SSD Disparity Estimation for Dynamic Stereo

E. Trucco
Department of Computing and Electrical Engineering
Heriot-Watt University
Riccarton, Edinburgh, SCOTLAND, EH14 4AS

V. Roberto, S. Tinonin, M. Corbatto
University of Udine
Via delle Scienze 206, I 33100 Udine, ITALY

Abstract

We analyse experimentally some subpixel-accuracy disparity and uncertainty estimators based on the SSD method, frequently used in stereo and motion analysis. We identify key inadequacies, and introduce new, practical algorithms. We discuss results and performance tests, and demonstrate the effective use of the new estimators in a complete, working system reconstructing dense depth maps using dynamic stereo. The system achieves good accuracy (average percentage errors smaller than few percents) and reliable uncertainty discrimination without any expensive smoothing or regularisation of disparity maps, adopted commonly.

1 Introduction

This paper analyses the estimation of subpixel-accuracy disparity and its uncertainty in the context of dynamic stereo for computing dense depth maps [3, 4, 7, 9, 10]. Careful solutions to this problem, supported by adequate solutions to resampling and drop ins/outs, allow us to achieve good reconstructions without expensive smoothing or regularisation of the disparity map [3, 5, 8]. Our experimental testbed is a complete dynamic stereo system based on a Kalman filter (KF). In the context of dynamic stereo, *pixel-based methods* for computing dense disparity maps have been reported in [3, 5, 7], and *feature-based methods*, leading to sparse disparity maps, in [4, 5, 9, 10]. The *squared-sum difference method* or SSD [1, 5] (Section 3) is a correlation-based algorithm [1, 3, 7, 5] of the former class; it can be cheap, effective and useful in many contexts. Here, we analyse experimentally the behaviour of the subpixel interpolation and quadratic-fit uncertainty estimator proposed in [5], identify key inadequacies, and formulate new algorithms (Section 3.1 and 3.2). We use the *same* module for estimating motion and stereo disparities and their uncertainties, guaranteeing maximum consistency of the uncertainty values fed to the fusion stage. Dynamic fusion based on KFs may take place in 3-D [3, 10] or at disparity level [5, 8]. We adopt the latter approach, thus taking the depth-from-motion framework proposed in [5] to dynamic

BMVC 1996 doi:10.5244/C.10.63

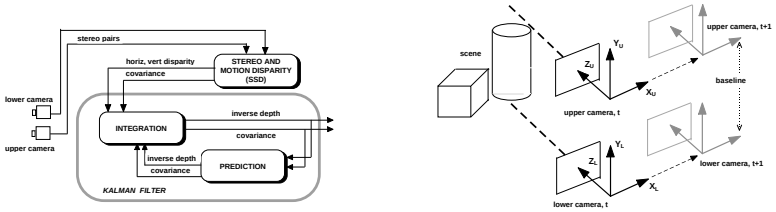

Figure 1: *system architecture (left) and setup geometry (right).*

stereo (Section 4). Section 5 presents some experiments with real sequences, quantitative results, and performance issues. Section 6 offers a short discussion of our work.

2 System architecture and assumptions

The architecture of our dynamic stereo system is illustrated in Figure 1 (left). A single module supplies horizontal (motion) and vertical (stereo) subpixel-accuracy disparities and their uncertainties, which are integrated by a KF to produce a depth map and its uncertainty. We chose the simple geometry in Figure 1 (right) to prototype an experimental testbed quickly. A sequence of stereo views is acquired by two parallel cameras translating rigidly along the x axis of the camera frames. The optical axes of the cameras are displaced along y, which implies orthogonal stereo and motion disparities and simplifies the mathematics. Our assumptions, found frequently in similar systems, are: *(a)* parallel stereo cameras [4, 9, 7]; *(b)* rigid, translational motion [4, 5, 8]; *(c)* translation parallel to the x axis of the camera frame [5, 8], and stereo baseline parallel to the y axis of the same frame (Figure 1, right); *(d)* only the diagonal elements of the state covariance matrix are stored [7, 5, 8] (all depth estimates independent [4, 9]); *(e)* independent stereo and motion observations.

