

Appearance Matching of Occluded Objects Using Coarse-to-fine Adaptive Masks

Jeff Edwards and Hiroshi Murase
NTT Basic Research Labs
Information Sciences Laboratory
Atsugi-shi, Kanagawa, Japan 243-01
{jledward,murase}@eye.brl.ntt.co.jp

Abstract

In this paper, we discuss an appearance matching technique for the interpretation of color scenes containing occluded objects. Dealing with occlusions is very difficult, and we have explored the use of an iterative, coarse-to-fine correlation-based method that uses hypothesized occlusion events to modify the scene-to-template similarity measure at run-time. Specifically, a binary mask is used to adaptively exclude regions of the template image from the correlation computation. At each iteration, these masks are adjusted based on higher resolution scene data and the occluding interactions between multiple object hypotheses. We present results which demonstrate the technique is reasonably robust over a large database of color test scenes containing objects at a variety of scales, and tolerates minor object rotations and global illumination variations.

1. Introduction

This paper addresses the difficult problem of *scene interpretation*, (i.e., the identification and location of objects) in the presence of strong occlusions. In general, a *geometric approach* to this problem attempts to match a 3-D object model to a set of geometric features extracted from the scene. Since such matching relies on local features such as edges and corners, it tends to be tolerant of occlusions. Examples include [2], [9], and [11]. Unfortunately, it seems that the applicability of the geometric approach is limited to very simplistic objects comprised of geometric features that are easy to both model and extract.

In contrast, *appearance-based approaches* model objects purely in terms of 2-D image features. Since the

scene-to-model matching process is performed directly in the image domain rather than in the domain of local geometric features, performance is not degraded by geometric complexity. Demonstrations of robust appearance matching of complex objects include [13], [14], and [7]. The disadvantage of such approaches is that object appearance is a global feature and is therefore very sensitive to occlusions. Rather than extending a geometric method to deal with complexity, we have chosen to investigate an extension of an appearance-based approach to deal with occlusions.

This paper is organized as follows: Sec. 2 describes the problem of interpreting occluded scenes. Sec. 3 outlines our approach and introduces its core concept: the *adaptive mask*. Sec. 4 illustrates the use of adaptive masks on two example scenes. Experimental results are presented in Sec. 5, followed by a discussion in Sec. 6.

2. Problem Description

Fig. 1a shows a typical scene of interest. The scene is assumed to contain M *target objects* against an arbitrary background; the M objects may strongly occlude each other. Given such a scene, we seek to estimate the location, scale, and relative depth order of each target object. Fig. 1b shows the template image (acquired off-line) associated with each target object.

2.1. Appearance-Based Image Spotting

It is the occluding interactions between objects that make interpretation of such scenes so difficult. Recent appearance matching techniques (e.g., [1] and [6]) have had great success with non-occluded objects. In the absence of occlusions, an arbitrarily complex object may be *spotted* by scanning the scene with a template and computing the scene-to-template correlation

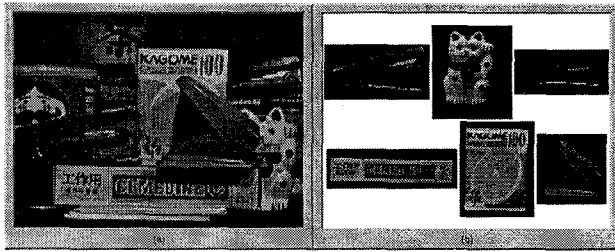


Figure 1. Typical scene and objects of interest. (a) Scene with occlusions and cluttered background. (b) Object templates: stapler2, cat, stapler1, glue box, juice, and stapler3.

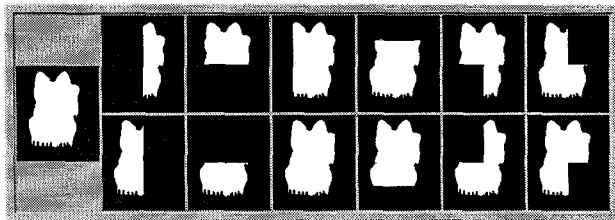


Figure 2. Thirteen “guessed” cat masks.

at each image location. A binary 2-D *mask* is applied to exclude all background pixels from the correlation. The mask is essentially a “silhouette” of the reference object in the template image.

