Project P419

Cooperation of Stereo and Motion for 3D Data Acquisition and Representation

C. Frigato [1], E. Grosso [1], G. Sandini [1], M. Tistarelli [1], & D. Vernon [2]

(1) Department of Communication, Computer and Systems Science
University of Genoa
Via Opera Pia 11a, I-16145 Genova, Italy

(2) Department of Computer Science
Trinity College
Dublin, Ireland

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.
In order to obtain a 3D 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 visuo-motor 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.
An "early vision" system is described to illustrate how this computational paradigm can be effected in a coherent and integrated environment: this system forms the basis by which the generation of this 3-D information is actively controlled and used by higher-level cognitive modelling sub-systems.

1. Introduction

One of the primary goals of project P419 is the integration of different visual modalities for the extraction of volumetric data. The rationale comes directly from the observation that visual information is intrinsically noisy and, often, ambiguous if considered in its single components (motion, texture, disparity, etc.).
Moreover it is now widely understood that almost all *computationally reasonable* algorithms that been proposed to acquire information about the three-dimensional shape of the "world" [1,2,3,4,5,6,7,8,9] suffer for errors and uncertainities, peculiar to each method. For example, the illumination condition is a weak point in deriving shape from shading and the computation of stereo disparity fails when the matching is performed on edges parallel to the epipolar lines10] . Another weak point of many depth estimation

algorithms has been the use of small amount of information (two or three images at most). This produces unreliable results very often requiring "high level reasoning" to be of some use. The robustness of depth estimation, on the other hand, can be increased by "simply" using more information taking into account the almost continuous (in space and time) nature of visual data.

2. Early processing

This continuous flow of information can, and very often will, provide contradictory and ambiguous measures because of the intrinsic approximation of each visual modality. So the problem is how to solve ambiguities and how to merge (integrate or fuse) contradictory measures. Let us suppose, for example, that the depth of an environmental point is measured differently by the stereo modality and by the motion one. Which one is the most accurate? How can we took advantage of both measures? Our approach is to compute, at each step of the processing level, not only the measure itself but also its reliability (or uncertainty) [11, 12, 13, 14] . The integration process starts from the independent measures performed by each modality and "combines" them according to each individual reliability. In order to use this approach for multimodal integration we must not only be able to measure uncertainty, but also to express the output of each modality in the same unit of measure.

The basic *assumptions* of our approach are:

- the visual process can be subdivided into a set of parallel processes each exploiting a single modality;
- the system has knowledge of (or can compute) its motor/proprioceptive data;
- each modality runs independently (and asynchronously) of the others.
- each modality provides a measure of uncertainty for each of its outputs;
- the output data as well as the uncertainty measure can be expressed in the same metric space (e.g. depth, time−to−impact etc.);

Given these assumptions the integration process can be seen as "simply" an average of the output of each modality weighted by uncertainty. In the example presented here, for example, the output of each modality is represented by a depth image and an uncertainty image. This image pair is used by the integration process, along with the measure of the position of the observer in space, to build and maintain an object centered "voxel" representation of the environment. A notion of an "uncertain world" is used in which each element (in our case a voxel) codes the ongoing probability of filled space. Note that this approach relies heavily on the paradigm of dynamic vision.

Given these assumptions it is clear that each modality can be studied and implemented independently of the others and that the synchronization of each modality is not necessary (each modality has its own pace and integration time). Moreover the volumetric object−centered representation acts as an accumulator of the incoming visual information (i.e. acts incrementally: if new information comes it is "simply" added, no matter what the past history has been). It is also worth noting that the output of each single modality *does not need* to be accurate as long as an estimate of the accuracy can be obtained.

In this schema different processes may have common inputs (e.g the contours extracted or the proprioceptive information about the position of the sensor(s)) but the basic requirements is that all the processes must end up by producing comparable representations of the world. This representation is represented by three distinct pieces of information: an iconic part (a depth image), a proprioceptive part (the position and orientation of the observer), and an uncertainty part (the reliability of each depth measure). We call the ensemble of these measures a *visual bas-relief*. The "integrator" continuously modify the volumetric representation according to the outputs of the different processes.

