Combination of Foveal and Peripheral Vision for Object Recognition and Pose Estimation

Mårten Björkman and Danica Kragic
Centre for Autonomous Systems & Computer Vision and Active Perception Lab
Royal Institute of Technology, Stockholm, Sweden, Email:{celle, danik}@nada.kth.se

***Abstract*— In this paper, we present a real-time vision system that integrates a number of algorithms using monocular and binocular cues to achieve robustness in realistic settings, for tasks such as object recognition, tracking and pose estimation. The system consists of two sets of binocular cameras; a peripheral set for disparity based attention and a foveal one for higher level processes. Thus the conflicting requirements of a wide field of view and high resolution can be overcome. One important property of the system is that the step from task specification through object recognition to pose estimation is completely automatic, combining both appearance and geometric models. Experimental evaluation is performed in a realistic indoor environment with occlusions, clutter, changing lighting and background conditions.**

I. Introduction

In service robot frameworks as the one considered here, a visual system is required for autononomous navigation, object manipulation and grasping which, for general environments, requires high degrees of flexibility and robustness. This paper considers the problems of vision based scene segmentation, object detection and recognition, object pose estimation and tracking using both monocular and binocular cues. Compared to other systems, our system consists of a large number of integrated processes, that have previously been considered independently or in relatively simple settings.

The presented vision system is heavily based on the *active vision* paradigm, [1]. Instead of just passively observing the world, it actively changes the viewing conditions such that the most accurate results are obtained, in relation to the task at hand. Our particular system consists of two pairs of stereo cameras: a peripheral camera set and a foveal one. Recognition and pose estimation can be done using either one of these, depending on the size of and distance to an observed object. From segmentation based on binocular disparities, objects of interest are found using the peripheral camera set, which then triggers the system to perform a saccade, moving the object into the centre of the foveal cameras. Thus a combination of a large field of view and high image resolution can be achieved, without sacrificing the performance.

Compared to one recent system, [2], our system i) uses both hard (detailed models) and soft modeling (approximate shape) for object segmentation, and ii) choice of binocular or monocular cues depending on the task. In addition, we believe that our approach can easily be used with different visual servoing configurations, i.e. eye-in-hand (foveal camera) and stand-alone camera (peripheral camera) thus offering a trade-off between accuracy and speed.

II. The system

Figure 1 shows a schematic overview of the basic building blocks of the system. These blocks do not necessarily correspond to the actual software components, but are shown in order to illustrate the flow of information through the system. For example, the visual front end consists of a several components, some of which are running in parallel and others hierarchically. On the other hand, action generation, such as initiating 2D or 3D tracking, is distributed and performed across multiple components.

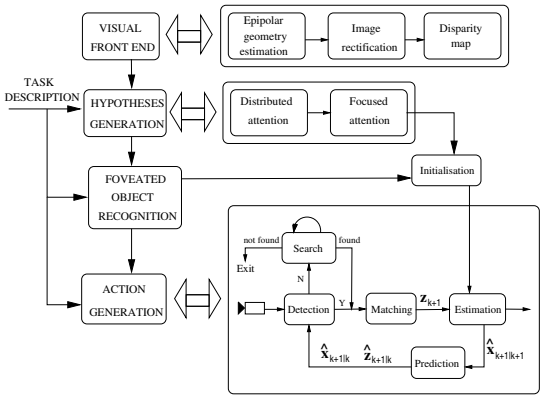

Fig. 1. Basic building blocks of the system.

The most important building blocks are:
1) The Visual Front-End is responsible for the extraction of visual information needed for figure-ground segmentation and other higher level processes.
2) Hypotheses Generation produces a number of hypotheses about the objects in the scene that may be relevant to the task at hand. The computations are moved from being distributed across the whole image to particular regions of activation.
3) Recognition is performed on selected regions, using either corner features are color histograms, to determine the relevancy of observed objects.
4) Action Generation triggers actions, such as visual tracking and pose estimation, depending on the outcome of the recognition and current task specification.

