Marr's framework revisited. *

Kenneth M. Dawson, Mairead Flanagan and David Vernon

University of Dublin, Trinity College, Dept. of Computer Science, Dublin 2, Ireland

Abstract.

A framework for the visual process is one of the major aims of current investigations in computer vision. Perhaps the most compelling framework developed to date is that of David Marr, in which vision is modelled as an information processing system premised on the modular processing of representations, sequentially developing representations of increasing structural richness and culminating in an explicit 3-D representation of the viewed world. This paper reexamines Marr's approach in light of an actual implementation which was based on that approach. The implementation lends strong support to Marr's scheme by demonstrating the potential of Marr's framework for accurately deriving 3-D structure which can then be used to reason about the scene (and in particular, to recognise objects).

The recognition technique is a new development which we have previously shown to be quite powerful and robust, using 3-D models derived from actively sensed range data.

1 Introduction.

David Marr made a significant contribution to the area of computer vision. His framework for the visual process [1] is one of the most compelling theories developed to date. This paper reexamines Marr's framework by presenting a system which was developed along the lines of Marr's framework. Note, though, that this paper is a summary of the work done to date, and that the reader is intended to make use of the references for further information (in particular see [2]).

In section 2, our implementation of Marr's framework is described. That implementation does not precisely match the framework described by Marr, but the central idea (i.e. that vision may be modelled as a modular and sequential information processing system) and many of the algorithms and representations are identical to those of Marr.

An explicit 3-D model is the result of Marr's framework (and of our implementation). However, it is not the culmination of processing for visually derived data as, in order for us to be able to reason about the world we must identify the structures, or objects, which we see. Hence, in order to give additional support to the usefulness of extracting an explicit 3-D model in section 3 we consider the recognition of the 3-D models derived using our implementation of Marr's framework.

Finally, in section 4 we conclude with a brief comment on the usefulness of Marr's approach in light of our implementation and development of his ideas.

* The research described in this paper has been supported by ESPRIT P419, EOLAS APT-VISION and ESPRIT P5363.

2 An experimental implementation of Marr's framework.

In order to implement Marr's framework a visual processing platform, *VIS a VIS* [2], was developed which gave both the processes and the representations equal importance. Each representation (in Marr's framework), along with any differences in our implementation of them, and details of the processes used to compute them are given in the sections which follow.

2.1 Experimentally obtaining the raw primal sketch.

The first stage in Marr's framework; i.e. the derivation of the Raw Primal Sketch (which is basically an abstract representation of the intensity discontinuities in an image); is extremely well detailed in the literature [1] and our implementation followed those theories directly. We employ a $\nabla^2 G$ edge detector with various values for the standard deviation of the Gaussian filter. The resultant zero-crossings are stored using boundary chain codes and are then analysed in order to identify spatial coincidence in multiple channels.

The raw primal sketch primitives are extracted from the resultant significant zero-crossings using simple approximation techniques (e.g. least-squared error). Overall the raw primal sketch was found to be a quite concise representation of the 'important' information in an image.

Finally, we diverge very slightly from Marr's framework. The depth information which may be derived for the significant intensity discontinuities (and which, in Marr's framework, is computed during the derivation of the $2\frac{1}{2}$D sketch) is associated with the primitives of the raw primal sketch and, hence, it is possible to employ this information during the grouping process. The computation of depth information is described in the section on the $2\frac{1}{2}$D sketch.

2.2 Experimenting with the full primal sketch.

Relatively little research has been done on this area of visual analysis (i.e. the grouping of features). The most significant work is that of David Lowe [3] and he has demonstrated the potential benefit of employing perceptual grouping techniques to reduce the computational complexity of the object recognition problem [4].

We have addressed this issue from the same perspective but also with a view to facilitating the derivation of 3-D structure [5]. The grouping processes operate on 2-D, and 3-D, raw primal sketch primitives in a recursive manner.

