Foveated Figure-Ground Segmentation and Its Role in Recognition

M. Björkman and J.O. Eklundh
Department of Numerical Analysis and Computer Science (NADA)
Royal Institute of Technology (KTH), Stockholm, Sweden
celle@nada.kth.se, joe@nada.kth.se

Abstract

Figure-ground segmentation and recognition are two interrelated processes. In this paper we present a method for foveated segmentation and evaluate it in the context of a binocular real-time recognition system. Segmentation is solved as a binary labeling problem using priors derived from the results of a simplistic disparity method. Doing so we are able to cope with situations when the disparity range is very wide, situations that has rarely been considered, but appear frequently for narrow-field camera sets. Segmentation and recognition are then integrated into a system able to locate, attend to and recognise objects in typical cluttered indoor scenes. Finally, we try to answer two questions: is recognition really helped by segmentation and what is the benefit of multiple cues for recognition?

1 Introduction

In an earlier report we presented a real-time vision system that is able to locate, foveate and recognise objects in cluttered indoor scenes [1]. In this paper we take a closer look at the problem of figure-ground segmentation and in particular its role in recognition. Unlike most other recognition systems, our goal is not to detect objects in images, but in scenes, which is a significantly more complicated task. A system acting in the real world needs to actively search the scene, constantly updating the fixation point, until a conclusion can be drawn on the existence of requested objects. Our particular system is equipped with four different cameras, a wide field binocular set for attention and a foveal one for recognition. A wide field is necessary for objects to be localised within reasonable time, while a high resolution is preferable for attended objects to be successfully recognised. The problem of combining these two requirements has previously been recognised by others and a number of more of less complicated solutions have been proposed [13, 8, 14].

The two processes, segmentation and recognition, are closely interrelated. Given the solution to one of the problems, the solution to the other can more easily be obtained. Unfortunately, robust segmentation has shown to be very difficult and most state-of-the-art recognition methods either ignore the problem [10] or perform segmentation as part of the recognition process [16, 9]. In our system we exploit the fact that we have a binocular setting and continuously segment the object currently in fixation. This will be explained in detail in Section 2. We then integrate figure-ground segmentation with the recognition system described in Section 3 and the attentional process previously presented in [1].

Figure 1: Two foveal image pairs obtained using an active vision system.

We evaluate the complete system based, not on the performance of individual components, but on its ability to do what it is intended for, that is finding and recognising objects in cluttered scenes. Thus we apply a systems oriented methodology, where methods are benchmarked based on their practical behaviours in the real world, instead of on artificial ground truth data in isolated settings. Since the quality of a method depends on its purpose, methods should be tested with the end purpose in mind. A particular segmentation method suitable for video coding for example, does not necessarily have to be successful if the goal is object recognition. A segmentation is rarely an end result in practical applications. In a series of experiments presented in Section 4 the benefit of segmentation for recognition will be analysed. What we try to answer is a relevant question. Given what we know of the limitation of current segmentation methods, does segmentation really benefit recognition? In this respect our study is different from earlier studies.

2 Foveated segmentation

Many recognition methods perform considerably better if attended objects are segmented from their backgrounds and surrounding objects. This is especially true for methods based on dense statistics collected in histograms. In this section we study the segmentation problem in the context of a binocular setting. Using an attentional system described in [1], the camera system automatically fixates upon objects of interest, before they are segmented and later recognised. We determine the foveated segmentation using a two-step approach, where the first step computes a dense disparity map. However, instead of simply thresholding this map, we statistically model the disparity results and globally solve a binary labeling problem, one label for foreground (**FG**) and another one for background (**BG**). Here we apply the Ising model. Since we are only interested in two labels, the labeling problem can be exactly solved by a single graph-cut [5].

The benefit of this approach is two-fold. By explicitly modeling the disparity results we are able to cope with a limitation that occurs in most modern disparity methods, i.e. the fact that the disparity range treated by the method should be at least as wide as the total range that appears in the images. For narrow field cameras this assumption can be difficult to satisfy. In our case the range of possible disparities is wider than the image itself, if the expected shortest distance is less than 0.8 meters. The six leftmost disparity maps in Figure 2 illustrate results from a couple of state-of-the-art methods on the stereo pairs in Figure 1, if disparities were falsely assumed to lie within a 48 pixel range around zero-disparity. The methods use here were based on belief propagation [17], cooperative stereo [19] and graph-cuts [6]. As seen in the images the number of false positives is substantial, even if 48 pixels are significantly more than what is the case in typical benchmarks [15].