Since appearance varies with view direction, a set of L templates (one for each sampled view direction) is correlated with the scene. The *parametric eigenspace* method, introduced in [7], performs these L correlations very efficiently by using the Karhunen-Loève transform to project the image onto a low-dimensional subspace. The “figure” (i.e. non-background) regions of each of the L masks (one mask per template image) are ANDed together to generate a composite mask. This composite mask is used to search for the target object, which may appear from any viewing direction in the space spanned by the L templates.



Figure 3. Scene at $\frac{1}{8}$, $\frac{1}{4}$, and $\frac{1}{2}$ resolution.

2.2. Dealing with Occlusion

This approach works well at spotting complex non-occluded objects in cluttered scenes for multiple view directions, and tolerates mild occlusions (perhaps 10% of object area.) But as occlusion increases, correlation ceases to reliably indicate object presence because the scene pixels in the occluded regions are uncorrelated with the template pixels; in a sense, these occluded regions add noise to the correlation computations.

One recent proposal uses *local* appearance matching in which small scene windows are correlated with small template windows [8]. The response from each window “votes” for a global scene interpretation. However, the robustness of previous appearance matching results stems largely from the use of *global* appearance rather than local windows. Another technique expands the scene using a set of basis functions that closely resemble the template image; the resulting matching process is more robust to occlusions compared with standard template matching [3].

In contrast, we have taken advantage of the global occluding interactions between scene objects to adaptively improve the correlation metric. Given the hypothesized location of two objects in a particular occluding configuration, one can adjust the correlation mask to take this information into account; i.e., one can attempt to “mask out” the (hypothesized) occluded regions in order to compute the scene-to-template correlation over only the non-occluded regions of the object. By ignoring the occluded pixels, it is hoped that this modified correlation metric will serve as a reliable indicator of object presence.

Unfortunately, this “masking out” of occlusions introduces a dilemma. In order to search the scene for a target object, we need a reliable indicator of object presence (e.g., a correlation-based metric that ignores occluded regions.) Yet to properly mask out these occluded regions, we must already know the objects’ locations so that we can determine which regions are similar to the template (i.e. not occluded), and which regions are grossly different from the template (i.e., occluded.) Since we have no *a priori* knowledge of these occluded regions, we cannot design a fixed mask to exclude them, such as the composite mask in [6]. To escape this circular dilemma, we use an *adaptive* mask that is adjusted at run-time based on scene data.

To study such adaptive masks in isolation from other complicating issues, we have made two simplifying assumptions. First, we restrict the M objects to appear from a single canonical viewing direction. Second, we restrict a target object to be occluded only by one or more of the other $M - 1$ target objects, rather than by



Figure 4. The cat template and the first of the $N=13$ cat masks at multiple resolution levels.

some unknown background entity.

3. The Adaptive Mask Concept

One way to avoid the dilemma discussed in Sec. 2.2 is to initially “guess” which regions of an object are occluded in the scene, and modify the mask accordingly so that these regions are excluded from the correlation computation. A search of the scene is then performed using N such masks, each of which corresponds to a different initial (“guessed”) occlusion hypothesis.

Fig. 2 shows a set of $N = 13$ initial “guessed” occlusion masks for the cat object. The first mask corresponds to a non-occluded cat. The remaining 12 masks each hypothesize a rectangular occluding entity at a certain location with respect to the cat.

Note that this set of masks is not expected to contain a “perfect match” with the actual occluded region, nor must the set span the space of all possible occlusion configurations. It is only necessary that at least one member of this set of N masks be a *reasonable approximation* to the occluded regions of the cat¹. Such a “correct” mask, when placed at the actual location of the object in the scene, should yield the highest correlation from among all N guessed masks, because it eliminates the most occluded, noise-inducing pixels from the correlation computation, while including the most non-occluded, information-bearing pixels.

3.1. Coarse-to-fine Search

The main problem with this approach is the computational burden of MN searches (one search for each of M objects, using each of N guessed masks.) Therefore, we perform the searches at a greatly reduced scene resolution. As noted in [5], the cost of template search increases proportional to the fourth power of resolution. This speed-up technique can be found in several previous works such as [10], [12], and [1].

