Machine Vision and Applications (1998) 10: 232–242

Machine Vision and
Applications

Coarse-to-fine adaptive masks for appearance matching of occluded scenes

J.L. Edwards, H. Murase

Media Information Recognition Research Group, NTT Basic Research Labs, Atsugi-shi, Kanagawa 243-01, Japan;
e-mail: {jledward,murase}@eye.brl.ntt.jp

Received: 21 November 1996 / Accepted: 14 October 1997

Abstract. In this paper, we discuss an appearance-matching approach to the difficult problem of interpreting color scenes containing occluded objects. We have explored the use of an iterative, coarse-to-fine sum-squared-error method that uses information from hypothesized occlusion events to perform run-time modification of scene-to-template similarity measures. These adjustments are performed by using a binary mask to adaptively exclude regions of the template image from the squared-error computation. At each iteration higher resolution scene data as well as information derived from the occluding interactions between multiple object hypotheses are used to adjust these masks. We present results which demonstrate that such a technique is reasonably robust over a large database of color test scenes containing objects at a variety of scales, and tolerates minor 3D object rotations and global illumination variations.

Key words: Object recognition – Occlusion – Appearance matching – Image similarity metrics – Coarse-to-fine search

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. There are essentially two broad approaches to this problem: geometry-based and appearance-based. In general, geometric approaches attempt to match a 3D object model to a set of geometric features extracted from the scene. Since the matching relies on *local* features such as edges and corners, geometric approaches tend to be tolerant of occlusions. Examples include Ansari and Delp (1990), Han and Jang (1990), Chaudhury et al. (1990), Ray and Majumder (1991), and Salari and Balaji (1991). Unfortunately, the applicability of the geometric approach tends to be limited to very simplistic objects comprised of geometric primitives that are easy to both model and extract.

In contrast, appearance-based approaches model objects purely in terms of their 2D appearance in images, and the scene-to-model matching process is performed directly in the image domain rather than in the domain of geometric features. Consequently, the performance of appearance-based approaches is essentially unaffected by object complexity. For example, human faces are quite complex geometrically, and so it comes as no surprise that, when Brunelli and Poggio (1993) directly compared the local feature approach against the appearance-matching approach, they found that the latter was significantly more robust. Additional demonstrations of robust appearance-matching of complex objects include Turk and Pentland (1991), Wiles and Forshaw (1993), Liu and Caelli (1988), and Murase and Nayar (1995a).

However, the disadvantage of the appearance-matching approach is that object appearance is a *global* feature and is therefore very sensitive to occlusions. Consequently, to interpret scenes containing complex occluded objects, one could consider extending a geometry-based approach to deal with complexity, or extending an appearance-based approach to deal with occlusions; we have chosen to investigate the latter course.

Examples of similar approaches to the problem include the *local* appearance-matching work of Ohba and Ikeuchi (1996), in which small scene windows are correlated with small template windows, and the "expansion-matching" method of Ben-arie and Rao (1993, 1994), in which the scene image is expanded using a set of basis functions that closely resemble the template image.

In contrast, this paper investigates the possibility of taking advantage of the global occluding interactions between multiple scene objects to adaptively improve a similarity measure based on sum-squared error (SSE) by "masking out" suspected occluded regions in the scene. The paper is organized as follows. Section 2 describes the problem of scene interpretation in the presence of occlusions. Section 3 outlines our approach and introduces its core concept: the *adaptive mask*. Section 4 demonstrates the approach on two example scenes and presents preliminary experimental results, followed by a discussion in Sect. 5.