The criteria which are considered in order to effect grouping are collinearity, curvilinearity, similarity (e.g. of edge segment strength), spatial proximity, equal spacing, symmetry and familiarity. These, in many cases, are competing criteria and yet it is necessary to determine a single interpretation (in order to allow continued sequential processing). Hence, each grouping criteria has an associated weight and each feature group has a degree of confidence associated with it. This weighting of criteria and groups could also, potentially, be driven by world knowledge.

At the first level of grouping the collinearity and curvilinearity criteria were found to be of greatest importance. At subsequent levels, the various similarity and spatial proximity criteria became more important in the formation of groups. However, it was not clear how many levels of recursion should be performed, although an obvious termination condition was the level at which no more valid groups could be formed.

Overall, the results were quite promising as the feature groups generated (mostly) corresponded to single physical objects/phenomena. Also, it was notable that applying the grouping processes to 3-D raw primal sketch primitives (as opposed to the 2-D primitives) gave particularly good results. For further details see [2] or [5].

2.3 The $2\frac{1}{2}$D sketch.

Computation of the $2\frac{1}{2}$D sketch requires the use of many different processes/algorithms; in particular derivation of depth information for the intensity discontinuities and interpolation of that depth information. Both of these issues are addressed in the subsections which follow.

Deriving depth information. Depth from camera motion, as employed within this research, uses nine viewpoints of a scene taken from different positions around an arc. The optical flow is computed from image to image, allowing both the reliable correlation of features (as indicated by the optical flow) and an increase in the resultant angular disparity (i.e. when triangulation is effected using a viewpoint from each end of the arc). In fact as long as the angular disparity between viewpoints is not too great it is possible to correlate the features from successive images (in the sequence) using a computational rather than an algorithmic technique; i.e. the time derivative [6].

Detailed descriptions of the depth from camera motion technique employed are given in [7,8,9,10] and an example is shown in Figure 1.

Interpolation of depth values. Computing depth from passively sensed data through the triangulation of particular features only allows depth to be calculated at those specific features. If it is desired to build a complete 3-D surface or volumetric model, it is necessary to interpolate in some way between the computed depth values. Perhaps the best work in this field, to date, is that of Grimson [11] in which he employs Laplacian and quadratic variation functionals in order to determine the best fitting surface for sparse range data. However this technique assumes that the depth information provided pertains to a single continuous surface.

The intention is to address this segmentation (of surfaces/objects) using the grouping processes described in the previous section. However, at present, the grouping processes are not sufficiently well developed to allow reliable segmentation into physical surfaces. Hence, a simpler mechanism, planar interpolation, was employed which results in correct surfaces for single polyhedral objects (only).

The result of the interpolation is a dense depth map (see Figure 1) from which the remainder of Marr's $2\frac{1}{2}$D sketch can be generated (i.e. the needle diagram). However, it is important to bear in mind the significant amount of research which will be required if the interpolation is to be made generally valid. The segmentation process will need to be made much more powerful, and other issues (such as the identification of occluding contours) will have to be addressed.

2.4 Deriving a 3-D model.

The final task which must be addressed is the extraction of a 3-D model. The two main possibilities which are considered for such a model are usually *volumetric models* and *surface*