Fig. 3 shows the scene of Fig. 1a at three coarser resolutions. Fig. 4 shows a cat template and mask at the

¹By “reasonable”, we mean that the residual occluded scene regions *not* excluded by the mask are small enough that correlation will yield a robust indicator of object presence.

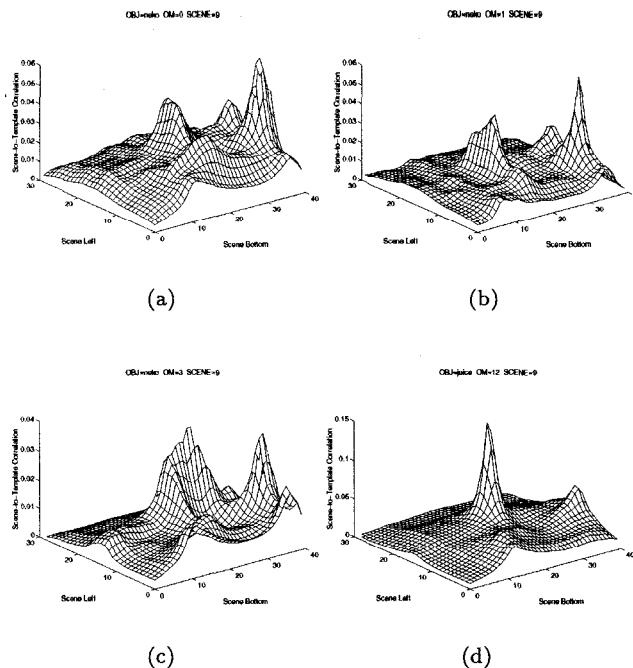


Figure 5. Correlation maps from coarse resolution version of Fig. 1a. (a) Using first cat mask of Fig. 2 (no occlusion.) (b) Using second cat mask (left half occluded.) (c) Using ninth cat mask (top half occluded.) (d) Using thirteenth juice mask (bottom right occluded.)

four resolutions. Each resolution level is generated via Gaussian filtering followed by dyadic downsampling. Thus, the original 320×240 pixel images were searched at a reduced resolution of 40×30 pixels.

Although coarse resolution search is very fast, it provides much less image information for drawing conclusions about object presence, location, scale, etc., and thus increases the possibility of mistakes. So we must *verify* the object hypotheses using the information-rich high-resolution image data.

3.2. Hypothesize and Verify

The hypothesize and verify procedure begins with MN coarse resolution searches. The result is zero or more location hypotheses for each of M target objects².

These hypotheses are verified or rejected based on high resolution image data. Due to the coarseness of the search, the hypothesized locations contain a great

²In our experiments, a hypothesis was generated at each image point where correlation was both a local maximum and above a threshold.

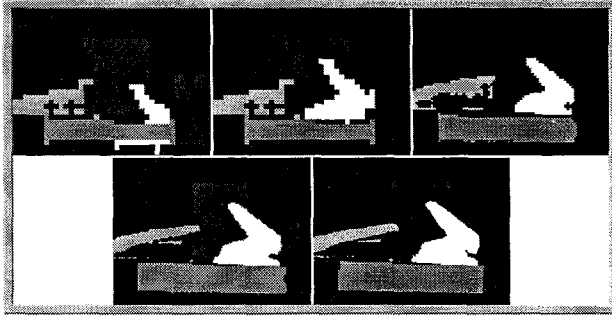


Figure 6. Coarse-to-fine hypothesis evolution.

deal of spatial uncertainty: each pixel in a 40×30 searched image maps to a $8 \times 8 = 64$ pixel neighborhood in the original image. So if we were to jump immediately to the full resolution image in order to verify a hypothesis, it is likely that the coarse-level hypothesized object locations will be significantly perturbed from their true locations, and spatial mismatches of even a few pixels are known to substantially degrade the reliability of correlations (see, for example, [8].)

So in effect, we must perform a fine-resolution search by spatially *perturbing* each coarse-level hypothesis within its range of uncertainty, and compute fine resolution correlations at each perturbed location. Such a search can be more efficiently performed by increasing resolution from coarse to fine *in stages*. At each stage, a medium-resolution search is performed over a small region (say 2×2), rather than a high-resolution search over a large region (say, 8×8 .)

