Multi-View Stereo Reconstruction and Scene Flow Estimation with a Global Image-Based Matching Score

Jean-Philippe Pons (jean-philippe.pons@certis.enpc.fr) and
Renaud Keriven (renaud.keriven@certis.enpc.fr)
Odyssée Laboratory, ENPC, Marne-la-Vallée, France

Olivier Faugeras (olivier.faugeras@sophia.inria.fr)
Odyssée Laboratory, INRIA, Sophia-Antipolis, France

Abstract. 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 global measure of image similarity, chosen depending on imaging conditions and scene properties. Contrarily to existing deformable surfaces methods, which integrate a matching measure computed independently at each surface point, our approach computes a global image-based matching score between the input images and the predicted images. The matching process fully handles projective distortion and partial occlusions. Neighborhood as well as global intensity information can be exploited to improve the robustness to appearance changes due to non-Lambertian materials and illumination changes, without any approximation of shape, motion or visibility. Moreover, our approach results in a simpler, more flexible, and more efficient implementation than in existing methods. The computation time on large datasets does not exceed thirty minutes on a standard workstation. Finally, 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 motion estimation algorithm on a very challenging multi-view video sequence of a non-rigid scene.

Keywords: stereovision, non-rigid 3D motion, scene flow, registration, prediction error, reprojection error, variational method, global image-based matching score, cross correlation, mutual information, non-Lambertian surface, level sets.

1. Introduction

1.1. Problem statement

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* [1] [36], from multiple video sequences. In this case, the input data

[1] The scene flow should not be confused with the optical flow, which is the two-dimensional motion field of points in an image. The optical flow is the projection of the scene flow in the image plane of a camera.

are a two-dimensional array of images, in which each row is a multi-view stereovision dataset for a given time instant, and each column is a video sequence captured by a given camera. Combining stereovision and scene flow allows to build a spatio-temporal model of a dynamic scene. Once such a model is available, some novel virtual views of the scene can be generated by interpolation across space and time [35].

Stereovision and scene flow estimation both require to match different images of the same scene, in other words to find points in different cameras and in different frames corresponding to a same physical point. Once the correspondence problem is solved, the shape and the three-dimensional motion of the scene can be recovered easily by triangulation. Unfortunately, the correspondence problem is a very difficult task in computer vision because a scene patch generally has different shapes and appearances when seen from different points of view and at different times. To overcome this difficulty, most existing stereovision and scene flow algorithms rely on unrealistic simplifying assumptions that disregard either/both shape/appearance changes.

1.2. Common photometric and geometric assumptions used for shape and motion estimation

The oldest and most naive assumption about the photometric properties of a scene is *brightness constancy*: corresponding pixels are assumed to have the same color. This only applies to strictly Lambertian objects and requires a precise photometric calibration of the cameras. Yet this assumption is still popular in the stereovision literature. It motivates the multi-view photo-consistency measure used in voxel coloring [28], space carving [19], and in some deformable surfaces methods [6, 20]. Similarly, the variational formulation of [32] relies on square intensity differences. In a later paper [31], 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.

This assumption is also present in many methods for scene flow estimation, through the use of the spatio-temporal derivatives of the input images [39, 4, 21]. Due to the 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, matching measures embedded in stereovision and scene flow algorithms have to aggregate neighborhood intensity 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 called the *fronto parallel*

hypothesis: the retinal planes of the cameras are identical and the scene is an assembly of planes parallel to them. This assumption can still be found in recent work [20, 12]. To minimize the impact of projective distortion, these authors compute the stereo discrepancy of a scene patch with its most front-facing cameras only. However, this approximation is questionable in most camera setups.

Some methods go beyond this hypothesis by taking into account the tangent plane to the object [8, 13, 6, 9], or by using adaptive matching windows [14, 25]. More generally, most techniques trade robustness to realistic photometric conditions for an approximation of shape and motion in the computation of the matching measure. As a result, the robustness of the matching process is uncertain in the parts of the scene that do not verify these approximations. For example, using fixed matching windows for stereo correspondence leads to an oversmoothing of depth discontinuities. Similarly, using a tangent plane approximation to compute the matching measure as in [8, 13, 6, 9], even if the tangent plane at nearby points does not have to be the same, is not relevant in the regions of high curvature of the objects.

1.3. Previous work on multi-view complete stereovision

Doing a complete review of the stereovision area is out of the scope of this article. We limit ourselves to the methods that allow to obtain a complete reconstruction of a scene from a high number of input views. The methods in which the geometry is represented by one or several depth maps or disparity maps are not of interest here, because they only yield partial models of the scene. Several such models can be fused at post-processing, but anyway these methods cannot handle visibility globally and consistently during the estimation. For sake of completeness, let us mention two recent important works in this category: the graph cuts method of [16] and the PDE-based method of [32]. The interested reader can also refer to [26] for a good taxonomy of dense two-frame rectified stereo correspondence algorithms.

Thus, in the following, we focus on multi-view complete stereovision methods. These methods fall into two categories: the *space carving* framework and the *deformable surfaces* framework.

1.3.1. *Space carving*

In the space carving framework [19], the scene is represented by a three-dimensional array of voxels. Each voxel can be labeled empty or occupied. When the algorithm starts, all voxels are occupied. Then the volume is traversed in an adequate order. If a voxel is not consistent with all the input images, it is relabeled empty. The order of the traversal is important because the visibility of the voxels is taken into account in the consistency test. In an earlier method called voxel coloring [28], there was a constraint on the

placement of the cameras, and the algorithm required only a single pass. Space carving handles arbitrary camera configurations but is a little more expensive computationally.

