3-D Object Reconstruction Using Stereo and Motion

ENRICO GROSSO, GIULIO SANDINI AND MASSIMO TISTARELLI

Abstract — **The extraction of reliable range data from images is investigated, considering, as a possible solution, the integration of different sensor modalities. Two different algorithms are used to obtain independent estimates of depth from a sequence of stereo images. The results are integrated on the basis of the uncertainty of each measure. The stereo algorithm uses a coarse-to-fine control strategy to compute disparity. An algorithm for depth-from-motion is used exploiting the constraint imposed by active motion of the cameras. To obtain a three-dimensional (3-D) description of the objects, the motion of the cameras is purposively controlled, as to move around the objects in view, while the direction of gaze is kept still toward a fixed point in space. This egomotion strategy, which is similar to that adopted by the human visuomotor system, allows a better exploration of partially occluded objects and simplifies the motion equations. The algorithm has been tested on real scenes, demonstrating a low sensitivity to image noise, mainly due to the integration of independent measures. An experiment, performed on a real scene containing several objects, is presented.**

I. Introduction

ONE OF THE PRIMARY goals of early vision is that of extracting volumetric measures about the observed objects in a scene from a continuous flow of visual information. In humans this task is accomplished using many different sources of information coming from multiple sensor modalities.

So far many methods have been proposed to acquire information about the three-dimensional (3-D) structure of the world, with the aim of building a feasible and handy representation of the environment [1]–[9].

It is our opinion that, at present, all computationally reasonable algorithms for range estimation suffer from errors and uncertainities peculiar to each method. For example, the illumination condition is a weak point in deriving shape from shading [10], [11] and the computation of stereo disparity fails when the matching is performed on edges parallel to the epipolar lines [3], [12]. A possible solution is to select and to integrate, according to a reliability measure, the results obtained from different information sources to obtain a unique representation of the environment [13], [14].

In this paper we present an example of the integration of range data computed from stereo matching and optical flow, ending with a 3-D (volumetric) representation of the solids in view. The experimental setup is based on a pair of cameras, with a coplanar optical axis directed toward a common fixation point, moving around an object and tracking an environmental point [15], [16] (i.e., the movement is performed keeping the fixation point still).

Along with the measure of depth, an explicit estimation of uncertainty is carried out. This uncertainty value, transformed into a *reliability* map and associated with the corresponding *depth* map, is used, along with the instantaneous position in space and the geometry of the stereo pair (i.e., *proprioceptive information*), to update the volumetric representation of the environment continuously [14], [17].

In fact, in spite of the great deal of information, obtained from the different viewpoints, it is clear that from a *bas relief* (i.e., a depth image) only a partial description of the 3-D shape can be derived. To complete and refine this information it is necessary to "move around," exploring the environment actively [18]. This is true not only because occluded objects can become evident from different viewpoints, but also because during the motion depth information can be derived from motion parallax. The tracking strategy, adopted to drive the egomotion, allows the active inspection of the environment and of the objects in the scene from different viewpoints.

In principle, the continuous flow of information, represented as continuously changing depth images derived from stereo measures and motion parallax, needs to be cast into an incremental representation. This casting process acts like an accumulator where only the "new" information changes the current description whereas the redundant (or duplicate) information does not affect it.

In our approach the casting process makes use of a geometric description of the world in terms of a 3-D array of voxels. The visual *bas relief* computed from each viewpoint is used to update this volumetric description.

During the integration process, the reliability map is used to weight the depth measure with respect to the

Manuscript received September 20, 1988; revised March 30, 1989. This work was supported in part by grants from ESPRIT (Project P419) and from the Italian National Council of Research and in part by an ELSAG SpA fellowship.

The authors are with the Department of Communication, Computer and Systems Science, University of Genoa, Via Opera Pia 11a, 1-16145 Genova, Italy.

IEEE Log Number 8930369.

0018-9472/89/1100-1465$01.00 ©1989 IEEE

accumulated one. Each voxel of the volumetric accumulator stores, at each instant of time, the measure of probability of "empty space." It is worth noting that this analogic geometric representation of the environment is certainly not sufficient for high-level processing (for example, recognition); on the other hand its primary use is to help the accumulation of depth information. More "complete" geometric description, including also surface information, can be derived from the voxel representation and logically linked to it.

II. Acquisition of Range Data

The outline of the experiment described in this paper is presented in Fig. 1. The images were acquired from four positions in space (east, west, south, and north), moving a stereo pair around a set of objects while tracking a fixed point on the surface of the central object. The distance of the cameras from the fixation point was kept constant during the movement (a circular trajectory) and equal to 103 cm.

From each position in space eleven stereo pairs were acquired. The displacement between successive position is 4° along the circular trajectory, around a vertical axis centered on the fixation point. The total angular span, between the first and last image of each sequence, is, consequently, 40°.

On the eleven stereo pairs, the third was used to compute depth from stereo, while all the remaining images were used in the motion algorithm to compute the depth map, also relative to the third image.

All the images were acquired at 256×256 pixels with 8 bits of resolution in intensity, with two CCD cameras (COHU 4713) and a VDS 7001 Eidobrain image processing system.

A. Disparity Extraction and Edge Matching

The first part of the algorithm is based on the computation of the cross correlation between corresponding square patches of the stereo pair; the images over which the cross correlation is performed are obtained by convolving the originals with a Laplacian of Gaussian operator and representing only the sign of the filtered images [19], [3]. The estimation of correspondence is performed, using a coarse-to-fine approach, in three successive steps. At each step the correlation is computed at a different spatial frequency band (i.e., filtering the images with a $\nabla^2 G$ mask of different size) going from low to high spatial frequencies; as a consequence also the size of the correlation patches decreases at each step (it is directly proportional to the size of the mask). At the end of each step, a measure of disparity is obtained which is successively refined during the following steps.

Finally, the maximum precision in disparity is achieved, performing an explicit edge matching between the zero crossings of the right and left image, extracted at the highest resolution scale. This process is performed using the disparity value computed by the three-step cross-correlation procedure. In particular, starting from an edge point in the left image, the corresponding contour point in the right image is searched in a neighborhood of the disparity computed during the previous phase. The amplitude of the search space is determined by the amplitude of the $\nabla^2 G$ mask used to filter the image at the higher resolution scale.

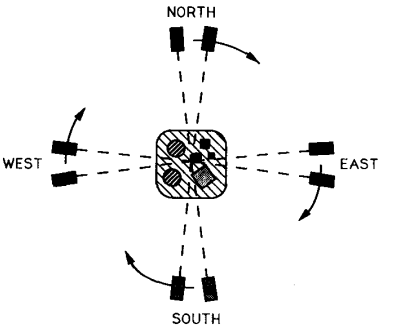


Fig. 1. Schematic representation of experimental setup. Square in middle represents tray with objects on top; initial positions of stereo cameras, with respect to objects, are indicated as north, south, east, and west. Arrows indicate direction of movement of cameras acquiring sequences.

The correlation values computed in a region with uniform shade (i.e., lacking in significant edges) have a low reliability and can produce many errors on the final results. For this reason the correlation is performed only over the regions of higher contrast, which correspond to the image areas whose energy, measured on the left image convolved with a Laplacian or Gaussian operator, is greatest.

Overlap constraints, with a threshold on the minimum energy value, limits the overall number of patches used to perform the correlation. In practice, the patches can be positioned directly on the edges, while the slope of the zero crossings (extracted from the convolved images) is used as the sorting value for the selection of the best image patches.

The correlation measure is weighted using measures of slope and spatial orientation of the gradient computed over the filtered images. As a consequence the disparity, identified by the peak of the correlation function, is also a function of local orientation. The value of correlation is used as a reliability factor [20], [21].

In Fig. 2, the four stereo pairs used in the experiment are presented. In Fig. 3 the result of the correlation process is shown, performed on the first stereo pair filtered with a $\nabla^2 G$ mask with standard deviation σ equal to 8, 4, and 2 pixels; the grey level is proportional to the disparity of the patches. The map presented in Fig. 3(a) represents the final disparity obtained from the regional part of the algorithm.