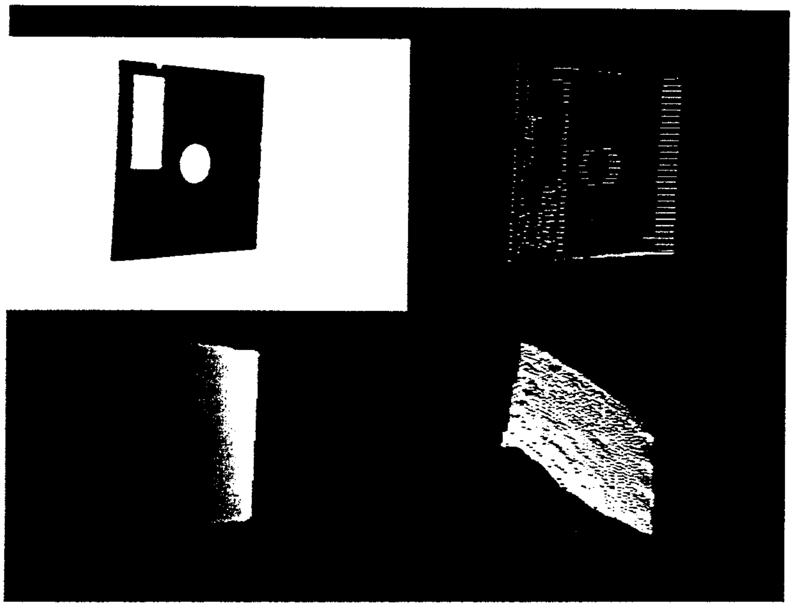


Fig. 1. The third image in a sequence of nine images is shown (top left), along with the optical flow determined for each of the significant zero-crossings (top right), the depth map interpolated (bottom left) and a side view of the derived 3-D model.

based models. Selecting a representation must take into account many issues, such as accessibility, scope, uniqueness, stability, and sensitivity [12]. Although Marr seemed to favour the volumetric model [1], that model suffers from severe problems in terms of accessibility. Hence a surface based model was employed in this research, where the primitives of the model were all simple planar surface patches. This representation is readily accessible, and satisfies the scope and sensitivity criteria. However it most definitely does not provide unique descriptions of objects, nor is it a stable representation. To overcome these problems a new method of object recognition was developed, *implicit model matching*, which is briefly described in the next section.

Deriving the planar surface based model was done directly from the dense depth map which is produced by the processes which have already been described. 3-point seed surfaces [13] are instantiated between any three depth points which are within the sampling of each other. These seed surfaces are normally then merged into larger surfaces (which correspond to actual surfaces in the world), but for the object recognition technique developed this was not necessary (although it was found necessary to smooth the surface normals of the surface patches using a local smoothing function). An example is shown in Figure 1.

3 Object recognition; after Marr's framework.

The development of a 3-D model from passively sensed data is not, by any means, the culmination of processing required for visually derived data. It is necessary to reason with these models in order to allow navigation and also, more importantly, to allow cognition of the world in which the system is based. This understanding of the world can only come about through the association of previously known information with information sensed from the environment (e.g. the 3-D models). In that way the system will recognise the objects in it's environment and be able to consider them in terms of their function, hence facilitating the performance of useful and intelligent actions.

The domain of object recognition from image data is an immense one, and is a subject which is receiving progressively more attention. While it would seem that recognition should be facilitated through the derivation of as much information/structure as possible, most researchers have found that it is necessary to make a compromise as the processes for deriving complex models (e.g. 3-D surface based models) from passively sensed data are not sufficiently robust. In fact, in many ways, the research community seems to be accepting that recognition should be performed using 2-D images features. However researchers such as David Lowe (who has developed one of the most impressive works, to date, in 2-D feature based recognition [4]) have more been cautious in this regard, stating only that *the role of depth recovery in common instances of recognition has been overstated.* Recognising 3-D structure through the comparison of 3-D models is not easy, though, even when employing models derived from actively sensed range data.

In order to demonstrate that Marr's framework for vision has any validity, it would seem essential to show that it is possible to recognise 3-D models derived using the framework. This, then, would demonstrate a potential connection between Marr's visual framework and the subsequent cognitive processes.

We have developed such a technique; i.e. for the recognition of 3-D objects through the comparison of 3-D models. This technique, *implicit model matching* [14,15,16], compares object models through the use of secondary representations (i.e. representations which are derived from the 3-D models). The technique is described very briefly in this paper and interested readers are directed to the cited references.

