Generation of 3D Dense Depth Maps by Dynamic Vision

An Underwater Application*

José Santos-Victor
João Sentieiro
E-mail : d2760@beta.ist.rccn.pt
CAPS/ISR - Instituto Superior Técnico
Av. Rovisco Pais 1, 1096 Lisboa Codex, PORTUGAL

Abstract

This paper presents a dynamic 3D Vision system that is able to estimate dense depth maps from an image sequence. The depth maps computed at each time instant are used in an Extended Kalman filtering structure, that integrates all depth measurements over time, reducing uncertainty. Results with images acquired by an underwater camera, are presented.

1 Introduction

During the last decade, an increasing interest has been devoted to research and development activities in the area of autonomous systems, capable of performing complex tasks in unknown environments, without human intervention. These systems must be able to build and maintain internal models of the observed world. Fields of application range from mobile robots, to autonomous vehicles, space robots, AUVs (autonomous underwater vehicles), etc...

In this paper we address the problem of depth extraction by using an image sequence acquired by a moving camera (with known motion), in an unknown environment. A stochastic approach is adopted to model the existence of uncertainty in every depth estimate, and well established techniques, like Kalman filtering, are used to reduce the uncertainty of the estimated depth maps, over time.

This approach differs from the one presented in [1], both in the matching stage, where geometric constraints arising from the camera motion are included, and in the filtering stage, where a different formulation of the state space model for the motion/observation equations, leading to a clearer formulation of the kalman filtering structure.

The 3D Vision System described in this paper, comprises three major processing modules: the matching process, the regularization procedure and the Kalman filtering stage.

*This work has been supported in the context of the MOBIUS project, of the EEC MArine Science Technology (MAST) programme. The authors wish to thank Thomson CSF-LER and Thomson Sintra ASM, for providing the images for the underwater application, and Prof. Takeo Kanade for the valuable comments made on this work.

BMVC 1992 doi:10.5244/C.6.14

In the matching process, a correlation-like method, based on the *Sum of Squared Differences* method (SSD) [1, 2], is used. To improve the disparity estimate, prior knowledge of the camera motion is considered, as a geometric constraint (epipolar line), and sub-pixel resolution is achieved by interpolating the SSD error surface.

Noting the ill-posed nature of the matching process, a regularization stage is formulated, to constrain the desired disparity field to smooth solutions. In this way, the noise levels of the disparity estimates can be reduced, and new estimates can be calculated to fill in the areas, wherever the matcher was unable to produce estimates.

Kalman filtering is used in the final stage of this work, to integrate multiple disparity measurements over time, in order to produce a more reliable depth estimate [1, 3].

By integrating, over time, several measurements, one can benefit from the advantages of having closely spaced image view points (which simplifies the matching problem, but usually degrades the depth estimates precision) without jeopardizing the precision of the depth estimates computation.

In the following sections, the models involved in the depth from motion algorithms, are described. Results obtained with a sequence of synthesized images and with real underwater images are presented. Finally, we draw some conclusions and some future directions of research are pointed out.

2 Model

This section is devoted to the study and description of all the models involved in the 3D dynamic vision system. First, the dynamics associated to the position of a 3D point relatively to a moving camera referential is derived. Then, the camera model is introduced, and used to obtain the equations that describe the apparent motion induced in the image plane. Finally a model for the uncertainty affecting the system, is established.

Consider a camera moving with relation to a fixed point in the space. Let {C} be a cartesian coordinate system (frame) attached to the camera. Let $\vec{\omega}$ and $\vec{T}$, be the angular and translational velocities of the camera with relation to a fixed referential, and let $\vec{P}$ be the position vector of a fixed point in the space. The velocity of P with relation to {C} (rigid body motion) is described by the following differential equation [4]:

$$\frac{d\vec{P}}{dt} = -\vec{T} - \vec{\omega} \times \vec{P} \tag{1}$$

To determine how this motion is perceived in the image plane, we have used a *pinhole* camera model [4, 6]. According to this model, the location of a given image pixel, (x, y), is determined by the perspective projection in the image plane, of the corresponding 3D point, $[X\ Y\ Z]^T$:

$$x = \frac{X}{Z} \qquad y = \frac{Y}{Z} \tag{2}$$

By using the camera model in equation (1), and eliminating Z, we finally obtain a new equation set, that expresses the apparent motion (velocity field)

induced in the image plane by the actual camera motion [3, 5, 6][1]:

