Recognition of Objects in the Real World from a Systems Perspective

Jan-Olof Eklundh, Mårten Björkman
CVAP, NADA, KTH, Stockholm

Based on a discussion of the requirements for a vision system operating in the real world we present a real-time system that includes a set of behaviours that makes it capable of handling a series of typical tasks. The system is able to localise objects of interests based on multiple cues, attend to the objects and finally recognise them while they are in fixation. A particular aspect of the system concerns the use of 3D cues. We end by showing the system running in practice and present results highlighting the merits of 3D-based attention and segmentation and multiple cues for recognition.

1 Introduction

A crucial problem in cognitive vision concerns recognition and categorisation of objects. Despite its simple formulation this problem entails many deep issues. For instance, it is far from obvious what constitutes an object. This depends as much on the seeing agent and what it is doing as on the world itself. However, even if the objects and classes of objects are given, as is the case in much of the recognition literature, there are a number of remaining problems. Indeed, there exists a number of sophisticated methods that give high performance both on single objects [5, 18] and objects in real, cluttered scenes [19, 15, 8, 13]. A careful scrutiny of these methods shows two things. First, they generally assume that the relevant objects subtend a considerable part of the visual field and often that they are centrally located in the image. Secondly, they do not use 3D information, even implicitly.

The first assumption implies that there is enough information in the image to identify the objects to begin with and that search is no serious problem. The second one is based on the generally accepted idea that recognition can be performed without resort to 3D cues. However, a question is how realistic these assumptions are for an artificial "seeing system" operating in a real environment. When we humans enter a room or look at, say, a table from a few meters distance, we can locate objects of interest over our full visual field. We can then shift our gaze to fixate on them and in such a way obtain foveal resolution on features and parts of relevance. This process includes both attention and figure-ground segmentation and uses 2D appearance as well as 3D structure, e.g what is far and what is near and what belongs together in 3D. In a sense these processes provide views and cues similar to those assumed by many existing recognition algorithms. The question we raise in this work is if processes of the same type will enable an artificial vision system to perform recognition in the real-world and do so more appropriately than existing algorithms, and hence form a basis for cognitive visual processing in the real world.

In this vein we have developed a system capable of locating and recognising objects in real scenes. It does so by sequentially attending to different objects, segmenting them from ground in a wide field of view, in particular using 3D cues from binocular disparities or motion, and then foveating and fixating them to get a view suitable for recognition. Technically, the system is implemented on a stereo head provided with two binocular camera pairs, a wide field pair for attention and a foveal one for recognition. We consider the case in which the head and the objects being studied remain static. It has been shown in earlier work [4, 16] that observer or object motion provide strong cues to figure-ground segmentation. However, in many applications the objects to be recognised are in fact static and motion cues can not be exploited. Thus the scenario studied here is typically harder than if motion were present.

The emphasis of the paper lies on the systems aspects and how to obtain a system that functions in the real world. Most techniques used are known from previous work. However, there are issues that only become apparent, if one tries to integrate different components to create a fully operational system running in a closed loop with reality. Robustness and continuous operation is then often preferable to accuracy, if one has to choose. For a more thorough discussion on implementational considerations, see [3].

2 The system and its tasks

It is possible to determine a set of capabilities that a visual system of the type we have in mind needs to have, capabilities that are more or less independent of the tasks the system is supposed to perform. The system needs the ability to segregate physical objects from each other. Different parts of the visual field need to be grouped so that the extent and location of objects can be determined. This is traditionally done by various image grouping and segmentation techniques. In this work we rely heavily on 3D cues. Humans who seem to perform segmentation in the real world without effort obviously have access to such cues. In the presented system we especially use binocular information. To relate binocular information to distances a complication is that the camera