3.1 Implicit Model Matching.

The basic problem of rigid object recognition is to establish the correspondence between an object model, which is *viewed*, and a particular *known* object model, and the computation of the associated pose. The majority of object recognition techniques match *known* object models with viewed instances of objects through the comparison of model primitives (e.g. edges). However, it is extremely unlikely that the model primitives extracted will be identical to those of a model which is *known a priori*. In order to overcome that problem it is possible to employ secondary representations, such as the Extended Gaussian Image (or EGI) [17], although using the EGI has proved difficult [18].

The technique of implicit model matching introduces a new, more powerful, but similar idea. It employs *several* secondary representations which allow the problem to be considered in terms of sub-problems. Initially, orientations of *known* objects which may potentially match the *viewed* object are identified through the comparison of visible surface normals, for each possible orientation of each *known* object (where 'each possible orientation' is defined by the surface normals of a tessellated sphere). Potential orientations are then fine tuned by correlating 1-D histograms of specific components of surface orientations (known as directional histograms). The object position is estimated using approximate knowledge of the physical configuration of the camera system, and fine tuned using a template matching technique between needle diagrams derived from the *known* and *viewed* models. Finally, using normalised correlation, each hypothesis is evaluated through the comparison of the needle diagrams.

At each stage in the generation, tuning and verification of hypotheses only comparisons of the various secondary representations are employed. The central concept behind implicit model matching is, then, that 3-D object models may be reliably compared through the use of secondary representations, rather than (or, more properly, as well as) by comparison of their component primitives. Additionally it is important to note that object pose may be determined to an arbitrarily high degree of accuracy (through the fine-tuning stages), although initially only a limited number of views are considered.

3.2 Experimental Results.

Testing the recognition technique was performed using randomly selected views of the known (CAD) models, models derived from actively sensed range data, and finally 3-D models derived using our implementation of Marr's framework. The results described here (i.e. those using 3-D models derived using Marr's framework) are intended to demonstrate both the potential of Marr's approach and to test the recognition technique in the presence of incorrect data.

Both fabricated and 'real' test objects were considered, and all objects were (simple) polyhedral structures. Also, only scenes with single objects were considered. These constraints were necessary due to the simplifications which were used during the generation of the $2\frac{1}{2}$D sketch (i.e. ignoring the issues of segmentation, and employing planar interpolation).

The database of known objects is shown in Figure 2, and then two examples of deriving and recognising structure are given.

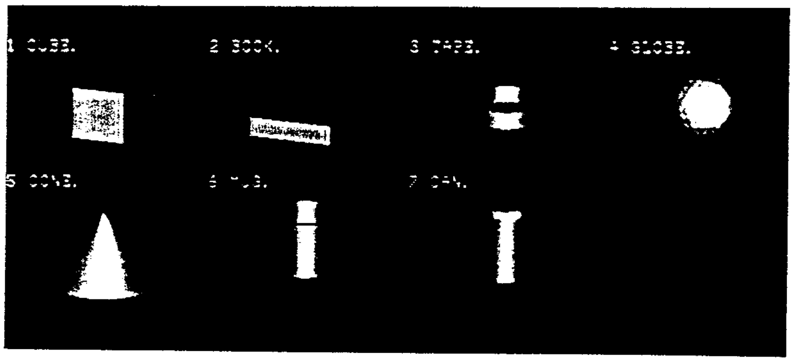


Fig. 2. The database of known object models.

The first example is the book, in Figure 3. The 3-D model appears qualitatively, at least, to be quite accurate (see Figure 4) with the main exception being the region of the model corresponding to the title of the book. Nevertheless, the recognition algorithm is successful (see Figure 5).