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \frac{1}{Z}\begin{bmatrix} -1 & 0 & x \\ 0 & -1 & y \end{bmatrix}\vec{T} + \begin{bmatrix} xy & -(1+x^2) & y \\ (1+y^2) & -xy & -x \end{bmatrix}\vec{\omega} \quad (3)$$

$$\dot{Z_{(t)}} = (-\omega_y x + \omega_x y)Z_{(t)} - T_z \quad (4)$$

Uncertainty (the simplified camera model, errors associated to the camera motion and parameters, the image acquisition process, etc) is modeled by additive white gaussian noise in equations (3,4), and the final model is obtained by approximating derivatives in equation (3, 4) with forward differences (with sampling period τ): [2]

State equation :

$$Z_{[t+\tau]} = a_{[t]}\, Z_{[t]} + b_{[t]} + \eta \quad (5)$$

Observer equation (measurements):

$$\vec{\mathbf{d}} = \begin{bmatrix} x_{[t+\tau]} - x_{[t]} \\ y_{[t+\tau]} - y_{[t]} \end{bmatrix} = \mathbf{C}_{[t]}\frac{1}{Z_{[t]}} + \mathbf{D}_{[t]} + \mu \quad (6)$$

where η is a zero mean gaussian random variable with variance r, μ is a zero mean gaussian random vector with covariance matrix Q and $\{\mu, \eta\}$ are independent. The terms $a_{[t]}$, $b_{[t]}$, $\mathbf{C}_{[t]}$ and $\mathbf{D}_{[t]}$ depend on the motion parameters and the image plane coordinates of a given pixel [7].

3 System Description

The complete system structure is shown in figure 1. At every sample instant, a new image is acquired by the moving camera and a new depth map is computed. A matching algorithm is applied to each pair of successive images, to compute the disparity vector field and uncertainty estimates. This vector field is the input to a regularization stage, used to decrease the noise levels and fill in unavailable estimates. Every new observation (regularized disparity vector) is finally used in a Kalman filtering module to update an estimated depth map resultant from previous measurements, thus reducing the uncertainty over time.

3.1 Matcher

To determine the displacement of each image pixel, induced by the camera motion in the static environment, we have used a correlation based technique, derived from the Sum of Squared Differences (SSD) method [2, 3].

To evaluate potential match candidates, p'_t and $p'_{t+\tau}$, from images acquired at time t and $t+\tau$, it is assumed that homologous points have similar gray-levels. Hence, the sum of the squared gray level differences, for pixels inside

[1] For simplicity, we have not explicitly written the different variables as time functions (e.g. $x(t)$ instead of x), as it should be clear from the context.

[2] Note that the image coordinates in equation (6) correspond to a virtual camera and the actual image pixel coordinates are obtained using the *intrinsic* camera model parameters [4]. See [7] for details.