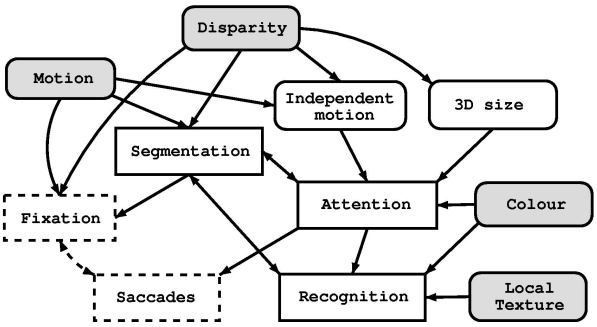

Figure 1: The flow of information

system has to be calibrated in some sense. Since the system operates continuously calibration should also be performed on-line. Another process that might benefit from 3D cues is attention. In order to extract the relevant information from the continuous flow of image data, attention can be used to highlight regions that stand out in relation to its surroundings. Motion is a particularly strong cue for attention. We have included motion in earlier work [1], but as mentioned above we are here considering static objects and therefore mainly use binocular stereo information. When an object is searched for, the attentional process may guide the camera system to analyse the most likely parts of the scene first. What is discriminant in search is related to the task at hand. At the same time there is a need for a bottom-up system, independent on task and objects, so that salient patterns and unexpected events trigger the system and the observer can react accordingly. Without knowing exactly what to expect, the system needs to consider every possible cue. Strategies based on multiple cues become a necessity. The processes described are in turn based on a set of low level processes. These include control of the oculomotor system in terms of fixation and saccading and tracking of moving objects, and of course also a set of algorithms for extracting information from the images.

To summarise, our system has a set of basic visual processes, fixation and attention processes, and modules for object recognition and figure-ground segmentation. These processes work in parallel, at different time frames, and share information through asynchronous connections. The flow of information through the system is summarised in Figure 1. The task here is to locate and recognise certain learned objects in cluttered real world scenes. Hence, attention is driven by requests for specific objects, which invoke cues characteristic to these objects. It could of course also be driven by other top-down requests, or bottom-up by scene properties and events. Given the attentional cues, the wide field cameras trigger the foveal cameras to fixate, possibly after a rapid gaze shift (these cameras are continuously striving to fixate on some point in the world). Ideally, fixation is shifted to and kept on one of the objects of interest, which then can be recognised, possibly after it has been segmented from its background. It is worth noting that, since fixation is kept on objects, the recognition module normally has some time to "understand" what specific object it is looking at.

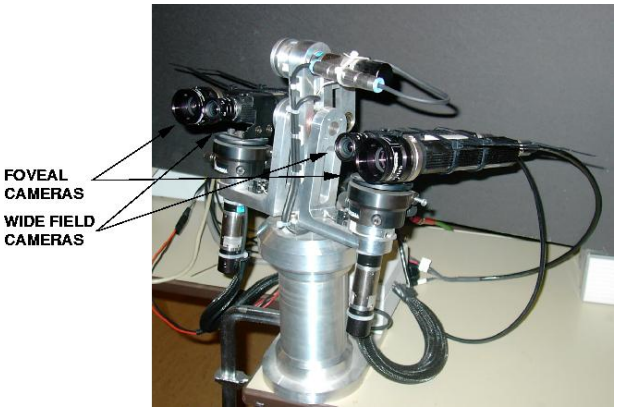

Figure 2: The Yorick stereo head [20].

In the work reported here we consider recognising exemplars of 3D static objects. That is relevant in several applications of e.g. robotics and also makes the use of 3D information for attention crucial. There is nothing in the system architecture preventing us from dealing with classes of objects. However, our present learning and recognition mechanisms are not suited for this more generic case. We now give a short summary of the most important characteristics of the system, followed by an account of some experiments on the intended tasks. A more detailed technical description of the individual components can be found in [3].

Figure 3: A table-top scene (left) with disparities estimated using sums of absolute differences (right).

3 Stereopsis

Our system has two sets of stereo cameras, with focal lengths 6 mm and 28 mm respectively. The wide field cameras guide the foveal ones so that recognition can be performed foveally, while objects of interest are found in a wider field. This is just one particular realisation,but the methods proposed could be applied to other set-ups.[1] We are currently working on a con-

[1]A similar system based on zoom-lenses has previously been demonstrated by Green and Nelson [10]. However, with foveation based on zooming, the system will not be able to react to peripheral changes in the scene, while an object is foveated.

figuration in which the two stereo pairs are separated by as much as a meter. Dense disparities, calculated from sums of absolute difference correlations [12], are used to slice up the scene in depth to aid figure-ground segmentation. These disparities need to be related to actual metric distances. Hence the epipolar geometry has to be determined, which in our system is done using corner features [11] and a non-linear method based on the optical flow constraint [14]. Moreover, the two camera sets need to be calibrated externally against each other. Since the overlap between the visual fields can never be larger than the foveal fields and the foveal fields are narrow in our case, we use affine geometry for this purpose. For details of the external calibration and the continuous updating of the epipolar geometry we refer to [3]. An example of a disparity map in the wide field can be seen in the right image of Figure 3. Darker regions are located closer to the observer, whereas white areas show texture-less regions for which no disparities could be determine.