Correspondence to: J.L. Edwards

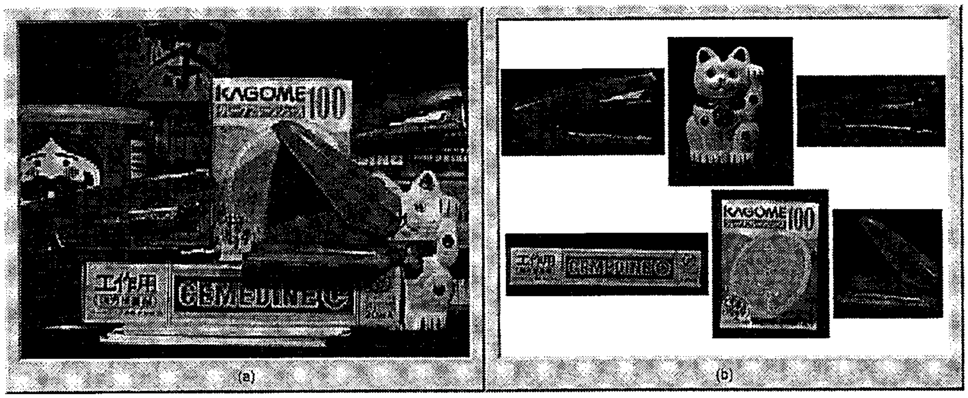


Fig. 1a, b. Typical scene and objects of interest. **a** Scene with occlusions and cluttered background. **b** Object templates: `stapler2`, `cat`, `stapler1`, `glue box`, `juice`, and `stapler3`

2 Problem description

Figure 1a shows an example of the class of scenes that we are interested in interpreting. The scene is assumed to contain L *model objects* set against an arbitrarily complex background. The L objects may or may not occlude each other. Given such an input scene, the goal is to estimate the location, scale, and relative depth order of each object. Figure 1b shows the reference templates (acquired off-line) associated with each of the $L=6$ model objects in a particular database.

2.1 Appearance-based image spotting

The occluding interactions between the L objects is why the interpretation of scenes such as Fig. 1a is so difficult. In fact, recent appearance-matching techniques, based on global object appearance, have had great success when dealing with non-occluded objects. In the absence of occlusions, an object of arbitrary complexity may be *spotted* by scanning the scene with a reference template and computing the scene-to-template SSE at each image location.

Let the N-dimensional vector $\mathbf{v}$ represent the ordered pixel values (normalized such that $||\mathbf{v}||_2=1$) of a particular model object obtained from a reference template, where N is the number of model object pixels (background pixels[1] that do not correspond to the actual model object are not included in $\mathbf{v}$.) Likewise, let the N-vector $\mathbf{u}(i,j)$ represent the values of a corresponding ordered set of *scene* pixels, where (i,j) denotes a particular column and row displacement (i.e., a location) in the scene image. Then the similarity of vectors $\mathbf{v}$ and $\mathbf{u}(i,j)$ will provide evidence regarding whether or not an instance of the model object exists in the scene at location (i,j). This similarity can be measured by the SSE $E^2(i,j)$:

$$E^2(i,j)=\sum_{k=1}^{N}(v_k-u_k(i,j))^2\ , \qquad (1)$$

where $u_k(i,j)$ and v_k represent the kth pixel values in $\mathbf{u}(i,j)$ and $\mathbf{v}$, respectively. The value $\min\limits_{(i,j)} E^2(i,j)$ can be compared with a threshold[2] to decide the presence or absence of the model in the scene, and if present, then a reasonable estimate for the location of the object in the scene would be:

$$(i^*,j^*)=\arg\min_{(i,j)} E^2(i,j)\ . \qquad (2)$$

Indeed, if one were to generate an *SSE map* by plotting $E^2(i,j)$ for each location (i,j) in the scene, one would expect to see a global minimum at the actual location of the model object (assuming it is present.) Since we are dealing with color images, it should be noted that the individual pixel differences in Eq. 1 are actually summed squared differences across the three RGB color channels:

$$\begin{aligned}v_k-u_k(i,j)=&\ |v_k^R-u_k^R(i,j)|^2+|v_k^G-u_k^G(i,j)|^2\\&+|v_k^B-u_k^B(i,j)|^2\ .\end{aligned} \qquad (3)$$

Examples of this basic approach include Anisimov and Gorsky (1993) and Murase and Nayar (1995b).

2.2 Dealing with occlusion

This approach works very well at spotting complex non-occluded objects in cluttered scenes, and is generally tolerant to mild occlusions (covering perhaps 5–10% of the object.) However, as the degree of occlusion increases, the metric $E^2(i,j)$ quickly ceases to be a reliable indicator of object presence because a large fraction of the scene pixels (in the occluded regions) will no longer have any statistical correlation with the template pixels. For example, let $\mathbf{y}$ be an N-dimensional binary vector (an *occlusion indicator function* such that

$$y_k=\begin{cases}1 & \text{if scene pixel } u_k(i,j) \text{ is not occluded}\\ 0 & \text{if scene pixel } u_k(i,j) \text{ is occluded}\end{cases} \qquad (4)$$