The second benefit of the presented approach is the speed. Global optimisation is done

Figure 2: Disparity maps calculated using belief propagation (left), cooperative stereo (centre-left), graph-cuts (centre-right) and local area based correlation (right) for the image pairs shown in Figure 1.

only in the second step, which is fast as only two different labels need to be considered, the foreground and background label. The disparities themselves are less relevant. In the first step we use a simplistic method based on area based correlation [7], which is easier to model than more complex alternatives. Even if this results in a somewhat lower density, the results after the second step are very similar independently on the disparity method used. This leads to a total difference in speed of at least an order of magnitude.

2.1 Disparity modeling

The rightmost two images in Figure 2 show some results obtained with area based correlation, including a left-right consistency check. The number of false positives, that appear at points outside the range considered, is not much higher than those of the more complex methods. To compute the final segmentation we need to model the behaviour of these results. From the disparity method, each point $\mathbf{p}$ is given a state $\mathbf{L_p}$, that can be either textureless **TL**, mismatch **MM** or any disparity $\mathbf{d}_k$ within the range of considered disparities. An image point is textureless if the variance in the local area is insufficient. A mismatch is indicated by a failure in the left-right consistency check performed following correlations. Remaining points get states corresponding to the disparities of highest correlation between the left and right camera images.

From these states foreground probabilities, $Pr(\mathbf{X_p} = \mathbf{FG} \mid \mathbf{L_p})$, can be computed using Bayes' rule and the expected distributions of mismatches and textureless regions. The prior foreground probability $Pr(\mathbf{FG})$ is given by the estimated projected size, which is determine by the attentional system prior to fixation. For the background, disparity values are assumed to be uniformly distributed over the whole range. The same is assumed for foreground points, but only within a range equivalent to the expected size of the requested object. Since the considered range is significantly narrower than the full range, most mismatches can be expected in the background. We further expect textureless areas to be somewhat more common in the background. For the experimental results in this paper we use the following expected frequencies:

$$\begin{array}{lll} Pr(\mathbf{TL} \mid \mathbf{FG}) = 0.20 & Pr(\mathbf{MM} \mid \mathbf{FG}) = 0.10 & Pr(\{\mathbf{d}_k\} \mid \mathbf{FG}) = 0.70 \\ Pr(\mathbf{TL} \mid \mathbf{BG}) = 0.25 & Pr(\mathbf{MM} \mid \mathbf{BG}) = 0.40 & Pr(\{\mathbf{d}_k\} \mid \mathbf{BG}) = 0.35 \end{array}$$

Figure 3: Patches coloured by average foreground priors (left). Foreground segments found using graph-cuts on patches of pixels (middle) and on individual pixels (right).

Without an extensive set of ground truth examples, we cannot derive any more precise frequency estimates. However, through a series of experiments with different combinations of frequencies, we concluded that the exact values are not as important as the relative differences between foreground and background frequencies.

2.2 Patch-wise segmentation

Since there are only two possible labels, **FG** and **BG**, we can solve the global optimisation problem using a single graph-cut. Each pixel can be represented by a node that has one link to a source and another one to a drain, with capacities determined by the priors, $Pr(\mathbf{FG} \mid \mathbf{L_p})$. Neighbouring nodes (assuming 4-neighbours) are further linked with capacities reflecting the likelihood of two neighbours belonging to different classes. It can be shown that these capacities are given by the negative log-likelihoods of corresponding probabilities. Typically these likelihoods are estimated from similarities between neighbouring points, which in our system is done using the image brightness values, $\mathbf{I_p}$. This is reasonable, since depth discontinuities often coincide with high contrast edges.

Instead of assigning labels pixel-wise, we use image patches and assign the same label to all pixels within a patch. Thus only one network node is needed per patch. The patches are found using watershedding on gradient magnitudes. The scene is oversegmented, such that all depth discontinuities hopefully coincide with some patch edge. The two leftmost images in Figure 3 illustrate segmentations obtained using this procedure, with colouring based on average foreground priors. As can be seen, regions end up being undetermined, shown in grey, if they are either located outside the depth range considered or without enough texture. The prior of a particular patch is given by the set of states, $\{\mathbf{L}_{\mathbf{p}_i}\}$, that corresponds to the image points within the boundary of the patch. Since all points belong either to the foreground or the background, the patch prior can be written as

$$Pr(\mathbf{FG} \mid \{\mathbf{L}_{\mathbf{p}_i}\}) = \frac{\prod_i Pr(\mathbf{X}_{\mathbf{p}_i} = \mathbf{FG} \mid \mathbf{L}_{\mathbf{p}_i})}{\prod_i Pr(\mathbf{X}_{\mathbf{p}_i} = \mathbf{FG} \mid \mathbf{L}_{\mathbf{p}_i}) + \prod_i Pr(\mathbf{X}_{\mathbf{p}_i} = \mathbf{BG} \mid \mathbf{L}_{\mathbf{p}_i})}.$$