4 Attention and scene search

These processes lie at the heart of our model. We will therefore describe them in some detail. The aim of the attentional process is to deliver a set of hypotheses of where a particular requested object might be located in the scene and guide the foveal system towards the most likely areas. This process is in the current experiments based on two kinds of information, 3D size and hue. These cues are object specific and relatively insensitive to varying viewing conditions. Since calculations are performed across the whole viewing space, they need to be fast enough in relation to possible changes in the scene. Incorrect hypotheses are acceptable as long as the correct one will eventually be attended to.

Since the projected size of an object depends on its depth and filtering is performed in image space, the 3D space is divided into a number of layers, that are processed one at a time. Each layer represents points that are located within a certain depth range and the width of such a range is determined by the size of the requested object. Since a particular object might be split between two different layers, the layers are three times overlapped. From the disparities we get a binary map of points for each depth layer. The maps are weighted by a hue saliency map, explained below, and blob-like features are detected using differences of Gaussians (DoG) [17]. The sizes of the Gaussian filters are set such that the largest responses are generated from blobs of the requested size in 3D. The peaks are finally recorded as the most salient regions within the observed scene.

Hue saliency is used in order to enhance blobs of a particular requested hue. It is computed from correlations of hue histograms. Local histograms are computed around each pixel and the corresponding saliency is determined in relation to the hue of the requested object, using normalized cross-correlation of hue histograms. Figure 4 shows two different examples of such hue saliency maps.

A new gaze direction is determined from the peaks of the DoG filtered images. Among the largest peaks, one will be selected as the new gaze direction. The gaze is then shifted through a saccade. In order to cope with situations in which a requested object does not generate the largest peak, either due to occlusions or nontypical lighting conditions, random noise equivalent to about 20% of a largest peak, is added to the peaks prior to selection of the largest one. Thus the requested object will eventually generate the strongest peak, so that it can be foveated and recognised, even though it might take a couple of saccades. For the gaze to be directed towards a region of interest as seen by the foveal cameras, not by the wide field ones where regions are localised, affine transfers of hypothesised image positions are made.

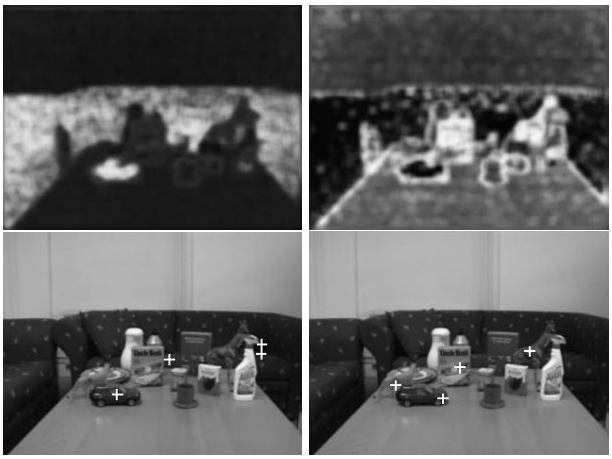

Figure 4: Hue saliency maps (first two images) when a blue car and orange package are sought. The four largest peaks in each case are shown below.

Inhibition on return is implemented by recording previous gaze directions in relation to the orientation of the observer itself. The radii of observed objects are also recorded and the system is prevented from immediately returning to directions within these radii. While the system is running the radii are gradually reduced, allowing the system to return to the same direction after a long enough period. Following each gaze change, the observed object is recognised, as will be described in Section 6. This is repeated until the requested object is found or the system times out, typically after about five saccades. In a more generic situation it could have been another task that ended the search. However, no other tasks have been included in these experiments.

5 Fixation

The foveal cameras are always striving towards fixation, i.e. the optical axes intersect on a physical object somewhere in the scene. Fixation is especially important for a moving observer, but estimation of cues such as shape and motion is always simplified if the object of interest is in fixation [7]. With the object in the centre of the field of view, the amount of overlap between the left and right cameras is as large as possible. Thus fixation is desirable for a binocular recognition system, since the amount of available relevant data as well as the quality of data is maximized.