3 Disparity and its uncertainty

A subpixel-accuracy disparity map and its uncertainty are computed by the same SSD algorithm for both motion and stereo. The pixel-accuracy disparities of (x, y) are

$$\Delta x_M = \min_{W_x}\{e_x\} = \min_{W_x}\left\{\int\int [I_t(x-\Delta x+\lambda, y+\eta) - I_{t-1}(x+\lambda, y+\eta)]^2 d\lambda d\eta\right\}$$

$$\Delta y_M = \min_{W_y}\{e_y\} = \min_{W_y}\left\{\int\int [I_L(x+\lambda, y-\Delta y+\eta) - I_U(x+\lambda, y+\eta)]^2 d\lambda d\eta\right\}$$

where e_x and e_y are error functions, I_L and I_U are the stereo images from the lower and upper camera respectively (Figure 1, right), W_x,W_y are small image regions (typically 5×5 or 7×7 in our experiments), and t, $t-1$ are successive time instants.

3.1 Subpixel interpolation

Detailed experimental analysis of correlation-based algorithms achieving subpixel precision through unweighted, quadratic interpolation around the minimum of e [5, 8] shows that these methods can lead to inconsistent results owing to the asymmetry of e around its minimum. We tested quadratic interpolation on a number of controlled image pairs, created by shifting segments of a real image by known quantities (up to 10 pixel, using 5x5 and 7x7 correlation masks), and adding increasing amounts of Gaussian noise (up to $\sigma = 30$ with 8-bit images). We found that *(a)* even without noise, subpixel disparities were dispersed around the true values (standard deviation about ± 0.15 pixel), which was caused by the asymmetry of e around its minimum; *(b)* with noise, disparities were not distributed uniformly, and tended to cluster around integer values. Problem *(a)* is solved by weighted interpolation schemes in which the importance of a point is inversely proportional to e, e.g. along x

$$\Delta x_{subpix} = \frac{\frac{\Delta x_M - 1}{e(\Delta x_M - 1)} + \frac{\Delta x_M}{e(\Delta x_M)} + \frac{\Delta x_M + 1}{e(\Delta x_M + 1)}}{\frac{1}{e(\Delta x_M - 1)} + \frac{1}{e(\Delta x_M)} + \frac{1}{e(\Delta x_M + 1)}}$$

where Δx_{subpix} is the subpixel-precision disparity estimate. In our tests, we used both 3-pixel and 5-pixel neighbourhoods of the pixel-precision minimum Δx_M. Figure 2 (top) illustrates the behaviour of weighted schemes, showing that problem *(b)* is still present. To rectify *(a)* and *(b)*, we formulated a new interpolation scheme which increases the symmetry as e_{min} decreases, e.g. along x:

$$\left\{ (\Delta x_M - 1, e(\Delta x_M - 1)), (\Delta x_M, e(\Delta x_M) - \frac{\alpha}{1 + e(\Delta x_M)}), (\Delta x_M + 1, e(\Delta x_M + 1)) \right\}$$

where α is a constant. Typical, experimental disparity distributions showing the satisfactory behaviour of our scheme with respect to *(a)* and *(b)* are shown in Figure 2 (bottom) (for the case $\Delta x_M = 6$ true disparity). Figure 3 summarises some of our accuracy tests: for instance (Figure 3(b)), with noise level $\sigma = 10$, 75% of the pixels have errors less or equal to 5%. The results are very satisfactory, given that no disparity regularisation, image magnification [5], or expensive smoothing/fitting [3] is used.