We model the probability of a discontinuity between two neighbouring points, $\mathbf{p}_i$ and $\mathbf{p}_j$,

Figure 4: Typical segmentation results.

from different patches as

$$Pr(\mathbf{X}_{\mathbf{p}_i} \neq \mathbf{X}_{\mathbf{p}_j}) = \exp\left(\frac{a}{b + \sqrt{|\mathbf{I}_{\mathbf{p}_j} - \mathbf{I}_{\mathbf{p}_i}|}}\right).$$

Here $|\mathbf{I}_{\mathbf{p}_j} - \mathbf{I}_{\mathbf{p}_i}|$ is the gradient magnitude and, a and b are chosen so that the probability is 10% if the gradient magnitude is zero and 80% for maximum gradients. Instead of summing up the log-likelihoods of points along an edge between two patches, we use the product of the length and the minimum negative log-likelihood. The reason is to prevent discontinuities from being introduced along edges that occur as a result of shading.

The segmentation we get after applying graph cut to the images in Figure 1 can be seen in the middle of Figure 3. A number of additional examples obtained using the same approach are shown in Figure 4. Later in Section 3 we will see that these segmentations will significantly improve the recognition results, compared to if no segmentation were used at all. If we would model each pixel separately, without grouping pixels into patches, we would get the results to the right in Figure 3. Here we see that the foreground segments tend to expand and cover parts of the background. Since a discontinuity boundary belongs to the occluder, the foreground will easily dominate an occluded background, if there is not enough texture in the background to suggest otherwise. By grouping pixels into patches these opposing interpretations are more easily resolved.

3 Multi-cue recognition

Our recognition system consists of two separate modules, one based on scale and rotation invariant SIFT features and the other on colour histograms. Each time the attentional process directs the cameras towards a new fixation point, the recognition system determines whether the object that was requested can be detected in the new foveal views. The two cues were chosen since they are more or less complementary. When objects contain a sufficient amount of image texture, SIFT features have shown to be reliable for object recognition [10]. On the other hand, colour histograms only work if objects contain a small, but distinct, set of colours. In order to cope with varying illumination, we use a colour constant model of Gevers & Smeulders [4]. These are integrated in a framework based on a 2D support vector machine (SVM), with dimensions given by the normalised

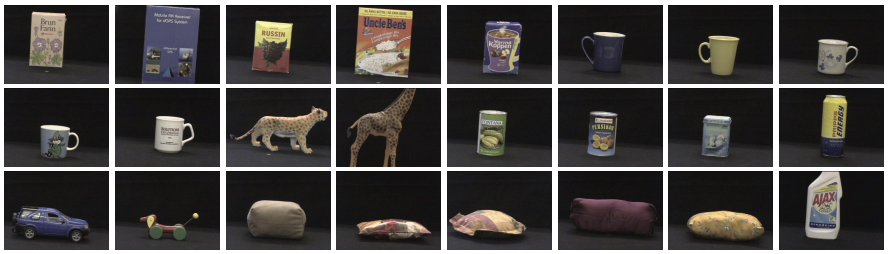

Figure 5: Objects used for recognition experiments: *Farin*, *GPSBox*, *Raisin*, *UncleBen*, *VarmKopp*, *BlueCup*, *YellCup*, *FlowCup* (first row); *MuminCup*, *MyCup*, *Tiger*, *Giraffe*, *MangoCan*, *PeachCan*, *BoatCan*, *Pripps* (second row); *BlueCar*, *DogCar*, *BrownCush*, *FlowCush*, *RoundCush*, *Violett*, *YellCush* and *Ajax* (third row).

detection scores of each cue. Thus the integration of different cues is soft, in that each cue is computed independently and only fused to deliver a final composite detection score.