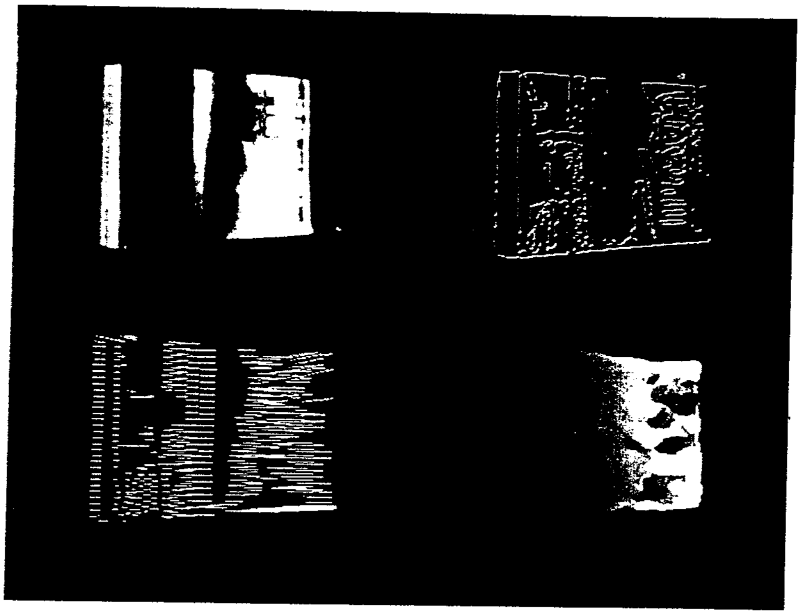


Fig. 3. The third image in a sequence of nine images (top left), its associated 'significant' zero-crossing (top right), the optical flow determined between the third and eighth images (bottom left) and the resultant smoothed, interpolated depth map (bottom right).

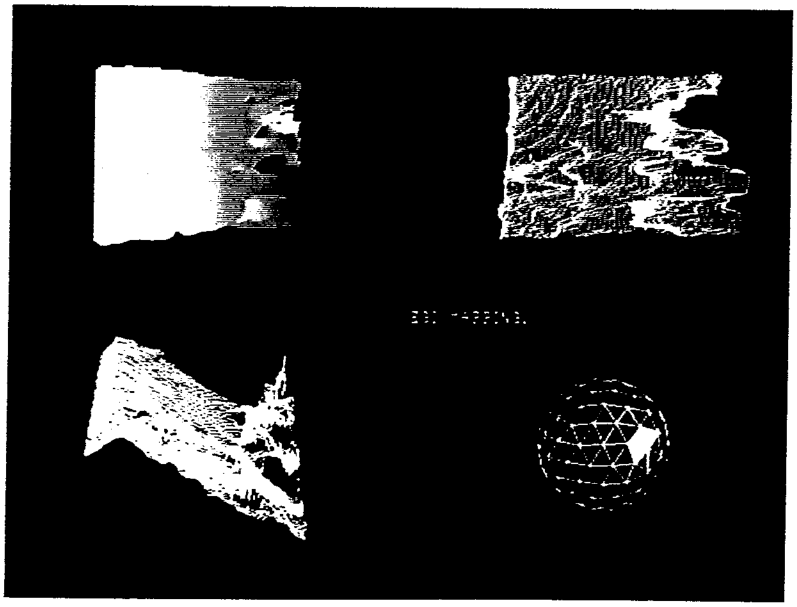
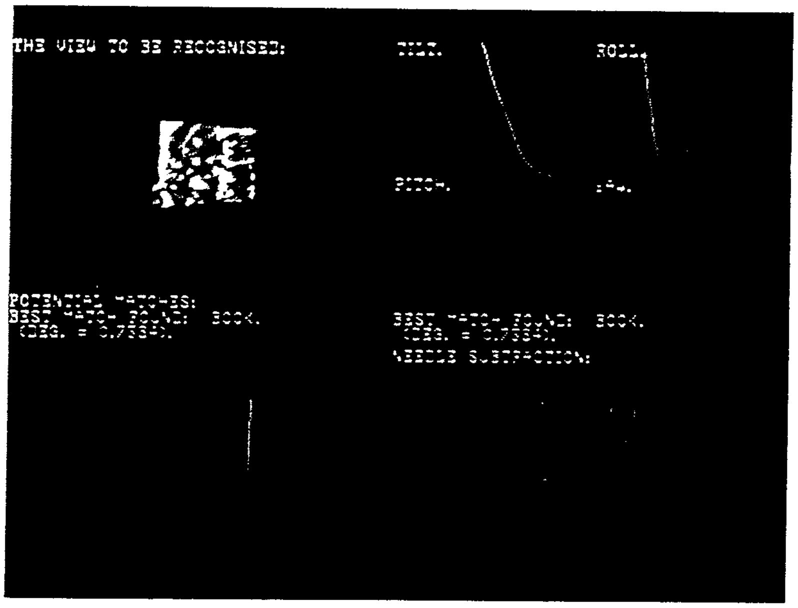