Due to the complexity of the software system, it was partitioned into a number of smaller modules that communicate through a framework built on a interprocess communication standard called CORBA (Common Object Request Broker

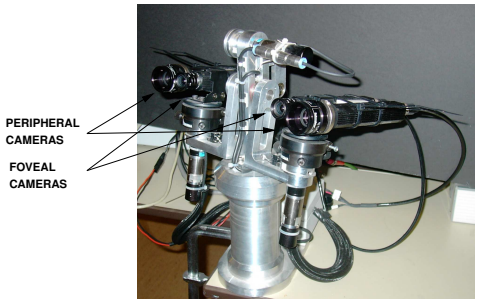

Fig. 2. The Yorick stereo-head.

Architecture).The current system consists of about ten such modules, each running at a different frame rate. The lowest level frame grabbing module works at a frequency of 25 Hz, while the recognition modules is activated only upon request.

The experimental evaluation has been performed on a 1.2 GHz dual Athlon MP computer running under the Linux operating system. The binocular stereo-head used is shown in Figure 2. This stereo-head, known as Yorick, [14], has four mechanical degrees of freedom; neck pan and tilt, and pan for each camera in relation to the neck. The head is equipped with two pairs of Sony XC999 cameras, with focal lengths 28 mm (14° FOV) and 6 mm (60° FOV) respectively. Even if this is just one particular system, the software components may easily be changed to fit similar systems, since the knowledge of the stereo-head is limited to the intrinsic camera parameters and the length of the baseline.

III. Visual front-end

For an autonomous robot operating in a dynamic world where unexpected events may occur, it is essential to provide a constant flow of reliable data from the surrounding environment. In order to extract metric information, e.g. sizes and distances, about objects observed by the robot, the presented system relies heavily on binocular information. The reason for using multiple cameras is the fact that it simplifies the problem of segmenting the image data into different regions representing objects in a 3D scene. This is often referred to as *figure-ground segmentation*. In cluttered environments figure-ground segmentation is particularly important and difficult to perform and commonly the reason for experiments being performed in rather sparse, simplified environments. This is especially true in scenarios such as ours, where a robot arm is to be transported to the vicinity of an object.

Since the field of view of a typical camera is quite limited, binocular information can only be extracted from those parts of the 3D scene that are covered by both cameras' field of view. In order to make sure that an object of interest is foveated by both cameras, the head is able to actively change gaze direction and vergence angle, i.e. the difference in orientation between the two cameras. However, since our final goal is robotic object manipulation and grasping, we have also integrated a number of monocular visual algorithms in the system, such as those involving pose estimation and tracking.

A. Epipolar geometry

The presented system uses binocular disparities to perform figure-ground segmentation and guide the robot towards potential objects of interest. If the epipolar geometry is known it is possible to relate these disparities to actual metric distances. Instead of relying on the motor counters of the stereo-head, the epipolar geometry is estimated continuously from image data alone. The reason for this is that small disturbances such as vibrations and delays introduce significant noise to the estimation of the 3D structure. In fact, an error of just one pixel leads to an error in depth of several centimeters on a typical manipulation distance.

The epipolar geometry is estimated robustly using Harris' corner features, [6]. These corners are extracted and matched between images using normalized cross-correlation. The orientation of the baseline $(t_x, 0, t_z)$ measured in the first camera frame, vergence angle ω_y, relative tilt ω_x and rotation around the optical axes ω_z are sought using a model of the disparities,

$$\begin{pmatrix} dx \\ dy \end{pmatrix} = \begin{pmatrix} (1+x^2)\,\omega_y - y\omega_z \\ xy\,\omega_y + \omega_x + x\omega_z \end{pmatrix} + \frac{1}{Z}\begin{pmatrix} t_x - xt_z \\ -yt_z \end{pmatrix}, \quad (1)$$

where Z is the unknown depth of a point at image position (x, y). Nonlinear optimization is performed using a combination of RANSAC [7] for parameter initialisation, and M-estimators for improvements. In our previous work we have experimentally shown that this optical flow based model, [13], is more robust than the essential matrix in the case of binocular stereo heads, [8], even if the essential matrix leads to a more exact description of the epipolar geometry, [12].