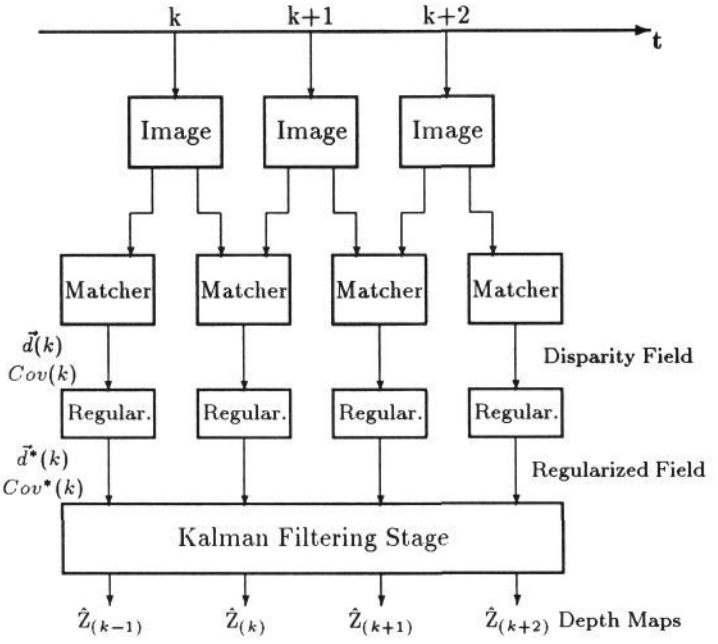

Figure 1: Block Diagram of the Vision System

windows centered in p'_t and $p'_{t+\tau}$, can be used to quantify the "likelihood" of a matching decision [1].

Moreover, the location of a pixel $p'_{t+\tau}$, homologous to p'_t, is constrained to a line, *the epipolar line*, in the image at $t+\tau$, determined by the camera motion and camera parameters [7]. By introducing the prior knowledge of the epipolar line, we define the *Extended Sum of Squared Differences* (ESSD) criterion that measures gray level compatibility and simultaneously penalizes deviations from the epipolar line:

$$ESSD(u,v,x,y) = \sum_{\beta}\sum_{\alpha} \phi_{(\alpha,\beta)} \left[I_{(t,\alpha,\beta)} - I_{(t+\tau,\alpha+u,\beta+v)}\right]^2 + \lambda_{ep}\, d^2_{ep}(x,y,u,v) \quad (7)$$

where $I_{(t,x,y)}$ designates the pixel (x,y) of the image acquired at time t, $d_{ep}(x,y,u,v)$ is the distance of the matching candidate pixel, $I_{(t+\tau,x+u,y+v)}$, to the epipolar line, and $\phi_{(\alpha,\beta)}$ is a weighting function. The contribution of the prior knowledge to the final cost functional[3] is quantified λ_{ep}.

In [1] the disparity vector is estimated by fitting a quadratic surface to a neighbourhood of the minimum SSD point, while in [3] two one dimensional quadratic curves are fit, both in the x and y directions. Due to the increased robustness to noise and simplicity, we adjust two one dimensional quadratic curves in each direction, in a neighbourhood of the minimum ESSD point:

$$q(u) = au^2 + bu + c \quad (8)$$

[3] Since the computation of the epipolar line is based on the camera and motion parameters, the value of λ_{ep} should be related to the uncertainty associated to these parameters.

where a, b and c are estimated from the data. The minimum point can then be determined, with sub pixel resolution, yielding the optimal disparity estimate:

$$\hat{u}_{opt} = -\frac{b}{2\,a} \tag{9}$$

whenever the coefficient a is different from 0.

The uncertainty of the estimate is related to the shape of the ESSD error surface [1, 2, 3]. In [2], the uncertainty is a function of the SSD surface curvature along the principal axis, and in [1] error propagation techniques in the SSD cost functional, were used. In each direction, we will approximate the estimate variance by [3]:

$$\sigma_u^2 = \left(\frac{d^2 q(u_{opt})}{du^2}\right)^{-2} q(u_{opt}) \tag{10}$$

The first term of (10) expresses the decrease of the uncertainty with the increase of the error surface curvature, while the second term is a normalizing factor dependent of the minimum ESSD surface value.

3.2 Regularization

Many visual reconstruction processes, aimed at recovering 3D information by processing 2D image data, are inverse, ill-posed problems (eg. the estimation of disparity fields between successive images) [8, 9].

Using the framework of regularization, ill-posed problems can be reformulated as variational principles, by introducing *a priori* knowledge about the solution. Standard Tikhonov regularization, uses stabilizing functionals to restrict the space of admissible solutions to smooth functions.