3.2 Estimating uncertainty

Local quadratic fit has been suggested as the basis for uncertainty estimation with SSD algorithms [1, 5], but again the irregular shape of the error functions around the minimum leads to inconsistent estimates. We analysed the experimental behaviour of the quadratic estimator suggested in [5] (based on the reciprocal of the quadratic coefficient of the interpolating quadric) in the controlled conditions described in Section 3.1. The tests were performed with variously textured images (rather uniform to strong textures). The results indicated that *(a)* noise can result in low uncertainties associated with significantly wrong disparities, and *(b)* uncertainty estimates can decrease as the noise level increases, in both highly and marginally textured regions. This is because the noise can drive the quadratic fit to produce a very narrow (low uncertainty) function centered around a wrong

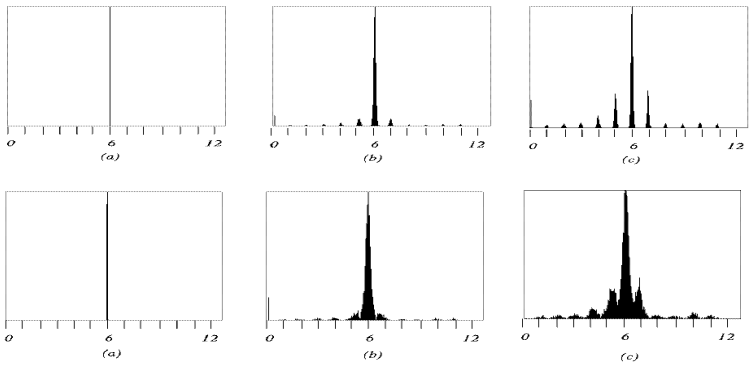

Figure 2: *Experimental distributions of subpixel disparity estimates around exact value 6, with additive Gaussian noise: $\sigma = 0$ (a), 10 (b), 20 (c). Top: weighted interpolation (5-pixel neighbourhood of Δx_M). Bottom: our scheme.*

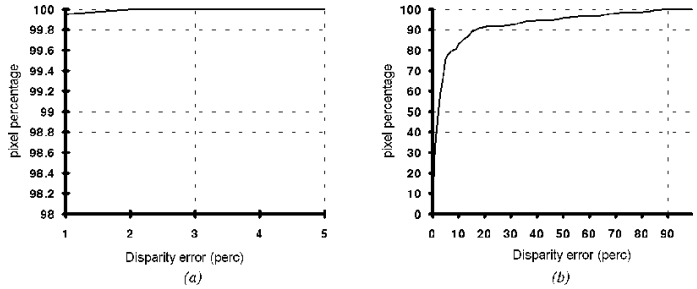

Figure 3: *Percentage of pixels (y axis) with disparity error (in percent) less than x value (x axis); synthetic images with (a) no noise, (b) Gaussian additive noise of $\sigma = 10$ (b).*

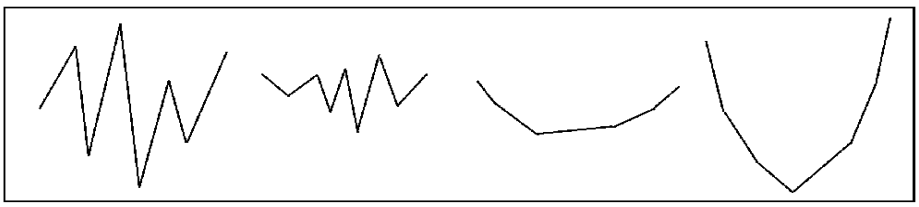

Figure 4: *Qualitative error profiles associated with decreasing uncertainty (left-right).*

minimum. To design a correct estimator, we targeted three criteria: *(1)* minimum errors increase, on average, with the level of noise; *(2)* the larger the SSD error, the more jagged the local shape of e; *(3)* the larger the local peak-to-peak variation of e around the minimum, the more unreliable the estimates. So a satisfactory estimator must consistently attach decreasing uncertainties to the qualitative error profiles in Figure 4, from left to right. We quantified the criteria in three coefficients, tested their behaviour individually, and combined them in a complete estimator. The absolute SSD error $\|e\|$ served as the first coefficient. Irregularity was quantified by a simple measure

$$i = \frac{0.5(\sum_{k \in I} \mid s(k) - s(k-1) \mid) - 1}{2h}$$

where h spans an interval $I = [\Delta x_{min} - h + 1, \Delta x_{min} + h]$ around the extremum Δx_{min}, and

$$s(k) = \begin{cases} -1 & k = 0 \\ sgn(e(k) - e(k-1)) & k \neq 0 \end{cases}$$

where e is the SSD error function. The local variation was quantified by

$$v = \sum_{k \in I} \left[\frac{e(k) - e(k-1)}{\max_{w \in I} \{e(w)\} - \min_{w \in I} \{e(w)\}} \right]^2$$