The space carving framework suffers from several important limitations. First, it makes hard decisions. Once a voxel is carved away, it cannot be recovered. And if one voxel is removed in error, further voxels can be erroneously removed in a cascade effect. This limitation is partially alleviated by the probabilistic space carving method [3]. Second, in the original space carving algorithm, the photo-consistency test derives from a brightness constancy constraint, and the choice of the global threshold on the color variance is often problematic. Recently, there have been some attempts to relax these photometric constraints [34, 38]. The robustness to calibration errors is also addressed in [18]. Third, the voxel-based representation disregards the continuity of shape, which makes it very hard to enforce any kind of spatial coherence. As a result, space carving is very sensitive to noise and outliers, and typically yields very noisy reconstructions.

1.3.2. *Deformable surfaces*

These methods inherit from the active contour method pioneered in [15]. Here, contrarily to the space carving framework, the formulation is continuous and has a geometric interpretation. The unknown scene is modelled by a two-dimensional surface, and scene reconstruction is stated in terms of an energy minimization. An initial surface, positioned by the user, is driven by a partial differential equation minimizing an energy functional.

The earlier and most inspiring work in this category is the level set stereovision method of [8]. In this work, the stereovision problem is formulated as a minimal surface approach, in the spirit of the geodesic active contours method [5]. In other words, the energy functional is written as the integral on the unknown surface of a data fidelity criterion. This criterion is the normalized cross correlation between image pairs. The surface evolution is implemented in the level set framework [22]. On the one hand, the implicit representation offers numerical stability and the ability to handle topological changes automatically. On the other hand, it is quite expensive computationally, even with a narrow band approach. Several variations to this approach have been proposed: an implementation with meshes [6], the addition of 3D points data [6, 20] and of silhouette information [12, 20], and an extension to spatio-temporal scenes [9]. More original is the method proposed in [13] to cope with non-Lambertian scenes. This method can estimate both the shape and the non-Lambertian reflectance of the scene. It outputs a geometric and photometric model which allows to predict the appearance of novel views. The surface deformation is driven by the minimization of the rank of a radiance tensor.

These deformable surfaces methods share several limitations. First, in all these methods, the matching measure is computed independently at each surface point, then these quantities are integrated on the surface. The matching measure at a point relies on a local approximation of the surface, either by a fronto-parallel plane [20, 12] or at best by the tangent plane [8, 13, 6, 9]. Moreover, the visibility of the whole neighborhood is assumed to be the same as the reference point. For example, in [8], the cross correlation between two slanted matching windows is computed without taking into account the eventual partial occlusions of the windows. The primary purpose of these assumptions is a simplification of the modelling and of the resulting computations. They are clearly not valid in real-world scenes, which typically include many occlusions, depth discontinuities and sharp angles. Thus, these simplifying assumptions make the robustness of the matching process on real data very uncertain.

Second, all these methods follow a minimal surface approach. One drawback of this approach is that data fidelity and regularization are mixed. As a result, it is difficult to tune the regularizing behavior. A good discussion of this topic can be found in [29]. The authors show in some numerical experiments that the results of [13] can be further improved by integrating the matching measure on the images rather than on the surface.

Third, they lack flexibility in the choice of the matching criterion. Photo-consistency [6, 20], the normalized cross correlation [8, 9, 12] and lastly the radiance tensor [13] have been considered. These matching measures are hard-wired in their respective method and cannot be upgraded to cope with different imaging conditions.

Finally, the dependency of the matching measure on the surface normal leads to a complex implementation. It requires to handle matching windows of different shapes or a tessellation of the tangent plane, at each surface point. It also results in a very complex minimizing flow involving second-order derivatives of the matching score [10, 30]. More precisely, the energy has the following form:

$$E(S) = \int_S g(\mathbf{x}, \mathbf{N})\, d\mathbf{x}\ , \tag{1}$$

where S denotes the surface and g the matching measure. The gradient writes

$$\nabla E(S) = [\nabla g \cdot \mathbf{N} + 2gH - \mathrm{div}_S(g_{\mathbf{N}})]\, \mathbf{N}\ , \tag{2}$$

where $\mathbf{N}$ is the normal, H is the mean curvature of the surface, $g_{\mathbf{N}}$ is the derivative of the matching measure with respect to the orientation of the tangent plane and div_S is the intrinsic divergence operator on the surface.

The computation of the last term of equation (2) is tricky, time-consuming and unstable, and, to our knowledge, all authors have resigned to ignore it.

1.4. Previous work on scene flow estimation

Three-dimensional motion estimation from multiple video sequences has long been limited to rigid or piecewise-rigid scenes or parametric models. The problem of computing a dense non-rigid three-dimensional motion field from multiple video sequences has been addressed only recently. Two types of methods prevail in the scene flow literature.

The first family of methods [39, 4, 21] relies on the spatio-temporal derivatives of the input images. As pointed out in [36], estimating the 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 [1]. In [4, 21], several samples of the spatio-temporal derivatives are combined in order to overconstrain the scene flow, whereas in [39], the aperture problem is solved by complementing 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, these differential methods apply mainly to slowly-moving Lambertian scenes under constant illumination.

In the second family of methods [36, 39], the optical flow is computed independently in each camera, then these estimations are combined to get the scene flow. This approach is not optimal since it disregards the consistency between the different corresponding optical flows. Moreover, the noise in the different optical flows and the bias introduced by the heuristic spatial smoothness constraints alter the scene flow in an unpredictable manner.

1.5. Motivations of our approach

In this article, we propose a common variational framework for multi-view complete stereovision and scene flow estimation which overcomes most of the limitations listed above. 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 [33] 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.

Contrarily to existing deformable surfaces approaches, which compute a matching measure independently at each surface point and integrate these quantities on the surface, or on the image domain as in [29], our approach computes a global image-based matching score between the input images and the predicted images. The matching process fully handles projective distortion and partial occlusions. Neighborhood as well as global intensity in-

formation can be exploited to improve the robustness to appearance changes, without any approximation of shape, motion or visibility.