To determine a regularized function $\mathcal{U}$, using a set of data $\mathcal{D}$, we define an error function $\Psi_d(\mathcal{D},\mathcal{U})$, to measure the compatibility of the proposed solution and the data, a stabilizing functional $\Psi_p(\mathcal{U})$ that quantifies the smoothness conditions on the desired solution, and search for the $\mathcal{U}^*$ that minimizes the following cost criterion [10, 11].

$$\Psi(\mathcal{U},\mathcal{D}) = \Psi_d(\mathcal{U},\mathcal{D}) + \lambda\Psi_p(\mathcal{U}) \tag{11}$$

The choice for both functionals Ψ_p and Ψ_d guarantees that, under certain weak conditions, a solution for the optimization problem exists [2]. For the regularization of the displacement/disparity vector field, a *thin membrane* [2, 8, 10] stabilizing functional was used:

$$\Psi(\mathcal{U},\mathcal{D}) = \lambda\sum_{x,y}(\vec{u}-\vec{d})^T Q^{-1}(\vec{u}-\vec{d}) + \int\int trace\,\{(\nabla\vec{u})(\nabla\vec{u}^T)\}dxdy \tag{12}$$

$$Q = \begin{bmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_v^2 \end{bmatrix} \tag{13}$$

where $\vec{u}(x,y) = [u(x,y)\;\; v(x,y)]^T$ is the regularized solution, σ_u^2 and σ_v^2 are the variances of the x and y components of the measured disparity vectors, ∇

is the gradient operator and λ quantifies the relative weight of the fitness to data term, in the global cost functional.[4]

The domain of the surface $\vec{u}(x,y)$ is usually discretized using either the finite differences method or the finite element method [2, 4, 8]. Applying finite element analysis, as proposed by Terzopoulos [8], to the functionals involved in the minimization problem we obtain:

$$\Upsilon(\mathcal{U},\mathcal{D}) = \sum_{x,y} \lambda(\vec{u}-\vec{d})^T Q^{-1}(\vec{u}-\vec{d}) + \| \vec{u}_{(x+1,y)} - \vec{u}_{(x,y)} \|^2 + \| \vec{u}_{(x,y+1)} - \vec{u}_{(x,y)} \|^2 \quad (14)$$

To minimize (14), the Gauss-Seidel relaxation algorithm is used, and $\vec{u}$ is obtained iteratively at each point as a function of the values of its neighbours (similar equations are obtained for u and v):

$$u^{n+1}_{(x,y)} = \bar{u}^n_{(x,y)} + \frac{\lambda\sigma_u^{-2}}{1+\lambda\sigma_u^{-2}} (u^0_{(x,y)} - \bar{u}^n_{(x,y)}) \quad (15)$$

where $u^0_{(x,y)}$ is the measured x disparity at pixel (x,y) and $\bar{u}_{(x,y)}$ is the local mean value, given by:

$$\bar{u}_{(x,y)} = (u_{(x+1,y)} + u_{(x-1,y)} + u_{(x,y+1)} + u_{(x,y-1)})/4 \quad (16)$$

To determine the uncertainty associated to the regularized disparity field, the uncertainty values must be updated, as the regularization procedure imposes changes in the original field. Unfortunately, as the regularization iterations proceed, $\bar{u}^n$ will depend on an increasing number of samples, thus hardening the calculation of the uncertainty. For computation efficiency, we have adopted the following simpler expression:

$$var[u^{n+1}] = \frac{\alpha_m^2 var[\bar{u}^n] + \alpha_0^2 var[u^0]}{\alpha_m^2 + \alpha_0^2} \quad (17)$$

3.3 Kalman Filtering - Recursive Estimation of Depth Maps

We will now see how to use Kalman filtering theory, to combine different disparity measurements over time, yielding a more reliable single depth estimate.

Using the state space model framework, we can formulate an estimation problem, to determine the value of $Z_{[t]}$, based on the noisy observations $\vec{d}_{[t]}$. Depth is estimated independently at each pixel, while spatial dependencies are embodied in the regularization process.

This state estimation problem can be conveniently dealt with, using Kalman filtering techniques. Since the observation equation is non-linear in the state variable, the discrete-time Extended Kalman Filter (EKF) [12] must be used. The filtering equations comprise a prediction step and an update/filtering step. In the prediction stage, the future values of depth and associated uncertainty are predicted, based on past information and on the dynamic model:

[4] Whenever the measured variable $\vec{d}$ is unavailable, λ will be set to zero.