When a saccade has been executed towards a region of interest, a number of rapid vergence shifts are performed gradually decreasing the disparities around the observed region. Since the depth is approximately known prior to the saccade, a rough estimate of the vergence angle is available. However, due to the large difference in focal lengths, a small error in the wide field cameras can be considerably larger in the foveal ones. Hence, simple image correlation or gradient based methods are hardly sufficient during the first few critical frames following a saccade. The affine fundamental matrix between the images is instead estimated using corners from the two foveal images and the data set is pruned from false matches, that would otherwise affect the quality of the vergence estimate. Disparities are finally measured along the epipolar lines and the vergence angle is updated accordingly, after being processed by a Kalman filter.

To aid recognition the system attempts to segment the observed foreground object from its background and neighbouring objects, using a disparity map generated from the foveal images. However, as is described in [3] disparities are very hard to determine in the foveal view. It may still work successfully on well textured objects. Examples are shown in Figures 5 and 6. Objects are segmented by applying a Mean Shift algorithm [6] on the reconstruction points available from the computed disparities. A cluster center in 3D is found and points close to the center are assigned to the foreground. The results are finally processed by a series of morphological operations.

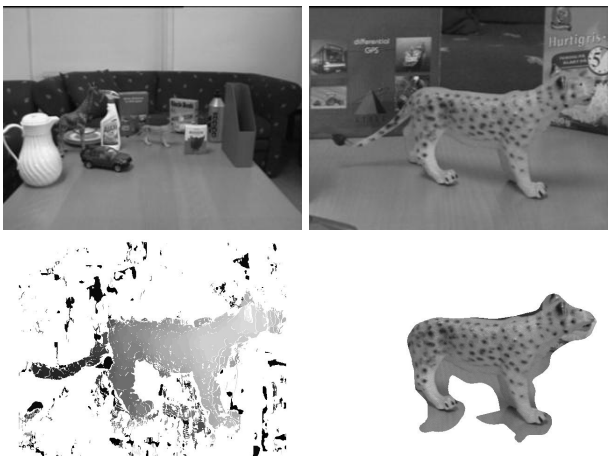

Figure 5: The wide field and foveal left images of a table-top scene, while fixating on the leopard object. The last two images show the disparity map and the resulting mask.

6 Recognition

The task that we used to test the presented system was to automatically detect previously learned objects in a scene. Object hypotheses are delivered by the attentional system, presented in Section 4. Once at fixation an object is automatically segmented from its background. The purpose of the segmentation process is to yield more distinct detection scores in situations when the scene is heavily cluttered. If the recognition system fails to detect the requested object at a particular fixation point, the system performs another saccade to a new fixation point, until the object has been found or the system times out. Since objects, if at all in the scene, are typically found within a couple of saccades, we use a limit of five saccades before a time out is signaled.

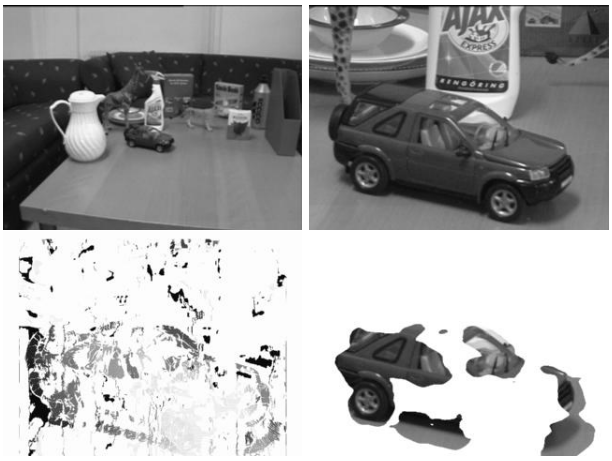

Figure 6: The wide field and foveal left images of a table-top scene, while attending to the blue car. The last two images show the disparity map and the corresponding mask.

We performed a series of experiments on 24 everyday objects viewed in 26 table-top scenes. 240 different search tasks were executed, each involving 5 saccades. As cues for recognition we used SIFT features [15] and colour histograms [9]. Every object was searched for 10 times, out of which 6 involved the object actually appearing in the scene. A total of 32 failures were observed, 25 true and 7 false ones. In none of these searches the requested object failed to be placed in the center of view, when it was in fact present. The foveated segmentation failed at 6 different occasions. A segmentation was regarded as successful if at least 2/3 of the segment covered the requested object and more than 1/3 of the object was covered by the segmentation. These results show that the loop of attention foveation and fixation functioned very well and gave desired output, even though the ensuing recognition process was not always successful.