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 *correlation ratio* [24], *mutual information* [37], 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. Also, any user-defined regularity constraint can be used.

Our method computes global matching scores on entire images from which projective distortion and semi-occluded regions have been removed, thereby avoiding the complex machinery usually needed to handle many matching windows of different shapes, or many tessellations of the tangent plane. The pixels used in the computation of the matching score are exactly the ones that are visible, judging from the current position of the surface. Moreover, the minimizing flow is much simpler than in [8, 13, 6, 9], in the sense that it only involves first-order derivatives of the matching score. This results in elegant and efficient algorithms.

Our scene flow method does not fall into the two existing categories. It works directly in 3D object space. It evolves a 3D vector field to register the input images captured at different times.

The rest of this article is organized as follows. In Section 2, we present our variational formulation of multi-view complete stereovision and non-rigid 3D motion estimation. In Section 3, we detail two particular similarity measures that can be used in our framework: normalized cross correlation and mutual information. Section 4 describes our implementation with level sets and graphics card hardware acceleration. Finally, in Section 5, 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 5). For example, in the case of stereovision, it corresponds to what is classically known as the reprojection error: 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 reprojected images will be to the corresponding input images, modulo noise, calibra-

tion errors, appearance changes and semi-occluded areas. This is the core principle of our approach. Although the expression "reprojection error" is more common in the stereovision literature, we use "prediction error" in the following, because it has the advantage of being relevant to both shape and motion estimation.

In our framework, both shape and motion estimation are formulated as a generic image registration task. This analogy is widely used in the context of rectified stereovision and optical flow. But it has fewer illustrations in multi-view stereo with arbitrary camera configurations, and it is definitely novel in the context of scene flow estimation. The registration task is entrusted to a global 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. Neighborhood as well as global intensity information can be used in this measure.

We incorporate the similarity measure and a regularization term in an energy functional. The regularization term is required to make the problem well-posed. It is application-specific. For example, it could be designed to preserve shape or motion discontinuities. Here we focus primarily on the design of the matching term and we propose a basic smoothing regularization term.

The exact minimization of our energy functionals is computationally unfeasible due to the huge number of unknowns. Indeed, *simulated annealing* is extremely slow in practice. The *graph cuts* method is a powerful energy minimization method which allows to find a global minimum or a strong local minimum of an energy. In the last few years, this method has been successfully applied to several problems in computer vision, including stereovision [16] and image segmentation [2]. However, it has a severe limitation: it cannot be applied to an arbitrary energy function [17], and, when applicable, is computationally expensive.

In our case, graph cuts are not applicable. A suboptimal strategy must be adopted. 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. We run the optimization on a series of smoothed and subsampled images.

2.1. STEREOVISION

In the following, let a surface $S \subset \mathbb{R}^3$ model the shape of the scene. We note $I_i : \Omega_i \subset \mathbb{R}^2 \to \mathbb{R}^d$ the image captured by camera i. In practice $d = 1$ for grayscale images or $d = 3$ for color images. The perspective projection performed by camera i is denoted by $\Pi_i : \mathbb{R}^3 \to \mathbb{R}^2$. Our method takes into account the visibility of the surface points. In the sequel, we will refer to S_i

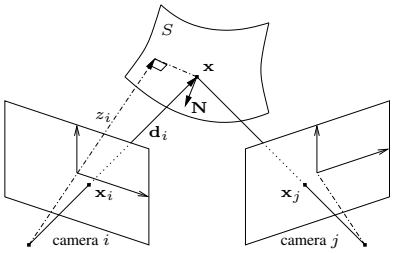

Figure 1. The camera setup and our notations.

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 reprojected 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:

$$\mathcal{M}(S) \;=\; \sum_i \sum_{j \neq i} \mathcal{M}_{ij}(S) \;, \tag{3}$$

$$\mathcal{M}_{ij}(S) \;=\; M|_{\Omega_i \cap \Pi_i(S_j)} \left(I_i \;,\; I_j \circ \Pi_j \circ \Pi_{i,S}^{-1} \right) \;. \tag{4}$$

Following [29], and for the reasons given in Subsection 1.3, we depart from the minimal surface approach. Our energy functional is the sum of a matching term computed in the images and of a user-defined regularization term. But our approach goes further than [29], in the sense that the matching process is global and completely image-based. In contrast, in [29], the matching measure was computed independently at each surface point, using an object-based tangent plane approximation, and later integrated on the image domain.

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. We neglect the variation related to visibility changes. Indeed, the latter would yield an additional term confined to the horizons of the surface in the different cameras. Hence, this term only has a codimension-two support and its influence would be considerably decreased by the regularity constraint. This technical assumption is commonly used in

the stereovision literature [8, 13, 6, 20]. Using the chain rule, we get

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

where $\mathbf{x}_i$ is the position in image i and $D.$ denotes the Jacobian matrix of a function. To guide the reader, we have indicated the dimensions of the different matrices appearing in the product.

When the surface moves, the reprojected 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 Section 3. Throughout this section, for sake of conciseness, we have omitted the images for which this derivative is evaluated. But the reader must be aware that the reprojected images, as well as the domains where the similarity measures are computed, change along the minimizing flow.