Prediction :

$$\hat{Z}(t/t-1) = a_{[t-1]}\hat{Z}(t-1/t-1) + b_{[t-1]} \quad (18)$$

$$\sigma^2_{Z(t/t-1)} = a^2_{[t-1]}\sigma^2_{Z(t-1/t-1)} + r \quad (19)$$

where $\hat{Z}(t/t-1)$ is the predicted depth value for time t, based on the data available up to time $t-1$, $\sigma^2_{Z(t/t-1)}$ is the corresponding variance and r is the variance of the motion equation noise (see Section 2).

At time t, a new disparity measurement is available, and the predicted depth can be updated. This is the filtering step procedure and the EKF uses a linearized version of the observation equation in a neighbourhood of the predicted value.

$$\vec{\mathbf{d}}_{[t]} \approx \vec{\mathbf{d}}^L_{[t]} = \mathbf{C}^L_{[t]} Z_{[t]} + \mathbf{D}^L_{[t]} + \mu \quad (20)$$

where $\mathbf{C}^L_{[t]}$, $\mathbf{D}^L_{[t]}$ are the coefficients of the linearized model [7]. The filtering step is given by

Filtering :

$$\mathbf{K}_t = \sigma^2_{Z(t/t-1)}(\mathbf{C}^L_{[t]})^T \left[\, \mathbf{C}^L_{[t]}\sigma^2_{Z(t/t-1)}(\mathbf{C}^L_{[t]})^T + \mathbf{Q}_t \,\right]^{-1} \quad (21)$$

$$\sigma^2_{Z(t/t)} = (1 - \mathbf{K}_t\mathbf{C}^L_{[t]})\sigma^2_{Z(t/t-1)} \quad (22)$$

$$\hat{Z}(t/t) = \hat{Z}(t/t-1) + \mathbf{K}_t \left(\vec{d}_{[t]} - \mathbf{C}^L_{[t]}\hat{Z}(t/t-1) - \mathbf{D}^L_{[t]} \right) \quad (23)$$

with $\hat{Z}(0)$ and $\sigma^2_{Z(0)}$ being the initial depth and depth uncertainty estimates, and $\mathbf{K}_t$ being the Kalman gain.

3.3.1 Warping the Depth Map

To complete the EKF analysis, there is still an additional problem to be solved, concerning the prediction step.

The predicted depth value $\hat{Z}(t/t-1)$ is obtained by applying the motion equation. However, this predicted depth value, no longer corresponds to the original pixel (x, y), since the x and y coordinates have also changed. As the predicted depth values does not coincide with image pixel positions, the depth at the grid point (x, y) must be inferred from these values. This problem can be approached in several ways like bi-linear or bi-cubic interpolation [1, 3].

We compute the depth estimate at (x, y) as a weighted average of the estimates falling within a 3x3 window centered in (x, y). The warped depth uncertainty is estimated using error propagation techniques:

$$Z_{(x,y)} = \frac{\sum_{i=1}^{n} d_i^{-2} Z_{(x'_i,y'_i)}}{\sum_{i=1}^{n} d_i^{-2}} \qquad \sigma^2_{Z(x,y)} = \frac{\sum_{i=1}^{n} d_i^{-4} \sigma^2_{Z(x'_i,y'_i)}}{(\sum_{i=1}^{n} d_i^{-2})^2} \quad (24)$$

where $Z_{(x'_i,y'_i)}$ stands for the predicted depth values, and d_i is the euclidean distance from (x'_i, y'_i) to (x, y).

4 Results

The 3D vision system was tested under a set of different synthetic and real conditions. The next section describes the results obtained using a sequence of synthetic images. The final results consist of tests with real underwater images.

4.1 Results with Synthetic Images

We started by assuming a scene structure composed by several rectangular patches placed at different distances from the camera, and we defined a brightness pattern that allows the generation of an image sequence. Figure 2 shows the last image of the sequence. The background is placed at 10m from the camera, the left rectangle at 5m and the right rectangular patch at 3.33m.