We ascertained experimentally that the behaviour of the last two measures was consistent with the criteria above. The final estimator combines the three contributions by expressing the variance of the disparity measurements as $\sigma^2 = (k * R)^2$, with $R = v(i - \tau) + q$. τ is chosen so that an irregularity greater than τ implies that uncertainty increases for increasing v (local variation), and *vice versa*. So, in the presence of the *same* high local variation, the rightmost profile of Figure 4 leads to low uncertainty, the leftmost one to high uncertainty. q ensures that R is not negative; its value depends on I and the range of variation of the parameters in the R expression. We achieved the best results over a large set of experiments with $q = 0.41$ and $\tau = 0.2$.

4 Integrating stereo and motion

All disparities and uncertainties are fed to a KF based on the measurement equation

$$\mathbf{D} = \mathbf{H}\mathbf{u} + \xi \tag{1}$$

where $\mathbf{u}$ is the $n^2 \times 1$ vector (assuming $n \times n$ images for simplicity) of the inverse depths $\frac{1}{Z(i,j)}$ of each pixel (i, j); $\mathbf{D} = \left[\Delta_m^T \mid \Delta_s^T\right]^T$ is the $2n^2 \times 1$ vector formed by the observed horizontal (motion) disparities Δ_m^T, and the vertical (stereo) disparities Δ_s^T; ξ is the usual white Gaussian noise of mean 0 and covariance $\mathbf{R}$ $(2n^2 \times 2n^2)$. The measurement equation for motion, in our assumptions, is

$$\Delta x = -T_x f \frac{1}{Z} + \rho = -T_x f u + \rho \tag{2}$$

where $\rho \sim (0, \sigma_\rho^2)$ is an additive Gaussian noise. For stereo (vertical disparity), the parallel-camera setup implies

$$Z = \frac{f_1(d_y f_2 + y_2 \Delta Z - f_1 + f_2)}{f_2 y_1 - f_1 y_2}$$

where f_1, f_2 are the cameras' focal lengths, d_y is the stereo baseline (distance between the optical centers), ΔZ the distance between the parallel retinas, and $y_2 - y_1$ the stereo disparity. Assuming calibration can achieve $\Delta Z \cong 0$, $f_1 \cong f_2$ within sufficient accuracy for our purposes, rearranging, and adding Gaussian noise $\eta \sim (0, \sigma_\eta^2)$, we obtain

$$\Delta y = -d_y f u + \eta$$

The variances $\sigma_\rho^2, \sigma_\eta^2$ are estimated as described in Section 3.2. The KF is expressed as [6]

$$\begin{aligned} \mathbf{u}(t_i^+) &= \left[\mathbf{P}(t_i^+)\mathbf{P}(t_i^-)^{-1}\right] \mathbf{u}(t_i^-) + \left[\mathbf{P}(t_i^+)\mathbf{H}(t_i)\mathbf{R}(t_i)^{-1}\right] \mathbf{D}(t_i) \\ \mathbf{P}(t_i^+) &= \left[\mathbf{P}(t_i^-)^{-1} + \mathbf{H}(t_i)^T \mathbf{R}(t_i)^{-1} \mathbf{H}(t_i)\right]^{-1} \end{aligned} \tag{3}$$

where $\mathbf{P}$ is the state covariance. The assumptions on geometry, independent stereo and motion measurements, and independent pixels, imply that $\mathbf{H}(t_i), \mathbf{R}(t_i)$ are block diagonal, and Eq.(3) can be written in terms of individual pixels:

$$\begin{aligned} u(t_i^+) &= \frac{P(t_i^+)}{P(t_i^-)} u(t_i^-) + P(t_i^+) \left[\frac{-fT_x}{\sigma_x^2(t_i)} \Delta_x(t_i) + \frac{-fd_y}{\sigma_y^2(t_i)} \Delta_y(t_i)\right] \\ P(t_i^+) &= \frac{P(t_i^-)\sigma_x^2(t_i)\sigma_y^2(t_i)}{\sigma_x^2(t_i)\sigma_y^2(t_i) + P(t_i^-)\left[\sigma_y^2(-fT_x)^2 + \sigma_x^2(-fd_y)^2\right]} \end{aligned}$$

