Motion – Stereo Integration for Depth Estimation

Christoph Strecha[1] and Luc Van Gool[1]

KU Leuven ESAT/PSI,
B-3001 Leuven, Belgium
{Christoph.Strecha, Luc.VanGool}@esat.kuleuven.ac.be
http://www.esat.kuleuven.ac.be/psi/visics

Abstract. Depth extraction with a mobile stereo system is described. The stereo setup is precalibrated, but the system extracts its own motion. Emphasis lies on the integration of the motion and stereo cues. It is guided by the relative confidence that the system has in these cues. This weighing is fine-grained in that it is determined for every pixel at every iteration. Reliable information spreads fast at the expense of less reliable data, both in terms of spatial communication and in terms of exchange between cues. The resulting system can handle large displacements, depth discontinuities and occlusions. Experimental results corroborate the viability of the approach.

1 Introduction

Stereo and shape-from-motion are among the dominant methods for 3D shape reconstruction. Both have been studied extensively in computer vision, but mostly separately. Nevertheless, the two methods have much in common and an integration could follow rather naturally while bringing added value, such as less sensitivity to occlusions and more robust convergence.

Similar as in the case of separate stereo or motion analysis, integrated approaches have used discrete features (e.g. points [1,2] or line segments [3]) or dense correspondences (e.g. based on correlation windows [4], spatio-temporal image gradients [5] or MRFs [6]). The method proposed here belongs to the latter category, as our extension to the PDE approach for optic flow by Proesmans *et al.* [7] yields dense correspondences. This approach was an interesting point of departure as it can handle large disparities and as it detects occlusions and flow discontinuities. 'Large disparities' may also be relatively small in absolute terms in that traditional optic flow methods tend to fail as soon as motions are large compared to the granularity of scene texture. Our integrated approach reinforces these advantages further, in contrast to earlier integrated approaches which either neglect the detection of occlusions or are limited to small motion displacements.

In our approach the occlusion detection is in fact part of a correspondence quality estimation scheme, that determines the relative influences of the stereo and motion cues. For instance, in cases where one but not the other suffers from

A. Heyden et al. (Eds.): ECCV 2002, LNCS 2351, pp. 170–185, 2002.

occlusion, the one least affected will take the upper hand. But this weighing scheme is reaches farther than depth discontinuity and occlusion detection, in that it guides the relative influences of both cues at every iteration and at every pixel during the evolution towards the solution. This sets our approach apart from earlier motion-stereo integration work.

Our method assumes a calibrated stereo rig (both the relative position of the cameras and their internal parameters are known beforehand), that is moved around with an unknown motion in a static environment. This motion is determined by the system. The parameters of the stereo rig are assumed to remain known during the motion. In our experiments, they have been kept fixed. If needed, one can do away with the rig calibration by using self-calibration methods for mobile stereo rigs [8]. The paper describes the basic integration procedure, where two subsequent stereo views are taken, i.e. four images in total. The extension to a whole stereo video can easily be made.

Of course, the combined use of multiple views for 3D reconstruction is not new. Bundle adjustment is a well-established technique to achieve exactly that. It improves both the 3D reconstruction and the camera calibration by exploiting the data provided by multiple cameras. However, whereas bundle adjustment basically takes the given feature correspondences for granted and tries to explain away visible deviations through adaptations to both camera parameters and 3D structure, the proposed approach is geared towards updating and coupling the correspondences themselves based on all the image data. It therefore plays a complementary role and acts at an earlier stage of the 3D reconstruction process.

The paper is organized as follows. Section 2 introduces some notations and explain the calibration of the stereo rig. Section 3 describes preparatory steps for the search of corresponding points within the single cues. Emphasis is on the introduction of a depth related parameterisation that facilitates the combination of information over multiple images and between cues. Section 4 discusses the integration of stereo and motion into a single scheme to extract depth and rig motion. Section 5 gives experimental results. Finally, section 6 concludes the paper.

2 The Stereo Setup

Our experimental setting consists of two video cameras mounted on a stereo rig at a distance of approximately 0.2 m. We choose the left camera center to be the Euclidean coordinate center. A 3D point denoted by $\mathbf{X} = (X, Y, Z, 1)^T$ is projected to left image coordinates $\mathbf{x}_l = (x_l, y_l, 1)^T$ and right image coordinates $\mathbf{x}_r = (x_r, y_r, 1)^T$ through[1] :

$$\lambda_l \mathbf{x}_l = \mathbf{K}_l[\mathbf{I}|0]\mathbf{X} \tag{1}$$

$$\lambda_r \mathbf{x}_r = \mathbf{K}_r[\mathbf{R}^T| - \mathbf{R}^T\mathbf{t}]\mathbf{X} \tag{2}$$

[1] In the following we will use the vector sign to describe non-homogeneous pixel coordinates $\boldsymbol{x}_{l,r} = (x_{l,r}, y_{l,r})$

where $\mathbf{K}_l$ and $\mathbf{K}_r$ denote the left and right camera matrices, $\mathbf{R}$ is the 3×3 rotation matrix specifying the relative orientation of the cameras and $\mathbf{t} = (t_x, t_y, t_z)^T$ is the translation vector between the two cameras. The camera calibration matrices $\mathbf{K}$ are described by:

$$\mathbf{K} = \begin{pmatrix} f & s & x_0 \\ 0 & af & y_0 \\ 0 & 0 & 1 \end{pmatrix} \tag{3}$$

where f denotes the focal length, (x_0, y_0) is the principal point, s is related to the pixel skew and a is the aspect ratio. When dealing with real cameras the above projection equation (1) will be perturbed by radial distortion of the lens. This can be modeled by the following equation [9]:

$$\begin{aligned} \boldsymbol{x}_u &= \boldsymbol{x}_d + \hat{\boldsymbol{x}}_d(k_1||\hat{\boldsymbol{x}}_d||^2 + k_2||\hat{\boldsymbol{x}}_d||^4) \\ \hat{\boldsymbol{x}}_d &= \boldsymbol{x}_d - \boldsymbol{c} \end{aligned} \tag{4}$$

where $\boldsymbol{x}_u$ is the unmeasurable undisturbed 2D point on the image plane and $\boldsymbol{x}_d$ the measured distorted point. We allow in this model the center of distortion $\boldsymbol{c}$ to be different from the image center. In our experiments the radial distortion was estimated and corrected ($\boldsymbol{x} = \boldsymbol{x}_u$).