A planar model is used to determine depth from stereo (see Fig. 4). In this case the depth of a world point with respect to the left camera is a function of six independent

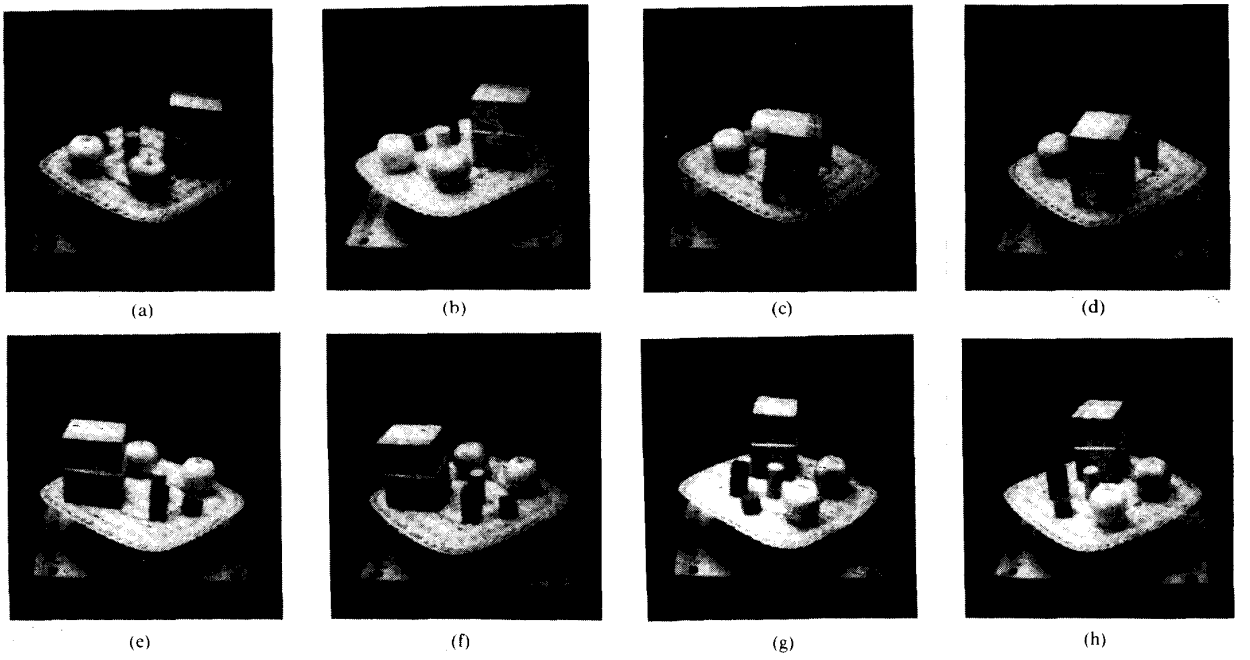


Fig. 2. Stereo pairs used in experiment. Resolution of images is 256×256 pixels.

parameters:

$$Z_s = Z_s(x_0, x_1, \theta, F, l, m) = \sqrt{F^2 + x_0^2}\,\frac{l[x_1 \sin\theta + F\cos\theta] + [F + m][F\sin\theta - x_1\cos\theta] + Fx_1}{[x_1 x_0 + F^2]\sin\theta + F[x_0 - x_1]\cos\theta}. \quad (1)$$

The four parameters θ, l, m, F (refer to the scheme depicted in Fig. 4) depend upon a calibration procedure, while x_0 and x_1 are two corresponding points in the left and right image, respectively (i.e., $x_1 = x_0 + d$, where d is image disparity).

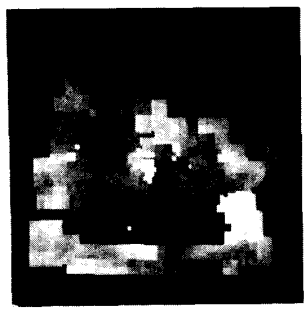
The depth map obtained from the first stereo pair is presented in Fig. 3(b) The depth values are computed at the edge points obtained from the convolved image at the highest resolution; the final depth map is computed, for all the image points, with a linear interpolation. Along with depth also the associated uncertainty measure is presented in Fig. 3(c); this measure reflects the reliability of the computed depth (see Section III-A).

B. Estimation of Depth from Motion

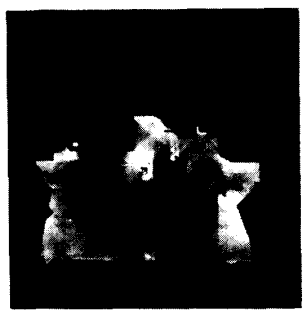
The estimation of the optic flow from an image sequence is based on a gradient technique in which proprioceptive knowledge of the egomotion parameters is used to constrain and solve, in closed form, the motion equations. The measurement is performed at the contour points obtained by filtering the images with a $\nabla^2 G$ operator and extracting the zero crossings [22]. The procedure for the computation of the optic flow is divided into the following steps:

- computation of the velocity component perpendicular to the local orientation of the contour; for each contour point the component $V^{\perp}$ is computed as the ratio between the time derivative and the local edge slope;
- computation of the true direction of motion.
- The computation of the direction of motion requires, in the case of general motion of the camera in a steady environment, the knowledge of at least seven variables related to the egomotion (six for displacements and rotations and the focal length). A reduction of the parameters required is achieved constraining the movement of the observer.

In this approach the motion of the camera was constrained as to keep the fixation point still during the motion around the objects in view. As a consequence the egomotion parameters that need to be measured are the distances of the camera from the fixation point (D_1 and D_2) measured at successive time instants, the rotation angles θ, ϕ, and ψ (measured for each sampled frame) and the focal length of the camera F (for explanation of the symbols refer to Fig. 5). From these parameters the direction of the flow due to the translation of the sensor $\vec{V}_t$ and the vector $\vec{V}_r$, due to the rotational part of motion, are computed.



(a)



(b)

(c)

Fig. 3. Results of stereo algorithm relative to single view (topmost pair of Fig. 2). (a) Results of regional correlation step; grey level codes image disparity (b) Depth map obtained by linear interpolation of contour values, obtained after edge matching refinement (depth is proportional to gray level). (c) Associated uncertainty (uncertainty is proportional to gray level).

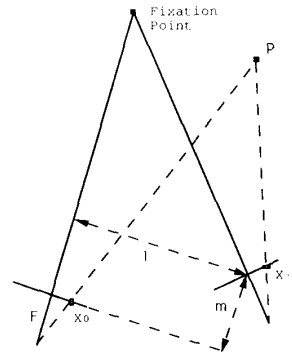
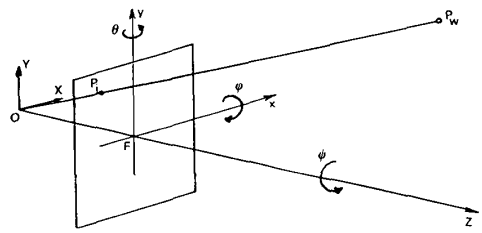


Fig. 4. Geometry of stereo setup.



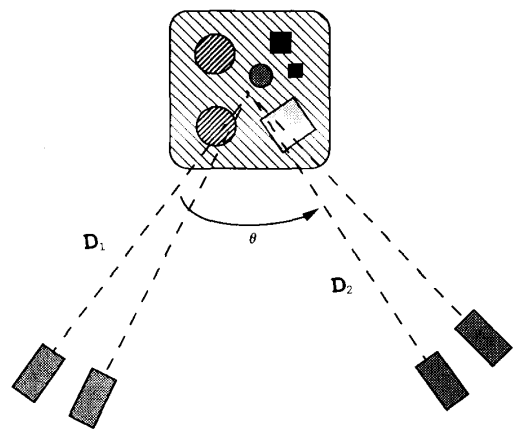


Fig. 5. Schematic representation of camera coordinate system with egomotion parameters used in motion algorithm.