Our implementation of SIFT differs slightly from the original. Instead of using differences of Gaussians for feature detection, we use scale invariant Harris' features [11]. When an object has been segmented from its background, as described in Section 2, features are detected within the segmented area. Each feature is matched to those of the object models and gives a vote to the most similar match. The total number of votes for each model is then counted. Unlike the original version, a detection score is computed as the quotient of votes on the requested object model and the total number of votes on all models, leading to a detection score between zero and one. The rational behind this approach is that only one object is assumed to be located in each segmentation. By comparing scores between models, instead of using an absolute measure, the scores will be more distinct in cases when there are few extracted features due to low texturing.

Colour histograms have been applied for recognition for many years. However, it has been questioned whether colours can ever be used for robust recognition [3]. There are primarily two problems associated with colour histograms. The first is the lack of robustness in cases of clutter and occlusions, which leads to a need for segmentation. The other problem is that of colour constancy. Assume (E_R, E_G, E_B) to be the illumination, which is typically unknown, and let the reflectance be given by (S_R, S_G, S_B). Then the measured colour can be approximated as $(R,G,B) = K\,(E_R S_R, E_G S_G, E_B S_B)$, where K is a factor that depends on the direction of incoming light. This factor and the luminance component of the illumination can be canceled out through a projection,

$$(r,g) = (\frac{R}{B}, \frac{G}{B}) = (\frac{E_R}{E_B}\frac{S_R}{S_B}, \frac{E_G}{E_B}\frac{S_G}{S_B}) = (E_r S_r, E_g S_g).$$

By observing that the illumination chromaticy (E_r, E_g) may be different between database and query images, it is clear that uniformly coloured objects cannot be recognised from measured colours alone. However, relative colours may still be robustly compared. Gevers & Smeulders [4] do this by computing the fraction of neighbouring colours. Given that (r_1, g_1) and (r_2, g_2) are the colours of two different pixels, we get an expression independent on coloured illumination,

$$(\frac{r_1}{r_2}, \frac{g_1}{g_2}) = (\frac{E_r S_{r_1}}{E_r S_{r_2}}, \frac{E_g S_{g_1}}{E_g S_{g_2}}) = (\frac{S_{r_1}}{S_{r_2}}, \frac{S_{g_1}}{S_{g_2}}). \quad (1)$$

Figure 6: Some scenes used for detection experiments.

Since (E_r, E_g) can be assumed to be approximately similar for all pixels in a given image, they are canceled out. From pairs of pixels, 2D histograms of relative colours can then be created for database and query images. Histograms are compared with histogram intersection [18], which has shown to be moderately robust to occlusions. In order to make comparisons less sensitive to scale changes, we collect data from pairs that are separated by four different distances, between 4 and 16 pixels. By collecting data in 8 different directions histograms are made more or less invariant to rotations.

4 Experimental results

The goal of the system demonstrated in this paper is to automatically detect objects in scenes, given the identify of a particular requested object. Initial hypotheses are delivered by an attentional system that uses a wide field camera set. Each hypothesis is visited in the order of saliency, which is calculated in relation to the hues and 3D size of the object being requested, so that the most similar image regions are visited first. At each fixation point, the object in fixation is segmented from its background prior to recognition in the foveal views, using the approach presented in Section 2. We do so to get more distinct detection scores when the scene is heavily cluttered. Since a requested object, when it is available, is typically found within a couple of saccades, we use a limit of five saccades before a time out is signaled. The cycle time between two saccades is currently about 2 seconds when the system is running on a 1.2 GHz Athlon CPU. Additional details on the attentional and fixation processes can be found in [1].

There are a number of reasons why the system might fail, some of which are critical. A notable feature of the system, however, is that similar locations may be visited more than once. If the system fails to successfully fixate upon the requested object the first time, a second trial may be permitted within a couple of saccades. This highlights the importance of evaluating the system based on its performance as a whole, rather than on the individual components. This was done by a large series of experiments using the set of objects in Figure 5. A database of SIFT feature and colour histograms was created, with objects viewed from 8 different directions. The experiments were performed in 26 table-top scenes similar to those in Figure 6. Note that the projective sizes of most objects are very small in comparison to what is typically required for recognition. This was the original reason for using two separate camera systems, a wide-field system for attention and a foveal one for recognition.

