# Attending, Foveating and Recognizing Objects in Real World Scenes

Mårten Björkman and Jan-Olof Eklundh
Computer Vision and Active Perception Lab
Royal Institute of Technology, Stockholm, Sweden
Email:{celle, joe}@nada.kth.se

**Abstract**

Recognition in cluttered real world scenes is a challenging problem. To find a particular object of interest within a reasonable time, a wide field of view is preferable. However, as we will show with practical experiments, robust recognition is easier if the object is foveated and subtends a considerable part of the visual field. In this paper a binocular system able to overcome these two conflicting requirements will be presented. The system consists of two sets of cameras, a wide field pair and a foveal one. From disparities a number of object hypotheses are generated. An attentional process based on hue and 3D size guides the foveal cameras towards the most salient regions. With the object foveated and segmented in 3D, recognition is performed using scale invariant features. The system is fully automised and runs at real-time speed.

## 1 Introduction

In recent years there has been a considerable interest in object recognition and categorization. Advanced methods have been developed that give high performance both on single objects [5, 19] and objects in real, cluttered scenes [14, 20]. A careful scrutiny of these methods show that they generally assume that the objects in question subtend a considerable part of the image and also are centrally located in the image. For a realistic "seeing system", such as a mobile robot provided with cameras or a wearable system used by humans such assumptions do not hold in general. When we enter a room or look at a table from a few meters distance, objects of interest can be located all over the full visual field that humans have. Whatever field of view the cameras of an artificial system has, it is therefore difficult to apply known recognition algorithms either because objects are too small in the images, or because they are outside the field of view.

In this paper we describe a system capable of locating and recognising objects in the real world. 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, and then foveating and fixating them to get a view suitable for recognition. The system is implemented on a stereo head system provided with two binocular camera pairs, a wide field pair for attention and a foveal one for recognition. 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. We consider the case in which the head and the objects being studied remain static. It has been shown in earlier work [4] 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 was present.

Some methods applied to our system, such as those used for recognition and calibration, are already known from studies of others. However, some issues will only become apparent, if one tries to integrate different components to create a fully operational system running in a real-time setting. Robustness is often preferable from accuracy, if one is forced to make a trade-off between the two. Novelty lies in the combination of wide field cameras for attention and foveal ones for recognition, the adoption of 3D size as a cue for attention and methods for foveal segmentation prior to recognition. Finally, through practical experiments we show the benefits of foveated recognition.

# 2 The role of stereopsis

Our system is based on a binocular stereo-head known as Yorick [22], shown in the left image of Figure 1. The head consists of two sets of stereo cameras, a wide field set (60° f.o.v.) and a foveal one (14° f.o.v.), with focal lengths 6 mm and 28 mm respectively. The goal of the wide field set is to guide the foveal one, such that recognition can be performed foveated, 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, given that the intrinsic parameters of the cameras and the length of the baseline are provided.

Since dense disparities are used for figure-ground segmentation and these need to be related to actual metric distances, the epipolar geometry of the wide field cameras has to be known. We derive the epipolar geometry from image data alone. While the system is running, the epipolar geometry is continuously updated using the following optical flow constraint [15]:

$$\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = \begin{pmatrix} (1+x^2)\alpha - y r_z \\ xy\alpha + r_x + x r_z \end{pmatrix} + \frac{1}{Z} \begin{pmatrix} 1 - xt \\ -yt \end{pmatrix},$$

where $\alpha$ is the vergence angle, $r_x$ and $r_z$ the relative tilt and rotations around the optical axes, and $t$ represents the direction of the baseline in the left camera frame. From the disparities $(\Delta x, \Delta y)$ of a set of corner features $(x, y)$, extracted using Harris' method [11], parameters are estimated using Gauss' method and a combination of RANSAC [8] and M-estimators to reduced the influence of outliers in the data set. Even if the above constraint only gives an approximation of the epipolar geometry, it has shown to more robust than the more popular fundamental matrix for our application [2].

## 2.1 Disparities

When the external camera parameters are known, a dense disparity map can be computed. and from this map a number of object hypotheses can be derived, as will be explained in Section 3. An example of a disparity map estimated using sums of absolute differences [13] can be seen in the right image of Figure 1. White areas represent points for which no disparity could be determined, while darker areas correspond to points located closer to the observer. The visual angular resolution is 11.4', which means a depth resolution of between 2 and 6 cm, depending on depth. The total disparity range in this example is about 48 pixels, that is considerably wider than typical benchmark scenes [21].


Figure 1: The Yorick stereo-head (left), a table-top scene (middle) with disparities estimated using sums of absolute differences (right).