B. Disparities

Since most efficient methods for dense disparity estimation assume the image planes to be parallel, rectification has to be performed using the estimated epipolar geometry before disparities can be estimated. The current system includes seven different disparity algorithms, from simple area correlation, [15] to more complicated graph-cut methods, [16]. The benefit of using a more advanced global method, is the fact that they often lead to denser and more accurate results. However, even if density is important, the computational cost of these methods makes them infeasible for our particular application. From our experiments we have concluded that denser results rarely justify the increased complexity of the global methods, and that simpler methods tend to be more robust in practice. The second image of Figure 4 shows an example of disparities calculated using sums of absolute differences.

In summary, the visual front-end of the presented system is responsible for delivering 3D data about the observed scene. Such information is extracted using a three-step process, which includes epipolar geometry estimation, image rectification and calculation of dense disparity maps. The generation of this data is done continuously at a rate of 6 Hz, independently of the task at hand and used by more high-level processes for further interpretation. Further information on this part of the system can be found in [9].

IV. Hypotheses generation

The purpose of this component is to derive qualified guesses of *where* a requested object might be located in the current scene. As mentioned above, this step is performed using the peripheral cameras, while the recognition is done foveated.

A. Distributed attention

Unlike focused attention, distributed attention works on the whole image, instead of being concentrated to a particular region. From binocular disparities a target region, that might represent an object of interest, is identified. Even if the current system is limited to disparities, it is straightforward to add additional cues, such as in the model proposed in [17]. Here, we have concentrated on disparities because they contain valuable information about object sizes and shapes. This is especially important in a manipulation task, where the color of an object might be irrelevant, whereas the size is not.

The only top-down information needed for hypotheses generation is the size of the requested object and a predefined depth range. The system sweeps through this range, sequentially considering subranges equivalent to the size of the object. For each subrange, a binary maps containing those points that are located within the range is created. The third image of Figure 4 shows such a map overlayed on-top of the corresponding left peripheral image. Initial hypotheses are then generated from the results of difference of Gaussians filters applied to the binary maps. The scales of these filters are set so as to maximize responses of image blobs representing objects of the requested size and the corresponding distances.

B. Focused attention

From the object hypotheses, a target region is automatically selected so that the gaze can be redirected and recognition performed using the foveal cameras. The system selects the hypothesis corresponding to the largest peak of the differences of Gaussians. Noise, equivalent to 20% of the largest peak, is added to the peaks prior to selection, in order to prevent the system from being stuck at local minima. When executed in a loop, multiple saccades can be executed until recognition finally verifies that the requested object has been found.

Since hypotheses are found in the peripheral camera frames and recognition is performed using the foveal ones, the relative transformations have to be known. These are found applying a similarity model to a set of Harris' corner features similar to those used for epipolar geometry estimation in Section III-A. The relative rotations, translations and scales are continuously updated at 2 Hz. Using this information the system translates the target positions into the foveal camera frames.

Before a saccade is executed the target position is refined in 3D. During a couple of image frames, a local high-resolution disparity map is calculated around the target area. A mean shift algorithm, [18], is run iteratively updating the position from the cluster of 3D points around the target position, represented by the disparity map. The maximum size of this cluster is specified using the size of the requested object.

V. Recognition

In the current system, two recognition modules are available: i) a feature based module based on Scale Invariant Feature Transform (SIFT) features, and ii) an appearance based module based on color histograms. Both of these methods are presented in some detail in the following sections. The choice of recognition algorithm is determined prior to task execution. We are currently evaluating the performance of both methods for different types of objects with the goal of integrating them into a more robust combined recognition process.