The direction of the flow field is obtained solving, in closed form, a set of non-linear equations, which incorporates the known egomotion parameters and the motion constraints:

- matching of corresponding contours of successive image pairs (instantaneous optic flow).

The optic flow computed at the previous steps is refined searching for the first zero crossing in the successive image along the computed direction. This kind of search is motivated by the fact that, capturing the images with a high sampling frequency and by the smoothing operated by the Gaussian filtering, it is unlikely to find more than one contour between corresponding edges.

A confidence measure is associated with the matched points which reflects the likelihood of the match to be correct. This measure is obtained comparing the edge orientation and slope at corresponding contour points (this topic is further discussed in Section III-A).

To achieve a sufficient range of velocity for distance computation, the instantaneous optic flows, resulting after the edge matching, are joined together, giving *a global optic flow*, relative to a part of the sequence.

The images in Fig. 6 are the first and last left images of one of the stereo sequences acquired for the experiment (refer to Section II. and Fig. 1 for more explanation of the acquisition procedure).

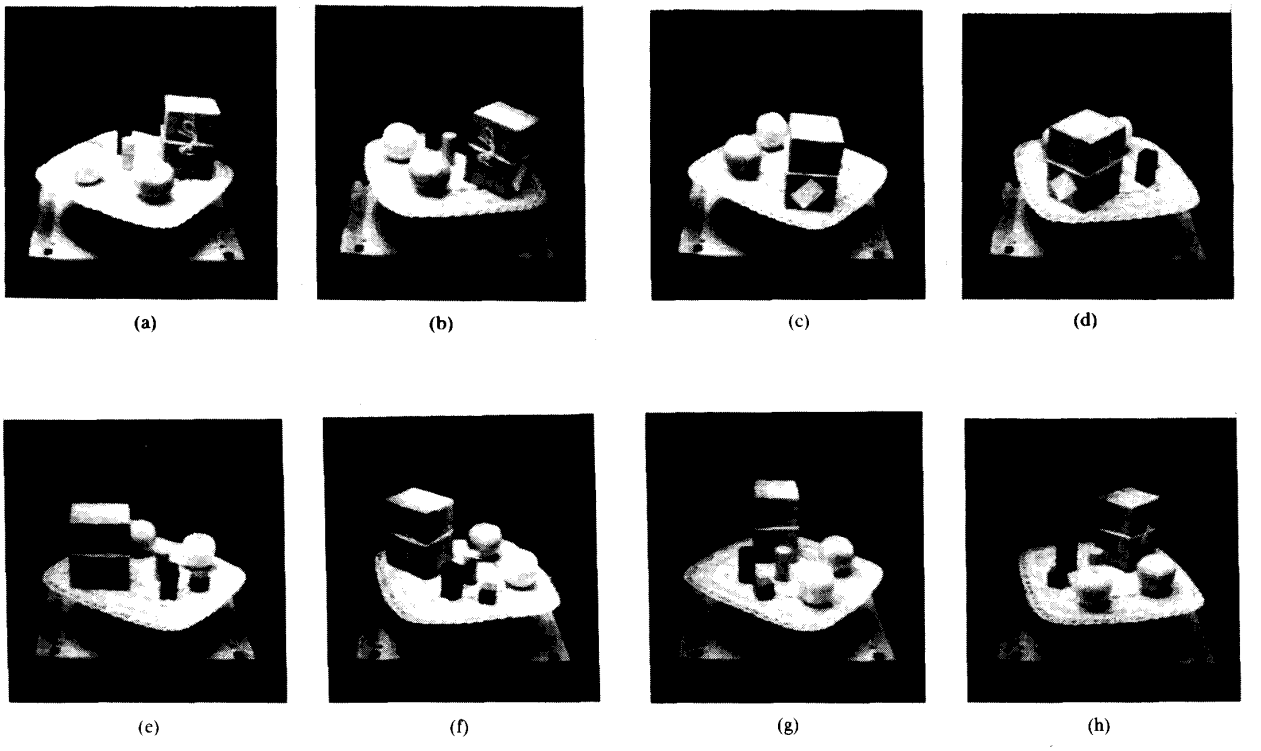


Fig. 6. First and last image of four sequences of 11 images used in experiment. Fixation point was kept still during motion of camera. Resolution of images is 256×256 pixels.

The zero crossings extracted from the first, third, fifth, seventh, and ninth left images are presented in Fig. 7(a) superimposed (the value of σ used is four pixels). The contours corresponding to image noise were eliminated using an hysteresis threshold on the average slope measured at the edge points [23]. A representation of the global optic flow is given in Fig. 7(b), the displayed vectors are evenly spaced along image contours and were obtained from the original flow by taking one vector every other vector.

The distance of the objects in the scene is determined from the optic flow and the known egomotion parameters

$$|Z_m + W_Z| = \frac{D_f W_Z}{|\vec{V}_t|}. \tag{2}$$

$\vec{V}_t$ is the component of the image velocity vector due to camera translation: it is computed by subtracting the rotational component $\vec{V}_r$ from the whole velocity $\vec{V}$; D_f is the displacement of the considered contour point from the focus of expansion (FOE) or the focus of contraction (FOC)[1]; W_Z is the velocity of the camera along the optic axis; Z_m is the distance of the world point from the camera along the direction of the optic axis.

The information acquired at the contour points is not a depth map; in the experiments we compute a dense depth map by interpolating the values at the contour points in a linear manner. The procedure described in this section has been applied to the optic flow of the analyzed sequence. The velocity of the camera W_Z and the position of the FOE were computed from the known egomotion parameters D_1, D_2, and the rotation angles θ and ψ (the camera did not rotate about the X axis).

In Fig. 7(c) the depth map of the analyzed scene is presented. It was obtained, for all image points, by a linear interpolation of the depth values computed at contour points. The gray level is proportional to depth. The uncertainty relative to the depth map shown in Fig. 7(c) is presented in Fig. 7(d); the intensity is proportional to the uncertainty of the measured depth.

III. Depth Uncertainty and Data Integration

Numerous sources of noise must be considered in processing images. Geometric distortions of the sensors and aliasing, in addition to the inevitable discretization of the image plane, are among the most important causes of errors. Also, stereo geometry and the egomotion parameters are among the potential sources of error. The reliability of these parameters is directly related to the accuracy of the measurement device.

[1]A *focus* in the optic flow represents the intersection of an imaginary straight line, along the direction of translation and passing through the *convergence point* of the optical system, with the image plane. In particular, in the case of egomotion, a FOE is produced if the camera moves towards the scene, then the velocity vectors are radiating from the focus; if the camera is moving away we have a FOC, and the velocity vectors collapse on it.

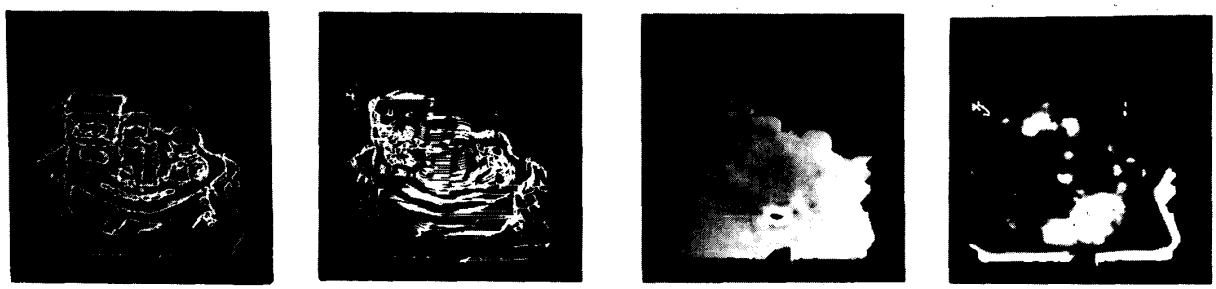


Fig. 7. (a) Zero crossings of four images of sequence. (b) Optical flow. (c) Depth map obtained by linear interpolation of depth values computed from optical flow. (depth is proportional to gray level). (d) Associated uncertainty (uncertainty is proportional to gray level).