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

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

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

$$\left.\frac{\partial \mathcal{M}_{ij}(S+\epsilon\,\delta S)}{\partial \epsilon}\right|_{\epsilon=0} = -\int_{S_i \cap S_j} \left[\partial_2 M(\mathbf{x}_i) DI_j(\mathbf{x}_j) D\Pi_j(\mathbf{x}) \frac{\mathbf{d}_i}{z_i^3}\right] \left[\mathbf{N}^T \delta S(\mathbf{x})\right] d\mathbf{x} \ .$$

In other words, the gradient of the matching term is

$$\nabla \mathcal{M}_{ij}(S)(\mathbf{x}) = -\delta_{S_i \cap S_j}(\mathbf{x}) \left[\partial_2 M(\mathbf{x}_i) DI_j(\mathbf{x}_j) D\Pi_j(\mathbf{x}) \frac{\mathbf{d}_i}{z_i^3}\right] \mathbf{N} \ , \qquad (5)$$

where $\delta_.$ 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

$$\frac{\partial S}{\partial t} = \left[-\lambda\, H + \sum_i \sum_{j \neq i} \delta_{S_i \cap S_j}\, \partial_2 M\, DI_j\, D\Pi_j \frac{\mathbf{d}_i}{z_i^3} \right] \mathbf{N}\ , \tag{6}$$

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

2.2. SCENE FLOW

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 $v^t : S^t \to \mathbb{R}^3$ 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.

$$\mathcal{F}(v^t) \;=\; \sum_i \mathcal{F}_i(v^t)\ , \tag{7}$$

$$\mathcal{F}_i(v^t) \;=\; M\left(I_i^t\ ,\ I_i^{t+1} \circ \Pi_i \circ (\mathrm{Id} + v^t) \circ \Pi_{i,S^t}^{-1} \right)\ . \tag{8}$$

Its gradient writes

$$\begin{aligned} &\nabla^T \mathcal{F}_i(v^t)(\mathbf{x}) = \\ &-\delta_{S_i^t}(\mathbf{x})\, \frac{\mathbf{N}^T \mathbf{d}_i}{z_i^3}\, \underbrace{\partial_2 M(\mathbf{x}_i)}_{1\times d}\, \underbrace{DI_i^{t+1}\left(\Pi_i(\mathbf{x} + v^t(\mathbf{x}))\right)}_{d\times 2}\, \underbrace{D\Pi_i(\mathbf{x} + v^t(\mathbf{x}))}_{2\times 3}\ . \end{aligned} \tag{9}$$

Here, 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 intrinsic Laplacian, often called the Laplace-Beltrami operator. The evolution of the scene flow is then driven by

$$\frac{\partial v^t}{\partial \tau} = \mu\, \Delta_{S^t} v^t + \sum_i \delta_{S_i^t}\, \frac{\mathbf{N}^T \mathbf{d}_i}{z_i^3}\, \left[\partial_2 M\, DI_i^{t+1}\, D\Pi_i \right]^T\ , \tag{10}$$

where τ is the fictitious time of the minimization, Δ_{S^t} denotes the Laplace-Beltrami operator on the surface, and μ is a positive weighting factor.

3. Some Similarity Measures

In this section, we present two similarity measures that can be used in our framework: normalized cross correlation and mutual information [37]. 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 [11] 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.

Note that the shape of Ω can be very complex. In practice, it corresponds to the domain where both an input image and an associated prediction are defined. In other words, semi-occluded regions are discarded from Ω.

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:

$$\lim_{\epsilon \to 0} \frac{M(I_1, I_2 + \epsilon\, \delta I) - M(I_1, I_2)}{\epsilon} = \int_\Omega \partial_2 M(I_1, I_2)(\mathbf{x})\, \delta I(\mathbf{x})\, d\mathbf{x}\ . \quad (11)$$

3.1. Cross correlation

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 [26]. Some authors alleviate this problem by using adaptive windows [14, 25]. 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.

Thus, we gather neighborhood information using convolutions by a Gaussian kernel of standard deviation σ. The local mean, variance, covariance and cross correlation of the two images respectively write

$$\mu_i(\mathbf{x}) = \frac{G_\sigma \star I_i(\mathbf{x})}{\omega(\mathbf{x})}\ , \quad v_i(\mathbf{x}) = \frac{G_\sigma \star I_i^2(\mathbf{x})}{\omega(\mathbf{x})} - \mu_i^2(\mathbf{x}) + \beta^2\ ,$$

$$v_{1,2}(\mathbf{x}) = \frac{G_\sigma \star I_1 I_2(\mathbf{x})}{\omega(\mathbf{x})} - \mu_1(\mathbf{x})\,\mu_2(\mathbf{x}) \;, \quad cc(\mathbf{x}) = \frac{v_{1,2}(\mathbf{x})}{\sqrt{v_1(\mathbf{x}) v_2(\mathbf{x})}} \;,$$

where ω is a normalization function accounting for the shape of the domain: $\omega(\mathbf{x}) = \int_\Omega G_\sigma(\mathbf{x}-\mathbf{y})\,d\mathbf{y}$. The β constant prevents the denominator from being zero. Beyond its numerical usefulness, this constant has a rigorous justification, as shown in [11]. It is related to the Parzen Gaussian kernel used to estimate the local joint probability distribution of the two images [23].

We aggregate the opposite of the local cross correlation to get a similarity measure corresponding to our needs:

$$M^{CC}(I_1, I_2) = -\int_\Omega cc(\mathbf{x})\,d\mathbf{x} \;. \tag{12}$$

The minimizing flow involved by our method includes the derivative of the similarity measure with respect to the second image. In this case, it writes

$$\partial_2 M^{CC}(I_1, I_2)(\mathbf{x}) = \alpha(\mathbf{x})\, I_1(\mathbf{x}) + \beta(\mathbf{x})\, I_2(\mathbf{x}) + \gamma(\mathbf{x}) \;, \tag{13}$$

where

$$\alpha(\mathbf{x}) = G_\sigma \star \frac{-1}{\omega\sqrt{v_1\, v_2}}(\mathbf{x}) \;, \quad \beta(\mathbf{x}) = G_\sigma \star \frac{cc}{\omega\, v_2}(\mathbf{x}) \;,$$

$$\gamma(\mathbf{x}) = G_\sigma \star \left(\frac{\mu_1}{\omega\sqrt{v_1\, v_2}} - \frac{\mu_2\, cc}{\omega\, v_2} \right)(\mathbf{x}) \;.$$

In practice, along the minimizing flow, the α, β, γ functions change slowly relative to I_1 and I_2. So, in our implementation, we update them only every ten iterations to reduce the computational burden.