Note that we perform this coarse-to-fine search for both hypothesis verification and reduction of location uncertainty. We can achieve tolerance to scale variations by perturbing hypothesized object *scale* at each stage as well. A similar method is described in [1].

3.3. Objective Function

Rather than rely strictly on masked scene-to-template correlation as the similarity metric used in the staged search, we modulate this metric by incorporating its equivalent, the sum-squared (L_2) error, into an *objective function* $C_h(i, j, T_{h,k})$. For the h th target object, $C_h(i, j, T_{h,k})$ is evaluated at each scene location (i, j) , and for each guessed mask $T_{h,k}$, with $k = 1, \dots, N$ and $h = 1, \dots, M$. A low value of $C_h(i, j, T_{h,k})$ corresponds to high similarity.

The objective function consists of the masked, area-normalized, scene-to-template L_2 error $E(i, j, T_{h,k})$ (computed over the three RGB channels) modified by a pair of soft top-down constraints represented by a *masking term* $P(T_{h,k})$ and a *scaling term* $Q(T_{h,k})$:



Figure 7. Final interpretation with hypothesized non-occluded regions in white, and occlusions in black. (a) Example I. (b) Example II.

$$C_h(i, j, T_{h,k}) = E(i, j, T_{h,k}) + \alpha P(T_{h,k}) + \beta Q(T_{h,k})$$

where $\alpha, \beta \geq 0$ are weighting coefficients. The term $P(T_{h,k})$ increases the cost of an adaptive mask $T_{h,k}$ proportional to the fraction of “masked out” pixels. The idea is to penalize hypotheses in which the object is heavily occluded, so as to force such hypotheses to compensate for their smaller (less reliable) set of supporting pixels with a smaller error $E(i, j, T_{h,k})$.

The scaling term $Q(T_{h,k})$ imposes a cost that varies inversely with the scale of the template mask $T_{h,k}$ (i.e., it favors larger-scaled over smaller-scaled templates.) The staged search has a tendency to converge to hypotheses with scale *slightly* less than the correct scale in the scene. This is because appearance variations within an object are typically less drastic than the appearance variations between object and background. Consequently, a scaled template that is slightly too large (and extends beyond the boundaries of the object into the background) will tend to have larger L_2 error than a template that is slightly too small but which fits within the boundaries of the scene object. The term $Q(T_{h,k})$ counteracts this tendency.

As the coarse-to-fine search progresses, the occluding interactions between the M objects are taken into account at each stage, in order to improve the quality of the initial “guessed” masks, and thus improve the quality of the similarity metric. If the hypothesized locations and scales of cat and juice are such that cat is occluding juice, then the initial juice mask can be replaced with a new mask based on the shape and hypothesized location and scale of cat (with respect to the hypothesized location and scale of juice.)

As resolution increases and hypothesized locations and scales become more precise, the occluding interactions between objects will result in the masks becoming better approximations to the true occluded regions. These improved masks are then used to compute L_2

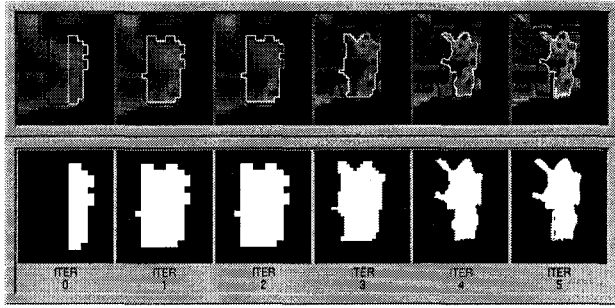


Figure 8. Evolution of the cat mask.

error at the next higher resolution level, and so on. The final goal is a globally consistent scene interpretation, verified at the original resolution, in which each object’s mask has converged to a very good approximation of that object’s actual occluded region.

It should be noted that the coarse resolution hypotheses will sometimes be incorrect due to sparse scene data. The computed L_2 error of such a hypothesis will become very large as resolution increases (i.e., as more information becomes available), and can be rejected at this time. The coarse-to-fine verification is then repeated using the next-most promising coarse-level hypothesis (from the initial scene search) as a replacement. In other words, *backtracking* must be incorporated into the coarse-to-fine search.