These errors should be taken into account estimating range from images. The problem can be faced by determining, for each parameter, an uncertainty region or a suitable probability distribution (Gaussian normal, triangular, etc.) [24]–[26].

Stochastic models of visual processes have already been proposed in the past [25], [27]–[29]. In particular, concerning 3-D vision, research has been mainly devoted to the refinement of depth through uncertainty measurements. Many researchers have identified Kalman filtering as a viable solution for this problem, as it explicitly incorporates a representation of uncertainty and allows incremental refinement of the measurement over time. Ayache and Faugeras [30] developed an elegant formalism (extended Kalman filter) to build and refine a 3-D representation of an observed scene. While it is very powerful, their system, which is based on the matching of sparse object features like points and lines, is not suitable to represent volumetric objects and their spatial occupancy, though it is most appropriate for path planning in robotic navigation. Matthies and Kanade [31], [26] applied the Kalman formalism for motion estimation from stereo images and for incremental depth measurement from known camera motion. Only two special cases of camera motion are considered, which, on the other hand, seem to maximize the accuracy in depth estimation (relative to other translational trajectories). However, the system computes only a depth map of the scene; a 3-D description of objects is not provided nor is an explicit integration between stereo- and motion-derived information performed.

Poggio and his associates at MIT [32], [33] followed another line of research, investigating random Markov fields (RMF's) as a tool for merging visual modalities, detecting discontinuities in image features, and incrementally refining a representation of the observed scene. This approach, first proposed by Geman and Geman [29], is general, as it can be applied to heterogeneous image measurements, and provides a nice representation for both dense feature maps and their discontinuities. Up to now the system has been limited to integration of visual maps without an explicit representation of volumetric objects.

In this paper we present a method for the integration of range data computed from stereo matching and optical flow, based on the incremental accumulation of dense depth maps, with their uncertainty, into a 3-D (volumetric) representation of the observed scene. The 3-D model of the objects is continuously updated according to a simplified version of Kalman filter.

In our scheme some quantities are directly measured (position and intensity of pixels, stereo and egomotion parameters) to compute the depth value. Starting from the uncertainty of the measured quantities, the goal is then to obtain an uncertainty measure of depth. A further purpose is to strengthen the estimate of the values of the uncertain quantities, for example, using repeated measures of the same object or integrating different sensor modalities. In the same way, the correct position of edge points depends upon the procedure of contour extraction, while the uncertainty in the disparity and velocity measures are determined according to the computational model.

A. Uncertainty Analysis

In Section II-B we addressed the computation of visual motion, aimed at determining depth maps from dynamic views of a static scene. The proposed approach is subject to errors due to the finite accuracy of the measured parameters and, particularly, to the computational scheme. To evaluate, and eventually reduce, the amount of errors in the computed parameters (optical flow and depth), a statistical analysis of the computational process has been performed. The basic idea is that of considering the measurement process as stochastic, where the state variables are Gaussian with known or measurable variance and mean values corresponding to their actual values.

A generic function $h(\cdot)$ of $s(i, j)$ variables

$$Z = h[s(1,1), \cdots, s(M-1, N-1)]$$

produces, by linear approximation, a Gaussian output statistic with mean and variance defined as follows:

$$\bar{Z} = h[\bar{s}(1,1), \cdots, \bar{s}(M-1, N-1)]$$

$$\sigma_Z^2 = \boldsymbol{J}\boldsymbol{V}\boldsymbol{J}^{\boldsymbol{T}} \tag{3}$$

where $\boldsymbol{V}$ is the covariance matrix for the sequence $s(i, j)$ and $\boldsymbol{J}$ is the Jacobian of the function $h(\cdot)$.

To estimate depth from motion, the unknown depth is expressed as a function of the known parameters, as stated

by (2):

$$Z_m = Z_m(x, y, V_{tx}, V_{ty}, W_Z, D_f, F) \tag{4}$$

where x and y are the coordinates of the considered point on the image plane, W_z is the component of camera velocity along the optic axis, which corresponds to the Z axis referred to the camera coordinate system, V_{tx} and V_{ty} are the components of the image velocity vector due to the translation of the camera, D_f is the position of the FOE with respect to the image point (x, y), and F is the focal length of the camera.

Considering (2), the distance D_f of a pixel P_i from the FOE can be expressed in the following way:

$$D_f = |\mathrm{FOE}_x - x, \mathrm{FOE}_y - y| = \frac{|FW_X - xW_Z, FW_Y - yW_Z|}{|W_Z|}$$

assuming that

$$N = \sqrt{[FW_X - xW_Z]^2 + [FW_Y - yW_Z]^2}\,; \tag{5}$$

then (2) can be rewritten as

$$Z_m = \frac{N}{|\vec{V}_t|} + W_z = \frac{M}{|\vec{V}_t|}$$

$$M = N + |\vec{V}_t| W_Z. \tag{6}$$

Moreover, assumed the tracking egomotion strategy, the camera translation is computed from the parameters D_1, D_2, ϕ, and θ (refer to Fig. 5):

$$W_X = D_2 \cos\phi \sin\theta$$
$$W_Y = D_2 \sin\phi$$
$$W_Z = D_1 - D_2 \cos\phi \cos\theta$$

The translational component of the image velocity is computed as difference between the optical flow and the rotational component of image velocity, which is computed from the rotation angles of the camera ϕ, θ, ψ. Hence the depth function Z_m results:

$$Z_m = Z_m(x, y, V_x, V_y, D_1, D_2, \phi, \theta, \psi, F).$$

Considering all the state variables as Gaussian and uncorrelated, the mean value of depth is assumed equal to Z_m, while its variance is expressed, using a linear approximation, as

$$\begin{aligned}\sigma_Z^2 = {} & N^2 \left[\frac{\partial}{\partial Vx}\frac{1}{|\vec{V}_t|}\right]^2 \sigma_V^2 x + N^2 \left[\frac{\partial}{\partial Vy}\frac{1}{|\vec{V}_t|}\right]^2 \sigma_{Vy}^2 \\ & + \left[\frac{\partial}{\partial x}\frac{N}{|\vec{V}_t|}\right]^2 \sigma_x^2 + \left[\frac{\partial}{\partial y}\frac{N}{|\vec{V}_t|}\right]^2 \sigma_y^2 + \left[\frac{\partial}{\partial F}\frac{N}{|\vec{V}_t|}\right]^2 \sigma_F^2 \\ & + \frac{1}{|\vec{V}_t|^2}\left[\frac{\partial}{\partial D_1} M\right]^2 \sigma_{D1}^2 + \frac{1}{|\vec{V}_t|^2}\left[\frac{\partial}{\partial D_2} M\right]^2 \sigma_{D2}^2 \\ & + \left[\frac{\partial}{\partial \theta} Z_m\right]^2 \sigma_\theta^2 + \left[\frac{\partial}{\partial \phi} Z_m\right]^2 \sigma_\phi^2 + N^2 \left[\frac{\partial}{\partial \psi}\frac{1}{|\vec{V}_t|}\right]^2 \sigma_\psi^2 \end{aligned} \tag{7}$$

where σ_Z^2 represents the variance of depth, σ_x^2, σ_y^2 represent the errors in the localization of the contour point, and σ_F^2 is the variance of the computed focal length of the camera expressed in pixels. σ_{D1}^2, σ_{D2}^2 and σ_θ^2, σ_ϕ^2, σ_ψ^2 are the variances of the known egomotion parameters of the camera (i.e., the distances of the camera from the fixation point D_1, D_2 and the rotational angles ϕ, θ, ψ). These variances depend upon the accuracy of the measurement devices, while the variance of the focal length is obtained from the characteristics of the imaging sensor (position of the image center, deviation of the optical axis, etc.). The variance of the pixel position $[x, y]$ corresponds to the error in the localization of the contour, due to the $\nabla^2 G$ filtering (which is assumed to be approximately equal to half the standard deviation σ of the mask [34]). A better approximation can be obtained computing the statistic of the image (see, for example, [35]).