3.2. MUTUAL INFORMATION

Mutual information is based on the joint probability distribution of the two images, estimated by the Parzen window method [23] with a Gaussian kernel of standard deviation β:

$$P(i_1, i_2) = \frac{1}{|\Omega|} \int_\Omega G_\beta\left(I_1(\mathbf{x}) - i_1\,,\, I_2(\mathbf{x}) - i_2\right)\, d\mathbf{x} \;. \tag{14}$$

We note P_1, P_2 the marginals:

$$P_1(i_1) = \int_{\mathbb{R}} P(i_1, i_2)\, di_2 \quad , \quad P_2(i_2) = \int_{\mathbb{R}} P(i_1, i_2)\, di_1 \;.$$

Our measure is the opposite of the mutual information of the two images:

$$M^{MI}(I_1, I_2) = -\int_{\mathbb{R}^2} P(i_1, i_2)\, \log \frac{P(i_1, i_2)}{P_1(i_1) P_2(i_2)}\, di_1\, di_2 \;. \tag{15}$$

Its derivative with respect to the second image writes [11, 7]:

$$\partial_2 M^{MI}(I_1, I_2)(\mathbf{x}) = \zeta(I_1(\mathbf{x}), I_2(\mathbf{x})) \ , \tag{16}$$

where

$$\zeta(i_1, i_2) = \frac{1}{|\Omega|} \, G_\beta \star \left(\frac{\partial_2 P}{P} - \frac{P_2'}{P_2} \right) (i_1, i_2) \ .$$

In our implementation, the ζ function is updated only every ten iterations.

4. Implementation Issues

We have implemented our method in the level set framework [22], 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 [27]. 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 (see *http://www.gpgpu.org*).

The only parameters inherent to our framework are the regularization coefficients λ and μ in equations (6) and (10). This being said, the similarity measure embedded in our method may have its own parameters: the size of the correlation window for cross correlation, the standard deviation of the Parzen kernel for mutual information, etc.

In all the experiments of Section 5, we have used a matching window with a standard deviation of 2 pixels ($\sigma = 2$) for cross correlation, and a Parzen kernel of variance 10 ($\beta^2 = 10$) for both cross correlation and mutual information.

5. Experimental Results

5.1. STEREOVISION

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

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

Name	#Images	Image size	#Pairs	Measure	Level set size	Time (sec.)
Hervé	2	512×512	2	MI	128^3	107
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	128x128x256	1831

We have used either cross correlation (CC) or mutual information (MI), with $\sigma = 2$ and $\beta^2 = 10$. 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.

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. The number of camera pairs used in each experiment is given in Table I.

The number of iterations is 600 for all datasets. However, in most of our experiments, the convergence is attained earlier, so the computation time could be reduced using an appropriate stopping criterion. The only exception is the "Hervé" dataset, where the rear part of the face, not visible from any of the two cameras, is only driven by a mean curvature motion, and has not yet converged after 600 iterations, causing a rounded shape behind the face instead of a join of minimal area.

In all our experiments, the regularizer is a mean curvature motion, and the initial surface is an approximate bounding box of the scene. Although this initial guess is very far from the objects, we manage to converge to the expected shape and to recover its concavities thanks to the coarse-to-fine strategy. We use four levels in the multi-resolution pyramid. So the level set size at the coarser resolution is 16^3 for most datasets and 16x16x32 for the "Bust" dataset.

We show our results in Figures 2, 3, 4, 5 and 6. For each dataset, we display some of the input images, the ground truth when available, then some views of

Figure 2. "Hervé" stereo pair and our results.

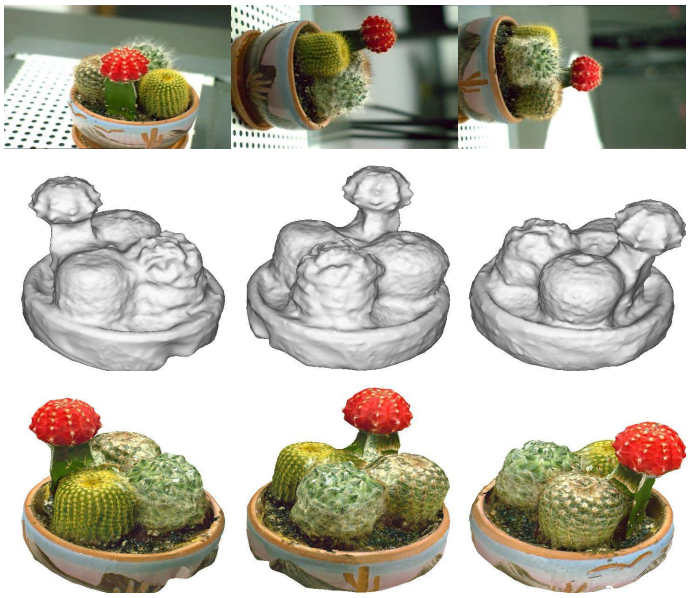

Figure 3. Some images from the "Cactus" dataset and our results.

the estimated shape, and finally the same views after reprojecting the texture coming from the most front-facing camera. Note that this texture-mapped representation does not aim at photorealism. In particular, it generates artifacts at the places where the source of the texture changes. It is only intended to show the validity of the output of our method for more sophisticated image-based rendering techniques.

Figure 4. Some images from the "Gargoyle" dataset and our results.

In all our experiments, the overall shape of the objects is successfully recovered, and a lot of details are captured: the eyes and the mouth of "Hervé", 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 saturated in some images. In Figure 7, we illustrate the multi-resolution evolution of the surface for the "Bust" dataset, starting from a coarse bounding box.

