TR-86-008

# 機械立體視覺

中央研究院資訊科學研究所

黃俊雄

中華民國75年12月

# 摘　要

機械立體視覺是非常的重要而有極廣大的應用。最大的用途在無人駕駛的車輛如坦克車或深海潛艦，或太空上之飛行車船。本篇文章主要是說明如何解決立體視覺問題。其解決原理是不斷轉動特別設計的照像機並利用兩組相機同時對物體上的一點聚焦。如此對應問題自然而解。由明暗度之變化來判定物體三度形狀，此問題可用 Haralick's sloped-facet model 〔1〕來解決。遮蓋問題可由檢查對應之點，線及邊界線或轉動相機從新對遮蓋線聚焦等方法來解決。此新提出的立體視覺系統應當對走動機器人有實際效用。本文之新構想是由數學轉至新出生嬰兒眼睛的活動而產生出來。最後本文討論由立體視覺至學習能力的培養而斷定立體視覺系統是建立一高功能視覺系統之必須該走的第一步。作者希望國內各單位專家們能合作共同朝此一方向努力。這第一步本文已詳細說了。最困難的是第二步：自動學習，由於視覺在人腦建立的觀念是具體的，不是抽象的，所以這方面的研究相信不致白費無成。

# MACHINE STEREO VISION

Jun S. Huang

Computer Vision Laboratory

Institute of Information Science

ACADEMIA SINICA

TAIPEI, TAIWAN, REP. OF CHINA

DECEMBER 22, 1986

ABSTRACT:

Machine stereo vision is very important and has wide applications. In this paper the stereo vision problem is solved by focusing two cameras at the same point and by moving cameras dynamically. The 3D shape from shading is solved by Haralick's sloped-facet model[1]. Occlusion problem is solved by checking feature points, lines and edges, or by moving cameras to refocus on the occluding line. This proposed system should work well enough for the mobil robot. The idea of this stereo vision system comes from the heuristics of eyes of a new born baby. This paper finally discusses the steps from stereo vision to learning and asserts that the stereo vision system is the necessary first step to build a powerful vision system.

# 1. INTRODUCTION

Machine stereo vision has wide applications, especially on mobil robots, and is the basic root for constructing a 3D vision system that can learn 3D shapes from images. The pioneer of using stereo vision on mobil robot is Moravec[2,3] at Stanford University and CMU. A list of related research papers is in references[4-9]. All of their methods use two parallel cameras to take images and try to resolve the correspondence problem on the two images. Evidently this will only work if the computer has stored enough object models. But how to store these object models? The only one way is by learning. Even if most of the object models are stored inside the computer, this kind of parallel camera formaton will not be able to solve the correspondence problem in regular textured images. However here I propose a complete different approach from the classical ones.

We know that the classical approaches all set up a two parallel cameras system as shown in Fig. 1. Besides parallelism the cameras are commercial ones that are composed of multiple lenses so that different object points with different depths can be focused clearly on the same sensor plane. Thus we are able to take clear pictures on wide viewing angles. From stereo images we select a meaningful point or an edge segment in the left image and try to find the corresponding point or edge in the right image. Obviously this procedure would not work for regular textured images and occluding objects. Hence we have