The prediction stage involves the following three substeps. **(1) Spatial propagation:** each pixel is propagated spatially along x according to Eq.(2). **(2) Covariance increment:** models the uncertainty increase caused by uncertain motion, calibration errors, and the like, by *exponential age-weighting* [6, 5]: $P(t_i^-) = \gamma P(t_{i-1}^+)$ where $\gamma > 1$ inflates the covariance heuristically. **(3) Resampling:** $x(t_i^-)$ is not, in general, an integer, and it is necessary to resample and interpolate depth and covariance maps. We tested experimentally six interpolation and resampling methods, linear and nonlinear (including the one in [5] and some original solutions), and designed a new resampling method which performed best with respect to *(a), (b)* and *(c)*. The new method is based on a careful characterisation of drop-ins/outs within an asymmetric neighbourhood of each pixel.

5 Testing the complete system

We mounted a camera (Digital Vision CCD-9-MICRO, 500×492) on an adjustable stand allowing accurate control of vertical translations. The stand could slide rigidly on a horizontal rail. Images were acquired by a DataCell S2200 framestore. All software was C, tested on both a 486 running Linux and a SPARC10. We used our implementation of Caprile and Torre's calibration algorithm [2]. Calibration

tests with synthetic and real images put the worst percentage errors at 1.8% for the focal length, and at 5% for the extrinsic parameters, quite adequate for our purposes. We tested the complete system with about twenty sequences, synthetic and real. Only two real experiments are reported here. Figure 5 (left) shows the first and last pair of a real, 9-pair stereo sequence (256x256, 8-bit images). Camera-object distances were 26cm for the horse cart and 35cm for the owl figure. The depth variation of the visible surfaces was less than 3cm. The calibrated focal length was 679 pixel. The camera pair translated by 5 mm between frames; the stereo baseline was 5 mm. Since the KF maintains uncertainty estimates, we could fix a desired confidence level and reject all pixels whose depth uncertainty exceeded that level. Figure 5 (right) shows the depth and variance maps after processing the first and last stereo pair. Notice the decrease in uncertainty (darkening of variance maps) and the increased number of accepted points. Figure 6 shows the depth map of points with confidence level greater than 90%, after processing the first pair (left), the last pair (middle), and applying a 7×7 median filter to the final result (right); notice the increased number of accepted points (denser map) between the left and middle/right images. Two 3-D views of the final depth map, after median filtering and simple surface interpolation to fill in gaps, are shown in Figure 7 (right). This result indicates that simple postprocessing suffices to achieve good range images. Figure 7 (left) is a histogram of the depths of the accepted pixels, showing that measurements are concentrated around the correct values. Notice that increasing noise levels reduce the number of accepted points but do not worsen the accuracy dramatically, as uncertainties reflect consistently the reliability of measurements. Figure 8 shows the first, middle and last pair of another 9-pair sequence (256×256, 8-bit images). The elephant figure, model car and wooden teddy bear were placed at about 25, 32 and 35 cm from the cameras respectively. Most points on elephant and bear were visible for a few frames only. Focal length, translation and stereo baseline were the same as in the previous experiment. Figure 9 (left) shows the histogram of the accepted depths after the middle (left) and last pair. Figure 9 (right) shows the median-filtered depth maps of the accepted points after the middle (left) and last pair. Figure 10 summarises the results of a series of tests with unoptimised C code on a 486 (33MHz) running Linux. Notice the good performance of integration and prediction even with unoptimised code.