4. Behavior of the Approach

In this section, we follow two scenes through the interpretation process to illustrate the algorithm. The set of $M=6$ target objects are shown in Fig. 1b.

4.1. Example I

In this example, Fig. 1a is recursively downsampled three times as shown in Fig. 3³. A coarse resolution search of the scene is then performed for each of $M=6$ objects and $N=13$ guessed masks (see Fig. 2). Fig. 5 shows scene-to-template correlation maps generated from different initial masks.

Figs. 5a shows a correlation map using the first mask of Fig. 2 (containing no occlusion.) The minimum correlation occurs at the correct location of cat in the lower right corner of the image (see Fig. 1a.) In Fig. 5b, the second mask of Fig. 2 (containing a left-half occlusion) was used to compute the correlations, and again the cat is correctly located, except that the peak is

³Note that the coarse-resolution versions of the M templates and their N guessed masks (e.g., Fig. 4), are generated off-line.

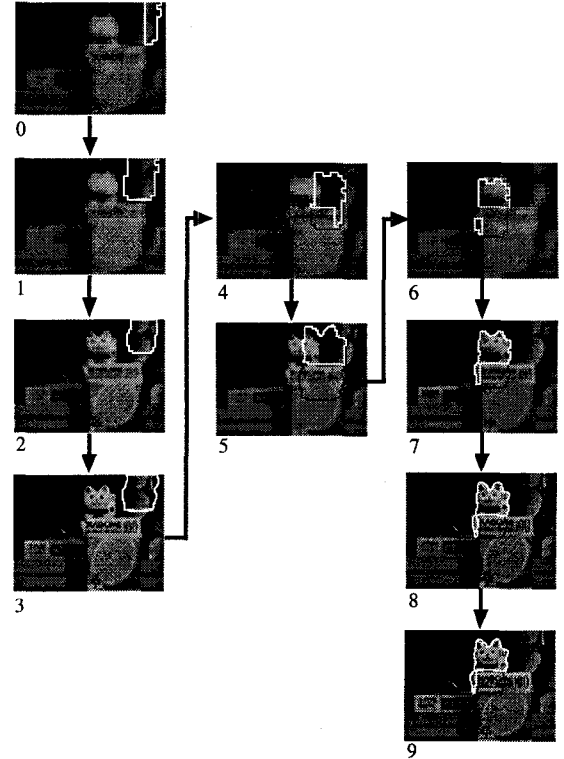


Figure 9. History of cat backtracking.

sharper and higher than in Fig. 5a. This is to be expected since the left portion of cat is occluded in the scene, so the second mask should yield a better cat detector than the first mask. Fig. 5c shows that the ninth mask of Fig. 2 (which incorrectly guesses the top half of cat to be occluded) results in an incorrect point of maximum correlation. Finally, Fig. 5d shows the sharp peak that results from the thirteenth guessed mask (not shown) associated with juice. This guess corresponds to an occluded bottom-right corner of juice, which happens to be true for this scene, so juice is correctly located in the center of the scene. These plots show that when the guessed masks are “correct” (i.e., similar to the occlusion situation in the scene), the resulting correlations become reliable indicators of object location. Note that incorrect local maxima may still exist and be incorrectly interpreted as likely object locations, hence the need for verification and backtracking.

After the MN coarse searches, verification is performed. For each object, a location hypothesis is selected that minimizes $C_h(i, j, T_{h,k})$ over all scene locations (i, j) , and all guessed masks $T_{h,k}$. These M hypotheses provide the starting point for the coarse-to-fine search discussed in Sec. 3.1. Fig. 6 shows the adaptive masks associated with these hypotheses as the

search progresses to full resolution.

Five of the six initial coarse-level object hypotheses were approximately correct; however, the location of `stapler3` (displayed as white in Fig. 6) was grossly in error. Consequently, at the first stage of verification, $C_h(i, j, T_h, k)$ for $h = \text{stapler3}$ exceeded a threshold. The search backtracked, and the incorrect hypothesis was replaced with the second-best (and correct) candidate. As verification proceeds through finer resolutions, residual ambiguities in object location and scale are resolved via the perturbation procedure, and the algorithm converges to the correct scene interpretation. Fig. 7a shows the adaptive masks at termination. Fig. 8 shows the evolving configuration of the cat mask.