for $k=1,\ldots,N$. Furthermore, let the non-occluded object pixels in the scene be indexed by the set $\mathsf{S}=\{k|y_k=1, k=1,\ldots,N\}$. Then we can rewrite Eq. 1 as:

$$E^2(i,j)=\sum_{k\in \mathsf{S}}(v_k-u_k(i,j))^2+\sum_{k\notin \mathsf{S}}(v_k-\eta)^2\ , \qquad (5)$$

where η denotes a random variable corresponding to the (unknown) distribution of scene pixel values across the entire population of possible scenes. So, in general, η will have no statistical correlation whatsoever with the corresponding

[1] This implies the existence of some simple (perhaps even manual) thresholding operation to label the template pixels as either "figure" or "background".

[2] Threshold values could be obtained from an analytical noise model or purely from experiments.

model pixel value v_k (because this kth scene pixel is occluded.) In other words, if $E^2(i,j)$ is the output of our "detector", then the first term in Eq. 5 can be thought of as the "signal", and the second term as the "noise". Since $(v_k - \eta)^2$ will tend to be much greater than $(v_k - u_k(i^*,j^*))^2$, it does not take very much occlusion to cause a dramatic degradation in the performance of such a "detector".

As occluded regions inject significant amounts of noise into the $E^2(i,j)$ computations, we will continue the analogy and attempt to "filter out" this noise. If we somehow had prior knowledge regarding the nature of the occluded region, we could attempt to "mask out" the regions that are known (or hypothesized) to be occluded, in order to compute SSE over only the non-occluded regions of the scene. By ignoring the occluded pixels, the signal-to-noise ratio associated with this modified SSE similarity measure can hopefully be improved enough to restore it as a reliable detector of object presence.

So let us define another N-dimensional binary vector **z** as an *occlusion mask*, such that

$$z_k = \begin{cases} 1 & \text{if scene pixel } u_k(i,j) \\ & \quad \text{is included during SSE computation,} \\ 0 & \text{if scene pixel } u_k(i,j) \\ & \quad \text{is ignored during SSE computation.} \end{cases} \tag{6}$$

Then in this case, the SSE computation in Eq. 1 will become:

$$E^2(i,j) = \sum_{k=1}^{N} z_k(v_k - u_k(i,j))^2 = \sum_{k \in \mathsf{T}} (v_k - u_k(i,j))^2 \ , \tag{7}$$

where $\mathsf{T} = \{k | z_k = 1; k = 1, \ldots, N\}$ is the set of non-masked pixels (i.e., the set of object pixels that are *believed* to be non-occluded and therefore included in the SSE computation.) If perfect a priori knowledge of the actual occlusion situation were somehow available, we could simply assign $\mathbf{z} = \mathbf{y}$ (and hence $\mathsf{T} = \mathsf{S}$) and eliminate all noise (i.e., occluded pixels) while keeping the full signal (i.e., the non-occluded pixels.) Unfortunately, such a priori knowledge of the scene occluded regions is never available.

3 The adaptive mask concept

Since we have no a priori information regarding the nature of the occluded regions, one course of action would be to generate a set $\mathsf{M} = \{\mathbf{z_h} | h = 1, \ldots, M\}$ of M initial *occlusion hypotheses* (i.e., "guesses".) A search of the scene could then be performed using these M occlusion masks, each of which will provide a different measure $E_h^2(i,j)$ of SSE:

$$E_h^2(i,j) = \sum_{k \in \mathsf{T}_h} (v_k - u_k(i,j))^2 \tag{8}$$

for $h = 1, \ldots, M$. When we apply these M different similarity measures $E_h^2(i,j)$ over the scene, we will obtain M different SSE maps. The "best" occlusion mask $\mathbf{z}_{h^*}$ (corresponding to SSE measure $E_h^2(i,j)$) will be the one which masks out the most "noise" (occluded pixels) and retains the most "signal" (non-occluded pixels.) Without providing a rigorous statistical analysis, it can be stated that a reasonable expectation is that the minimum-error triplet $(i^*, j^*; h^*)$, defined by

$$(i^*, j^*; h^*) = \arg \min_{(i,j;h)} E_h^2(i,j) \tag{9}$$

will provide us with the most likely position (i^*, j^*) of the model object, as well as the most likely occlusion situation $\mathbf{z}_{h^*}$ (at least from among the set of admissible occlusion masks $\mathbf{z}_h \in \mathsf{M}$.)