Finally, in Table II, we compare our results with the non-Lambertian stereovision method of [13] on the "Buddha" and the "Bust" datasets. We adopt the same shape error measure than in their work: the ratio between the volume of the symmetric difference between the estimated shape and the true shape and the volume of the true shape. The errors on the "Buddha" dataset are comparable. Our method performs significantly better than [13] on the "Bust" dataset. Moreover, visually, our reconstructions are slightly

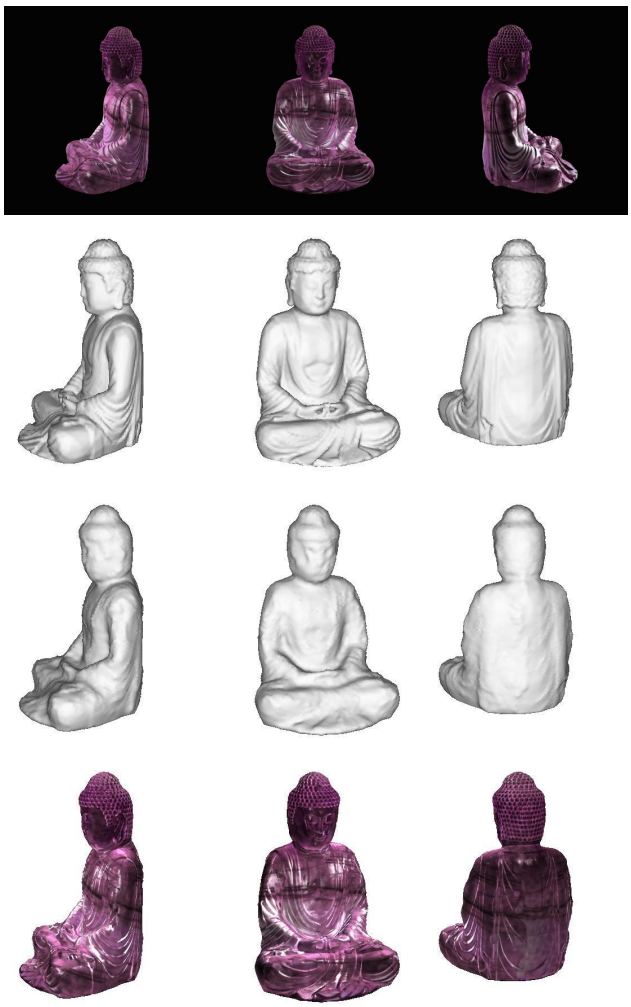

Figure 5. Some images from the "Buddha" dataset, ground truth and our results.

Table II. Quantitative comparison between our method and the non-Lambertian stereovision method of [13].

Method	Error on "Buddha"	Error on "Bust"
[13]	3.5 %	5.7 %
Our method	4.0 %	3.0 %

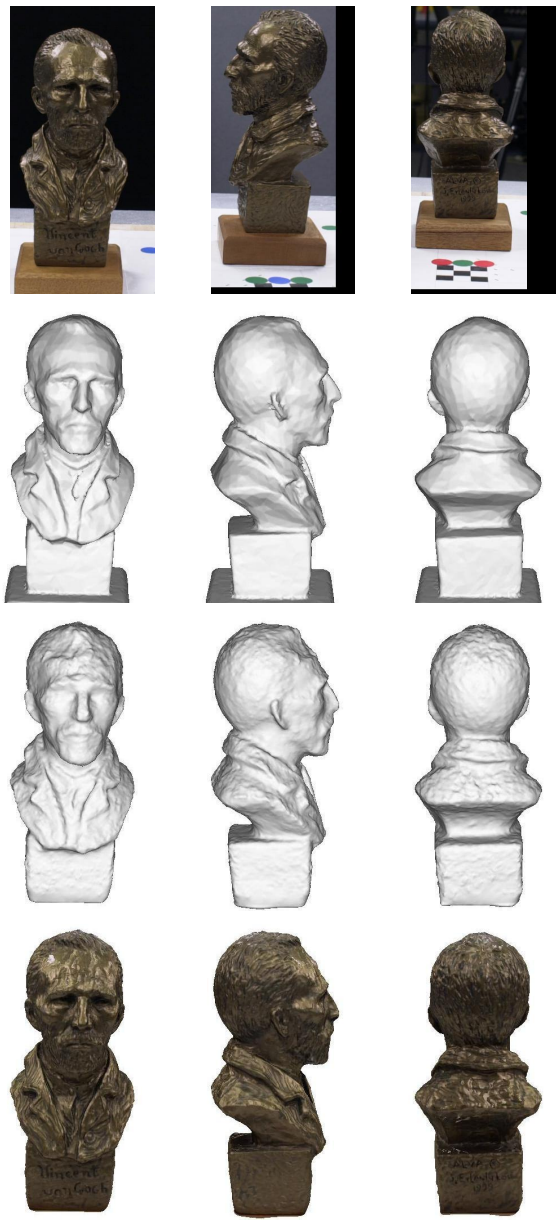

Figure 6. Some images from the “Bust” dataset, pseudo ground truth and our results.

more detailed. But above all, our computation time is considerably smaller. It does not exceed thirty minutes on a 2 GHz Pentium IV PC under Linux, versus several hours.

5.2. 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 Dr. Patrick Baker and Dr. Jan Neumann (University of Maryland) for benchmark purposes. This sequence shows a character (Pr. Yiannis Aloimonos) 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. We also compute the backward motion from $t + 1$ to t for the purpose of time interpolation.

In this experiment, we use cross correlation with the value of the parameters given in Section 4. The level set size is 128^3 and the number of levels of the multi-resolution pyramid is 4.

Figure 8 displays the first four frames of one of the input sequence and our estimation of shape and 3D forward 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.