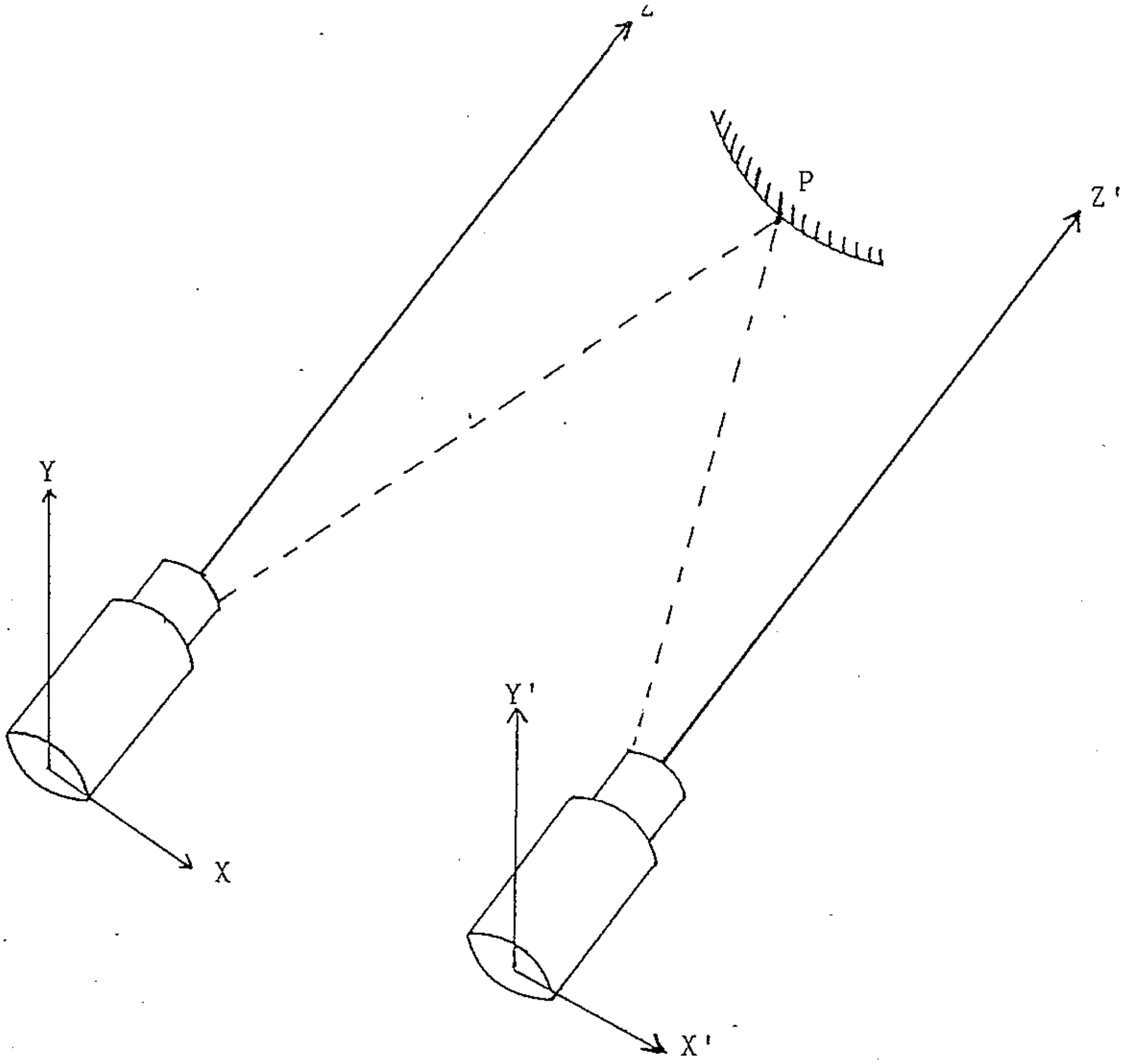

Fig. 1. Two parallel camera system.

to drop this classical approach and try some other ways. If you consider the biological eyes, the eye is composed of single lens that can be adjusted and turned very flexibly and quickly. So obviously we should study the advantages of single lens camera. The advantages are clear that a single lens camera is able to tell you whether an object point is focused exactly or not by using the lens aberration property, and also a single lens camera can be turned and refocused sharply. These abilities depend on the mechanical design. This is one of the directions of efforts we should do. If we use the property of exact focusing for a single lens to force two cameras focusing at the same object point, then the correspondence problem is resolved at this point. Hence we can try to match feature points around a small neighborhood of this object point in each image. If some feature points can not be matched exactly then we turn two cameras to refocus on these points and take images again. From these new actions we will be able to make conclusion whether these feature points are occluded or not.

One critical point to the success of stereo vision is the image sensor. So far as I know the commercial image sensor capability is far behind the human eye's: the resolution is low and the image signal is not stable and noise corrupted. If I take the same image at two different times by setting the environment unchanged, the two images taken will be far different in numerical values. One way to avoid this unstability is to take several images and average them. But human eyes take images continuously in time and the complexity is far beyond our imagination. Human eyes have very high resolution (about 11kx11k) and are quite stable to light sensing. The technology of making C.C.D. or C.I.D. sensor is still in progress and I suggest that the arrangement of the sensor elements be in circular form, dense in the center region, for the convenience of stereo vision analysis.