3.1 Selecting initial masks

Before we can begin to scan the scene with this set M of hypothetical occlusion masks, we must address the issue of how to choose this set, as well as the number of masks M that will be sufficient. Note that the space of possible occlusion masks is huge. Our six model objects each contain on the order of 5000 pixels; this implies the existence of approximately 2^{5000} possible occlusion situations for each model object! From this huge space, we may select only a few masks with which to test the scene.

Recall that the purpose of applying an occlusion mask $\mathbf{z}_h$ is to obtain a good approximation to the actual occluded regions **y** of the scene object. Furthermore, the fact that we are taking a minimum in Eq. 9 implies that our goal should be to select a set of masks M that minimize the probability that *none* of the M masks is a good approximation to **y**. If at least *one* selected mask is a good approximation to **y**, then it will yield a good object detector $E_h^2(i,j)$. A mask $\mathbf{z}_h$ can be considered a "good approximation" to **y** if the probability is high that, for any given pixel, $z_{h,k} = y_k$ (i.e., $z_{h,k}$ and y_k are both 1 or both 0.) In more formal terms, we seek a set of occlusion masks $\mathsf{M} = \{\mathbf{z}_h | h = 1, \ldots, M\}$ such that

$$\Pr\{(z_{h^*,k} \neq y_k) \geq \epsilon\} \leq \delta \ \ h^* = \arg \min_h \Pr\{z_{h,k} \neq y_k\} \tag{10}$$

and both δ and ϵ are arbitrarily small constants.

To obtain L sets of occlusion masks M_l (one set for each of the L model objects), we took a frequency interpretation of the probabilities in Eq. 10 and generated 1000 random, simulated, occluded scenes using the $L = 6$ model objects in our experimental database. For each model object, this scene data yielded 1000 points in an N-dimensional mask space, denoted by $\mathbf{y}_i, i = 1, \ldots, 1000$. Selecting a particular mask $\mathbf{z}_h$ can conceptually be thought of as placing a hypersphere of radius ϵ in this mask space, centered at $\mathbf{z}_h$. The random occlusion patterns $\mathbf{y}_i$ that happen to fall within this hypersphere will be well approximated by $\mathbf{z}_h$. One can continue to select masks (and place hyperspheres) until the fraction of the 1000 random mask points left "uncovered"[3] by any of the M hyperspheres is less than δ.

In our implementation, we select the set of (fixed) M masks that minimize δ for a fixed value of ϵ. We use a slight variation of the parametric clustering algorithm described in Fukunaga (1990) to select the M occlusion masks in a near-optimal manner. Figure 2 shows the resulting set $\mathsf{M}^{\mathtt{cat}}$ of 20 initial occlusion masks for the `cat` object, as generated by the clustering algorithm. Note that the first mask corresponds to a situation in which the `cat` is fully visible.

Of course these sets of masks cannot be expected to contain "perfect matches" with the actual occluded regions **y**.

[3] A random mask point $\mathbf{y}_i$ is "covered" by the hth mask's hypershere if the Hamming distance $||\mathbf{y}_i - \mathbf{z}_h||_1 \leq \epsilon$.

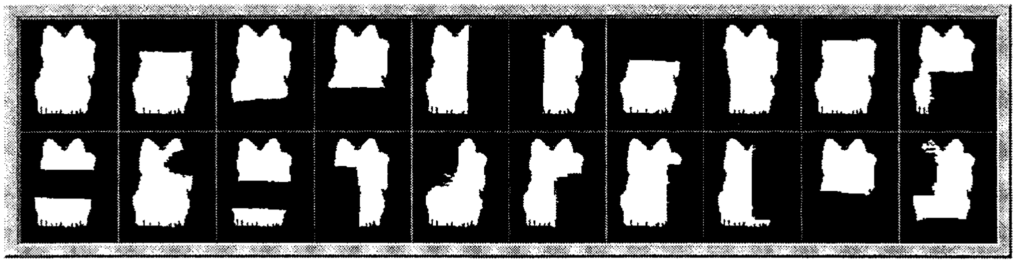


Fig. 2. Set $\mathsf{M}^{\mathsf{cat}}$ of 20 initial occlusion masks statistically generated for the `cat` object

However, since the mask sets are constructed in accordance with the constraint in Eq. 10, we will at least have probabilistic guarantees on their likelihood of approximating the actual occlusion situation in the scene.