An example of the system running in practice can be seen in Figure 7. The delay between two consecutive images is about a second, which means that most saccades are left out, since a saccade only lasts for a few tenths of a second. The first task is to find a grey bottle. Since the hue of this bottle is not particularly distinct, it finds the bottle first after three saccades. Initially it finds a white textured bottle of similar size and then an orange spotted toy animal. At both these fixation points the recognition system concludes that the wrong object has been found. Finally the grey bottle is reached and successfully detected. Next the blue car is requested. Due to its distinct colour it is immediately detected. However, due to the large number of specularities

Figure 7: Images taken while the system was involved in a series of search tasks.

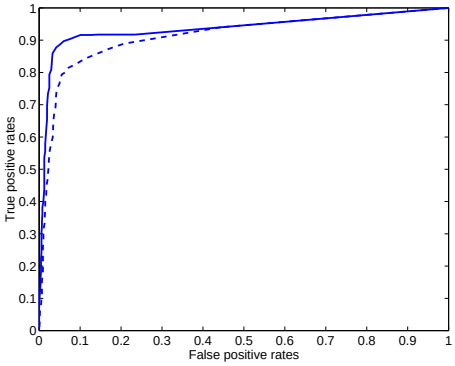

Figure 8: ROC curves for SIFT feature based object detection, with (solid) and without (dashed) foveated figure-ground segmentation.

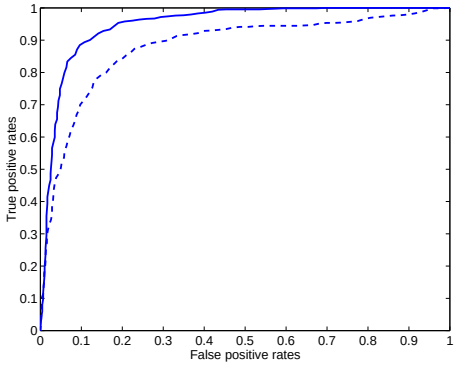

Figure 9: ROC curves for color histogram based object detection with segmentation based on disparities (solid) or rough size and position estimates (dashed).

it takes a while for the fixation process to settle. The following object to be found is a large blue box. Since the hue is much similar to that of the car and quick saccade is made to the back side of the car, until the box is found in the next saccade. The last object is a very distinct one, an orange textured box, which is found in a single saccade. The overlayed frames around the rectangular boxes indicate the computed pose. In another study these frames were used to initialize pose estimation for manipulation [2].

Due to the small projective sizes of observed objects, recognition directly in the peripheral view was not possible, at least not with the cues used here. Two other questions pertaining to our system architecture are: *Does segmentation really improve recognition performance*, and *It is of any benefit to use more than one cue?* To answer these questions we analysed the recognition system in isolation, using those saccades in which a physical object, known or unknown, was successfully segmented. This set constitutes 886 of the total 1200 saccades. The resulting foveal images were manually annotated with the identity of the segmented object. To the set we added an equal amount of false object identities, so as to analyze the false detection results. In Figure 8 two Receiver Operating Characteristic (ROC) curves can be seen illustrating the detection performance, with and without segmentation, when only SIFT features were used. The improvement due to segmentation and the use of a relative detection criterion is significant. Similar curves for detection based on color histograms can be seen in Figure 9, while Figure 10 shows the two cues combined using SVM. Here the 2D SVM was trained to deliver a detection hypothesis, given the detection scores of each respective method.

The experiments indicate that the answers to both questions are "Yes". This is not surprising, but the point is that our system realises such mechanisms in a successful manner.

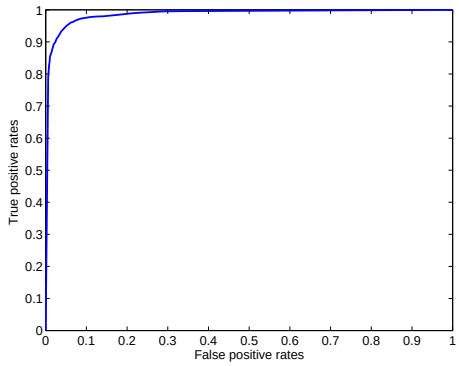

Figure 10: ROC curves when SIFT features and color histograms are combined for object detection using a 2D support vector machine.

7 Conclusions

In the paper we have discussed the prerequisites for a system capable of detecting, localising, recognising and also classifying objects in the real world. We have proposed a set of basic capabilities such a system might need and on the basis of these principles presented a system capable of performing such tasks operating continuously by watching and interacting with the real world. The system uses 3D cues in combination with object appearance both to find and to identify objects. By combining multiple cues we increase recognition performance. Future work concerns including still more cues for attention as well as recognition, and adaptation and learning at all levels. We conclude from our findings that this can be expected to give us even higher robustness and performance, and that it also will allow us to consider other tasks. In fact, using the system and its successors on other types and more complex tasks of perception and action is our main future direction of work.