The right side on the first line of (7) constitutes the part that makes explicit the dependency of depth uncertainty from the variance of the optical flow. This is the last factor to be determined to estimate the uncertainty of Z_m.

The model underlying the estimation of visual motion is quite complex, as it involves an initial differentiation to recover the component $V^\perp$ of velocity, followed by a computation of the direction of velocity from proprioceptive data (the egomotion parameters) and a final matching procedure to refine the estimation. The final flow of a long sequence is obtained accumulating the partial flow fields relative to image pairs.

An accurate analysis should take into account the propagation of the errors due to the matching and the accumulation processes [36]. As to the present paper the variance of the flow field is determined from these observations.

- Corresponding contour points should exhibit the same characteristics (such as edge slope and local orientation).
- The velocity of a straight edge segment cannot be determined if it is moving along a direction parallel to its orientation.
- The accuracy in the estimation of the component $V^\perp$ (which is used to compute the optical flow) is inversely proportional to the edge slope. In fact,

$$V^\perp = -\frac{\partial I^g/\partial t}{|\nabla I^g|} \Rightarrow E_V = \frac{\partial I^g/\partial t}{|\nabla I^g|^2} E_{\nabla I} - \frac{1}{|\nabla I^g|} E_{It}$$

  where I^g represents a zero crossing point of the image I filtered with the $\nabla^2 G$ operator, $|\nabla I^g|$ is the edge slope, E_V, $E_{\nabla I}$, and E_{It} are the measurement errors of the component $V^\perp$ of velocity, the edge slope, and the image time derivative $I_t = \partial I^g/\partial t$, respectively.

Taking these considerations into account, the variances of the x and y components of the flow field are determined in the following way:

$$\sigma_{Vx}^2 = V_x^2 \left[\frac{1}{|\nabla I^g|^2} + \Delta\eta^2 + \frac{1}{\Delta\omega^2}\right]\frac{1}{k}$$

$$\sigma_{Vy}^2 = V_y^2 \left[\frac{1}{|\nabla I^g|^2} + \Delta\eta^2 + \frac{1}{\Delta\omega^2}\right]\frac{1}{k}. \tag{8}$$

$\Delta\eta$ is the difference between the local orientation of the contour and the direction of the velocity vector $\vec{V}$, $\Delta\omega$ is the difference between the orientation of the corresponding contour points in the first and last frame of the sequence, and k is a normalizing factor bounding $[(1/|\nabla I^g|^2)+\Delta\eta^2+(1/\Delta\omega^2)](1/k)$ between zero and one.

Processing stereo images to estimate environmental depth involves the comparison of the gray-scale maps (to compute image disparity) and the use of external parameters relative to the acquisition sensors (to compute depth).

As in the case of depth-from-motion, the stereo algorithm provides an uncertain measure of distance:

$$Z_s = Z_s(l, m, \theta, F, x_0, d).$$

According to (3), the variance of depth is computed using a linear approximation:

$$\sigma_z^2 = \left[\frac{\partial Z_s}{\partial m}\right]^2\sigma_m^2 + \left[\frac{\partial Z_s}{\partial l}\right]^2\sigma_l^2 + \left[\frac{\partial Z_s}{\partial \theta}\right]^2\sigma_\theta^2 + \left[\frac{\partial Z_s}{\partial f}\right]^2\sigma_f^2 + \left[\frac{\partial Z_s}{\partial x_0}\right]^2\sigma_{x0}^2 + \left[\frac{\partial Z_s}{\partial d}\right]^2\sigma_d^2. \quad (9)$$

All parameters are considered stochastic Gaussian variables with known variance: among them the stereo pair parameters (θ, lm) are dynamically computed during image acquisition, whereas F, which is more properly a camera parameter, can be derived from camera–lens technical specifications. The uncertainty of these quantities depends on the accuracy of the calibration phase, and it is partially affected by external factors or noise (measurement errors). The position of the point x_0 and its uncertainty, depend upon the procedure of contour extraction.

Analyzing the case of egomotion, we have investigated the influence of the errors in the measured parameters (as well as of visual motion) in the estimation of depth. For the stereo algorithm we introduce some different considerations to determine the uncertainty of image disparity. An uncertainty estimate is possible, in this case, only in the presence of some precise simplifications and assumptions, as follows.

- An image is a spatially nonstationary random process: we can still use a simple description such as

$$f(i, j) = \bar{f}(i, j) + s(i, j), \qquad 0 \div M-1, 0 \div N-1.$$

- $s(i, j)$ is a high-frequency component which has stationary statistics and a mean value equal to zero. In more detail $s(i, j)$ are uncorrelated Gaussian variables in the range $(0 \div M-1, 0 \div N-1)$.

$$p(s) = k\exp\left[-\frac{1}{2}s^T\boldsymbol{R}_s^{-1}s\right].$$

- The correlation function, which depends on disparity d, is modeled in exponential form [3]. If a value of disparity d_0 corresponds to the peak of the correlation function, we can write

$$R(d) = A^2\exp[-\alpha|(d-d_0)|]. \quad (10)$$

A is the maximum intensity value of the images, while α determines the peak amplitude. The value of α can be easily stated considering that $R(d)$ should not have central amplitude greater than the amplitude of the convolution mask used for the last step of the pyramidal correlation process.

Under these assumptions it is possible to evaluate the disparity variance in relation to the statistical parameters used in the stereo algorithm. The correlation between two images I_1 and I_2, performed on a patch with dimension $H \times K$ shifted horizontally by τ can be expressed as

$$\begin{aligned} R(\tau, i, j) &= \frac{1}{HK}\sum_{m=0}^{H-1}\sum_{n=0}^{K-1} I_1(i+m, j+n)I_2(i+m+\tau, j+n) \\ &= \frac{1}{HK}\sum_{m=0}^{H-1}\sum_{n=0}^{K-1} \bar{I}_1(i+m, j+n)\bar{I}_2(i+m+\tau, j+n) \\ &\quad + S_1(i+m, j+n)S_2(i+m+\tau, j+n) \\ &\quad + \bar{I}_1(i+m, j+n)S_2(i+m+\tau, j+n) \\ &\quad + S_1(i+m, j+n)\bar{I}_2(i+m+\tau, j+n) \end{aligned} \quad (11)$$

where

$$E[R(\tau, i, j)] = \frac{1}{HK}\sum_{m=0}^{H-1}\sum_{n=0}^{K-1} \bar{I}_1(i+m, j+n) \cdot \bar{I}_2(i+m+\tau, j+n).$$

$R(\tau, i, j)$ is a function of $S_1(i+m, j+n)$ and $S_2(i+m+\tau, j+n)$ statistical variables for each m and n. The variance of $R(\tau, i, j)$ is obtained differentiating (11) with respect to S_1 and S_2:

$$\sigma_R^2(\tau, i, j) = \sum_{m=0}^{H-1}\sum_{n=0}^{K-1}\left[\frac{\bar{I}_1(i+m, j+n)}{HK}\right]^2 \cdot \operatorname{var} S_2(i+m+\tau, j+n) + \left[\frac{\bar{I}_2(i+m+\tau, j+n)}{HK}\right]^2 \cdot \operatorname{var} S_1(i+m, j+n)$$

where norm is a normalization factor for I_1 and I_2.

It is worth noting that considering images filtered with a $\nabla^2 G$ operator, as in the algorithm described in Section II-A, the variance of the correlation does not increase. In fact, as the $\nabla^2 G$ operator preserves the range of intensity values of the image, the filtering does not increase the variance of the image.