A. Feature Based

In a recent study, Mikolajczyk and Schmid tested a large number of interest point descriptors and their behaviors under scale and illumination changes, [19]. The most robust performance was obtained using the SIFT descriptor of Lowe, [20], which has been thus chosen for recognition in our system. The descriptor consists of local histograms of gradient directions and is invariant to scale and rotation. Some flexibility in translation is tolerated as well. Interest points locations are found using a series of difference of Gaussians on varying scales. For each SIFT feature extracted from the incoming image, the closest feature among those stored in the database is found and the corresponding model is given a vote accordingly. For matching, a metric based on cross-correlation is applied. The model that receives the most votes is selected for further consideration, as long as there are more than 10 such votes.

B. Appearance Based

In a previous study [5], we extended the recognition scheme based on the Cooccurrence Color Histograms (CCHs) that was originally proposed by Chang [10], to pose estimation. The histograms are used in a classical learning framework that facilitates a winner–takes–all strategy across scales. The major advantages of this two–step appearance based method are its robustness and invariance to scale and translation. The method is also computationally efficient since both recognition and pose estimation rely on the same object representation.

VI. Action Generation

After the image position of the object is available, object tracking may start depending on the given task. These tasks include for example visual servoing or target tracking.

A. 2D Tracking

Our 2D tracking system is based on integration of multiple visual cues (motion, colors and gradients) where *voting* is used as the underlying integration framework, [4]. Cues are fused using weighted super-position and the most appropriate action is selected according to a winner-take-all strategy. The advantage of a voting approach is that information of different cues can easily be combined without the need for explicit models as it is, for example, the case with Bayesian approaches. Lots of perceptual experiments support the idea that, when it comes to aspects of visual scenes, people most likely mention color, form and motion as being distinct.

Fig. 3. Objects used for experimental evaluation.

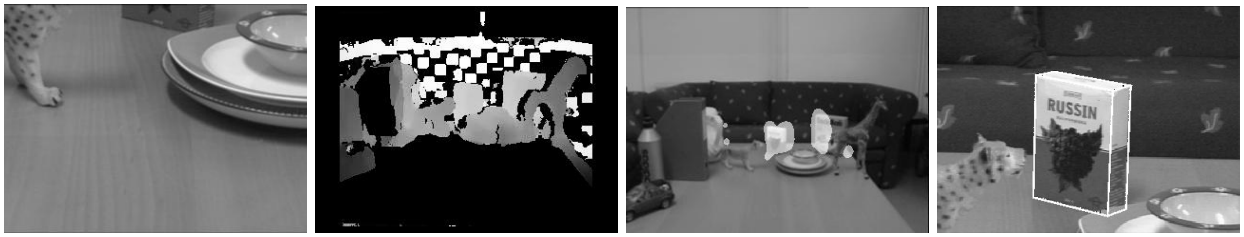

Fig. 4. Figure-ground segmentation and pose estimation. The first image shows the foveal image before a saccade has been issued. A disparity map can be seen in the second image with object hypotheses shown in the third. The last image show the pose of the recognised object being correctly estimated.

B. Pose Estimation for 3D Tracking

For robotic manipulation, it is usually required to accurately estimate the pose of the object to, for example, allow the alignment of the robot arm with the object or to generate a feasible grasp and grasp the object. There are three major steps in our model based tracking system: (1) *Initialisation* - the recognition modules presented in Section V are used here to provide an approximation to the current object pose, (2) *Pose Estimation* - the initialisation step is followed by a local fitting method that uses a geometric model of the object, and (3) *Pose Tracking* - the system provides a pose estimate of the object, if the object or the camera start to move. For pose estimation and tracking, the method proposed in [11] was extended as presented in [3] by integrating robust estimation and feature detection techniques.

