Note that this is a challenging experiment since potential errors are accumulated over many frames. Figure 5 shows some previews of this sequence. The movement of the jaw is successfully recovered.

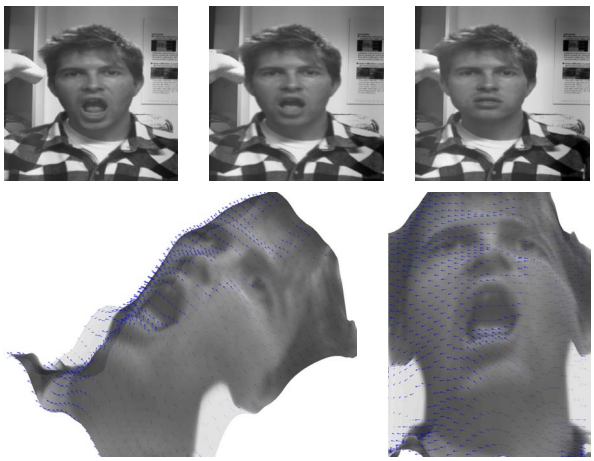

Figure 4. Some frames of the left input sequence (top) and some views of the estimated scene flow (bottom).

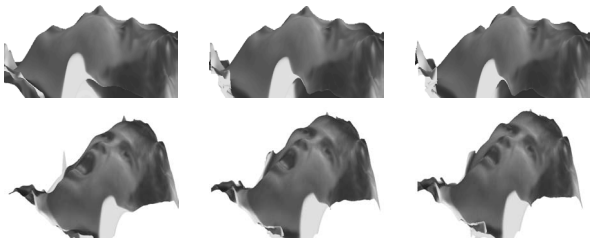

Figure 5. Preview of the motion-compensated 3D sequence.

5. Conclusion and future work

We have described a common variational framework for depth recovery and scene flow estimation from multiple calibrated video sequences. Our method avoids projective

distorsion by backprojecting the input images onto suitable surfaces and uses statistical similarity criteria to handle camera spectral sensitivity differences and illumination changes.

Our future work includes extending our framework to implicit surfaces defined by a level set function, and integrating shape and motion estimations in order to exploit their coherence and to improve their robustness and/or precision, as pointed out in [20]. We believe that this present work, by unifying stereo and scene flow estimation in the same coherent theoretical and computational framework, is a promising step towards this integration.

References

[1] L. Alvarez, J. Weickert, and J. Sànchez. Reliable estimation of dense optical flow fields with large displacements. *The International Journal of Computer Vision*, 39(1):41–56, Aug. 2000.

[2] J. Barron, D. Fleet, and S. Beauchemin. Performance of optical flow techniques. *The International Journal of Computer Vision*, 12(1):43–77, 1994.

[3] R. Carceroni and K. Kutulakos. Multi-view scene capture by surfel sampling: From video streams to non-rigid 3d motion, shape & reflectance. In *Proceedings of the 8th International Conference on Computer Vision*, pages 60–67, Vancouver, Canada, 2001. IEEE Computer Society, IEEE Computer Society Press.

[4] F. Dornaika and R. Chung. Stereo correspondence from motion correspondence. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 70–75, Fort Collins, Colorado, June 1999. IEEE Computer Society.

[5] 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, Mar. 1998.

[6] D. Garcia, J. Orteu, and L. Penazzi. A combined temporal tracking and stereo-correlation technique for accurate measurement of 3d displacements: Application to sheet metal forming. *Journal of Materials Processing Technology*, (125–126):736–742, Sept. 2002.

[7] G. Hermosillo. *Variational Methods for Multimodal Image Matching*. PhD thesis, INRIA, The document is accessible at ftp://ftp-sop.inria.fr /robotvis/html/Papers /hermosillo:02.ps.gz, 2002.

[8] G. Hermosillo and O. Faugeras. Dense image matching with global and local statistical criteria: a variational approach. In *Proceedings of CVPR'01*, 2001.

[9] T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(9):920–932, Sept. 1994.

[10] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging*, 16(2):187–198, Apr. 1997.

[11] s. Malassiotis and M. Strintzis. Model based joint motion and structure estimation from stereo images. *Computer Vision and Image Understanding*, 65(1):79–94, 1997.

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

[13] E. Parzen. On the estimation of probability density function. *Ann. Math. Statist.*, 33:1065–1076, 1962.

[14] J.-P. Pons, R. Keriven, O. Faugeras, and G. Hermosillo. Variational stereovision and 3D motion estimation with statistical measures. Research report, INRIA, 2002.

[15] L. Robert and R. Deriche. Dense depth map reconstruction: A minimization and regularization approach which preserves discontinuities. In B. Buxton, editor, *Proceedings of the 4th European Conference on Computer Vision*, Cambridge, UK, Apr. 1996.

[16] A. Roche, G. Malandain, X. Pennec, and N. Ayache. The correlation ratio as new similarity metric for multimodal image registration. In W. Wells, A. Colchester, and S. Delp, editors, *Medical Image Computing and Computer-Assisted Intervention-MICCAI'98*, number 1496 in Lecture Notes in Computer Science, pages 1115–1124, Cambridge, MA, USA, Oct. 1998. Springer.

[17] D. Scharstein and R. Szeliski. Stereo matching with nonliear diffusion. *International Journal of Computer Vision*, 28(2):155–174, June 1998.

[18] Y. Shi, C. Shu, and P. J.N. Unified optical flow field approach to motion analysis from a sequence of stereo images. *Pattern Recognition*, 27(12):1577–1590, 1994.

[19] C. Strecha and L. Van Gool. Motion - stereo integration for depth estimation. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *Proceedings of the 7th European Conference on Computer Vision*, pages 170–185, Copenhagen, Denmark, May 2002. Springer–Verlag.

[20] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. In *Proceedings of the 7th International Conference on Computer Vision*, volume 2, pages 722–729, Kerkyra, Greece, 1999. IEEE Computer Society, IEEE Computer Society Press.

[21] S. Vedula, S. Baker, S. Seitz, and T. Kanade. Shape and motion carving in 6d. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 592–598, Hilton Head Island, South Carolina, June 2000. IEEE Computer Society.

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

[23] W. Wang and J. Duncan. Recovering the three-dimensional motion and structure of multiple moving objects from binocular image flows. *Computer Vision and Image Understanding*, 63(3):430–446, 1996.

[24] Y. Zhang and C. Kambhamettu. Integrated 3D scene flow and structure recovery from multiview image sequences. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, South Carolina, June 2000. IEEE Computer Society.

[25] Y. Zhang and C. Kambhamettu. On 3d scene flow and structure estimation. In *Proceedings of CVPR'01*, 2001.