Posing $\operatorname{var} S_1(i+m, j+n) = \operatorname{var} S_2(i+m+\tau, j+n) = \sigma^2$ and bounding the range of $\bar{I}_1(i, j)$ and $\bar{I}_2(i, j)$ with A we obtain

$$\sigma_R^2(\tau, i, j) \leqq \sum_{m=0}^{H-1}\sum_{n=0}^{K-1}\frac{2\sigma^2A^2}{H^2K^2} = \frac{2\sigma^2A^2}{HK}.$$

In our case the side of the patch is $H = K = 4w$, hence

$$\sigma_R^2(i, j) < \frac{A^2\sigma^2}{8w^2} \tag{12}$$

which approximates the variance of the correlation between two images.

The variance of disparity is now computed inverting (10) and differentiating $d(R)$ with respect to R:

$$d = d_0 \pm \frac{\ln R - \ln A^2}{\alpha}$$

$$\sigma_d^2 = \left[\frac{\partial d}{\partial R}\right]^2 \sigma_R^2 = \frac{1}{\alpha^2 R^2}\sigma_R^2. \tag{13}$$

Substituting (12) in (13), we obtain

$$\sigma_d^2 < \frac{A^2\sigma^2}{R^2\alpha^2 8w^2} \tag{14}$$

where α is the width of the peak of the correlation function, R is the estimated maximum of the correlation function, A is the maximum intensity value of the image, w is the amplitude of the central lobe of the $\nabla^2 G$ mask used for the filtering of the stereo pair, and σ is the variance of the $s(i, j)$ process (which corresponds to the image noise).

B. Volumetric Integration

Both the stereo and motion algorithm described in the previous paragraphs provide a depth map of the scene from the same view point. The peculiar errors of each algorithm are coded in the uncertainty measure associated with the computed depth.

As explained in Section II, four depth maps and the associated uncertainty maps were computed for both stereo and motion. The integration of stereo and motion is performed by projecting into a 3-D voxel representation of space the stereo and motion *bas reliefs*, weighted with the relative uncertainty measures.

In other words, each partial representation embedded in the visual *bas reliefs* can be used to generate and/or update a full 3-D representation of the solids in view. As in the present paper, the 3-D integration is performed using the *bas reliefs* obtained from different viewpoints to *carve* a 3-D array of voxels representing the viewed space.

If the depth of an image point $\bar{p}$ has a variance σ_p, then we can suppose that along the line of sight crossing that image point the space is entirely empty for distances less than $(\bar{p} - \sigma_p)$ and vice versa: the space is full for distances greater than $(\bar{p} + \sigma_p)$. All the intermediate values represent uncertain values. In this way a sort of rind is associated with the depth map. The thickness of this rind is proportional to σ_p.

From a procedural point of view, the accumulation of incoming depth images is performed in the following way:

- a 3-D matrix of voxels is defined representing the "work space";
- the position of the observer with respect to the work space is computed (or known);
- a line is traced from the position of the observer through the depth image, and all the voxels crossed by the line are modified according to probability of empty space.

TABLE I

x_i Current Value Stored in the Matrix	y_{i+1} New Value of Occupation Probability	x_{i+1} Final Value of Occupation Probability
U	U	x_i
O	U	x_i
S	U	x_i
U	O	y_{i+1}
O	O	x_i
S	O	x_i
U	S	y_{i+1}
O	S	y_{i+1}
S	S	update probability according to (15)

U Unknown.
O Occluded.
S Seen.

To perform the actual accumulation, we must consider that a single view cannot carry information on the space which is not seen; this occurs for the objects outside the visual field and for occluded objects. In fact, we subdivide the space in three parts.

- *Seen space* (S): This portion of the work space lies inside the visual field and, at the same time, belongs to the rind $(\bar{p} - \sigma_p) < p < (\bar{p} + \sigma_p)$. This measure is different from zero in the zones where disparity information is present.
- *Occluded space* (O): It belongs to vision cone but lies behind the rind (i.e., at $p > (\bar{p} + \sigma_p)$.
- *Unknown space* (U): This is the space external to the vision cone.

Every time a new depth map is obtained, the voxel work space is updated in accordance with Table I (initially the array is set to unknown). The update is performed, voxel by voxel, projecting the new depth map into the existing volumetric representation. If a new depth map does not add any information (*unknown* probability), the voxel is not updated. Also, if the new value if *occluded*, the voxel is not updated, unless it was already unknown (fourth row of Table I); in this case we label the voxel as occluded.

It is worth noting the operational equivalence between the occluded and unknown space. On the other hand, the distinction is not redundant because the two situations have a very different meaning: the observer has a strong interest in the occluded space, which must be reduced as much as possible (the knowledge of occluded space can be used, for example, to drive exploratory strategies). On the contrary, unknown space represents "all the universe" and must be considered in particular situations only.

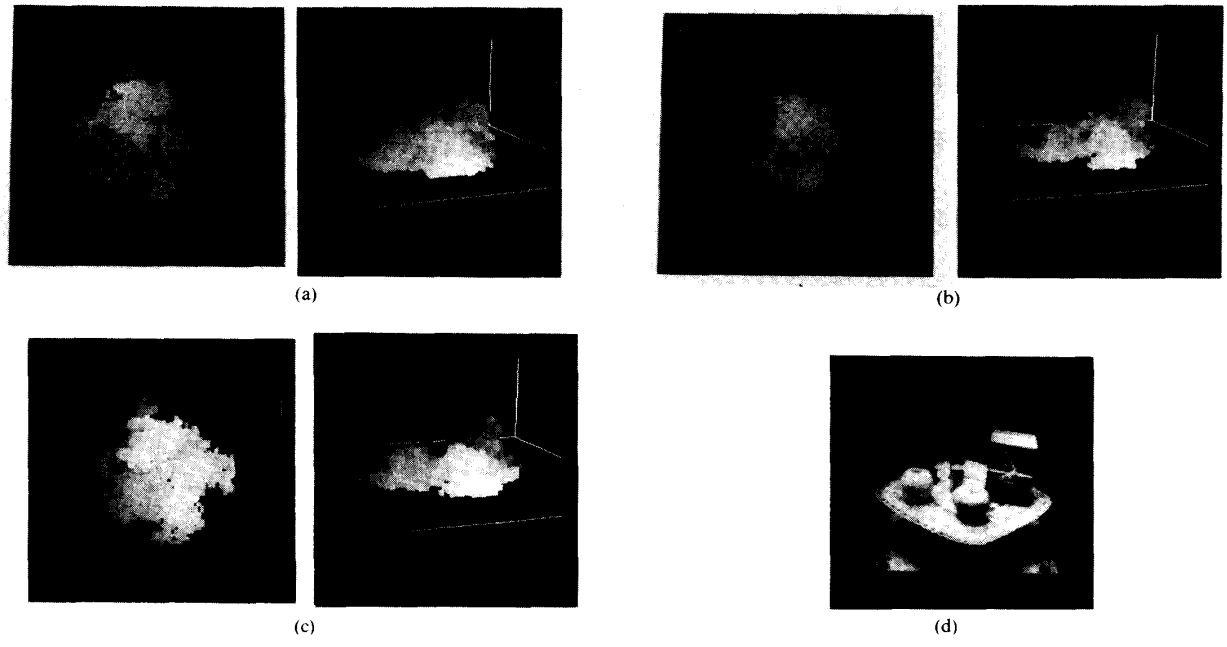


Fig. 8. 3-D volumetric integration of depth maps. Perspective representation of voxel-based accumulator. Gray-level codes distance (brighter means closer). (a) Integration of four depth maps derived from four stereo pairs of Fig. 2, (b) Integration of four depth maps derived from optical flow, (c) Integration of both stereo and motion-derived information. Left column: objects seen from above. Right column: object seen approximately from point of view of Fig. 8(d). Tall package is visible (brighter areas on top views) as are two apples. For this picture any information coming from below tray have been clipped (as consequence square base is not visible). (d) Original picture of scene observed from same point of view of perspective representation in (a), (b), and (c).