VII. Experimental evaluation

The presented system was considered as an integrated unit and its performance measured based on the behaviour of the complete system. The failure of one particular module does not necessarily mean that the whole system fails. For example, figure-ground segmentation might well fail to separate two nearby objects located on a similar distance, but the system might still be able to initiate pose estimation after recognition. A number of properties of the system have been evaluated, as will be described in more detail in the sections below.

For recognition, the set of objects shown in Figure 3 was considered. A database consisting of object models based on SIFT features and CCHs was created. Only one view per object was used for the SIFT models, while the CCHs were based on multiple views. At a later stage we hope to expand the database to include more models and views. Pose estimation was only considered for the first three box-like objects. For this purpose, the width, height and thickness of these objects had to be included to the database.

Due to the limitations of the database and the fact that the observed recognition scores did not significantly differ from those already published in [20] and [5], we have chosen not to include any additional quantitative results. However, observations have lead us to believe that recognition would benefit from CCHs and SIFT features being used in conjunction. For example, the blue car is rarely recognized properly using SIFT, since the most salient features are due to specularities. However, the distinct color makes it particularly suitable for CCHs, which on the other hand have a tendency of mixing up the tiger and the giraffe, unlike to recognition module based on SIFT features.

A. Binocular Segmentation and Pose Estimation

The first experiment illustrates the typical behavior of the system with binocular disparity based figure-ground segmentation and SIFT based recognition. Results from these experiments can be seen in Figure 4. The first image shows the left foveal camera image prior to the experiment. It is clear that a requested object would be hard to find, without peripheral vision controlling the gaze direction. However, from the disparity map in the second image the system is able to locate a number of object hypotheses, which are shown as white blobs ovarlayed on-top of the left peripheral camera image in the third image of the figure.

The recognition score for this example was 70%, measured as the fraction of SIFT features being matched to one particular model. Once an object has been recognised, pose estimation is automatically initiated. This is done using SIFT features from the left and right foveal camera images, fitting a plane to the data. Thus, it is assumed that there is a dominating plane that can be mapped to the model. The process is further improved by searching for straight edges around this plane. The last image show an example of this being done in practice.

B. Monocular CCH Recognition and Pose Estimation

Figure 5 shows an example of recognition and pose estimation based on monocular CCH. Here, object recognition and rotation estimation provide the initial values for the model based pose estimation and tracking modules. With the incomplete pose calculated in the recognition (first figure from the left) and orientation estimation step, the initial full pose

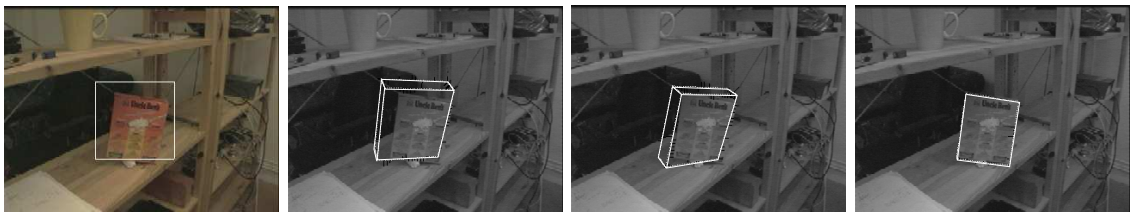

Fig. 5. From object recognition to pose estimation, (from left): i) the output of the recognition, ii) initial pose estimation, iii) after few fitting iterations, iv) the estimated pose of the object.

Fig. 6. The effect of imperfect segmentation on object localisation.

is estimated (second figure from the left). After that, a local fitting method matches lines in the image with edges of the projected object model. The images obtained after convergence of the tracking scheme is shown on the right. It is important to note, that, even under the incorrect initialization of the two other rotation angles as zero, our approach is able to cope with significant deviations from this assumption - here, the angle around the optical axis is more than 20°.