A 5x5 match window size was used, with $\hat{Z}(0) = 5m$, $\sigma^2_{Z(0)} = 25.0m^2$, $r = 1$ and $\lambda_{epip} = 5$. The final depth map is shown in figure 2, where depth is coded in gray level, darker points being closer to the camera. Figures 3 shows the first and final depth profiles along the x direction.

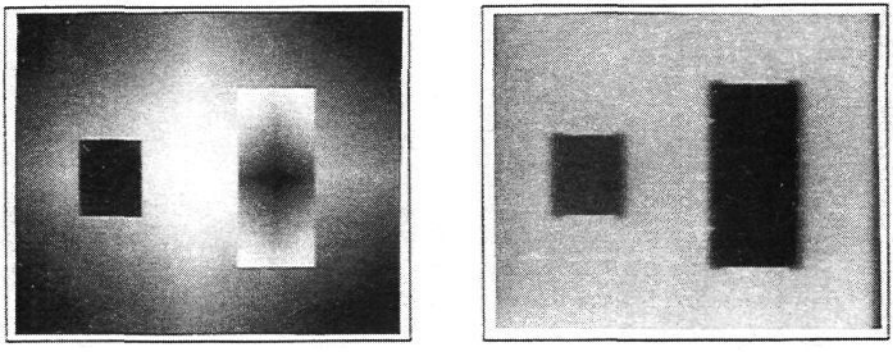

Figure 2: Last image of the synthetic sequence. Gray level coded depth map after the 10th iteration.

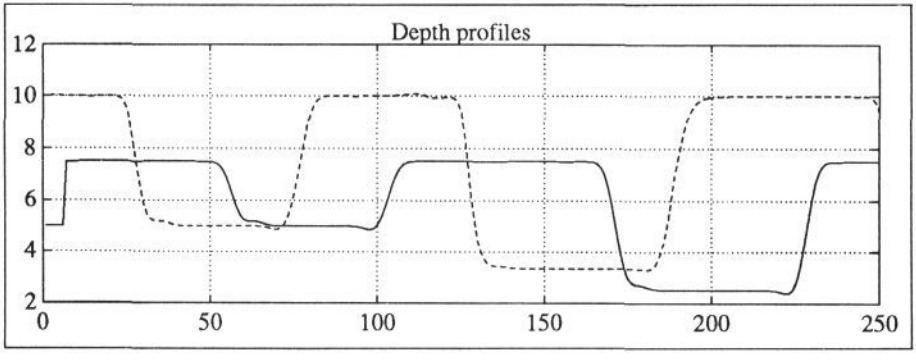

Figure 3: Depth estimate profile along the x direction (y = 128)

Although the synthesized images, are not corrupted with noise, they exhibit occlusion. The results obtained show a high level of precision. Notice the

improvement over time, as successive depth estimates are being combined, and the blurring effect in discontinuities. The blurring effect, due to the regularization procedure is more noticeable in the occluded areas, where the matcher fails, and the disparity field is filled in, by the regularization process.

4.2 Results with Real Underwater Images

In this section we present the results obtained using real underwater images acquired with a camera in a special tank. The camera is moving with downwards vertical speed of 4.4 cm/s.

The images were acquired at the video rate of 25 images/s. In order to use a suitable sampling period, we have assumed that all the objects in the scene were located at a distance from the camera in the range [1,6] meters, and to keep the disparity values within 1 and 12 pixels, the sampling period was chosen to be 0.4 seconds.

The *a priori* depth map estimate is 6 meters for every image pixel, with 0.01 m^2 variance. A 7x7 match window was used with $\lambda = 0.1$ and $\lambda_{epip} = 0$. Figure 4 shows the first and final image in the sequence, together with the perspective view of the corresponding depth maps.