3.2 Coarse-to-fine search

Гhe major difficulty with the approach outlined above is that performing a complete search of the scene for each of the M hypothesized occlusions masks would impose a very large computational burden. Therefore, matters of computational efficiency force one to perform both the search and the actual correlations at a greatly reduced scene resolution. As Burt (1988) noted, the computational cost of template search increases proportional to the fourth power of image resolution. So, there is a great incentive to reduce the image resolution at which the search is conducted[4]; this computational speed-up technique can be found in several previous works, such as Rosenfeld and Vanderbrug (1977), Sista et al. (1995), and Anisimov and Gorsky (1993).

Figure 3 shows the scene of Fig. 1a at three successively coarser resolutions. Similarly, Fig. 4a and b shows the template and an example mask associated with the `cat` object at the four resolutions. Each resolution level is generated via Gaussian filtering followed by downsampling (by a factor of two.) In our experiments, the SSE-based searches were 'nitially conducted at the coarsest of these four resolution ıevels (i.e., on a 40 × 30 array of pixels, rather than on the original 320 × 240 pixel image.)

Although a coarse resolution search is much faster, the trade-off is that it also provides much less image information (fewer pixels) for use in drawing conclusions about object presence, location, scale, etc., and thus significantly increases the probability of mistaken identification. Therefore, we must *verify* the hypothesized locations of the model objects using the much more information-rich high-resolution image data.

So the situation is as follows: a total of LM coarse-resolution SSE-based searches are conducted: one search for each of the L model objects, using each of the M hypothetical occlusion masks associated with that object. The result is zero or more promising hypotheses $(i^*, j^*; h^*)$ for each of the L model objects. In our experiments, a "promising" hypothesis was defined as a point (i', j') in the image where the SSE $E_h^2(i', j')$ for one of the occlusion masks $\mathbf{z}_h$ is both a local minimum and below an empirically established threshold.

This set of hypotheses must then be verified or rejected using higher resolution image data. In addition, due to the coarseness of the search, these location hypotheses contain a great deal of spatial uncertainty. Consider that each pixel in the 40 × 30 searched image maps to a 8 × 8 = 64-pixel neighborhood in the original 320 × 240 image. Thus, if we were to simply jump immediately to the full-resolution image in order to verify a hypothesis, it is quite likely that the coarse-level hypothesized object locations will be significantly perturbed from their true locations. Such spatial mismatches of even a few pixels are known to substantially degrade the reliability of SSE measurements (see, for example, Ohba and Ikeuchi 1996).

So, in effect, we must perform a fine-resolution search by spatially "perturbing" each coarse-level hypothesis $(i^*, j^*; h^*)$ within its range of resolution-induced uncertainty $(i^* + \delta i, j^* + \delta j; h^*)$, and compute fine-resolution SSE at each such perturbed location. Such a search can be much more efficiently performed by increasing the image resolution *in stages*[5] from coarse to fine, through each of the resolution levels shown in Figs. 3 and 4.

We perform such a coarse-to-fine, staged search for the purposes of both verifying and reducing the spatial uncertainty of the initial location hypotheses associated with each of the L model objects. In addition, we can achieve tolerance to scale variations by perturbing the hypothesized *scale* of each object at each stage as well, by computing multiple SSE values, using model templates $\mathbf{v}^\gamma$ (and occlusion masks $\mathbf{z}_h^\gamma$) scaled by different magnification constants[6] γ. A similar method was described by Anisimiov and Gorsky (1993).

3.3 Heuristic objective function

As discussed in the previous sections, the SSE metric $E_h^2(i, j)$ serves as our "object detector", and returns zero or more local minima $(i^*, j^*; h^*)$ for each model object; these minima are ranked by SSE value and become candidate hypotheses. However, it was noticed that experimental results could be

[4] Note that there exist a number of other speed-up methods for template matching, such as invoking probabilistic image models in order to perform the pixel-to-pixel correlation operations in a more intelligent sequence (see Margalit and Rosenfeld 1990). However, resolution reduction appears to be the most straightforward and powerful of these methods.