C. Robustness of disparity based figure-ground segmentation

As mentioned in Section IV, object hypotheses are found from a binary map of pixels located within a given depth range. There are some obvious disadvantages associated with such a procedure. First of all, an object might partially occupy the range, while parts of it extend beyond the range. This can be seen in the upper left image of Figure 6, while it does not occur in the second image on the same row. However, thanks to the target position refinement process, a saccade is issued to approximately the same location in both cases. This is shown in the last two images on the upper row, where the third image corresponds to the hypothesis found in the first image.

Another challenge occurs if two nearby objects are placed on almost the same distance, especially if the background lacks sufficient texture. Then the objects might merge into a single hypothesis, which is shown on the second row of Figure 6. In our experiments this seemed more common when a global disparity method [16] was used and is the reason why we normally use simple area correlation. The global optimisation methods tend to fill in the space between the two objects, falsely assuming that rapid changes in disparities are unlikely and thus should be suppressed. The right two images on the last row show that pose estimation might still be possible, even when hypotheses are merged, since the target position refinement will converge to either one of the two objects.

D. Robustness towards occlusions

In a cluttered environment, a larger fraction of objects are likely to be occluded. These occlusions affect most involved processes, in particular those of recognition and pose estimation. The first two images in Figure 7 show a scene in which the sugar box is partially occluded behind a bottle. In the first image, the recognition fails because not enough foveal features are available, while successful recognition and pose estimation is possible in the second image, as shown by the estimated pose in the third image. As is shown in the fourth image, a failure in the pose estimation does not necessarily mean that the results are useless, since the location of the object in 3D space might still be available.

E. Robustness of pose initialisation towards rotations

Since only one view per object was considered for SIFT based recognition, the sensitivity of the system to rotations was expected to be high. It is already known that for efficient recognition using these features, the relative orientation between test image and model ought to be less than about 30°, [20]. Likely because our model set only consisted of eight objects, our study indicated that slightly larger angles were in fact possible. In the image of Figure 8 an object was rotated about 40° and 60° respectively. The rise package was correctly

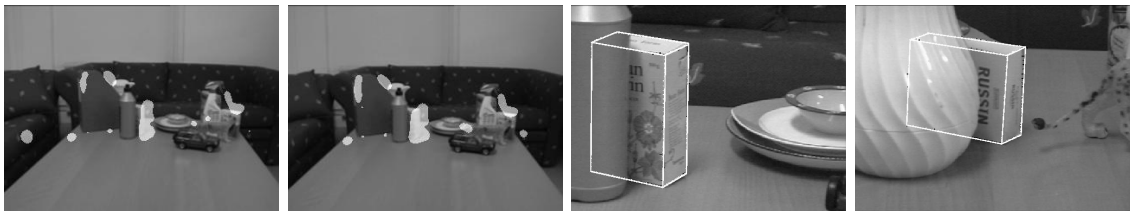

Fig. 7. The effect of occlusions on segmentation and pose estimation.

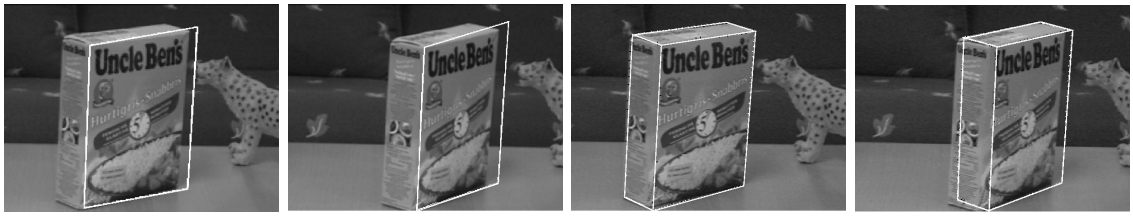

Fig. 8. The effect of large rotations on pose initialisation.

recognized at scores higher than 70%. However, the breakpoint turned out to be highly object dependent. For an object like the tiger, the breakpoint was as low as 20%.