4.2. Example II

Fig. 7b shows the final result superimposed over the original scene. In this case, verification backtracked several times before converging to a correct interpretation. Fig. 9 shows the cat mask at each iteration to illustrate the backtracking sequence; a total of 10 iterations were required prior to convergence.

Following coarse search, the initial hypothesized locations of `stapler2`, `glue box`, and `juice` are correct (within the spatial ambiguities inherent to coarse resolution search.) However, both the first- and second-best cat hypotheses are incorrect: the cat is initially hypothesized in the upper right corner of the scene. This incorrect hypothesis survives until the finest resolution level prior to rejection by the error threshold criterion. The search backtracks to the coarsest level, and the next (again incorrect) cat hypothesis survives to the second-finest resolution. Finally, the third (correct) cat hypothesis results in a correct convergence. This example shows that the absence of medium and high spatial frequencies at the coarsest level can lead to mistakes, and hence the need for verification.

5. Experimental Results

To evaluate the robustness of the adaptive masks, two experiments were performed, both using the objects displayed in Fig. 1b. Both the templates and test scenes were color images (displayed here in greyscale.)

5.1. Experiment I

The first experiment evaluated the robustness of the algorithm over a database of 50 occluded scenes. The scenes were generated by arranging the target objects in random occluded configurations against ran-

Object	Num. Instances			
	Total	Success	Mislocated	Missing
glue box	28	96%	0%	4%
stapler1	23	78%	9%	13%
cat	30	90%	10%	0%
juice	37	100%	0%	0%
stapler3	16	56%	0%	44%
stapler2	25	96%	4%	0%
TOTAL	159	89%	4%	7%

Table 1. Robustness results for 50 scenes.

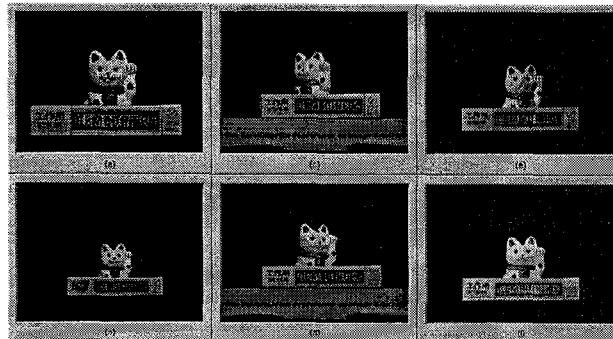


Figure 10. Test scenes. (a) and (b) Extremes of scale. (c) and (d) Extremes of 3-D rotation. (e) and (f) Extremes of illumination (for which successful scene interpretation was obtained.)

dom cluttered backgrounds. The average number of target objects present per scene was 3.2.

We ran the scene interpretation algorithm on each test scene. The results are summarized in Table 1. A total of 159 objects appeared in the 50 scenes; the algorithm correctly identified and located 142 of them, or 89%. In 6 cases (4%), the algorithm converged to an incorrect object hypothesis; in an additional 11 cases (7%), the algorithm rejected all object hypotheses although the object was actually in the scene.

5.2. Experiment II

In the second experiment, we explicitly investigated the robustness of the algorithm to variations in scale, 3-D object rotation, and global scene illumination.

In the first part of the experiment, a simple occluded scene was generated and tested 21 times. In each test, the scene was moved further from the camera to investigate performance subject to scale changes only. Figs. 10a and 10b show the two extremes of scale. In each of the 21 tests, the algorithm correctly determined the presence, location, and scale of cat and glue box.

Recall the simplifying restriction imposed in Sec. 2, in which we limit object appearance to a single view direction. In the second portion of the experiment, we

investigated the robustness of the adaptive masks to deviations from this restriction. The same simple configuration was used to generate 10 scenes in which the rotation of cat varied from -25° to $+20^\circ$ (with respect to cat's template image.) Figs. 10c and 10d show the two extremes of rotation. The algorithm succeeded 7 times, and failed 3 times, with the failures occurring at angles of -25° , -15° , and $+20^\circ$.

In the third portion of the experiment, tolerance to illumination variation was investigated. Fifteen scenes were generated in which illumination was varied by adjusting the position and intensity of spotlights. Figs. 10e and 10f show two such scenes.