Calibrating our system consists of extracting the internal camera parameters for each camera and the geometric relation between the two cameras. For the calibration we used a calibration box with 140 circles with known relative 3D coordinates. The center of each circle was computed by an ellipse fit to the Canny edge map. Assuming ellipses and using their centers is reasonable, since the circles are small compared to the image size (so that the ellipse centers correspond well to the projected circle centers and the radial distortion can be neglected). We used a stereo pair of the calibration box as well as one image of the box for each camera separately, with the box completely filling the whole image in the latter case. The last two images were added to get a better result for the radial distortion. For these calibration images we have 42 unknown parameters: the rotation and translation of the calibration box with respect to the first camera and the rotation + translation between the two cameras (12), the rotations and translations of the box when taking the separate, per camera images (12), 4 λ parameters, one calibration matrix per camera (10), and finally 2 radial distortion parameters per camera (4). We solve this nonlinear system by a Levenberg-Marquardt minimization.

3 Single Cue Correspondence Search

PDE based methods have been shown to give good results for stereo and optical flow correspondence search [10,7,11,12,13]. They are in general not dependent on preprocessing stages (feature point or line segment extraction) and provide directly a dense correspondence map. We use the PDE based approach proposed

by Proesmans *et al.* [7]. They propose a system of 6 coupled, non-linear diffusion equations that in effect yield not only the disparities but also discontinuity maps indicating depth discontinuities as well as parts visible to only one of the cameras (occlusions). The high number of equations is due to the symmetric exploitation of the two images: the system embarks on a simultaneous 1st-to-2nd and 2nd-to-1st image correspondence search (in the sequel referred to as forward and backward schemes). Another feature is the 'bi-local' nature of the differential computations, i.e. spatial and temporal derivatives at two different positions in the two images are combined. This has to do with the division of the disparity or motion displacement into a current estimate plus a residue, which gradually declines during subsequent iterations. The current estimate yields an offset between the points at which derivatives are taken in the two images. Working with residues allows to better linearize the problem for large disparities, an argument with the assumptions underlying the optical flow constraint equation. A similar strategy is followed here. In that respect our approach differs from others [5,6].

PDE methods often are expensive in terms of computation time, and may get stuck in some local minimum. Both problems can be reduced. Weickert *et al.* [14] used semi-implicit discretisation schemes to lower the computational effort. Multi-resolution techniques can help to find the global optimum [15,14]. The change to a semi-implicit and multi-resolution implementation of the original equations by Proesmans *et al.* [7] is a first modification to their system that we propose. A second is the use of epipolar geometry. In the original system epipolar geometry was not used. It is only along these lines that correspondence search proceeds, reducing the number of PDE's for correspondence search between single image pairs from 6 to 4.

The choice of a common parameter value for all corresponding points in the four images, where the paramtererisation runs along the epipolar lines and is directly related to depth, facilitates our integration of stereo and motion.

Depth parameterisation for stereo related views. We assume that the translation and rotation between the stereo cameras is known from precalibration. From eqs. (1, 2) the epipolar lines for the two cameras can be derived. It follows for corresponding image points $\mathbf{x}_l = (x_l, y_l, 1)^T$ and $\mathbf{x}_r = (x_r, y_r, 1)^T$:

$$\frac{\lambda_r}{Z}\mathbf{x}_r = \mathbf{K}_r\tilde{\mathbf{R}}\mathbf{K}_l^{-1}\mathbf{x}_l + \frac{\mathbf{K}_r\tilde{\mathbf{t}}}{Z}, \tag{5}$$

with $\tilde{t} = -\mathbf{R}^T\mathbf{t}$ and $\tilde{\mathbf{R}} = \mathbf{R}^T$. The stereo correspondence is divided into a component that depends on the rotation and pixel coordinate (according to the homography $\mathbf{H} = \mathbf{K}_r\tilde{\mathbf{R}}\mathbf{K}_l^{-1}$) and a depth dependent part that scales with the amount of translation between the cameras. For the left image the parameterisation of the corresponding point along the epipolar line in the right image is realized by $\boldsymbol{l}^r(\boldsymbol{x}_l, d)$ starting from the point $Z = \infty$ and going in the direction of the epipole $\mathbf{K}_r\tilde{\mathbf{t}}$ (see fig. 1). Points along the epipolar lines will be parameterised