Fig. 4. Three wireframe views and the EGI, of the model derived from the depth map shown in Figure 3.



THE VIEW TO BE RECOGNISED:

TILT.

ROLL.

PITCH.

YAW.

POTENTIAL MATCHES:
BEST MATCH FOUND: BOOK.
(DEG. = 0.7534).

BEST MATCH FOUND: BOOK.
(DEG. = 0.7534).
NEEDLE SUBTRACTION:

Fig. 5. Recognition of the model shown in Figure 4.

The second example was chosen in order to emphasise the areas which especially need to be addressed further (e.g. segmentation and interpolation). The roll of sellotape shown in uppermost left quadrant of Figure 6 results in a surface being interpolated, incorrectly, between the two sides of the roll (see the uppermost right quadrant). The recognition algorithm, not surprisingly, fails (see the lower left quadrant). However, if we could correctly segment the physical surfaces and objects (as in the lowermost right quadrant), the technique would succeed.

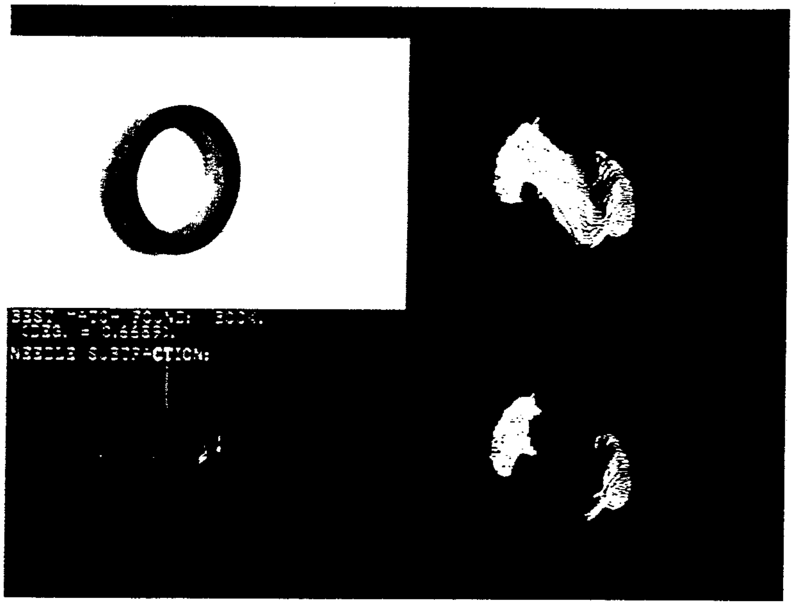


Fig. 6. A roll of sellotape (top left), the derived surface model (top right) from a slightly different perspective, a failure in terms of recognition (bottom left) due to the incorrect interpolation between the two sides of the roll, and finally (bottom right) a contrived version of the model in which interpolation was performed correctly (and on which the recognition technique succeeds).

While the object recognition experiments described in this paper obviously consider only extremely constrained scenes, the technique of implicit model matching has been found to work well with much more complex scenes (e.g. with models derived from actively sensed range data) [14,15]. Testing with models derived from passively sensed range data provides a evaluation of the robustness of the approach, and basic evidence in support of the use of 3-D viewed models in object recognition.