In five of the scenes, both cat and glue box failed to be identified. In the remaining ten scenes, including the two scenes in Fig. 10e and Fig. 10f, recognition was successful. It is difficult to show quantitative results for illumination tolerance; all that can be reported is that the algorithm succeeded in the presence of moderate illumination variations, such as those in Fig. 10, but failed when these variations became more pronounced.

6. Discussion

The main contribution of this work is the introduction and investigation of coarse-to-fine adaptive masks to address the degradation of appearance matching in the presence of occlusions. Hypothesized occlusions between multiple objects are used to perform run-time modification of the masks and their associated scene-to-template similarity metrics. The greatest strength, shared by all appearance matching approaches, is that performance is independent of object complexity.

The appearance matching core of the approach also yields a good degree of robustness to image noise and to reasonable amounts of variation in illumination, rotation, and scale; e.g., corruption of a few scene pixels will have only a minor effect on performance, unlike most geometric methods. The use of 2-D appearance templates as an object representation also allows model databases to be acquired via a "teach-by-showing" methodology.

Recall that in this work, the image spotting problem was restricted to a single view direction for each target object in order to study the occlusion issue in isolation. The next step will be to apply adaptive masks to the previous image spotting work of [6], in which objects may be viewed from a range of directions.

Another area of future research will be the development of a more sophisticated method of initial mask selection. Ideally, one would generate a stochastic occlusion model (using many thousands of simulated scene configurations) for a particular collection of objects.

Given such a model, an optimal (in some sense) set of initial masks could be generated.

We are also interested in improving upon L_2 correlation, which is by no means the optimal measure of similarity between two images. Recently proposed alternatives to L_2 correlation appear to improve performance for some classes of image [4].

Acknowledgements

The authors wish to thank Dr. T. Izawa and Dr. K. Ishii of NTT Basic Research Labs for their help and encouragement in conducting this research.

References

- [1] V. A. Anisimov and N. D. Gorsky. Fast hierarchical matching of an arbitrarily oriented template. *Pattern Recognition Letters*, 14:95–101, 1993.
- [2] N. Ansari and E. Delp. Partial shape recognition: A landmark-based approach. *IEEE Trans. on Pattern Analysis and Machine Intell.*, 12(5):470–483, 1990.
- [3] J. Ben-Arie and R. Rao. Optimal template matching by nonorthogonal image expansion using restoration. *Machine Vision and Applications*, 7:69–81, 1994.
- [4] M. Boninsegna and M. Rossi. Similarity measures in computer vision. *Pattern Recognition Letters*, 15:1255–1260, 1994.
- [5] P. J. Burt. Smart sensing within a pyramid vision machine. *Proc. of the IEEE*, 76(8):1006–1015, 1988.
- [6] H. Murase and S. Nayar. Image spotting of 3d objects using parametric eigenspace representation. In *The 9th Scandinavian Conf. on Image Analysis*, June 1995.
- [7] H. Murase and S. Nayar. Visual learning and recognition of 3-d objects from appearance. *Intl. Journal Computer Vision*, 14(1):5–24, 1995.
- [8] K. Ohba and K. Ikeuchi. Recognition of the multi-specularity objects for bin-picking task. In *Intl. Conf. on Intelligent Robots and Systems*, November 1996.
- [9] K. S. Ray and D. D. Majumder. Recognition and positioning of partially occluded 3-d objects. *Pattern Recognition Letters*, 12:93–108, 1991.
- [10] A. Rosenfeld and G. J. Vanderbrug. Coarse-fine template matching. *IEEE Transactions on Systems, Man, and Cybernetics*, 2:104–107, 1977.
- [11] E. Salari and S. Balaji. Recognition of partially occluded objects using b-spline representation. *Pattern Recognition*, 24(7):653–660, 1991.
- [12] S. Sista, C. Bouman, and J. Allebach. Fast image search using a multiscale stochastic model. In *Proc. Intl. Conference on Image Processing*, 1995.
- [13] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 586–591, 1991.
- [14] C. R. Wiles and M. R. B. Forshaw. Recognition of volcanoes on venus using correlation methods. *Image and Vision Computing*, 11(4):188–196, 1993.