The complete system was tested with a series of 240 search tasks, each task involving 5 saccades. In order to determine its weaknesses, we then analysed the cause of each failure. Every object was searched for 10 times, out of which 6 involved the object actually being located in the scene. In total 32 failures were observed, 25 true and 7 false ones. In no single search task the requested object failed to be placed in the centre of view within the time frame of 5 saccades. However, the system failed to properly fixate upon a requested object on 12 occasions. Besides *DogCar* (see Figure 5) these failures involved either the cushions or the cups. These objects are either textureless or have texture at too large a scale for corner features, required by the fixation process, to be extracted. The most difficult object turned out to be *BrownCush*, which is a brown uniformly coloured deformable object. Foveated segmentation failed on 6 occasions, all involving the cushions, except for *YellCup* that resulted in one failure. This can be explained by the fact that the cushions always lay flat on the table-top. Image deformations at the visible parts of the objects are thus large, which complicates stereo matching. This might be resolved by affine invariant matching, but only at the cost of higher computational complexity.

The false search failures, when an object was being mistaken for a requested one, were spread among 7 different objects. On one occasion the *Giraffe* object was mistaken for the *Tiger*, which is understandable, since their colours and patterns (sic!) are similar. The remaining errors were caused by single sporadic SIFT features being matched to the wrong object model, in combination with similarity in colour. A conclusion we can draw from these results is that more cues ought to be added for recognition, in particular cues suitable for uniformly coloured objects. We currently investigate the possibility of using contour segments for this purpose [16].

4.1 Recognition performance

We further analysed the recognition system in isolation, using those saccades in which a physical object, known or unknown, was successfully fixated. This set constitutes 886 of the total 1200 saccades. In the remaining cases it was hard to tell what was actually in fixation. The foveal images were manually annotated with the identity of the segmented objects. To this set we added an equal number of false object identities, so as to analyse the false detection results. To the left in Figure 7 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. However, there are still a number of objects that cause considerable problems. In about 9% of the searches very few SIFT features could be found. These objects are the same as those for which fixation fails, the cushions and the uniformly coloured cups.

Similar curves for detection based on colour histograms can be seen in the middle of the same figure. The curves show two different cases, when the segmentation mentioned in Section 2 was applied and when a rough segmentation based on the estimated size and position obtained from the attentional system was used instead. Without any segmentation whatsoever, the detection scores are practically useless. From the curvature of the ROC curves we see that SIFT features tend to be more discriminant than colour histograms, but fail completely if too few features can be extracted. However, for larger false positive rates, colour histograms are still preferable. The most difficult objects here are the blue ones, *GPSBox*, *BlueCup* and *BlueCar*. A possible explanation is that the blue colour channel typically is noisier than the red and green ones. Obviously uniformity in colour

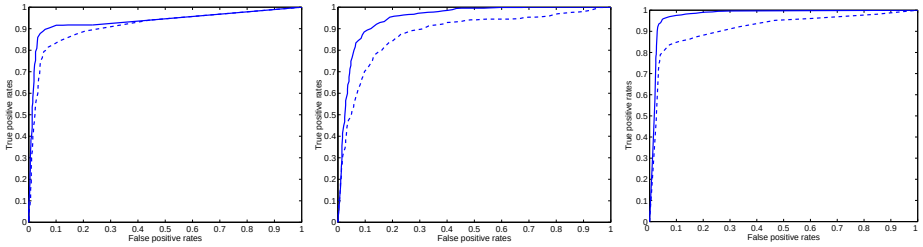

Figure 7: ROC curves for SIFT based (left), colour histogram based (middle) and combined (right) object detection, with (solid) and without (dashed) foveated segmentation.

does not disqualify colour histograms, even if the histograms we used only measure relative colours. It seems that, due to image noise, measurements are scattered in different ways depending on the colour, such that an object still can be recognised.

Since the problematic objects hardly overlap between the cues, one might expect that a combination of SIFT features and colour histograms could lead to better results. With the two cues combined using SVM, as described in the beginning of Section 3, the ROC curves to the right in Figure 7 were obtained. Cross validation was applied to ensure that training and test sets were kept separated. The results are good indeed and one might suspect that the similarity between the table-top scenes has affected the results. Nevertheless, a combination still seems to be recommended. More experiments in different environments have to be performed to draw any quantitative conclusions.

5 Conclusions

We have in this paper studied the problem of figure-ground segmentation in the context of a recognition system aimed for robotic tasks in the real world. In combination with an attentional process, earlier described in [1], our system is able to locate, attend to and recognise objects in cluttered scenes. Evaluation was done based on the performance of the complete system, not on the individual components. The reason for doing so is that if a single component fails, the system as a whole might still be functional. Instead of immediately finding a requested object, it may be found within a couple of saccades. This has led us to a methodology where methods are benchmarked with the final purpose in mind. We have earlier applied the same approach to pose estimation and manipulation on a robotic platform, but with another combination of wide-field and foveal cameras [2].