References

[1] M. Björkman and J.O. Eklundh, "A Real-Time System for Epipolar Geometry and Ego-Motion Estimation", in Proc. IEEE Computer Vision and Pattern Recognition, Vol. 2, Hilton Head, SC, pp. 506–513, Jun 2000.

[2] M. Björkman and D. Kragic, "Combination of Foveal and Peripheral Vision for Object Recognition and Pose Estimation," in Proc. IEEE Int'l Conf. Robotics and Automation, Apr. 2004.

[3] M. Björkman and J.O. Eklundh, "Vision in the Real World: Finding, Attending and Recognizing Objects," to appear in Int'l Journal Imaging Systems and Technology.

[4] K. J. Bradshaw, P. F. McLauchlan, I. D. Reid, and D. W. Murray, "Saccade and Pursuit on an Active Head/Eye platform," Image and Vision Computing, Vol. 12, No. 3, pp. 155–163, 1994.

[5] R. Brunelli and T. Poggio, "Face Recognition: Features Versus Templates," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 15, pp. 1042–1052, Oct. 1993.

[6] D. Comaniciu and P. Meer, "Mean Shift Analysis and Applications," in Proc. IEEE Int'l Conf. Computer Vision (ICCV 99), Kerkyra, Greece, pp.1197-1203, 1999.

[7] K. Daniilidis and I. Thomas, "Decoupling the 3D Motion Space by Fixation," in Proc. European Conf. Computer Vision (ECCV 96), Cambridge, UK, pp.I:685–796, 1996.

[8] L. Fei-Fei, R. Fergus, and P. Perona, " A Bayesian Approach to Unsupervised One-Shot learning of Object Categories," in Proc. IEEE Int'l Conf. Computer Vision (ICCV 03), Nice, France, pp. 1134–1141, Oct. 2003.

[9] T. Gevers and A. W. M. Smeulders, "A Comparative Study of Several Color Models for Color Image Invariants Retrieval", in Proc. Int'l Workshop Image Databases & Multimedia Search, Amsterdam, The Netherlands, August 1996, pp. 17–26.

[10] I. A. Green and R. C. Nelson, "Segmentation Propagation During a Camera Saccade," TR-766, Computer Science Dept, U. Rochester, Nov 2002.

[11] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," in Proc. Alvey Vision Conf., Manchester, UK, pp. 147–151, 1988.

[12] K. Konolige, "Small Vision Systems: Hardware and Implementation," Intl. Symp. Robotics Research, Salt Lake City, UT, pp. 203–212, 1997.

[13] B. Leibe and B. Schiele, "Interleaved Object Categorization and Segmentation," in Proc. British Machine Vision Conference (BMVC 03), Sep. 2003.

[14] H. Longuet-Higgins, "The Interpretation of a Moving Retinal Image," Philosophical Trans. Royal Society of London, B-208, pp. 385–397, 1980.

[15] D. G. Lowe, "Object Recognition From Local Scale-Invariant Features," in Proc. IEEE Int'l Conf. Computer Vision (ICCV 99), Kerkyra, Greece, pp. 1150–1157, Sep. 1999.

[16] A. Maki, P. Nordlund and J.O. Eklundh, "Attentional Scene Segmentation: Integrating Depth and Motion," Computer Vision Image Understanding, Vol. 78, No. 3, pp. 351–373, Jun. 2000.

[17] D. Marr and E. Hildreth, "Theory of Edge Detection," Proc. the Royal Society of London B, Vol. 207, pp.187–217, 1980.

[18] H. Murase and S. K. Nayar, "Visual Learning and Recognition of 3D Objects from Appearance," Int'l J. Computer Vision, Vol. 14, No. 1, pp 5–24, Jan. 1995.

[19] R. C. Nelson and A. Selinger, "A Cubist Approach to Object Recognition," in Proc. Int'l Conf. Computer Vision (ICCV 98), Bombay, India, pp. 614–621, Jan. 1998.

[20] P. M. Sharkey, D. W. Murray, S. Vandevelde, I. D. Reid and P. F. McLauchlan, "A Modular Head/Eye Platform for Real-time Reactive Vision," Mechatronics, Vol. 3, pp. 517–535, 1993.