6 Discussion

We have analysed experimentally SSD-based subpixel-accuracy disparity and uncertainty estimation methods, identified key inadequacies, and sketched new algorithms and their testing. We have illustrated the use of the algorithms in a complete dynamic stereo system which produces depth maps of similar accuracy as, and sometimes better than, those reported for comparable systems, but no regularisation or smoothing of disparity maps is used. The work has been supported by systematic testing throughout. Limitations concern mainly the geometric assumptions of Section 2; however, these were instrumental to prototype rapidly a complete system within which to test all algorithms. A challenging question is whether surface continuity could be modelled within the KF without increasing

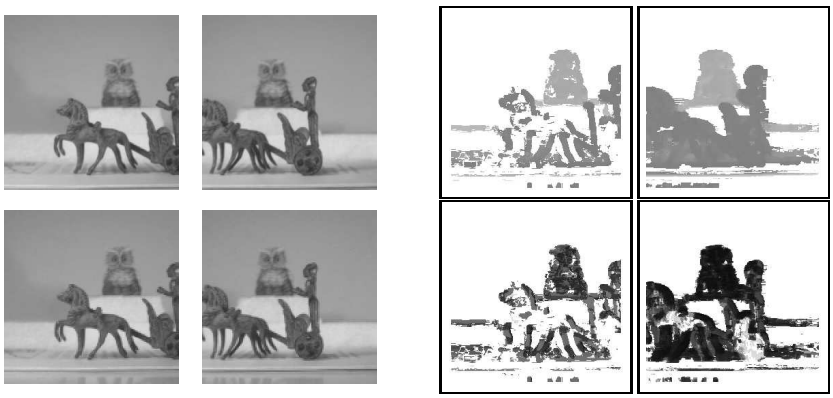

Figure 5: *Left block: first (left column) and last stereo pair of input sequence. Right block: depth (top) and variance maps after first (left column) and last pair.*

Figure 6: *Depth maps of accepted points (minimum confidence 90%). After first pair (left); after last pair (middle); the latter after median filtering.*

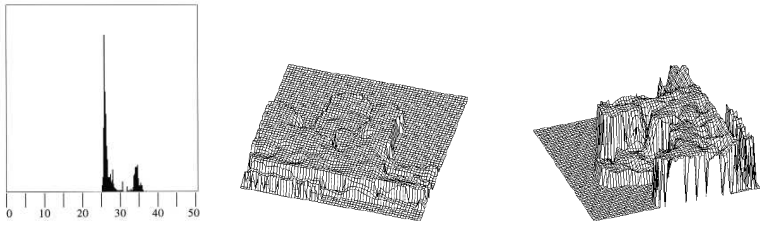

Figure 7: *histogram of estimated depths (distances in cm) and two 3-D views of final depth map of accepted points (median filtered).*