As can be seen in the first two images, larger rotations tends to be underestimated when the pose is initialised. However, these errors are still below what is required for the pose estimation to finally converge. The last two images show the estimated pose after a few initial iterations. Even at 60° the process will converge, but at a somewhat slower rate. For 40° and below, convergence is reach within a few frames.

VIII. Summary and Conclusions

In this paper, we have presented a real-time vision system that integrates monocular and binocular cues for figure-ground segmentation, object recognition, pose estimation and tracking using foveal and peripheral vision. One important property of the system is that all steps from task specification to pose estimation are completely automatic, combining both appearance and geometric models. Experimental evaluation, performed in a realistic indoor environment with occlusions, clutter and changing background conditions, shows the ability of the integrated system to perform tasks even in the cases where individual cues fail. Our current work investigates the importance of higher level, *a-priori* cognitive knowledge to guide the choice of algorithms depending on the task at hand and the benefits of combining complementary methods for recognition and figure-ground segmentation. In order to properly evaluate the performance of recognition methods, we intend to extend the database of available objects.

References

[1] D. H. Ballard, "Animate Vision," *Artificial Intelligence*, vol. 48, no. 1, Feb. 1991, pp. 57–86.

[2] S. Kim, I. Kim and I. Kweon, "Robust Model-based 3D Object Recognition by Combining Feature Matching with Tracking," *Proc. IEEE Int'l Conf. Robotics and Automation*, (ICRA 03), Taipei, Taiwan, 2003.

[3] D. Kragic and H. I. Christensen, "Confluence of Parameters in Model-Based Tracking," *Proc. IEEE Int'l Conf. Robotics and Automation*, (ICRA 03), Taipei, Taiwan, 2003.

[4] D. Kragic and H.I. Christensen, "Weak Models and Cue-Integration for Real-Time Tracking," *Proc. IEEE Int'l Conf. Robotics and Automation*, (ICRA 02), Washington, USA, 2002.

[5] S. Ekvall, F. Hoffmann and D. Kragic, "Object Recognition and Pose Estimation for Robotic Manipulation using Color Cooccurrence Histograms," *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems*, (IROS 03), Las Vegas, USA, 2003.

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

[7] M. Fischler and R. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Comm. of the ACM*, vol. 24, no. 6, 1981, pp. 381–395.

[8] M. Björkman and J-O. Eklundh, "Real-Time Epipolar Geometry Estimation of Binocular Stereo Heads," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, 2002, pp. 425–432.

[9] M. Björkman, *Real-Time Motion and Stereo Cues for Active Visual Observers*, doctoral dissertation, Computational Vision and Active Perception Laboratory (CVAP), Royal Inst. of Technology, Jun. 2002.

[10] P. Chang and J. Krumm, "Object Recognition with Colour Cooccurrence Histograms," *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (CVPR 99), Fort Collins, CO, 1999, pp. 498-504.

[11] T.W. Drummond and R. Cipolla, "Real-time Tracking of Multiple Articulated Structures in Multiple Views," *Proc. European Conf. Computer Vision* (ECCV 00), 2000, pp. 2:20-36.

[12] H. Longuet-Higgins, "A Computer Algorithm For Reconstructing a Scene From Two Projections," *Nature*, no. 293, 1981, pp. 133–135.

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

[14] 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, no. 4, 1993, pp. 517–535.

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

[16] V. Kolmogorov and R. Zabih, "Computing Visual Correspondence With Occlusions Using Graph Cuts," *Proc. IEEE Intl. Conf. Computer Vision* (ICCV 01), 2001, pp. II:508–515.

[17] L. Itti, C. Koch and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, Nov 1998, pp. 1254–1259.

[18] D. Comaniciu, V. Ramesh, and P. Meer. "Real-time Tracking of Non-Rigid Objects Using Mean Shift," *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (CVPR 00), San Francisco, CA, 2000, pp. 142–151.

[19] K. Mikolajczyk and C. Schmid, "A Performance Evaluation of Local Descriptors," *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (CVPR 03), Jun. 2003, pp. 257–263.

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