In fact, we tested eight different disparity methods, from simple correlations using sums of absolute differences to more complex global optimization methods [12]. Unfortunately, the latter have a number of weaknesses significant to our application. They all use the assumption that disparities tend to vary smoothly across the image plane. In the case of opposing interpretations they always choose the smoothest one. With smoothing the boarder between two foreground objects becomes less distinct, if the background lacks sufficient texture and the segregation between objects gets complicated. Thus we typically rely on the less complex methods, even if objects might be split into multiple hypotheses, when the texture is weak.

# 3 Visual scene search

The aim of the attentional process, that will be described in this section, 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 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. Distributions of hues are represented as histograms consisting of 128 entries and hue saliency is computed from correlations of such 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. Through the use of rotating sums, the computational cost is kept to a minimum.

Figure 2: Hue saliency maps (first two images) when a blue car and a orange rise package are sought. The four largest peaks in each case are shown to the right.

Figure 2 shows two different examples of such hue saliency maps. In the first case the blue car (see Figure 5) is requested and in the second case the orange rise package. For each case the four largest DoG peaks can be seen in the two images to the right. Note the three odd peaks in the first scenario. They belong to the blue sofa in the background and come as a result of foreground objects being fattened using the area based disparity method. Fortunately, these peaks are considerably lower than the peak of the blue car. The peaks in the second case are more natural, since the tiger and giraffe are both orange, a hue not much different from that of the rise package.

### 3.1 Directing gaze

As mentioned above the 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 rapid saccade. In order to cope with situations in which a requested object does not generate the largest peak, either due to occlusions or non-typical 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, such that is can be foveated and recognised, but it might take a couple of saccades. In order 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, an affine transfer of hypothesis positions is made.

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 recognized, as will be described in Section 5. This is repeated until the requested object is found or the system times out, typically after about five saccades.

## 4 Fixation

The foveal cameras are always striving towards fixation, i.e. the optical axes intersect on a physical object somewhere in the scene. Estimation of cues such as shape and motion is simplified when the object of interest is fixated [7]. With the object in the centre of the visual field, 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 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 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. This has lead us to the conclusion that 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.

## 4.1 Foveated Segmentation

Before recognition is taking place, the system attempts to segment the observed foreground object from its background and neighbouring objects, using a disparity map generated from the foveal images. Foveally, disparities are very hard to determine and the reason is the following. When the field of view is small and the range in depth is large, the number of possible disparities explode. In fact, most algorithms assume the disparity range to be around 16 to 32 pixels and are rarely evaluated for wider ranges [21].

For a system such as ours, where the baseline is 16 cm and the closest expected object is 100 cm in front, the disparity range is almost as wide as the width of the image itself. Fortunately, since the cameras are in fixation we know that, for the object we are interested in, correct matches can be found within a shorter range around zero disparity. However, for points whose correct matches are outside the considered range, the estimated disparities will not be reliable. The left-to-right consistency check [9] often applied to remove inconsistent matches will not suffice, since only a fraction of the whole disparity range is considered. A large number of false positives can thus be expected.

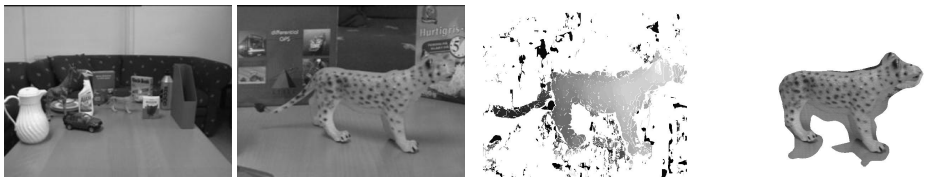

Figure 3: The wide field and foveal left images of a table-top scene, while fixating on the tiger object. The last two images show the disparity map and the resulting mask.

To identify these false matches we use a mean shift algorithm on the 3D data [6]. If the points for which disparities are available are considered as a cloud of 3D points, the cluster representing the objects of interest can be found as those points that lie within a rectangular window around the centre of the cloud. The size of this window is set to that of the requested objected size. Since the real size might be slightly different we gradually increase or decrease the size of the window until a radical change in point density is found. A blurred binary mask is finally generated from the points belonging to the cluster. We are currently looking at other possible methods of finding the cluster, since a rectangular window is hardly appropriate for all objects.

A good example of a textured object segmented foveally can be seen in Figure 3. The angular resolution here is 1.2', which is considerably finer than in the wide field case. In this particular example it corresponds a depth resolution of about 3 mm, resulting in the fine disparity gradient shown in the third image. It is also clear from the mask image to the right, that most false matches were successfully removed by the mean shift algorithm, but at the unfortunate cost of a lost tail. In Figure 4 a less successful example is shown. Due the specular surface of the plastic car, the segmentation has been seriously fragmented. In cases such as this, the system instead uses a rectangular shaped mask, centered as the same position, with a size set to that of the object that was originally requested.