## 2. MACHINE STEREO VISION

I propose a stereo vision system shown in Fig.2 where two lenses LENS 1 and LENS 2 are having centers located at L1 and L2 seperated at a distance 2b respectively. An object point P* in 3D space is focused at the same time by these cameras. That is, P* is projected to the two image centers. The choice of P* is on that the neighborhood of P* contains points, lines, or edges. A featureless region is not precisely focused but just roughly scanned over.

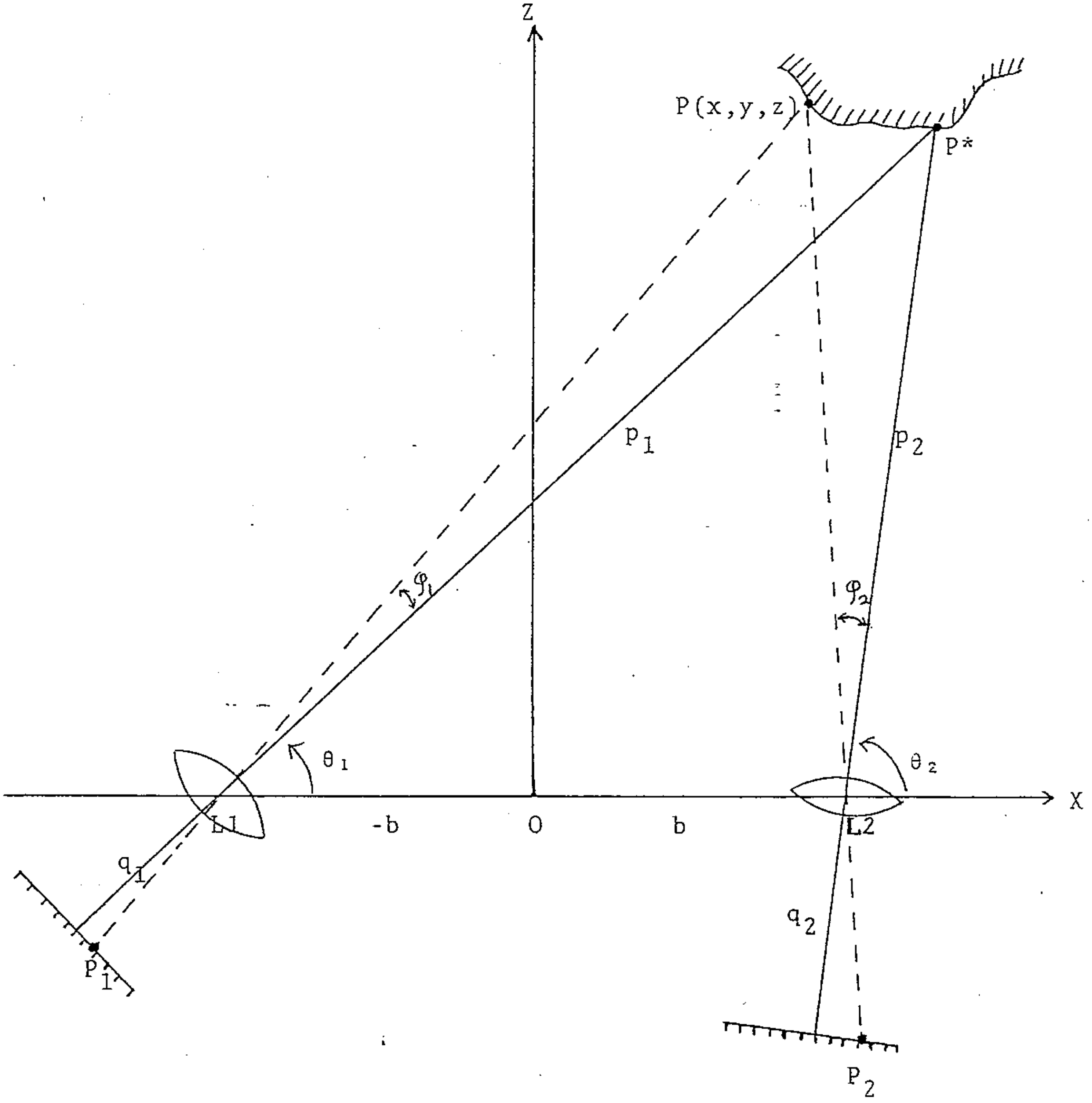

Fig.2. The proposed machine stereo camera system.