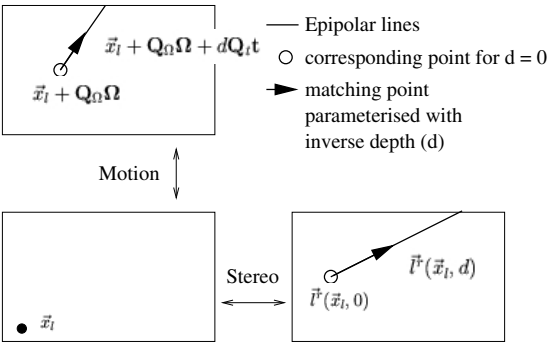

Fig. 1. *Parameterisation of the correspondence for a point $\boldsymbol{x}_l$ in the left image at the first time instance for stereo and motion related views*

with $d = \frac{1}{Z}$, i.e. a *depth related parameter*. In pixel coordinates this gives:

$$\boldsymbol{l}^r(\boldsymbol{x}_l, d) = \frac{\begin{pmatrix}\mathbf{H}[1]\mathbf{x}_l \\ \mathbf{H}[2]\mathbf{x}_l\end{pmatrix} + d\begin{pmatrix}\mathbf{K}_r[1]\tilde{\mathbf{t}} \\ \mathbf{K}_r[2]\tilde{\mathbf{t}}\end{pmatrix}}{\mathbf{H}[3]\mathbf{x}_l + dt_z} \tag{6}$$

$\mathbf{H}[i]$ is the 3-vector for the i th row of the homography $\mathbf{H}$ and similarly for $\mathbf{K}_r[i]$. This parameterisation differs from the one advocated by Alvarez *et al.* [12], which is less directly coupled to depth.

Depth parameterisation for motion related views. To make use of the epipolar constraint that is also present in motion related views (we assume a static environment) the rig motion has to be estimated. For that purpose we have extended the instantaneous motion model described in [16,17] to the general case of non-square and skewed pixels. Similar calculations provide us with the 6 motion parameters $\mathbf{p} = (\mathbf{\Omega}, \mathbf{t})^T = (\Omega_x, \Omega_y, \Omega_z, t_x, t_y, t_z)^T$ where the first 3 represent the rotation angles about the corresponding axes and the last 3 the translation components along these axes. The extraction of these motion parameters from motion correspondences is discussed in the next paragraph. The displacement for each pixel can then be expressed as a function of the depth related parameter d and the motion parameters:

$$\boldsymbol{u} = \mathbf{Q}_\Omega\mathbf{\Omega} + d\mathbf{Q}_t\mathbf{t} = \mathbf{Q}\mathbf{p} \tag{7}$$

$$\mathbf{Q}_\Omega = \begin{pmatrix} \frac{\tilde{x}\tilde{y}}{af} + s & -f - \frac{\tilde{x}^2}{f} + \frac{s\tilde{x}\tilde{y}}{af^2} & \frac{\tilde{y}}{a} - \frac{s\tilde{x}}{f} + \frac{s^2\tilde{y}}{af^2} \\ af + \frac{\tilde{y}^2}{af} & -\frac{\tilde{x}\tilde{y}}{f} + \frac{s\tilde{y}^2}{af^2} & -\tilde{x} + \frac{s\tilde{y}}{f} \end{pmatrix}, \mathbf{Q}_t = \begin{pmatrix} -f & -s & \tilde{x} \\ 0 & -af & \tilde{y} \end{pmatrix} \tag{8}$$

where $\tilde{x} = x - x_0$ and $\tilde{y} = y - y_0$ are the centered image coordinates. Fig. 1 shows this parameterisation, which is very similar to that of stereo. In the following we describe the motion correspondence by $\boldsymbol{m}(\boldsymbol{x}, d)$. For a pixel $\boldsymbol{x}_l \in I_1^l$ the corresponding pixel in image I_2^l (see fig. 1,2) parameterised by d is:

$$\boldsymbol{m}_2^l(\boldsymbol{x}_l, d) = \boldsymbol{x}_l + \mathbf{Q}_\Omega\mathbf{\Omega} + d\mathbf{Q}_t\mathbf{t} \tag{9}$$

Camera motion estimation. Following eq. (7), the motion displacements (correspondences) of the pixels and the motion parameters $\mathbf{p}$ can be extracted by an iterative process. Suppose we already have an estimate of the displacement vector $\boldsymbol{u}_0$. In the same vein as the offset + residual description of optical flow by Proesmans *et al.* [7], we split the displacement in a current estimate and a residual $\boldsymbol{u}_r = \boldsymbol{u} - \boldsymbol{u}_0$. The introduction of $\boldsymbol{u}_0$ is especially important for large displacements, for which we could not otherwise truncate the $O(\boldsymbol{u}^2)$ terms from the Taylor expansion of $\boldsymbol{u}$. Assuming Lambertian surfaces and hence identical intensities for corresponding pixels, we have

$$I_1(\boldsymbol{x}) = I_2(\boldsymbol{x} + \boldsymbol{u}_0 + \boldsymbol{u}_r) = I_2(\boldsymbol{x} + \boldsymbol{u}_0) + \frac{\partial I_2(\boldsymbol{x} + \boldsymbol{u}_0)}{\partial \boldsymbol{x}} \boldsymbol{u}_r \tag{10}$$

Setting $\boldsymbol{u}_r = \mathbf{Q}\mathbf{p}_r = \mathbf{Q}\mathbf{p} - \boldsymbol{u}_0$ and assuming that $\mathbf{Q}$ is known (it is because the depths at each pixel (d_0) obtain a value at each iteration in our final system), we look for those $\mathbf{p}$ that minimize the integral:

$$\int\limits_{\boldsymbol{x} \in \Omega} \left(I_1(\boldsymbol{x}) - I_2(\boldsymbol{x} + \boldsymbol{u}_0) - \frac{\partial I_2(\boldsymbol{x} + \boldsymbol{u}_0)}{\partial \boldsymbol{x}} (\mathbf{Q}\mathbf{p} - \boldsymbol{u}_0) \right)^2 d\boldsymbol{x}$$

This yields

$$\begin{aligned} &\mathbf{A}\mathbf{p} = \mathbf{b} \\ &\mathbf{A} = \sum_{\boldsymbol{x} \in \Omega} \mathbf{Q}^T I_{2x}^T I_{2x} \mathbf{Q}, \quad \mathbf{b} = \sum_{\boldsymbol{x} \in \Omega} \mathbf{Q}^T I_{2x}^T \left(I_1(\boldsymbol{x}) - I_2(\boldsymbol{x} + \boldsymbol{u}_0) + I_{2x} \boldsymbol{u}_0 \right) \end{aligned} \tag{11}$$

where

$$I_{2x} = \frac{\partial I_2(\boldsymbol{x} + \boldsymbol{u}_0)}{\partial \boldsymbol{x}} \tag{12}$$

and where we sum over some image domain Ω. This will later include all pixels with sufficiently good confidence scores for their correspondences.