In the following we present an example of the integration of range data computed from stereo matching and optical flow, ending with a 3D (volumetric) representation of the solids in view. The rationale for this choice derives from the fact that many visual tasks require the building of a volumetric representation. The most common, yet very general, is navigation where most of the tasks can be based on a map of free space but also most manipulation tasks require "simple" volumetric representations (can this object be grasped? How and where it is best grasped? Can these objects be stacked? and so on)

2.1. Estimation of depth from stereo

The experimental set-up is based on a pair of cameras, with 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).

The *stereo matching* algorithm follows a *coarse to fine* approach [17] and is 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. The estimation of point-to-point correspondence is performed, using a coarse-to-fine approach, in three successive steps. At the end of each step a measure of disparity is obtained which is successively refined during the following 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).
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. The disparity value computed from the regional correlation is used to limit the search space of edge points. The correlation measure is weighted using measures of slope and spatial orientation of the gradient computed over the filtered image.
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 [18].

A planar model is used to determine depth from stereo. In fig. 1, 4 stereo pairs are presented; they have been acquired moving a pair of stereo cameras around a set of objects, tracking a fixed point on the surface of the central object. The distance from the fixation point in space was kept constant during the egomotion and it is about 103 centimeters. The angular displacement between successive image pairs in fig. 1 is 40 degrees

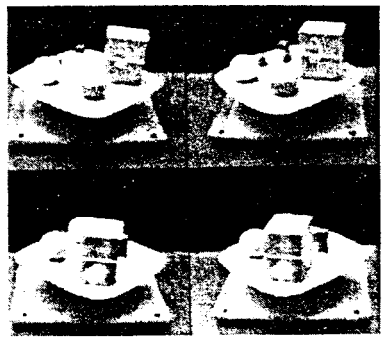

Figure 1. Stereo pairs used in the experiment. The resolution of the images is 256x256 pixels.

The images are 256x256 pixel with eight bits of resolution in intensity.
The depth map obtained from the first stereo pair is presented in fig. 2a. 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. 2b; this measure reflects the reliability of the computed depth.

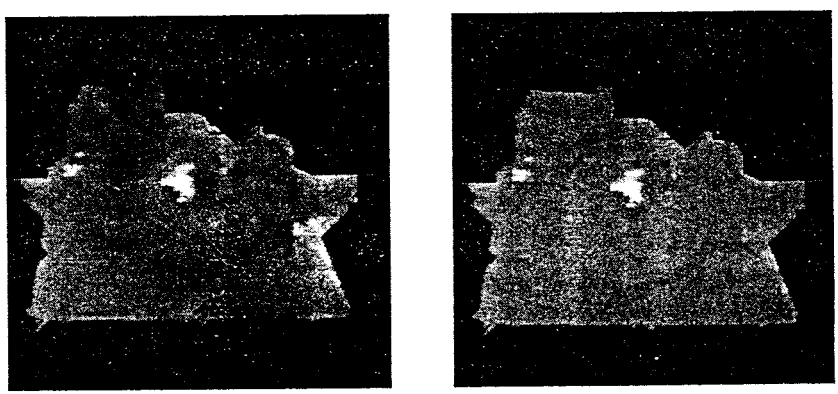

Figure 2. Results of the stereo algorithms relative to a single view (the topmost pair of Fig. 1). a) depth map obtained by linear interpolation of the contour values. b) the associated uncertainty (uncertainty is directly proportional to gray level).

2.2. Estimation of depth from motion

As far as motion analysis is concerned, a constrained egomotion strategy has been exploited, namely, the tracking of an environmental point during motion. The extraction of depth information is performed along the contours (zero−crossings) extracted from the image sequence. In order to take advantage of the continuous nature of flow information and at the same time to be able to compute depth with a reasonable accuracy, the algorithm first performs a "tracking" phase (to provide a sufficiently long base for triangulation) and then computes depth.

The algorithm is performed by means of the following steps [19]:

► Computation of the velocity component perpendicular to the local orientation of the contour $v^{\perp}$.

► Computation of the *true direction of motion.*

► Matching of corresponding contours of successive image pairs (*instantaneous optic flow*).

A confidence measure is associated to 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.