The requirements for this vision system are:

1. Each camera has a single lens having equal focal length f, and they are seperated at a distance 2b. A very high quality image sensor, such as C.C.D. or C.I.D. circular array elements, is attached with it.
2. In order to get high precision of focusing f should be long enough, say 300mm. The focusing formula is $\frac{1}{p} + \frac{1}{q} = \frac{1}{f}$, where p is the object distance and q is the image distance.
3. Each camera can rotate on X-Z plane and the rotating center is at the lens center : L1 and L2 for left and right cameras respectively. Each rotating axis is perpendicular to the line $\overline{L1L2}$
4. Besides rotating on X-Z plane both cameras can rotate simultaneously on the Y-Z plane with rotating axis $\overline{L1L2}$ ( Y axis is perpendicular to X-Z plane).
5. A supercomputer or an image processor array like LIPP is responsible for parallel computing.
6. An optical computing system with image transducers, FFT, EDGE filter, etc., is required for real time focusing and other computings.

Please note that extension to four or eight cameras system is better since if four cameras are focusing at the same object point the precision of focusing can be quite high by checking each camera's object distance.

The analysis of an object point P having coordinate (x,y,z) in

3D space and having been projected on two camera images goes as follows:

Let the projected point image be $P_1$ with $(x_1,y_1)$ coordinate in the left image, and be $P_2$ with $(x_2,y_2)$ coordinate in the right image. Since the two image planes are not in parallel the epipolar line, the line connecting $P_1$ and $P_2$ will not be parallel to X-axis. Without loss of generality, let the focused point P* be in X-Z plane as shown in Fig. 2. The rotating angles of two lenses are $\theta_1$ and $\theta_2$. Let the object distances for two lenses be $p_1$ and $p_2$ respectively, and the two image distances be $q_1$ and $q_2$ respectively. Then

$$\tan(\theta_1 + \varphi_1) = \frac{z}{b + x} \quad \text{and} \quad \tan\varphi_1 = \frac{x_1}{q_1} ,$$

$$\tan(\theta_2 + \varphi_2) = \frac{-z}{b - x} \quad \text{and} \quad \tan\varphi_2 = \frac{x_2}{q_2} .$$

Please note that $\varphi_1$ and $\varphi_2$ may be negative if P is located to the right of P*. Since $\theta_1$, $\theta_2$, $x_1$; $x_2$, $q_1$, $q_2$, b are all known, we can solve the above equations for x and z. The results are

$$z = 2b/(1/\tan(\theta_1 + \varphi_1) - 1/\tan(\theta_2 + \varphi_2)),$$

$$x = b(\tan(\theta_1 + \varphi_1) + \tan(\theta_2 + \varphi_2))/(\tan(\theta_2 + \varphi_2) - \tan(\theta_1 + \varphi_1)$$

The relationship between $y_1$ and $y_2$ in two point images is

$$\frac{y}{\sqrt{(b + x)^2 + z^2}} = \frac{y_1}{\sqrt{x_1^2 + q_1^2}}$$

and

$$\frac{y}{\sqrt{(b - x)^2 + z^2}} = \frac{y_2}{\sqrt{x_2^2 + q_2^2}} ,$$

which yield to

$$\left(\frac{y_1}{y_2}\right)^2 = \frac{(b - x)^2 + z^2}{(b + x)^2 + z^2} \cdot \frac{x_1^2 + q_1^2}{x_2^2 + q_2^2}.$$

Hence from this equation we can determine the epipolar line. Now given $(x_1,y_1)$ in the left image we search the corresponding point $(x_2,y_2)$ in the right image along the epipolar line (actually a curve):

$$y_2 = y_1 \cdot \sqrt{\frac{(b + x)^2 + z^2}{(b - x)^2 + z^2} \cdot \frac{x_2^2 + q_2^2}{x_1^2 + q_1^2}}$$

with variables $(x_2,y_2)$ where x and z are functions of $x_2$ listed in the above equations. The representation of these equations may be better expressed in terms of polar coordinate system for circular sensor arrays, not rectangular arrays.