4 Integration of Motion and Stereo

Our integration of stereo and motion takes the form of a single system of equations. This system yields dense scene depths, their discontinuities and occlusions, and the motion of the stereo rig. The evolving functions for inverse depth and motion are initialized to zero.

For the given image configuration (see figure 2) pixels in an image can typically obtain a depth through both stereo and motion based reconstruction. Exceptions are those that fall victim to occlusions in one or both of these cues. These two depths ought to be the same, and hence corresponding pixels in the two other images are expected to obtain equal epipolar parameter values (d for stereo = d for motion). Armed with these expected equalities (see fig 2) and exploiting the stereo and the motion companion views, the inverse depth d is

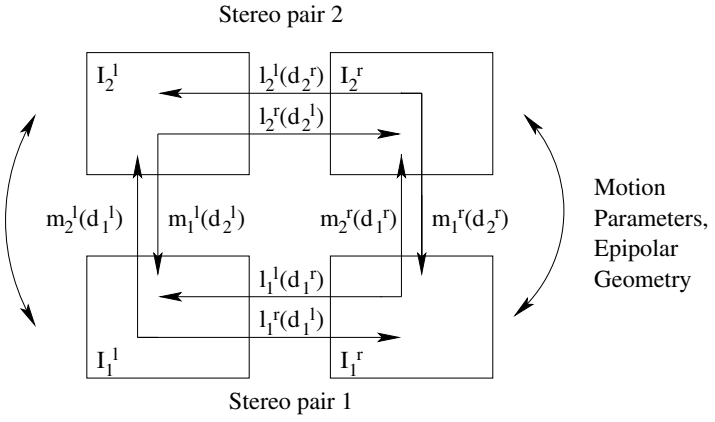

Fig. 2. *Stereo disparity relations: $l(d)$, $m(d)$ are the stereo and motion correspondences eq. 6 and 9*

solved through the iteration of a system of coupled, non-linear diffusion equations, à la Proesmans *et al.* [7]. As a matter of fact, such a system is solved for each of the four images, but their systems are coupled. In the case of image I_1^l the system takes the following form (at each pixel):

$$\begin{aligned}
\frac{\partial d_1^l}{\partial t} &= div(\delta(c^{s,m})\nabla d_1^l) \\
&- \frac{\gamma(c^s)}{\gamma(c^s)+\gamma(c^m)}\lambda I_{\boldsymbol{x}}^s(I_{\boldsymbol{x}}^s(d_1^l - d_{10}^l) + I_t^s) \\
&- \frac{\gamma(c^m)}{\gamma(c^s)+\gamma(c^m)}\lambda I_{\boldsymbol{x}}^m(I_{\boldsymbol{x}}^m(d_1^l - d_{10}^l) + I_t^m) \\
\frac{\partial c^s}{\partial t} &= \rho\nabla^2 c^s + 2\alpha(1-c^s)|C^s| \\
\frac{\partial c^m}{\partial t} &= \rho\nabla^2 c^m + 2\alpha(1-c^m)|C^m|
\end{aligned} \tag{13}$$

As mentioned, these are variations on Proesmans *et al.*'s [7] equations and the reader is referred to that reference for a detailed description. The superscripts $()^{m,s}$ are related to the motion or stereo pair, resp. The expressions I_t^m, I_t^s replace the temporal derivative of intensity in the traditional optical flow constraint, an adaptation due to the formulation in terms of residual motions, $I_{\boldsymbol{x}}^m$, $I_{\boldsymbol{x}}^s$ are spatial derivatives of intensity, taken here along the epipolar lines. The definitions for $\boldsymbol{x}_l \in I_1^l$ are:

$$\begin{aligned}
&I_t^m = I_1^l(\boldsymbol{x}_l) - I_2^l(\boldsymbol{m}_2^l(\boldsymbol{x}_l, d_{10}^l)), I_t^s = I_1^l(\boldsymbol{x}_l) - I_1^r(\boldsymbol{l}_1^r(\boldsymbol{x}_l, d_{10}^l)) \\
&I_{\boldsymbol{x}}^m = \frac{\partial I_2^l(\boldsymbol{m}_2^l(\boldsymbol{x}_l, d_{10}^l))}{\partial d}, I_{\boldsymbol{x}}^s = \frac{\partial I_1^r(\boldsymbol{l}_1^r(\boldsymbol{x}_l, d_{10}^l))}{\partial d},
\end{aligned} \tag{14}$$

Fig. 3. *Two stereo pairs (left and right images) at the first (bottom (I_1^l, I_1^r)) and the second (top (I_2^l, I_2^r)) time instance*

where $\boldsymbol{l}_1^r(\boldsymbol{x}_l, d_{10}^l)$ and $\boldsymbol{m}_2^l(\boldsymbol{x}_l, d_{10}^l)$ are the positions of the corresponding points for stereo and motion using the previous estimate d_{10}^l of the parameter value. Every pixel in image I_1^l also has confidence measures c^s and c^m for its stereo and motion correspondences, resp.. Values close to 0 mean high confidence, values close to 1 low confidence.