The most interesting case occurs when the new visual *bas relief* carries a new occupation probability for a voxel which labels it as *seen*: if it was unknown or occluded, then it is set to seen with the new probability value; if it was already seen, then the probability value is updated according to the new information. In the latter case, a simplified version of Kalman filter is used to update the voxel probability:

$$\begin{aligned} x_{i+1} &= x_i + k_i[y_{i+1} - x_i] \\ k_{i+1}^{-1} &= k_i^{-1} + 1 \\ k_0^{-1} &= 0 \qquad x_0 = \text{unknown}. \end{aligned} \tag{15}$$

K_i is the Kalman gain factor, x_i is the current probability of the considered voxel, x_{i+1} is the updated voxel probability, and y_{i+1} is the probability (inverse normalized uncertainty) of newly computed depth map. This expression achieves the arithmetic average among all considered probability values, updating the uncertainty associated to full space. However, it is required to store the k_i coefficient for each voxel, doubling the memory necessary to maintain a comprehensive description of the work space.

An example of integration is shown in Fig. 8; in Fig. 8(a) the integration of the four depth maps derived from stereo is shown. In Fig. 8(b) the result of the same procedure is shown for the motion-derived depth maps. In Fig. 8(c), finally, the integration of both stereo and motion-derived depth images is shown. The probability of occupation is, for the three images of Fig. 8, set to 0.35. As it can be noticed, the approximation is better in Fig. 8(c) than in both Fig. 8(a) and (b). In Fig. 9 the result obtained raising the value of probability of filled space is shown: raising the voxel probability, the volume of the solid is reduced, until the achievement of a nucleus is certainly full. Notice that the filled space increases if the certainty measure is lowered.

IV. Conclusion

The fusion of different sensor modalities constitutes one of the hot topics in today's computer vision. The main advantage of the integration of multiple sensorial outputs is the robustness of the overall measurement process; moreover, the precision of the measures can be considerably improved. This is especially important when dealing with 3-D vision and its applications in robotics, like obstacle avoidance or automatic vehicle guidance, etc. In those and other cases, it is dangerous to produce erroneous measures, while it is necessary to know their reliability.

In this paper we have presented two algorithms for recovering the 3-D structure of objects from stereo matching and motion parallax, as examples of visual processes that can independently compute environmental depth. The stereo algorithm is characterized by robustness, due to regional correlation, and precision achieved with the final edge matching. The structure-from-motion algorithm takes

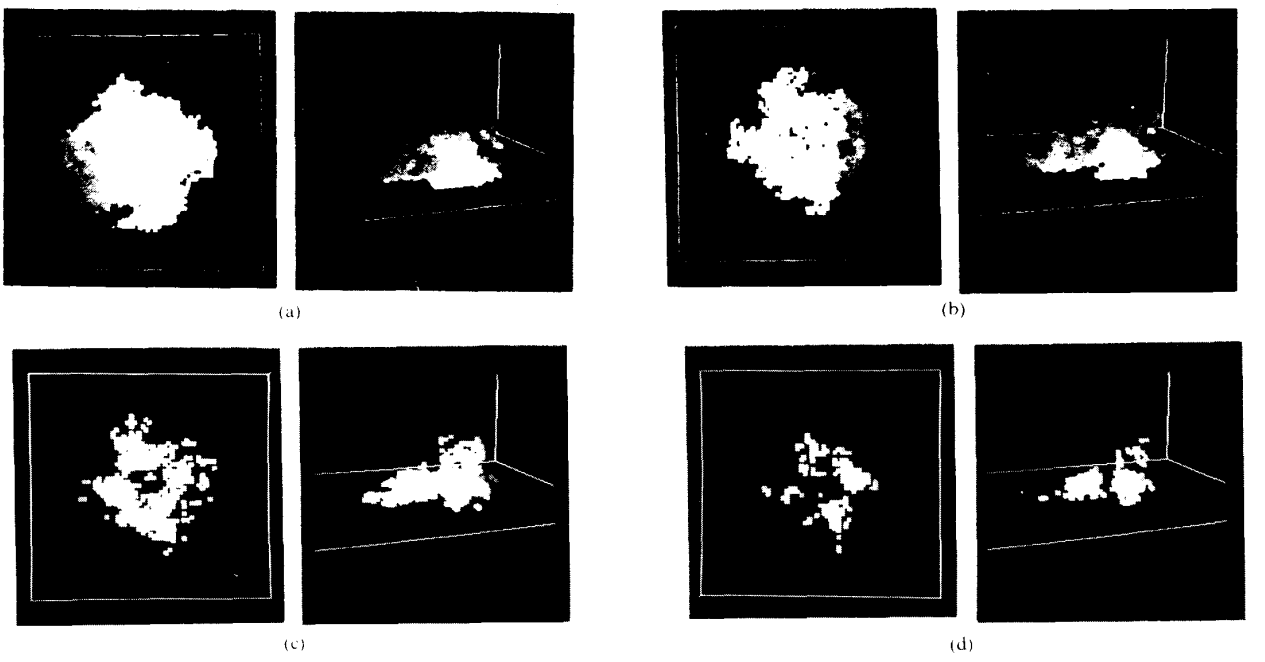


Fig. 9. Same representation as in Fig. 8. Results obtained by raising probability of filled space. From top to bottom probability changes from 0.15 to 0.75. For higher levels of probability tray disappears and only some voxels around package and two apples (i.e., largest objects) remains. Left column: objects seen from above. Right column: object seen approximately from point of view of Fig. 8(d).

advantage of a particular motion strategy of the observer, which allows the simplification of the motion constraint equations. The optic flow of a long sequence (obtained by matching successive image pairs) is used to compute depth. The depth is computed, in both cases, at the contour points; a dense depth map is obtained performing a linear interpolation of the depth values.

Both the stereo and motion algorithms provide a visual *bas relief* (a depth map) from several viewpoints, while the camera is moving around. Different measures are combined, continuously updating a volumetric description of the scene. We used a volumetric representation to model the environment: a cube of voxel, in which the *bas reliefs* are projected (with a ray-casting procedure) from the different viewpoints, carving the shape of the objects.

Each measure is characterized by an uncertainty value: the *bas reliefs* are projected into the work space weighted with their probability. To compute the uncertainty relative to each visual modality, a stochastic model of the processes has been developed in which all the parameters involved are assumed to be uncorrelated Gaussian variables with known variance and mean values corresponding to the measured values. The depth maps are integrated into the working space using Kalman filtering to update the volumetric representation.

In the experiment presented, the great advantage of multisensor integration is evident, as the final representation is more precise than that obtained using each single measurement process.

In the present paper we used only two visual processes to compute environmental depth, but the outputs of any other shape_from... algorithm or of ranging devices like ultrasound or laser range finders, could be added into the volumetric description of the world. For the future, we are developing a technique to segment the volumetric representation, isolating single objects, and to transform the voxel-based description to one based on the superficial characteristics and other geometric features of the observed scene.

REFERENCES

[1] Y. F. Wang, M. J. Magee, and J. K. Aggarwal, "Matching three-dimensional objects using silhouettes," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, no. 4, pp. 513–518, 1984.

[2] L. Massone, P. Morasso, and R. Zaccaria, "Shape from occluding contours," presented at the SPIE Symp. Intelligent Robots and Computer Vision, Cambridge, MA, November 4–8, 1984.

[3] H. K. Nishihara, "PRISM: A practical real-time imaging stereo matcher," *Opt. Eng.*,vol. 23 no. 5 pp. 536–545, 1984.

[4] D. Marr, *Vision*. San Francisco CA: Freeman, 1982.

[5] K. Prazdny, "Egomotion and relative depth map from optical flow," *Biol. Cybern.*, vol. 36, pp. 87–102, 1980.

[6] D. T. Lawton, "Processing translational motion sequences," *CVGIP*, vol. 22, pp. 116–144, 1983.