Through the edge matching a sort of *instantaneous optic flow* is determined, relative to each pair of frames. In order to achieve a sufficient range of velocity for distance computation, the instantaneous optic flows are joint together obtaining a *global optic flow*, relative to a part of the sequence, from image T to T+N, where N is the time span that must be considered for

The images in fig. 3 are the first and last left images of a sequence of 11 stereo pairs, acquired moving the camera tracking a fixed point in space. The rotation performed between successive positions is about 4 degrees. It is worth noting that the stereo pair processed by the stereo algorithm is the 3rd of the sequence.

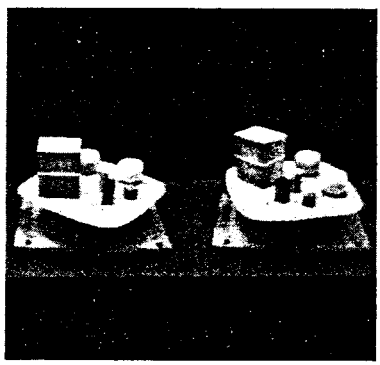

Figure 3. First and last image of the four sequences of 11 images used in the experiment. The fixation point was kept still. The resolution of the images is 256x256 pixels.

A representation of the global optic flow is given in fig. 4, the displayed vectors are evenly spaced along image contours and were obtained from the original flow by taking one vector every 2.

The last step of the motion algorithm is the determination of the distance of the objects in the scene. The computation is done straightforwardly from the knowledge of the optic flow, by a triangulation between corresponding points of the first and last frame of the interval within the sequence.

In fig. 5a 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. 5a is presented in fig. 5b.

3. Integration of stereo and motion

In spite of the great deal of information contained in each measure, obtained from the different view points, it is clear that from a "bas−relief" (i.e. a depth image) only a partial description of 3D shape can be derived. In order to complete and refine this information it is necessary to "move−around", exploring actively the environment [20] . This is true not only because occluded objects can become evident from different point of views 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 view−points.
In principle, the continuous flow of information, represented as continuously changing depth images derived from stereo measures and motion parallax, needs to be casted into an incremental representation. This casting process acts like an accumulator where only the "new" information changes the current description whereas the redundant information does affect it.

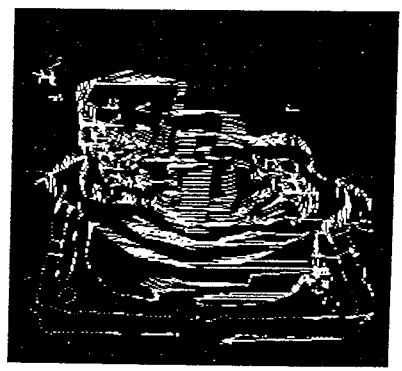

Figure 4. optical flow.

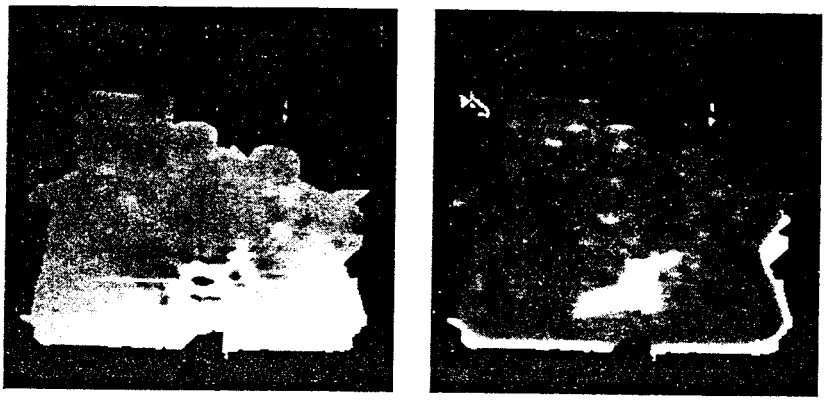

Figure 5. a) depth map obtained by linear interpolation of the depth values computed from the optical flow. b) the associated uncertainty (uncertainty is directly proportional to gray level).