Figure 4: 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.

# 5 Recognition

Mikolajczyk & Schmid [18] recently presented a study on various feature descriptors for object recognition. They analysed robustness under various changes, such as illumination and scale changes. The SIFT descriptor of Lowe [16] turned out to be one of the more robust descriptors in this study and has thus been applied for recognition in our particular system. The strength of using an existing method is that it has already been analysed elsewhere, which makes our results easier to interpret. In the work presented here, we had no intention to come up with any new methods for recognition per se, but to show results from a known method applied to a typical real world scenario, in conjuction with the attentional framework presented in Section 3.

Our implementation of SIFT differs only slightly from the original. Features are detected using differences of Gaussians and for each such feature, a scale and rotation invariant gradient histogram is created, representing the appearance of the local neighbourhood around the feature. A training image typically consists of a couple of hundred such 128-dimensional SIFT features. During recognition similar features are extracted from the test image and for each feature the nearest neighbour is found among those of the object models. Each successfully matched feature is equivalent to a vote on a particular model and view. An object is finally considered as detected if the maximum number of votes for a particular view of the object is higher than some predefined threshold. The identity of an unknown object is given by the model with the highest number of votes.

For the experiments presented in Section 6, the objects shown in Figure 5 were considered, with SIFT features extracted from 8 canonical views per object. The ROC curves from these experiments were generated varying the above mentioned threshold. With 21 objects in total, the number of features is about 24 thousands. To speed up recognition we arrange the model features in binary search trees for approximate nearest neighbour

look-ups using the Best Bin First strategy [1]. To suppress the Curse of Dimensionality, which complicates searches in very sparse spaces, the search trees are constructed based on only the 24 most dominating dimensions, found using principle component analysis. However, the metric used for feature matching still uses all 128 dimensions.


Figure 5: Objects used for the evaluations.

# 6 Experiments

The experiments presented in this section were performed using the stereo-head shown in Figure 1 and a 1.2 GHz dual Athlon MP machine running under the Linux operating system. The four Sony XC999 cameras deliver one CVBS signal each. Frame capture is done using a Leutron PicProdigy board, which is able to grab from all four cameras at 25 Hz. Software components were implemented as separate modules communicating through a CORBA based framework. The attentional process, including wide field figure-ground segmentation, runs at a rate of about 6 Hz. The bottleneck of the system is presently the recognition process, which takes about one second. However, recognition is only required upon demand and does not have to be executed continuously, unlike attention and fixation. Before recognition, right after a saccade, we let the fixation process stabilise during a couple of frames. Thus the total cycle time between two saccades, during which the system looks at and recognises an object, is about two seconds.

## 6.1 A visual search experiment

In order to analyse the stability of the system, a large number of experiments were conducted. Results from one such experiment are presented here in some detail. A movie was recorded and images from this movie are shown in Figure 6. The delay between two consecutive images is here about a second, which means that most saccades are left out, since a saccade only lasts for a few tenths of a second.

The system is initially given the command to locate the grey bottle (fifth object on last row of Figure 5). Due to the lack of a well defined hue, this object is somewhat difficult, especially when occluded. A gaze direction is found through the attentional system and a saccade is executed. After fixation has stabilised, the object at the new location is correctly recognised as the white textured bottle (sixth on last row). Another saccade is generated and the gaze is shifted towards an object that is later identified as the tiger (sixth of first


Figure 6: Images taken while the system was involved in a series of search tasks.

row). After a final shift in gaze, the grey bottle is detected and the system is ready for a new search task.

The next object to be located is the plastic blue car. Due to its distinct colour it is easily found by the attentional system. However, since fixation relies on corner features and most features here suffer from either specularities or transparencies, it takes a while for the process to stabilise and the blue car to be finally detected. The last two requested objects, the blue box and the orange rise package, are easily found and recognised by the system. Disregard the boundaries around those box-like objects in Figure 6 that received the highest SIFT feature scores. In another series of experiments these boundaries were exploited for pose estimation and tracking initialisation, using the presented system [3].

## 6.2 Wide field versus foveal detection

In the introduction we claim that objects are easier to recognise if they are foveated in the center of the images and subtend a considerable part of the visual field, especially when segmentated from the background. However, for objects to be found and recognised foveally, the system relies on an attentional process that controls gaze. With such a process there are more reasons why the system might fail, than in the case of a single wide field camera overlooking the scene. The question is whether the possible improvements in recognition really motivate the increased complexity of the overall system.