The first equation governs the evolution of the depth related parameter d_1^l. The first term is an anisotropic diffusion term. It blocks diffusion (smoothing) from places with a lower confidence in their correspondences. In a typical iteration process most places start out with low confidences, but at the end low confidence values (cs close to 1) tend to cluster near discontinuities and occlusions. More explanation is given later. The second and third terms impose optical flow constraints on the stereo and motion correspondences, resp. Similarly as with Proesmans *et al.*'s system, the formulation is in terms of residual displacements rather than complete displacements. The rationale is as said: the correspondence search becomes more amenable to linearization. The two terms are weighted with factors. The weight of the second term increases with the confidence in the stereo correspondence of that pixel, giving more importance to the depth ($1/d$) suggested by the stereo cue. The third term acts similarly, but for the motion cue. The function γ in these weighting factors is given by:

$$\gamma(c) = \exp\left(-\frac{c^2}{k}\right) \quad . \tag{15}$$

They are high for high confidences (i.e. low $c^{m,s}$). The relative confidences put in the stereo vs. the motion correspondence guide the relative influence that these cues get in the correspondence search. These relative values can change over time (over iterations) and from one pixel to the next.

The evolution of the confidence measures $c^{m,s}$ in the second and third equations is driven by the vectors C^m and C^s. They measure the difference between forward and backward flow in the stereo and motion direction, respectively. In the normal case of a pair of corresponding points, the extracted displacements for the first with respect to the second – forward flow – and the second with respect to the first – backward flow – are equal but of opposite sign. Hence, when summed the two displacements will cancel each other out. The C's represent these vector sums. Large $|C|$ yield values of c that are close to one – its maximal value – whereas at other places c tends to zero. Hence, c actually quantifies the inconsistency between the forward and backward flows. Its restriction to the interval $c \in [0, 1]$ is realized by the factor $(1 - c)$. The interested reader may note that we actually simplified the 2nd and 3rd equation with respect to those proposed by Proesmans *et al.* [7] without loss of performance. A separate inconsistency measure c^m and c^s is extracted for the motion and the stereo pair, resp. The use of this forward-backward regularization scheme is - different from usual optical flow techniques - important here. Since it is not only used to block smoothing but also to weight stereo and motion according to it's ability to extract depth especially near occlusion. However, other regularizers could be added with small contribution.

Although inconsistency values are calculated for all pixels separately, their surroundings matter as the 2nd and 3rd equations also contain a diffusion term. Referring to the first equation, its anisotropic diffusion coefficient $\delta(c^{m,s})$ is controlled by the inconsistencies (c^m, c^s). The role of this coefficient is to prevent smoothing across depth edges or incorrect displacements being spread towards neighbouring pixels. Since the corresponding pixel can be occluded in one neighboring image but not in the other, information exchange for the image pair with good correspondences should persist. This is realized by taking the maximum of $\gamma(c^m)$ and $\gamma(c^s)$. Subsequently, these maximal values are normalized as to sum to one over the 4 neighbors of a pixel and the result is called $\delta(c^{s,m})$. The reason behind this normalization is to make sure that diffusion does not stop completely in all directions simultaneously, as this can keep the system from evolving towards the solution at the early stages.

The overall system of equations that is solved, consists of the forementioned 3 equations for each of the four images. Additional to this an update of the motion parameters $\mathbf{p}_{l,r}$ for the left and right camera is done after each iteration. For the calculation we combine forward and backward motion since they are related by $(\mathbf{R}^{12}, \mathbf{t}^{12}) = (\mathbf{R}^{21^T}, -\mathbf{R}^{21^T}\mathbf{t}^{21})$. This is reflected by the following relation

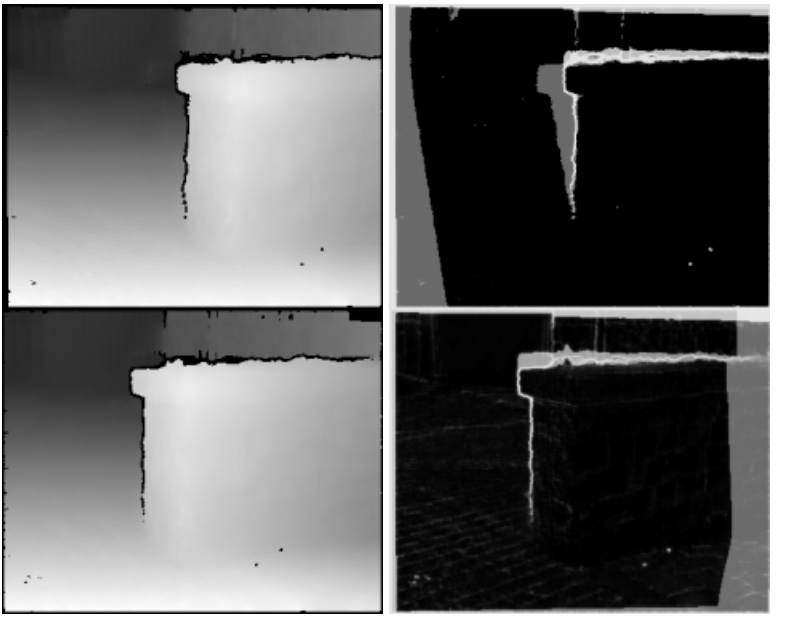

Fig. 4. *Left: Depth map (black pixels on the edge of the little wall indicate sharp discontinuities); Right: Inconsistency, hight confidence pixels in black, low confidence pixels in white (discontinuities). Dark gray corresponds to low confidence regions for stereo only (occlusions), light gray for motion only*

between the instantaneous motion vectors:

$$\begin{pmatrix} \Omega_x^{12} \\ \Omega_y^{12} \\ \Omega_z^{12} \\ t_x^{12} \\ t_y^{12} \\ t_z^{12} \end{pmatrix} = \mathbf{G} \begin{pmatrix} \Omega_x^{21} \\ \Omega_y^{21} \\ \Omega_z^{21} \\ t_x^{21} \\ t_y^{21} \\ t_z^{21} \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & -\Omega_z^{12} & \Omega_y^{12} \\ 0 & 0 & 0 & \Omega_z^{12} & -1 & -\Omega_x^{12} \\ 0 & 0 & 0 & -\Omega_y^{12} & \Omega_x^{12} & -1 \end{pmatrix} . \tag{16}$$