In our approach the casting process makes use of a geometric description of the world in terms of a 3D array of voxels. The partial information acquired from each view point is used to update this volumetric description.
It is evident that a correct integration of different measures must be based on an estimate of "reliability" associated to each measure. For this reason the stereo and motion algorithms provide, besides depth, an estimate of uncertainty used to "weight" the current depth measure with respect to the accumulated one. At each instant of time each voxel of the accumulator stores a 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 (like, for example, recognition) on the other hand its only use is to help the accumulation of depth information and moreover, more "complete" geometric description, including also surface information, can be derived from the voxel representation and logically linked to it. As to the present paper the 3D integration is performed using the bas-reliefs, obtained from different view points and modalities, to *carve* a 3D array of voxels representing the viewed space.

If the depth of an image point $\overline{p}$ 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 $(\overline{p} - \sigma_p)$ and vice versa the space is full for distances greater than $(\overline{p} + \sigma_p)$. All the intermediate values represent uncertain values. In this way a sort of rind is associated to 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 3D 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.

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* (the portion of the work space lying inside the visual field and, at the same time, belonging to the rind $(\overline{p} - \sigma_p) < p < (\overline{p} + \sigma_p)$), *Occluded space* (belongs to vision cone but it is lying behind the rind (i.e. at $p > (\overline{p} + \sigma_p)$, and, finally, *Unknown space* (the space external to the vision cone).

The voxel work−space is updated in accordance with its current status and the computed value (initially the array is set to unknown). For example the status of an "unseen" voxel is changed by whatever value is computed. The same is true for "occluded" voxels. If the voxel is in the "seen" status it holds the probability of being filled. This probability is changed according to:

$x_{i+1} = x_i + k_i(y_{i+1} - x_i)$

with: $k_{i+1}^{-1} = k_i^{-1} + 1$

$k_0^{-1} = 0 \qquad x_0 = U$

Where x_i is the current voxel value, y_{i+1} is the computed value, and x_{i+1} is the final value of occupation probability.

For integration purposes occluded and unknown space are equivalent. On the other hand the distinction is not redundant because the two situations have a very different meaning. The observer has a strong interest for the occluded space which must be reduced as much is 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.

A further consideration concerns the equation used to update a seen voxel in the presence of new significant information (last line of the above table). This expression, in Kalman form, achieves the arithmetic average among all considered values; 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. 6; This figure represents the result of the integration of 8 depth maps, 4 of them derived from motion and 4 from stereo. The probability of occupation is set to 0.5 .

4. Implementation in a well−structured *early vision* environment

As we mentioned in the introduction, the thrust of project 419 is the integration of different visual modalities; although very important, stereo and motion are but two of these visual cues. Other modalities include the detection of intensity discontinuities and low−level grouping processes. Indeed, stereo and motion require these prior processes. One result of the work of P419 has been the development of an integrated early vision environment VIS [21,22]. VIS provides two key features:

(i) to allow a user to build systems of images (hierarchically organized in pyramids according to their resolution), to archive them, dearchive them, and to exploit the early vision capabilities − edge detection, stereopsis, motion, and the generation of the Raw Primal Sketch; and

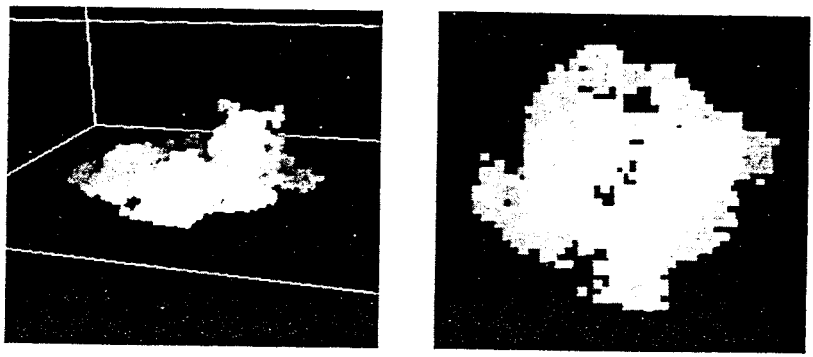

Figure 6. 3D Volumetric integration of the depth maps. Perspective representation of the voxel−based accumulator. The gray level codes distance (brighter means closer). Left column: the objects seen from above; right column: the object seen approximately from the point of view of Fig. 4. The tall package is visible (brighter areas on the top views) as are the two apples. For this picture any information coming from below the tray has been clipped (as a consequence the square base is not visible).