We performed a series of experiments involving the objects shown in Figure 5 and a couple of additional objects for which no models exist. Groups of objects were arranged in 18 different scenes, similar to the one shown in Figure 2, with moderate variations in backgrounds and illuminations. Within these scenes 168 different search tasks were executed, with a 50% probability of the requested object actually being present in the scene. For each search task, the gaze was shifted towards five different fixation points before determining whether a requested object had been detected or not. The diagrams in Figure 7 show the ROC curves from these experiments.

The left graph represents searches using the wide field left camera image in full resolution for recognition. Even if the requested object is in fact present in the scene, only the large scale features are found. These features are so few in comparison to the total

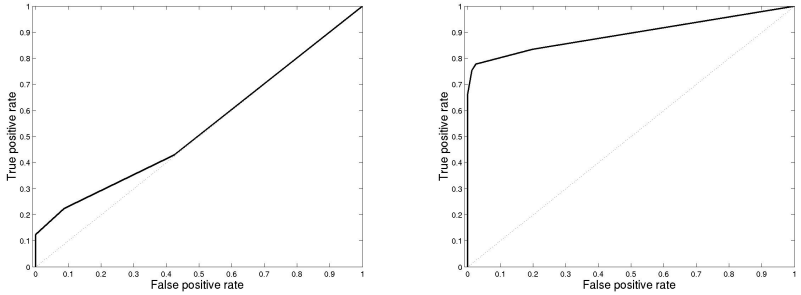


Figure 7: ROC (Receiver Operating Characteristics) curves when recognition is performed based on the wide field (left) and foveal (right) camera images.

number of features, which means that they are hard to separate from occasional outliers. The graph to the right shows the results using the foveal left camera images, with figure-ground segmentation performed automatically. For increased computational speed, only half the resolution was used in this case. From the results it is clear that even if the attentional process may fail and even if objects may be unsuccessfully segmented, detection is easier if recognition is done using the foveal camera images. A deeper analysis on why detection sometimes fails will be presented in a subsequent paper.

# 7 Conclusions

In this paper we have presented a real-time recognition system, capable of actively locating, attending to and recognising objects in a real world scene. A wide field attentional process guides the foveal system, so that objects can be recognised foveated in the centre of the images, even in cluttered environments where many object candidates are available. We have also shown that it is far easier to detect a requested object using this approach, rather than using a single view covering the whole scene.

The fact that we only use one cue for recognition has its limitations. Some objects do not contain enough texture to be detected at all. The false negatives of Figure 7, involve five objects in particular; the textureless cup, the blue car and the three cushions on the second row of Figure 5. We believe other cues need to be added for the overall robustness to be improved. Contour and luminance based cues are likely candidates. This opens up the question of how to integrate cues of radically different nature. The system presented here might provide a useful framework for such studies to be carried out, without the limitations of a given training set and predetermined features of choice.

In the future we intend to perform more comprehensive experiments in recognition and categorization. More cues for attention as well as recognition will be considered, including adaptation and learning at all levels to obtain higher robustness and performance. We hope to speed up the foveation-recognition loop and reach a frequency of about 1 Hz, permitting the system to work within more dynamic environments.

# References


[1] J. S. Beis and D. G. Lowe, “Shape Indexing Using Approximate Nearest-Neighbour Search in High-Dimensional Spaces,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (CVPR 97), pp. 1000–1006, Jun. 1997.

[2] 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, pp. 425–432, Mar. 2002.

[3] M. Björkman and D. Kragic, “Combination of Foveal and Peripheral Vision for Object Recognition and Pose Estimation,“ *IEEE Int’l Conf. Robotics and Automation*, Apr. 2004.

[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,” *Proc. IEEE Int’l Conf. Computer Vision* (ICCV 99), pp.1197-1203, Kerkyra, Greece, 1999.

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

[8] 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, pp. 381–395, 1981.

[9] P. Fua, “A Parallel Stereo Algorithm That Produce Dense Depth Maps and Preserves Image Features,” *Machine Vision Applications*, Vol. 6, No. 1, pp.35–49, 1993.

[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,” *Proc. Alvey Vision Conf.*, Manchester, UK, pp. 147–151, 1988.

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

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

[14] B. Leibe and B. Schiele, “Interleaved Object Categorization and Segmentation,” *Proc. British Machine Vision Conference (BMVC 03)*, Sep. 2003.

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

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

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

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

[19] 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.

[20] R. C. Nelson and A. Selinger, “A Cubist Approach to Object Recognition,” *Proc. Int’l Conf. Computer Vision* (ICCV 98), Bombay, India, pp. 614–621, Jan. 1998.

[21] D. Scharstein and R. Szeliski, “A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms,” *Int’l. J. Computer Vision*, Vol. 47, No. 1, pp.7–42, 2002.

[22] 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.