Now we come to see the correspondence problem. Given precise focusing of two cameras we have taken two images where textures and features around the center are clear and graduately falling unclear when away from the center. To solve the correspondence problem, we must utilize three important facts: principle of continuity[10], shape from shading[11] and camera mobility. The principle of continuity says that nearby points match nearby points unless there is occlusion which will give sharp edges in general. Shape from shading says that the graduate changes in gray level corresponds to the curved surface from which we can determine the light direction in heuristic ways. The whole procedure is listed as follows:

1. Use statistical theory of edge detection developed by Huang and Tseng[12] to detect points, lines, edges and changes in gray levels, by using Haralick's sloped-facet model.
2. Select points, line segments and edge segments around a small neighborhood of the left image center and find the corresponding points, lines and edges in the right image by searching along epipolar lines around the image center.
3. Graduately enlarge the neighborhood of the left image center and get more features. Then try to search the corresponding features in the right image.
4. When matching edge segments, try to check whether Haralick's model is the same structures in both images.
5. If some points or lines can not be matched by principle of continuity then there is occlusion. Resolve the occlusion by matching nearby edge segments (enlarge the neighborhood). If this does not work then rotate both cameras to a new focused point P* in the unresolved region.
6. If a region is featureless, we may skip it and try to rotate both cameras to focus along the edges enclosing it. If we succeed in matching these edge segments, then we can use Haralick's sloped-facet model to describe the curved surface and roughly estimate the light direction[13]. One way to achieve this is to match singular points described by Huang and Tseng[14] so that one can determine whether the surface is convex or concave. Once light direction is estimated check

it with other light direction estimates in other regions whether they are consistent. If not consistent, some surfaces may have to be changed from convex to concave or vice versa.

7: Shape from shading: A convenient way of doing this analysis is to describe a surface consisting of patches of quadratic surfaces, and set up relationship between coefficients of this model and the coefficients of Haralick's sloped-facet model. How good is this relationship depends on experiments on different surface materials. From these experiments we should be able to estimate the lighting characteristics such as uniform lighting or point lighting or lighting directions. We can set up a look-up-table or dictionary containing these information. My suggestion on this difficult problem is simple, heuristic and experimental whereas Horn's gradient space approach[11] is too complicated to be implemented. A typical example is that if the computation of Haralick's sloped facet model for gray level image yields a constant (i.e. slopes are near zero in statistical sense) then the corresponding 3D surface should be a 3D plane. If the slopes have large absolute values then the corresponding 3D surface should have high curvature values. The negative sign or plus sign of slope values will give the clue of the light direction.

8. For a mirror like or tansparent surface unless we have background edges or some features lying on the surface that can be matched in stereo sense, then we would not be able to detect the presence

of this surface. Sometimes we are able to utilize the light scattering and diffraction, or mirror image's distortion, to detect its presence. However this ability depends much on the human experience which is heavily related to learning.

## 3. FROM STEREO VISION TO LEARNING

Once we have set up a stereo vision system we can teach computer to see and to learn understanding 3D objects just like a little baby's seeing and learning. Objects should be simple at the beginning and then graduately become complex. A new born baby can see only objects nearby and after long time learning she can see things farther. At the same time her both eyes graduately change from cross looking to parallel looking. This is because she has learned the characteristics of objects and has set up proper object models in her brain and mind. Hence I consider this is the right way to solve vision problem : the first step is to build a stereo vision system and the second step is to teach computer to learn to see.