With a large series of on-line experiments we have shown that recognition based on both SIFT features and colour histograms benefits from figure-ground segmentation, even if segmentation is done automatically and sometimes fails. In our opinion this is a more relevant result, than results assuming segmentation to be of ground truth quality, an assumption hard to satisfy in practical applications. In the future we will spend more efforts on recognition in particular. Feeding back information from recognition to segmentation has already been tested, using the system presented in [9], but we have so far been unable to show any improved recognition scores due to such a feed-back. For our system to successfully work in more general environments, recognition ought to be complemented with additional cues, e.g. shape based cues like contour segments. Finally, learning and adaptation should be done on all levels of the system, as well as in-between levels.

References

[1] M. Björkman and J-O. Eklundh, "Attending, Foveating and Recognizing Objects in Real World Scenes", in *Proc. British Machine Vision Conference* (BMVC04), Sep 2004.

[2] D. Kragic, M. Björkman, H.I. Christensen and J.O. Eklundh, "Vision for Robotic Object Manipulation in Domestic Settings", *Robotics and Autonomous Systems* (to appear).

[3] B. Funt, K. Bernard and L. Martin, "Is Machine Colour Constancy Good Enough?", in *Proc. European Conf. on Computer Vision* (ECCV98), pp. 445–459, 1998.

[4] T. Gevers and A.W.M. Smeulders, "Color Based Object Recognition", *Pattern Recognition*, Vol. 32, No. 3, pp. 453–464, 1999.

[5] D.M. Grieg, B.T. Porteous and A.H. Seheult, "Exact Maximum A Posteriori Estimation for Binary Images", *Journal of Royal Statistical Society - B*, vol. 51, no. 2, pp. 271–279, 1989.

[6] V. Kolmogorov and R. Zabih, "Computing Visual Correspondence with Occlusions via Graph Cuts", in *Proc. Int'l. Conf. on Computer Vision* (ICCV01), Vancouver, pp. II:508–515, 2001.

[7] K. Konolige, "Small Vision Systems: Hardware and Implementation", in *Proc. Int'l. Symposium on Robotics Research*, Salt Lake City, Utah, pp. 203–212, Oct. 1997.

[8] Y. Kuniyoshi, N. Kita, K. Sugimoto, S. Nakamura, T. Suehiro, "A Foveated Wide Angle Lens for Active Vision", in *Proc. Int'l Conf. on Robotics and Automation* (ICRA95), Nagoya, Aichi, Japan, vol. 3, pp. 2982–2988, 1995.

[9] B. Leibe and B. Schiele, "Interleaved Object Categorization and Segmentation", in *Proc. British Machine Vision Conference* (BMVC03), Sep. 2003.

[10] D.G. Lowe, "Object Recognition From Local Scale-Invariant Features," in *Proc. IEEE Int'l Conf. Computer Vision* (ICCV99), Kerkyra, pp. 1150–1157, Sep. 1999.

[11] K. Mikolajczyk and C. Schmid, "Indexing Based on Scale Invariant Interest Points," in *Proc. IEEE Int'l Conf. Computer Vision* (ICCV01), pp. 525–531, Jul. 2001.

[12] K. Mikolajczyk and C. Schmid, "A Performance Evaluation of Local Descriptors," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (CVPR03), pp. 257–263, Jun. 2003.

[13] G. Sandini and V. Tagliasco, "An Anthropomorphic Retina-like Structure for Scene Analysis", *Computer Graphics and Image Processing*, vol. 14, no. 3, pp. 365–372, 1980.

[14] B. Scassellati, "A Binocular, Foveated, Active Vision System", Tech. report, MIT AI Memo 1628, Mar. 1998.

[15] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithm", *Int'l Journal of Computer Vision*, vol. 47, no. 1, pp. 7–42, May 2002.

[16] A. Selinger and R. Nelson, "A Perceptual Grouping Hierarchy for Appearance-Based 3D Object Recognition", *Computer Vision and Image Understanding*, 76(1), pp. 83–92, 1999.

[17] J. Sun, H-Y. Shum and N-N. Zheng, "Stereo Matching Using Belief Propagation", in *Proc. European Conf. Computer Vision* (ECCV02), pp. 510–524, May 2002.

[18] M.J. Swain and D.H. Ballard, "Color Indexing", *Int'l Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[19] C.L. Zitnick and T. Kanade, "A Cooperative Algorithm for Stereo Matching and Occlusion Detection", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 7, pp. 675–684, Jul. 2000.