We use our results to generate time-interpolated 3D sequences of the scene. To synthesize images at intermediate time instants, we can either use the previous shape and texture warped by the forward motion, or the next shape and texture warped by the backward motion. Ideally the two should coincide exactly, but of course this is never the case in practice. As a consequence, we linearly interpolate between forward and backward extrapolated images to guarantee a smooth blending between frames. In return it causes "crossfading" artifacts in some places where forward and backward extrapolation significantly diverge.

We display a short excerpt of such a time-interpolated sequence in Figure 9. Note the progressive opening and closing of the mouth. Please see the accompanying video and the *Odyssée Lab* web page for more results.

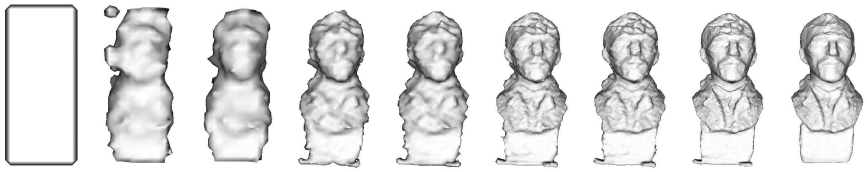

Figure 7. Multi-resolution shape evolution for the “Bust” dataset.

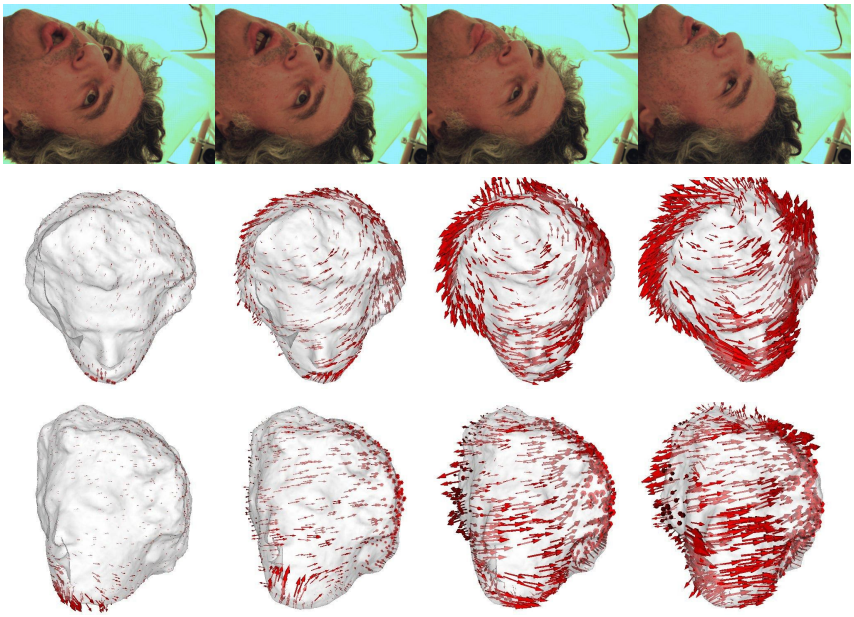

Figure 8. First images of one sequence of the “Yiannis” dataset and our results.

Figure 9. An excerpt of the time-interpolated 3D sequence for the “Yiannis” dataset.

6. Conclusion and Future Work

We have presented a novel method for multi-view stereovision and scene flow estimation which minimizes the prediction error using a global image-based matching score. We adequately warp the input views and we register the resulting distortion-free images with a user-defined image similarity measure, which can include neighborhood and global intensity information. No approximation of shape, motion or visibility is made in the matching process.

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 and we have synthesized time-interpolated 3D sequences.

Our future work includes a hardware implementation of our stereovision method with graphics processor units to further reduce the computation time, and the fusion of shape and motion estimations in order to exploit their redundancy.

Acknowledgements

We would like to thank Pr. Kyros Kutulakos for providing us with the "Cactus" and "Gargoyle" datasets, and Dr. Jan Neumann for his support on the "Yiannis" dataset.

References

1. Barron, J., D. Fleet, and S. Beauchemin: 1994, 'Performance of Optical Flow Techniques'. *The International Journal of Computer Vision* **12**(1), 43–77.
2. Boykov, Y. and V. Kolmogorov: 2003, 'Computing Geodesics and Minimal Surfaces via Graph Cuts'. In: *International Conference on Computer Vision*, Vol. 1. pp. 26–33.
3. Broadhurst, A., T. Drummond, and R. Cipolla: 2001, 'A Probabilistic Framework for Space Carving'. In: *International Conference on Computer Vision*, Vol. 1. pp. 388–393.
4. Carceroni, R. and K. Kutulakos: 2002, '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.
5. Caselles, V., R. Kimmel, and G. Sapiro: 1997, 'Geodesic Active Contours'. *The International Journal of Computer Vision* **22**(1), 61–79.
6. Duan, Y., L. Yang, H. Qin, and D. Samaras: 2004, 'Shape Reconstruction from 3D and 2D Data using PDE-Based Deformable Surfaces'. In: *European Conference on Computer Vision*, Vol. 3. pp. 238–251.
7. Faugeras, O. and G. Hermosillo: 2004, 'Well-posedness of two non-rigid multimodal image registration methods'. *Siam Journal of Applied Mathematics* **64**(5), 1550–1587.