4 Conclusions.

An experimental implementation of Marr's framework for the visual process was presented in this paper, culminating in the automatic recognition of the 3-D models which were derived by the system. Overall, the results demonstrate the immense potential of Marr's representational framework, in particular as it appears that by improving the various processes it should be possible to further develop the capabilities of systems based on this framework.

However, much further research on these various processes (in particular, segmentation and interpolation) is required. Additionally it should be evident that Marr's framework must be combined with many other approaches to vision in order to build a truly general-purpose vision system. For instance, recognition of objects in 2-D pictures (e.g. paintings) is impossible using Marr's basic framework.

In conclusion, though, it would seem that Marr's basic approach warrants much further attention.

References

1. D. Marr, "Vision," W.H. Freeman and Co. - 1982.
2. D. Vernon and G. Sandini (editors), "Parallel computer vision - The VIS a VIS System". Ellis Horwood, 1992.
3. D.G. Lowe, "Perceptual Organisation and Visual Recognition," Kluwer Academic Publishers 1985.
4. D.G. Lowe, "Three-Dimensional Object Recognition from Single Two-Dimensional Images," Artificial Intelligence, Vol. 31, No. 3, pp.355-395, March 1987.
5. M. Flanagan, "Generation of Grouped Three-Dimensional Raw Primal Sketches," M.Sc. thesis, Dept. of Computer Science, Trinity College, Dublin 2, Ireland, April 1991.
6. B.K.P. Horn, and B.G. Schunck, "Determining Optical Flow," Artificial Intelligence, Vol.1̃7, No.1̃, pp.185-204, 1981.
7. G. Sandini, V. Tagliasco, and M. Tistarelli, "Analysis of Object Motion and Camera Motion in Real Scenes," Proceeding of the IEEE International Conference on Robotics and Automation, San Francisco, pp.627-632, 1986.
8. G. Sandini, and M. Tistarelli, "Active Tracking Strategy for Monocular Depth Inference Over Multiple Frames," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-12, No. 1, pp.13-27, January 1980.
9. D. Vernon and M. Tistarelli, "Range Estimation of Parts in Bins using Camera Motion," SPIE Vol. 829 - Applications of Digital Image Processing X, pp.258-266, 1987.
10. D. Vernon and M. Tistarelli, "Using Camera Motion to Estimate Range for Robotic Part Manipulation," IEEE Robotics and Automation, Vol.6, No.5, pp.509-521, October 1990.
11. W.E.L. Grimson, "From Image to Surfaces: A Computational Study of the Human Early Visual System," MIT Press, Cambridge, Massachusetts, 1981.
12. D. Marr and H. Nishihara, "Representation and Recognition of the Spatial Organisation of three dimensional structure," Proceedings of the Royal Society of London, Series B, Vol.200, No.1140, 1978. pp.269-294.
13. O.D. Faugeras, and M. Herbert, "The representation, recognition and locating of 3-D objects," *International Journal of Robotics Research* , Vol. 5, No. 3, pp.27-52, Fall 1986.
14. K.M. Dawson, "Three-Dimensional Object Recognition through Implicit Model Matching," Ph.D. thesis, Dept. of Computer Science, Trinity College, Dublin 2, Ireland, 1991.
15. K.M. Dawson and D. Vernon, "Model-Based 3-D Object Recognition using Scalar Transform Descriptors," Proceedings of the conference on Model-Based Vision Development and Tools, Vol. 1609, SPIE - The International Society for Optical Engineering, November 1991.
16. K.M. Dawson and D. Vernon, "3-D Object Recognition using Passively Sensed Range Data," Proceedings of the Second European Vision Conference on Computer Vision (ECCV 2), May 1992.
17. B.K.P. Horn, "Extended Gaussian Images," Proceedings of the IEEE, Vol.72, No.12, pp.1671-1686, December 1984.
18. P. Brou, "Using the Gaussian Image to Find Orientation of Objects," The International Journal of Robotics Research, Vol.3, No.4, pp.89-125, Winter 1984.