Using eq. (11), the common motion vector $\mathbf{p}^{12}$ is estimated from:

$$\begin{bmatrix} \mathbf{A}^{12} \\ \mathbf{A}^{21}\mathbf{G} \end{bmatrix}_{6\times 12} \mathbf{p}^{12} = \begin{bmatrix} \mathbf{b}^{12} \\ \mathbf{b}^{21} \end{bmatrix}_{1\times 12} \tag{17}$$

The input to this (remember eq. 12 and eq. 8) contains the current estimate of the inverse depth. As described by eq. 11 we sum over appropriate domains. These contain all pixels with confidence values c^s *and* c^m that are below some threshold (0.15 in our experiments). In other words, only pixels with high confidence for both (motion and stereo) correspondences contribute to the extraction of the rig's motion.

A coupling between the left and right motion parameters is therefor realised by the coupling of the depth values. Additional constraints between left and

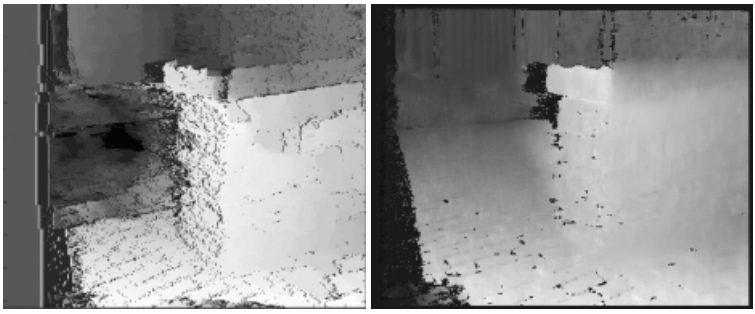

Fig. 5. *Pure stereo results; Left: depth map for the dynamic path algorithm; Right: depth map for left the algorithm described in section 3, black pixels have low confidence (e.q. occlusion).*

right camera motions can be imposed in the case of a fixed stereo rig. We intend to relax the condition that the stereo rig has to be fixed, only requiring it to be calibrated at the start and, hence, have not used this additional constraint at this point. The total system consists of 14 equations. First, for the four images the 3 equations (13) yielding the depths and inconsistency measures (discontinuities and occlusions) are iterated one in turn (12 equation). Then, with the resulting update for the depth values, a new estimate of the cameras motion is calculated(2 equations of type 17). This results in new epipolar constraints, that are fed back to the 12 equations for the next iteration. As mentioned we used Weickert's [14] semi-implicit discretisation for the sake of fast convergence.

5 Experimental Results

We tested the method on real images taken by two Sony DCR-TRV900E cameras with an image size of (360×288). We discovered a serious radial distortion and corrected for it in a preprocessing stage. The inverse depths and the motion vector (d and $\mathbf{p}$) were initialized with zero and estimated in a coarse-to-fine manner over 6 pyramid levels. Figure 3 shows the 4 input images.

For the sake of comparison, fig 5 gives the results for the determination of depth based solely on the initial stereo pair. The image on the left is the depth map obtained from correspondences obtained with our dynamic path search algorithm [18]. The middle and right images show the results of the forward and backward schemes of our modified Proesmans *et al.* [7] algorithm. It yields better results than the other method, but still shows some gaps and part of the depth discontinuities remain undetected.

The results for the left and right cameras of the integrated system are shown in figure 4. One can see on the left the inverse depth maps (d_1^l (top), d_1^r (bottom)) for the images I_1^l and I_1^r and on the right the corresponding inconsistency maps c^s, c^m for these images (confidences, with bright meaning low and dark high). Comparing these results with the results from single stereo (see figure 5), we can

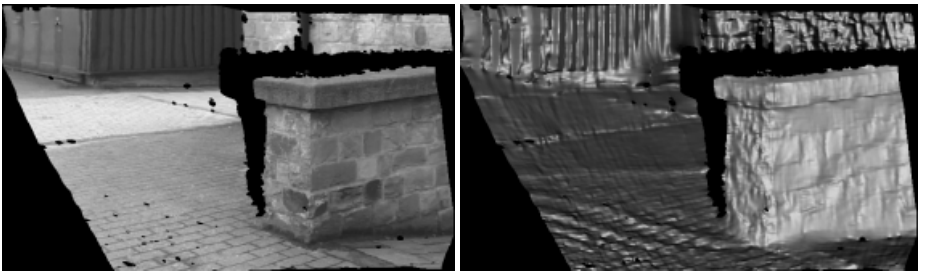

Fig. 6. *3D model from image I_1^l; left: model right with added texture*

conclude that the system is better able to assign good depth values to pixels that have no correspondence in the stereo partner image. The left part of the scene and the occluding parts behind the wall got their depth from the upward motion of the left camera. The occlusions to the motion / stereo partner are marked (light gray / dark gray) in the inconsistency maps of fig. 4. The same can be seen in the depth map for the right image I_1^r: the occluded part on the right of fig. 3 got its depths from the motion system. In conclusion, depth is recovered of more points than in the single stereo case. This is not surprising as we used more images. Nevertheless, this result shows the success of this motion-stereo integration scheme, as a system that is able to decide whether stereo or motion cues should be used for 3D estimation.