[7] T. D. Williams, "Depth from camera motion in a real world scene," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-2, no. 6, pp. 511–516, 1980.

[8] T. M. Strat and M. A. Fischler, "One eyed stereo: A general approach to modeling 3-D scene geometry," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-8, no. 6, pp. 730–741, 1986.

[9] M. Brady, "Artificial intelligence and robotics," MIT A. I. Laboratory, A. I. Memo 756, Cambridge, MA, Feb. 1984.

[10] B. K. P. Horn, "Understanding image intensities," *Artificial Intell.*, vol. 8, 1977.

[11] K. Ikeuchi and B. K. P. Horn, "Numerical shape from shading and occluding boundaries," *Artificial Intell.*, vol. 17, 1981.

[12] G. Sandini, M. Straforini, and V. Torrre, "3-D reconstruction of silhouettes," in *Proc. 4th Intl. ROVISEC*, London, UK, 1984, pp. 173–182.
[13] G. Sandini and M. Tistarelli, "Recovery of depth information: Camera motion as an integration to stereo," in *Proc. Workshop on Motion: Representation and Analysis*, Kiawah Island Resort, May 7–9, 1986. pp.39–43.
[14] P. Morasso, G. Sandini, and M. Tistarelli, "Active vision: Integration of fixed and mobile cameras," in *NATO ARW on Sensors and Sensory Systems for Advanced Robots*. Berlin, Germany: Springer-Verlag, 1986, pp. 449–462.
[15] A. Bandopadhay, B. Chandra, and D. H. Ballard, "Active navigation: Tracking an environmental point considered beneficial," in *Proc. Workshop on Motion: Representation and Analysis*, Kiawah Island Resort, May 7–9, 1986, pp. 23–29.
[16] G. Sandini, V. Tagliasco, and M. Tistarelli, "Analysis of object motion and camera motion in real scenes," in *Proc. IEEE Intl. Conf. Robotics & Automation*, San Francisco, CA, Apr. 7–10, 1986, pp. 627–633.
[17] G. Sandini, P. Morasso, and M. Tistarelli, "Motor and spatial aspects in artificial vision," in *Proc. 4th Intl. Symp. Robotics Research*, Aug. 1987 Cambridge, MA: MIT Press, pp. 351–358.
[18] A. Bandopadhay, J. Y. Aloimonos, and I. Weiss, "Active vision," *Int Comput. Vision*, vol. 1, no. 4, pp. 333–356, Jan. 1988.
[19] K. Ikeuchi, H. Nishihara, B. K. P. Horn, P. Sobalvarro, and S. Nagata, "Determining grasp points using photometric stereo and the PRISM binocular stereo system," *Int. J. Robotics Res.* vol. 5, no. 1, pp. 46–65, 1986.
[20] C. Frigato, E. Grosso, and G. Sandini, "Integration of edge stereo information," DIST-Univ. of Genoa, Esprit P419 Tech. Rep. TKW1-WP1-DI3, 1987.
[21] ____, "Extraction of 3-D information and volumetric uncertainty from multiple stereo images," *Proc. ECAI*, Műnchen, Germany, Aug. 1988, pp. 683–688.
[22] D. Marr and E. Hildreth, "Theory of edge detection," *Proc. Roy. Soc. London*, Ser. B, no. 207, pp. 187–217, 1980.
[23] G. Sandini and V. Torre, "Thresholding techniques for zero crossings," in *Proc. Winter 85 Topical Meeting Machine Vision*, Incline Village, NV. 1985, pp. ThD5-1–ThD5-4.
[24] M. A. Snyder, "Uncertainty analysis of image measurements," in *Proc. DARPA Image Understanding Workshop*, 1987.
[25] L. Matthies and S. A. Shafer, "Error modeling in stereo navigation," *IEEE Robot. Automation.*, vol. RA-3. no. 3, pp. 239–248, June 1987.
[26] L. Matthies and T. Kanade, "Using uncertainty models in visual motion and depth estimation," in *Proc. 4th Int. Symp. Robotics Research*, Santa Cruz, CA, Aug. 1987, pp. 120–138.
[27] J. Marroquin, S. Mitter, and T. Poggio, "Probabilistic solution of ill-posed problems in computational vision," *J. Amer. Statis. Assoc.*, vol. 82, no. 397, pp. 76–89, Mar. 1987.
[28] J. L. Marroquin, "Deterministic Bayesian estimation of Markov random fields with applications in computer vision," in *Proc. Int. Conf. Computer Vision*, Washington, DC, 1987.
[29] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, 1984.
[30] N. Ayache and O. D. Faugeras, "Maintaining representations of the environment of a mobile robot," *Proc. 4th Int. Symp. Robotics Research*, August 1987. Cambridge, MA: MIT Press, pp. 337–350.
[31] L. Matthies and T. Kanade, "The cycle of uncertainty and constraint in robot perception," in *Proc. 4th Int. Symp. Robotics Research*, Aug. 1987. Cambridge, MA: MIT Press, pp. 327–336.
[32] T. Poggio, "The MIT vision machine," in *Proc. DARPA Image Understanding Workshop*, Cambridge, MA, Apr. 1988, pp. 177–198.
[33] E. Gamble and T. Poggio, "Integration of intensity edges with stereo and motion," MIT A.I. Laboratory, Boston, MA, A.I. Memo 970, Feb. 1984.
[34] A. Huertas and G. Medioni, "Detection of intensity changes with subpixel accuracy using Laplacian–Gaussian masks," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-8, no. 5, pp. 651–664, Sept. 1986.
[35] L. Matthies, R. Szeliski, and T. Kanade, "Kalman filter-based algorithms for estimating depth from image sequences," Carnegie–Mellon Univ., Pittsburgh, PA CMU-RI-TR-88-1, 1988.
[36] M. Tistarelli and G. Sandini, "Uncertainty analysis in visual motion and depth estimation from active egomotion," in *Proc. IEEE/SPIE Intl. Conf. Applications of Artificial Intelligence VII*, Orlando, FL, Mar. 28–30, 1989.

Enrico Grosso was born on November 29, 1963, in Serravalle, Italy. He received the degree in electrical engineering in 1987 from the University of Genoa. He is a Ph.D. student at the Dipartimento di Ingegneria Elettrica of the University of Palermo.

Since 1985 he has been working at the "Dipartmento di Informatica, Sistemistica e Telematica" of the University of Genoa on the topic of artificial vision with particular emphasis on stereo analysis and 3-D reconstruction from multiple views.

Giulio Sandini was born on September 7, 1950 in Correggio, Italy. He received a degree in electrical engineering in 1976 from the University of Genoa, Italy.

Since 1976 he has worked on models of the visual system and on electrophysiology of the cat visual cortex at the Laboratorio di Neurofisiologia del CNR di Pisa. In 1978 and 1979 he was a Visiting Scientist at the Harvard Medical School in Boston, developing a system for topographic analysis of brain electrical activity. Since 1980 he has worked on Image Processing and Computer Vision particularly in the areas of low-level vision and feature extraction, at the Department of Communication, Computer and Systems Science of the University of Genoa, where he is currently a Associate Professor.

Massimo Tistarelli was born on November 11, 1962 in Genoa, Italy. He received a degree in electronic engineering from the University of Genoa, Italy. He is currently a Ph.D student.

Since 1984 he has worked on Image Processing and Computer Vision at the Laboratory of Robotic of the Department of Communication, Computer and Systems Science of the University of Genoa.

In 1986 he was a Research Assistant at the Department of Computer Science of Trinity College, in Dublin, developing a system for the analysis of image data, mainly aimed at the investigation of low-level visual processes.

In 1989 he was a visiting scientist at Thinking Machines Co., developing parallel algorithms for dynamic image processing on the Connection Machine. His research interests include Robotics, Artificial Intelligence, Image Processing and Computer Vision particularly in the area of three-dimensional and dynamic scene analysis.