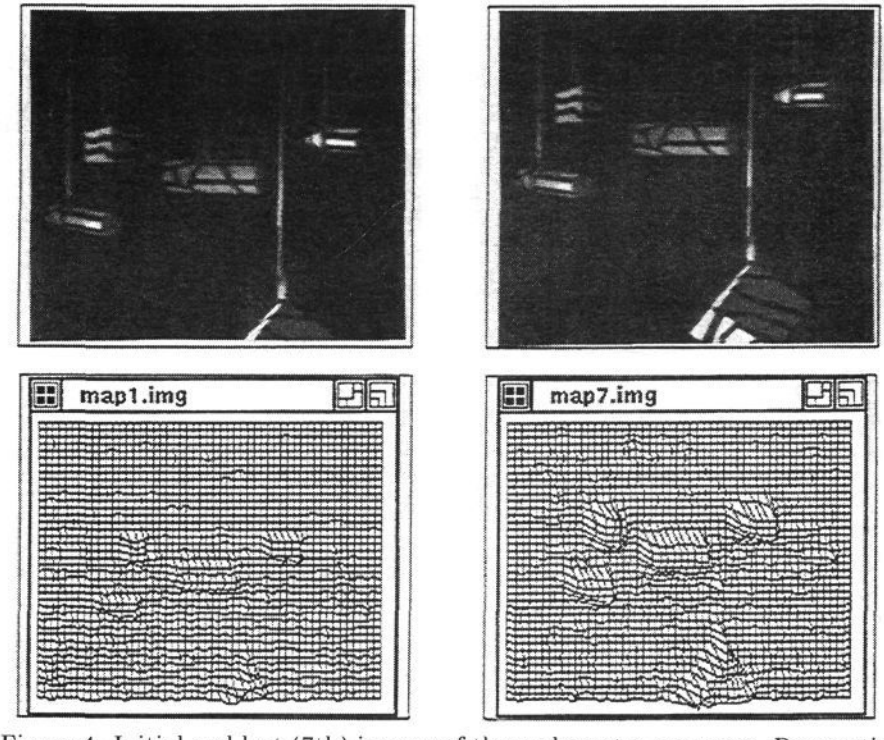

Figure 4: Initial and last (7th) images of the underwater sequence. Perspective of the first and final 3D depth maps

5 Conclusions

We presented a depth from motion vision system designed to compute dense depth maps from an image sequence, acquired by a moving camera. The system is based on a state space description of the depth from motion problem and models the existence of several uncertainty sources. A matching procedure including the epipolar constraint was defined, and regularization is used to filter the disparity vector field. Finally, depth measurements are combined over time, using Kalman filtering, to reduce uncertainty.

The system can be applied to a large number of problems in robotics, where the estimation of the scene structure may be found useful. A particular application related to underwater robotics was presented. Results obtained with synthetic images were presented, together with results with real images acquired in an underwater environment. In both cases the scene structure was recovered with remarkable accuracy.

References

[1] L. Mathies, T. Kanade, and Szelisky R. Kalman filter-based algorithms for estimating depth from image sequences. *Int. J. of Computer Vision*, 4(3):209–238, 1989.

[2] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *Int. J. of Computer Vision*, 4(2):283–310, 1989.

[3] J. Heel. Dynamic motion vision. In *Proc. of the DARPA Image Understanding Workshop*. Morgan-Kaufman Publishers, May 1989.

[4] B.K. Horn. *Robot Vision*. M.I.T.Press, 1986.

[5] B. Horn and B. Shunck. Determining optical flow. *Artif. Intell.*, 17:185–203, 1981.

[6] D. Ballard and C. Brown. *Computer Vision*. Prentice-Hall, London, 1982.

[7] J. Santos-Victor and J. Sentieiro. A Dynamic 3D Vision System. Technical report, Instituto Superior Técnico, February 1992. Ref. Mobius/rpt/02/92 JASV/JJSS.

[8] D. Terzopoulos. Regularization of inverse visual problems involving discontinuities. *IEEE Trans. on Pattern Anal. and Machine Intell.*, 8(4):413–424, July 1986.

[9] M. Bertero, T. Poggio, and V. Torre. Ill-posed problems in early vision. *Proceedings of the IEEE*, 76(8):869–889, 1988.

[10] R. Szeliski. Bayesian modeling of uncertainty in low-level vision. *Int. J. of Computer Vision*, 5(3):271–301, 1990.

[11] R. Szeliski. *Bayesian Modeling of Uncertainty in Low-Level Vision*. PhD thesis, Carnegie Mellon University, 1988.

[12] Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, 1970.