A second observation is that the integrated system yields better defined depth discontinuities (bright values in the inconsistency map on the right). Moreover, where disparities get large, pure stereo (fig. 5) had problems with local minima. This can be seen at the lower left part of the image, where the recurrent stone texture contains a lot of possibilities for wrong matches, and also near the top right part of the wall, with hardly any texture in the direction of the epipolar lines (nearly horizontal). These problems are alleviated by the integrated approach since the epipolar lines for both cues are typically not parallel and therefore contain complementary information. Pixels in white in fig. 4 are those where we have discontinuities with respect to both (stereo and motion) correspondences. These points form the anisotropic diffusion coefficient $\delta(c^{m,s})$.

Fig. 6 and 7 show views of the 3D model. In the last figures we added the simplified ground truth for the scene (measured with a ruler). The depth discontinuities in the 3D model correspond to the discontinuities $\delta(c^{m,s})$. The discontinuities are an explicit part of the output and are not extracted *post factum* through edge detection in the depth map, as e.g. in [5]. They are of a better quality that way, certainly in scenes like this one with several planes almost orthogonal to the image planes. All these planes yield high gradients in the depth map and tend to yield many spurious 'discontinuities' through depth edge detection. The 3D reconstruction is quite precise. For instance, the stone structures of the small wall close to the cameras come out well. This precision points at

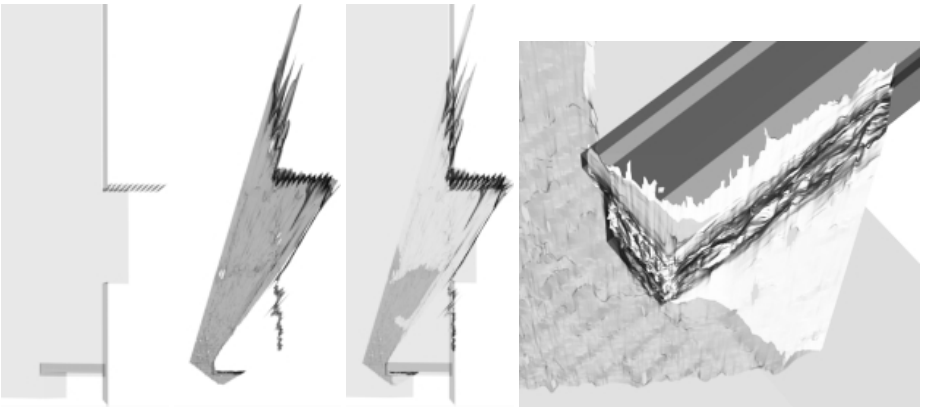

Fig. 7. *Views of the 3D model from image I_1^l; ground truth top view, model top view, combined ground truth and data; right: zoom on the wall in front with added ground truth*

the subpixel accuracy of the correspondences. For this stereo rig at a distance of 2m a match ± one pixel corresponds to a depth change of ± 2.5cm. We did not retrieve such detail when all correspondences were retrieved with our dynamic path search algorithm [18] The whole scene also lines up well with the ground truth. However, there are still some mistakes in the reconstruction. Some pixels at the back wall appear to be more to the front than they should. These pixels didn't have a correspondence in the stereo partner. Another mistake is the connection between the wall on the right and the back wall. There the motion inconsistency measure c^m was not able to find the discontinuity (see also fig. 4). The result was obtained after 10 minutes (image size 360×288, PC 700 MHz).

In fig. 8 we show the result for an indoor scene. By comparing the result with the result from single stereo in fig. 8(bottom) for both time instances one can see that it would not be possible to achieve the result of the combined approach by combining the two depth maps from single stereo. Especially the part behind the table is wrong in both single stereo depth maps. This said, in this case also the combined approach could not prevent that some errors occur. The depth discontinuity between the table cloth and the floor remains largely undetected. This is due to the fundamental problem that the floor is essentially untextured.

The computation time for this example was less than 4 minutes. This time again included the computation on all pyramid levels and the extraction of the full depth, inconsistencies and motion parameters for all 4 images up to image size 160×120.

6 Summary and Conclusions

We have proposed a scheme to integrate depth extraction from stereo and motion. A precalibrated stereo rig was moved with an unknown motion. The point

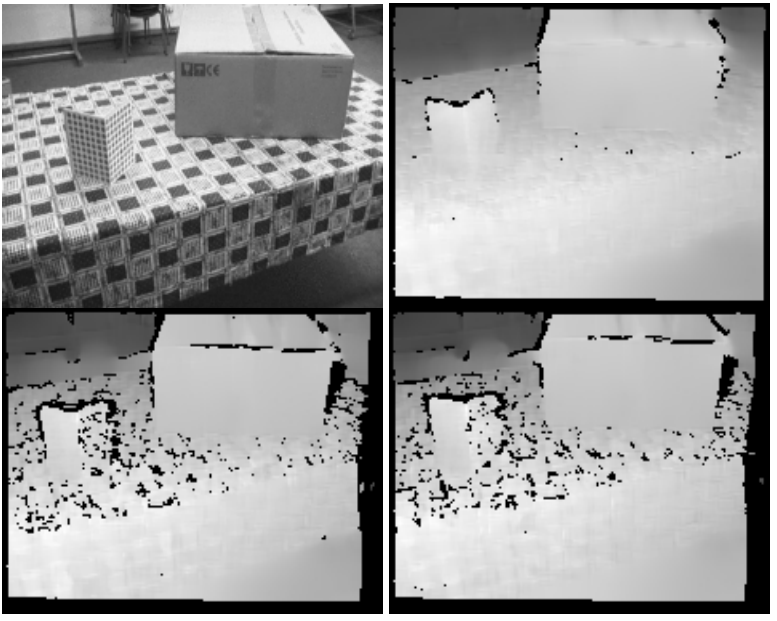

Fig. 8. *Top right: Depth map with the proposed approach; Bottom: Depth maps for the indoor scene from single stereo at first (left) and second (right) time instance*