(ii) a sophisticated internal structure for "typing" the image according to the manner in which its data was generated (e.g. a stereo disparity image and a motion velocity image) exploiting representations appropriate to the image (i.e. arrays for 2−D iconic intensity data; chain codes for zero−crossing contours augmented by, e.g., strength, disparity, velocity, depth measures; and trees for hierachically nested convolution regions).

Furthermore, the interdependencies between these representations is explicitly represented and the user has access to utilities to peruse and analyze these "logical links". In the particular instance of motion and stereo integration, the image data are logically represented as an ensemble of sets of image pyramids: each pyramid contains just one level, i.e. a single image. Each set comprises the images of the stereo pair and the associated processed images used in the computation of the optical flow and the stereo disparity. Since the coarse−to−fine strategy of computing stereo disparity described above requires filtering in three distinct spatial frequency bands for each stereo view, the set comprises: two intensity (the stereo pair), six convolution images (σ_1, σ_2, σ_3,for both stereo views), six region images, six zero−crossing contour, slope, orientation images, a zero−crossing stereo disparity image, and a zero−crossing velocity image. The ensemble represents the motion sequence, i.e. the collection of each (processed) stereo view−point. What is significant is that the ensemble also incorporates logical linkages between the individual sets (in particular, between the individual velocity images) to explicitly represent the image−to−image, and contour−to−contour, tracking described in the previous section. Finally, a depth image is incorporated to represent the computed depth along each zero−crossing contour. One can interpolate between these values to compute the depth to all points in the visible surface; this is represented in the so−called range image. Current developments are

directed at integrating the volumetric representations in VIS.

5. Use of the 3−D information by higher level cognitive sub−systems

Low−level early vision is not, of course, an end in itself. The information generated at this stage must necessarily feed the higher level cognitive modelling sub−systems and, indeed, these cognitive systems must control the low−level processing. At the current stage of development, VIS facilitates this two−way flow of control and information through the use of its interpretive control language: VISICL. Cognitive modelling systems, typically running an alien host computer generate appropriate VISICL programs, transfer them to the VIS host where they are automatically executed, and the resultant information is then transferred back to the alien cognitive modelling system. At present, the information which is tansferred is a modified version of Marr's Raw Primal Sketch, i.e. a description of the scene in terms of lines (comprising many straight edge segments and delimited by terminations). Bar and blob primitives are also generated but, as yet, these are not communicated to the alien system. The edge segments in this instance are identified by an initium and a terminus; two points specified, not just in 2−D, but by 3−D coordinates. Since these edge segments are generated from the zero−crossing contours, the third dimension can be directly extracted from the depth map derived from the stereo and motion ranging process. In summary, VIS is a tool which places a sophisticated low−level computer vision environment at the disposal of a vision scientist and facilitates integration of low−level visual cues. VIS is also a tool which allows the integration of the complete early vision system with the high−level cognitive modelling systems. Finally, it is worth noting that VIS is a portable tool and presently supports six host environments (Unix, VMS, Xenix, DOS, Genix, and transputers) with drivers for six different types of framestore devices (Imaging Technology PCVISION and FG100, VDS Eidobrain, Datacube, Vicom, and X−Windows). A commercial version of VIS, now available, supports a sub−set of these capabilities and features.

References

1. Y. F. Wang, M. J. Magee, and J. K. Aggarwal, "Matching Three−Dimensional Objects Using Silhouettes", *IEEE Trans. PAMI* PAMI−6 No.4 pp. 513−518 (1984).
2. L. Massone, P. Morasso, and R. Zaccaria, "Shape from Occluding Contours", *S.P.I.E Symposium "Intelligent Robots and Computer Vision"*, , Cambridge − USA(November 4−8, 1984).
3. H.K. Nishihara, "PRISM: a practical real−time imaging stereo matcher", A.I. Memo 780, MIT A.I. Laboratory, Boston, Mass. (May, 1984).
4. D. Marr, "", in *Vision*, Freeman and Co., San Francisco (1982).
5. K. Prazdny, "Egomotion and Relative Depth Map from Optical Flow", *Biol. Cybernetics* **36** pp. 87−102 (1980).