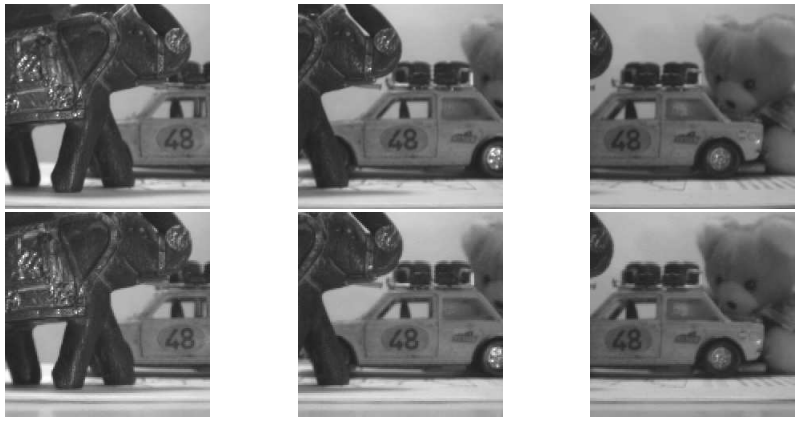

Figure 8: *First (left column), middle, and last stereo pair of input sequence.*

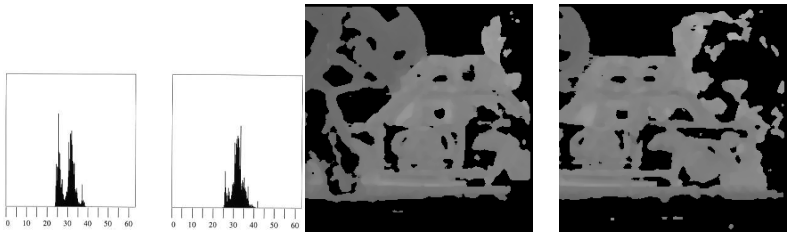

Figure 9: *Left: histogram of estimated depths (distances in cm) after middle (left) and last pair. Right: final depth map of accepted points, median filtered, after middle (left) and last pair.*

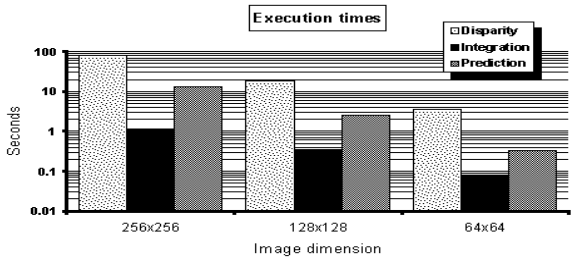

Figure 10: *CPU times (seconds) for disparity, fusion and prediction and different image sizes, averaged over several runs on a 486 running Linux.*

memory and time complexity unacceptably. Another point is that our uncertainty estimator does not really produce Gaussian variances [5]; although experimentation indicates clearly that this does not affect adversely either accuracy nor create low uncertainty for wrong points, we plan further theoretical investigation of the issue. Adaptive SSD windowing and general motion are further extensions planned.

Acknowledgements

We thank Alessandro Verri for useful discussions and comments. This work was partially supported by a British Council-MURST/CRUI grant, and by the Italian National Research Council (CNR), grant 95.00516.CT12 "Sistemi Percettivi Distribuiti".

References

[1] P. Anandan: *A Computational Framework and an Algorithm for the Measurement of Visual Motion*, Int. Journ. Comp. Vis. vol 2, 1989, pp. 283 – 310.

[2] B. Caprile and V. Torre: *Using Vanishing Points for Camera Calibration*, Int. Journ. Comp. Vis. vol 4, 1990, pp. 127 – 140.

[3] R. Koch: *3-D Surface Reconstruction from Stereoscopic Image Sequences*, Proc. IEEE Int. Conf. Comp. Vis. ICCV95, Cambridge(MA), 1995, pp. 109 – 114.

[4] L. Li and J. H. Duncan: *3-D Translational Motion and Structure from Binocular Image Flow*, IEEE Trans. Patt. Anal. Mach. Int. vol PAMI-15, 1993, pp. 657 – 667.

[5] L. Matthies, T. Kanade and R. Szeliski: *Kalman Filter-Based Algorithms for Estimating Depth from Image Sequences*, Int. Journ. Comp. Vis. vol 3, 1989, pp. 209 – 236.

[6] P. S. Maybeck: Stochastic models, estimation, and control, vol 1, Academic Press, New York, 1979.

[7] J-Y Shieh, H. Zhuang and R. Sudhakar: *Motion Estimation from a Sequence of Stereo Images: a Direct Method*, IEEE Trans. Syst. Man Cyb. vol 24, 1994, pp. 1044 – 1053.

[8] R. Szeliski: *Bayesian Modelling of Uncertainty in Low-Level Vision*, Int. Journ. Comp. Vis. vol 5, 1990, pp. 271 – 301.

[9] A. P. Tirumalai, B. G. Schunk and R. C. Jain: *Dynamic Stereo with Self-Calibration*, IEEE Trans. Patt. Anal. Mach. Int. vol PAMI-14, 1992, pp. 1184 – 1189.

[10] Z. Zhang and O. D. Faugeras: *3-D Motion Computation and Object Segmentation in Long Sequences of Stereo Images*, Int. Journ. Comp. Vis. vol 7, 1992, pp. 211 – 241.