I have sketched the architecture of a stereo vision system. But to teach computer to learn to see seems too difficult to solve. However since we are dealing with concrete concepts, not abstract concepts, this learning research should bear some fruitful results in the near future. Heuristic approach suggests that two mechanisms must be solved. One is extracting concrete concepts from 3D images such as lines, corners and shapes, etc. The other is to store and manipulate these concepts efficiently. We can ask the system to move cameras to trace the interesting boundaries and regions so that precise and related information can be gathered and passed to analyzer, which may be Hough like transforms, statistical modelings, and informatic classification trees, etc. Once concrete concepts are extracted, a graph representation is generated for each observed object. How to store and link these graphs efficiently is crucial to the success of the vision system. A machine LIPP proposed by Linköpin University[15] is suitable for this task. However the development of an efficient subgraph isomorphism parallel algorithm is also crucial[16] where we can use the algorithm to match an unknown object against what is stored inside the machine. I believe this is the key points we should do.

Finally I want to address some words about the airplane development and make some predictions on machine vision research. Centuries before, human were eager to be able to fly like birds and at the beginning of this century human was finally able to fly freely in the sky. Strangely, the airplane invented by human is not like birds flapping the wings and stay freely on trees. But it is just big heavy machine and can not stay at any place but only at those specific run ways. The reason for these phenomena is that human is much heavier than birds and from laws of physics we can calculate how large a force is needed to lift the airplane and the man. This force is proportional to the speed of air passing through both wings. Hence scientists and engineers were naturally heading toward developing a powerful but light weighted engine so that the airplane can get better speed. This logic is very important because if people in the nineteenth century knew this logic they would'nt try many fruitless flying experiments. Similarly we look at the development of computer vision system; human eyes are very powerful and complicated and the brain is highly intelligent. We can make some specific vision systems with limited power for some special applications. But to develope a vision system with full power capability we must head toward developing: 1). A powerful image sensor with very high resolution, low distortion and noises, sensitive but stable to wide spectrum. 2). Fast mechanisms to rotate both cameras as pointed out in this paper. 3). Very large memory and very large number of processors with complicated interconnections. 4). Intelligent analysis and learning to process vast image data, and this is the most difficult challenge to vision researchers.

# REFERENCES

1. R. M. Haralick, Edge and region analysis for digital image data, Computer Graphic and Image Processing 12, 60-73,1980.
2. H. P. Moravec, Visual mapping by a robot rover, Proc. 6th Int. Jt. Conf. Artificial Intelligence, Tokyo, Japan, 598-600, 1979.
3. ______________, The Stanford cart and the CMU rover, Proceedings of the IEEE, Vol. 71, 872-884, 1983.
4. M. D. Levine, D. A. O'Handley and G. M. Yagi, Computer determination of depth maps, Computer Graphic and Image Processing 2, 131-150, 1973.
5. Y. Yakimovsky and R. Cunningham, A system for extracting three - dimensional measurements from a stereo pair of TV cameras, Computer Graphic and Image Processing, 7 195-210, 1978.
6. R. D. Arnold and T. O. Binford, Geometric constraints in stereo vision, SPIE Vol. 238--Image Processing for Missile Guidence, 1980.
7. S. T. Barnard and M. A. Fischler, Computational stereo, Computing Surveys, Vol.14, 553-572, 1982.
8. W. H. Tsai, S. I. Gao and N. Y. Yeh, 3D object recognition by line features using binocular images, Information Science and Engineering, Vol. 2, 99-123, 1986.
9. A. M. Waxman and J. H. Duncan, Binocular image flows: step toward stereo-motion fusion, IEEE Trans. Pattern Analy. Mach. Intell., 715-729, 1986.
10. D. Marr, Vision, San Francisco, CA, Freeman Co., 1982.
11. B.K.P. Horn and M.J. Brooks, The variational approach to shape from shading, Computer Vision, Graphics and Image Processing 33, 174-208, 1986.
12. J. S. Huang and D. H. Tseng, Statistical theory of edge detection, Submitted to Computer Vision. Graphics and Image Processing, 1986.

13. A.P. Pentland, Finding the illumination direction, J. Optical Soc. Amer. 72, 1982.

14. J. S. Huang and D. H. Tseng, Statistical detection and estimation of signal change with application to image processing, Proceeding of National Computer Symposium, Taiwan, Rep. of China, 1985.

15. M.J.B. Duff, Computing Structures For Image Processing, Academic Press, London, 1983.

16. G. Flower, R. Haralick, G. Gray, C. Feustel and C. Grinstead, Efficien graph automorphism by vertex partitioning, Artificial Intelligence 21, 245-269, 1983.