of departure of our integration was a PDE scheme for the extraction of correspondence between pairs of images, as introduced by Proesmans *et al.* [7]. It has the advantage that it detects discontinuities in the disparities or motion fields as well as occlusions. It can also deal with large displacements and it yields good precision.

We have modified this scheme in a number of important ways. First, we have shortened the processing time substantially by using Weickert *et al.*'s [14] semi-implicit discretisation, by restricting the search along epipolar lines, and by applying a multi-resolution approach. This reduces the time from hours to minutes. The latter alteration also resulted in higher robustness, e.g. against ambiguities in the case of periodic textures. Secondly, we have adapted this scheme for the integration of motion and stereo. On the one hand, we have introduced a very direct coupling that was guided by the dynamic, relative weighing of both schemes at every pixel and at every iteration. This allows our method to really get the best from both cues. On the other hand, we have used inverse depth as a common parameter for all correspondences, which facilitated the direct combination of matches coming from the different cues and which directly yields the valuable depth of points.

We plan to extend this work in a number of directions. Stereo videos rather than two subsequent stereo pairs will be processed. Static scenes will be replaced by scenes with independently moving objects. It is there that having stereo vision really pays off, as the scene is always static as far as the stereo image pairs are

concerned (one moment in time). The stereo rig will be made variable, so that the vergence and focal lengths can be changed. Finally, we plan to improve the correspondence search under variable lighting conditions and in the presence of specular reflections.

Acknowledgment. The authors gratefully acknowledge support by K.U.Leuven GOA project 'VHS+' and EU IST project 'CogViSys'.

References

[1] F.Dornaika and R.Chung: Stereo correspondence from motion correspondence. *CVPR*, vol. 1, pp. 70–75, 1999.
[2] P.K.Ho and R.Chung: Stereo-motion with stereo and motion in complement? *PAMI*, vol. 22, no. 2, pp. 215–220, 2000.
[3] Z.Zhang, Q.T.Luong, and O.Faugeras: Motion of an uncalibrated stereo rig: Self-calibration and metric reconstruction. *IEEE Trans. Robotics and Automation*, vol. 12, no. 1, pp. 103–113, feb 1996.
[4] R.Mandelbaum, G.Salgian, and H.Sawhney: Correlationbased estimation of ego-motion and structure from motion and stereo. *ICCV*, pp. 544–550, 1999.
[5] G.P.Stein and A.Shashua: Direct estimation of motion and extended scene structure from a moving stereo rig. *CVPR*, pp. 211–218, 1998.
[6] G.Sudhir, S.Banerjee, R.Bahl, and K.Biswas: A cooperative integration of stereopsis and optic flow computation. *J. Opt. Soc. Am. A*, vol. 12, pp. 2564, 1995.
[7] M.Proesmans, L.Van Gool, E.Pauwels, and A.Oosterlinck: Determination of optical flow and its discontinuities using non-linear diffusion. *ECCV*, vol. 2, pp. 295–304, 1994.
[8] A.Zisserman, P.A.Beardsley, and I.D.Reid: Metric calibration of a stereo rig. *In Proc. IEEE Workshop on Representation of Visual Scenes*, pp. 93–100, 1995.
[9] G.P.Stein: Lens distortion calibration using point correspondences. *CVPR*, pp. 143–148, 1997.
[10] J.Shah: A nonlinear diffusion model for discontinuous disparity and half-occlusions in stereo. *CVPR*, pp. 34–40, 1993.
[11] M.Proesmans, E.Pauwels, and L.Van Gool: *in Geometry-driven diffusion in computer vision*, ed. Bart M. ter Haar Romeny Kluwer, 1994.
[12] L.Alvarez, R.Deriche, J.Sanchez, and J.Weickert: Dense disparity map estimation respecting image discontinuities: pde and scalespace based approach. Tech. Rep. RR-3874, INRIA, 2000.
[13] J.Weickert and C.Schnörr: Variational optic flow computation with a spatio-temporal smoothness constraint. Tech. Rep. 15, University of Mannheim, 2000.
[14] J.Weickert, B.M.ter Haar Romeny, and M.A.Viergever: Efficent and reliable schemes for nonlinear diffusion filtering. *IEEE Transactions on Image Processing*, vol. 7, no. 3, pp. 398–410, 1998.
[15] D.Marr and T.Poggio: A theory of human steropsis. *Proc. Royal Soc. B*, vol. 204, pp. 301–328, 1979.
[16] D.J.Heeger and A.D.Jepson: Subspace methods for recovering rigid motion I:algorithm and implemenation. *IJCV*, vol. 7, no. 2, pp. 95–117, 1992.
[17] M.Irani: Multi-frame optical flow estimation using subspace constraints. *ICCV*, pp. 626–633, 1999.

[18] G.Van Meerbergen, M.Vergauwen, M.Pollefeys, and L.Van Gool: A hierarchical stereo algorithm using dynamic programming. *IEEE Workshop on Stereo and Multi-Baseline Vision*, pp. 166–174, 2001.