8. Faugeras, O. and R. Keriven: 1998, 'Variational Principles, Surface Evolution, PDE's, Level Set Methods and the Stereo Problem'. *IEEE Transactions on Image Processing* **7**(3), 336–344.
9. Goldlücke, B. and M. Magnor: 2004a, 'Space-time Isosurface Evolution for Temporally Coherent 3D Reconstruction'. In: *International Conference on Computer Vision and Pattern Recognition*, Vol. 1. pp. 350–355.
10. Goldlücke, B. and M. Magnor: 2004b, 'Weighted Minimal Hypersurfaces and Their Applications in Computer Vision'. In: *European Conference on Computer Vision*, Vol. 2. pp. 366–378.
11. Hermosillo, G., C. Chefd'hotel, and O. Faugeras: 2002, 'Variational methods for multimodal image matching'. *The International Journal of Computer Vision* **50**(3), 329–343.
12. Hernández Esteban, C. and F. Schmitt: 2004, 'Silhouette and Stereo Fusion for 3D Object Modeling'. *Computer Vision and Image Understanding* **96**(3), 367–392.
13. Jin, H., S. Soatto, and A. Yezzi: 2005, 'Multi-view Stereo Reconstruction of Dense Shape and Complex Appearance'. *The International Journal of Computer Vision* **63**(3), 175–189.
14. Kanade, T. and M. Okutomi: 1994, 'A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **16**(9), 920–932.
15. Kass, M., A. Witkin, and D. Terzopoulos: 1987, 'Snakes: Active Contour Models'. *The International Journal of Computer Vision* **1**(4), 321–331.
16. Kolmogorov, V. and R. Zabih: 2002, 'Multi-camera Scene Reconstruction via Graph Cuts'. In: *European Conference on Computer Vision*, Vol. 3. pp. 82–96.
17. Kolmogorov, V. and R. Zabih: 2004, 'What Energy Functions Can Be Minimized via Graph Cuts?'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(2), 147–159.
18. Kutulakos, K.: 2000, 'Approximate N-View Stereo'. In: *European Conference on Computer Vision*, Vol. 1. pp. 67–83.
19. Kutulakos, K. and S. Seitz: 2000, 'A Theory of Shape by Space Carving'. *The International Journal of Computer Vision* **38**(3), 199–218.
20. Lhuillier, M. and L. Quan: 2003, 'Surface Reconstruction by Integrating 3D and 2D Data of Multiple Views'. In: *International Conference on Computer Vision*, Vol. 2. pp. 1313–1320.
21. Neumann, J. and Y. Aloimonos: 2002, 'Spatio-Temporal Stereo Using Multi-Resolution Subdivision Surfaces'. *The International Journal of Computer Vision* **47**, 181–193.
22. Osher, S. and J. Sethian: 1988, 'Fronts Propagating with Curvature-Dependent Speed: Algorithms Based on Hamilton–Jacobi Formulations'. *Journal of Computational Physics* **79**(1), 12–49.
23. Parzen, E.: 1962, 'On Estimation of a Probability Density Function and Mode'. *Annals Mathematical Statistics* **33**, 1065–1076.
24. Roche, A., G. Malandain, X. Pennec, and N. Ayache: 1998, 'The correlation ratio as new similarity metric for multimodal image registration'. In: W. Wells, A. Colchester, and S. Delp (eds.): *Medical Image Computing and Computer-Assisted Intervention-MICCAI'98*. Cambridge, MA, USA, pp. 1115–1124.
25. Scharstein, D. and R. Szeliski: 1998, 'Stereo Matching with Nonlinear Diffusion'. *The International Journal of Computer Vision* **28**(2), 155–174.
26. Scharstein, D. and R. Szeliski: 2002, 'A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms'. *The International Journal of Computer Vision* **47**(1), 7–42.

27. Segal, M., C. Korobkin, R. van Widenfelt, J. Foran, and P. Haeberli: 1992, 'Fast Shadows and Lighting Effects Using Texture Mapping'. *Computer Graphics* **26**(2), 249–252.
28. Seitz, S. and C. Dyer: 1999, 'Photorealistic Scene Reconstruction by Voxel Coloring'. *The International Journal of Computer Vision* **35**(2), 151–173.
29. Soatto, S., A. Yezzi, and H. Jin: 2003, 'Tales of Shape and Radiance in Multi-View Stereo'. In: *International Conference on Computer Vision*, Vol. 2. pp. 974–981.
30. Solem, J. and N. Overgaard: 2005, 'A Geometric Formulation of Gradient Descent for Variational Problems with Moving Surfaces'. In: *International Conference on Scale Space and PDE Methods in Computer Vision*.
31. Strecha, C., R. Fransens, and L. Van Gool: 2004, 'Wide-baseline Stereo from Multiple Views: a Probabilistic Account'. In: *International Conference on Computer Vision and Pattern Recognition*, Vol. 2. pp. 552–559.
32. Strecha, C., T. Tuytelaars, and L. Van Gool: 2003, 'Dense Matching of Multiple Wide-Baseline Views'. In: *International Conference on Computer Vision*, Vol. 2. pp. 1194–1201.
33. Szeliski, R.: 1999, 'Prediction Error as a Quality Metric for Motion and Stereo'. In: *International Conference on Computer Vision*, Vol. 2. pp. 781–788.
34. Treuille, A., A. Hertzmann, and S. Seitz: 2004, 'Example-Based Stereo with General BRDFs'. In: *European Conference on Computer Vision*, Vol. 2. pp. 457–469.
35. Vedula, S., S. Baker, and T. Kanade: 2002, 'Spatio-Temporal View Interpolation'. In: *ACM Eurographics Workshop on Rendering*. pp. 65–76.
36. Vedula, S., S. Baker, P. Rander, R. Collins, and T. Kanade: 2005, 'Three-Dimensional Scene Flow'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(3), 475–480.
37. Viola, P. and W. M. Wells III: 1997, 'Alignement by Maximization of Mutual Information'. *The International Journal of Computer Vision* **24**(2), 137–154.
38. Yang, R., M. Pollefeys, and G. Welch: 2003, 'Dealing with Textureless Regions and Specular Highlights: A Progressive Space Carving Scheme Using a Novel Photo-consistency Measure'. In: *International Conference on Computer Vision*, Vol. 1. pp. 576–584.
39. Zhang, Y. and C. Kambhamettu: 2001, 'On 3D Scene Flow and Structure Estimation'. In: *International Conference on Computer Vision and Pattern Recognition*, Vol. 2. pp. 778–785.