Acknowledgments: A fellowship from ELSAG SpA is gratefully acknowledged by Dr. E. Grosso.

6. D. T. Lawton, "Processing Translational Motion Sequences", *CVGIP* 22 pp. 116–144 (1983).

7. T. D. Williams, "Depth from Camera Motion in a Real World Scene", *IEEE Trans. PAMI* **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. PAMI* **PAMI–8 No. 6** pp. 730–741 (1986).

9. M. Brady, "Artificial intelligence and robotics", A.I. Memo 756, MIT A.I. Laboratory, Boston, Mass. (February, 1984).

10. G. Sandini, M. Straforini, and V. Torre, "3D Reconstruction of Silhouettes", *Proc. 4th Intl. ROVISEC*, pp. 173–182 , London, UK(1984).

11. G. Sandini and M. Tistarelli, "Recovery of Depth Information: Camera Motion as an Integration to Stereo", *Proc. of "Workshop on Motion: Representation and Analysis"*, pp. 39–43 , Kiawah Island ResortIEEE Computer Society, (May 7–9, 1986).

12. E. Grosso, G. Sandini, and C. Frigato, "Extraction of 3D Information and Volumetric Uncertainty from Multiple Stereo Images", *Proc. of ECAI–88 in press*, (1988).

13. G. Sandini, C. Frigato, E. Grosso, and M. Tistarelli, "Multimodal Integration in Artificial Vision", *NATO ARW–workshop on "Robots with redundancy: Design, Sensing and Control"*, (Salo June 1988).

14. P. Morasso, G. Sandini, and M. Tistarelli, "Active Vision: Integration of Fixed and Mobile Cameras", *NATO ARW on Sensors and Sensory Systems for Advanced Robots*, , Berlin HeidelbergSpringer–Verlag, (1986).

15. A. Bandopadhay, B. Chandra, and D. H. Ballard, "Active Navigation: Tracking an Environmental Point Considered Beneficial", *Proc. of "Workshop on Motion: Representation and Analysis"*, pp. 23–29 , Kiawah Island ResortIEEE Computer Society, (May 7–9, 1986).

16. G. Sandini, V. Tagliasco, and M. Tistarelli, "Analysis of Object Motion and Camera Motion in Real Scenes", *Proc. IEEE Intl. Conference on "Robotics & Automation"*, pp. 627–633 , San FranciscoIEEE–CS, (April 7–10, 1986).

17. G. Sandini, "Extraction of Regional Stereo Information", TK1–WP2–DI1, DIST– University of Genoa – Esprit Project 419, Genoa (1986).

18. C. Frigato, E. Grosso, and G. Sandini, "Integration of edge stereo information", Esprit P419 Tech. rep. TKW1–WP1–DI3, DIST–University of Genoa (1987).

19. G. Sandini and M. Tistarelli, "Integration of Edge Motion Information", TK1–WP1–DI2, DIST– University of Genoa – Esprit Project 419, Genoa (1986).

20. A. Bandopadhay, J. Y. Aloimonos, and I. Weiss, "Active Vision", *International Journal of Computer Vision* 1 – **No 4** pp. 333–356 , Boston, Mass.Kluwer Academic Publishers, (jan 1988).

21. G. Sandini and D. Vernon, "Tools for Integration of Perceptual Data ", pp. pp. 855–856 in *ESPRIT'86: Results and Achievements*, Elsevier Science Publishers

B.V. (North–Holland) (1987).

22. D. Vernon and G. Sandini, "VIS: A Virtual Image System for Image Understanding", *Software: Practice and Experience* **18**, No. 5 pp. 395–414 (1988).

ESPRIT '88

Putting the Technology to Use

Proceedings of the 5th Annual ESPRIT Conference
Brussels, November 14–17, 1988

Edited by

COMMISSION OF THE EUROPEAN COMMUNITIES
Directorate-General TELECOMMUNICATIONS,
INFORMATION INDUSTRIES and INNOVATION

Part 1

1988

NORTH-HOLLAND
AMSTERDAM • NEW YORK • OXFORD • TOKYO