[5] At each stage, a small-area, medium-resolution perturbed search is performed over a small region (say 2 × 2 low-resolution pixels), rather than a high-resolution search over a large (say, 8 × 8 high-resolution pixels) region.

[6] In our experiments, γ took values of 0.7, 0.8, 0.9, 1.0 (no scaling), 1.1, 1.2, and 1.3.

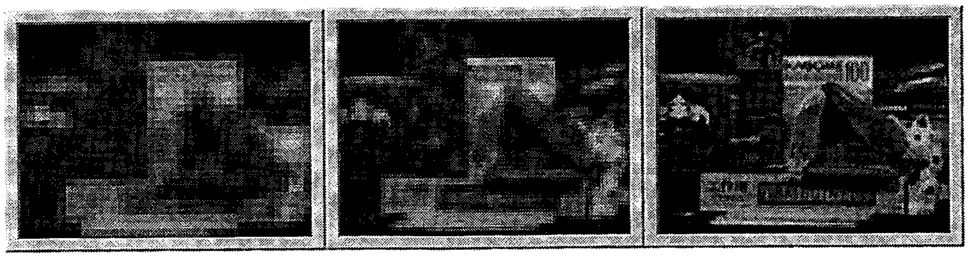


Fig. 3. Scene of Fig. 1a at $\frac{1}{8}$, $\frac{1}{4}$, and $\frac{1}{2}$ resolution

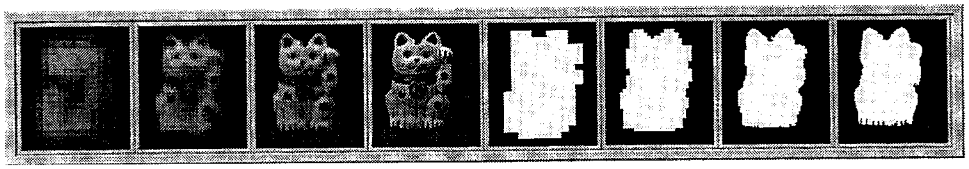


Fig. 4a, b. Multiple resolution levels. **a** The `cat` model template $\mathbf{v}^{\mathtt{cat}}$. **b** The mask $\mathbf{z}_1^{\mathtt{cat}}$

improved somewhat if the hypothesis ranking was based not only on the raw, bottom-up SSE values, but also on some simple top-down heuristics. Therefore, when deciding which hypotheses to verify, we form our rankings based on an *objective function* $\mathscr{C}(i,j,h)$, which combines the raw $E_h^2(i,j)$ output with a *masking-area term* $P(\mathbf{z}_h)$ term and a *scaling term* $Q(\gamma_h)$:

$$\mathscr{C}(i,j,h) = E_h^2(i,j,h) + \alpha P(\mathbf{z}_h) + \beta Q(\gamma_h)\ , \qquad (11)$$

where α and $\beta \geq 0$ are weighting coefficients which allow the influence of the masking-area and scaling terms to be tuned for a specific set of objects. Hypotheses $(i^*, j^*; h^*)$ with lower values of $\mathscr{C}(i,j,h)$ are more favorable.

The purpose of the masking term $P(\mathbf{z}_h) = (N - ||\mathbf{z}_h||_1)/N$ is to increase the cost penalty on an adaptive occlusion mask $\mathbf{z}_h$ as the fraction of "masked- out" pixels increases. The purpose is to impose a penalty on hypotheses that ignore large numbers of scene pixels (because they are presumably occluded), and hence force such hypotheses to compensate for their smaller and less reliable set of support pixels with smaller error $E_h^2(i,j)$ over their non-masked regions.

The scaling term $Q(\gamma) = (1-\gamma)/\gamma$ imposes a negative cost for magnified hypotheses, a positive cost for dilated hypotheses, and zero cost for non-scaled hypotheses (recall from Sect. 3.2 that γ refers to the factor of magnification used for scale perturbation during the coarse-to-fine search.) The justification is that the staged search has a tendency to converge to object hypotheses that are otherwise correct but are scaled *slightly* smaller than the actual object in the scene (usually less than 5%.) This is because the appearance variations within an object are generally less drastic than the appearance variations between the object and the background. Consequently, a scaled model template that is slightly too big (and therefore extends beyond the boundaries of the scene object into the background) will tend to have larger SSE than a template that is slightly too small (but which at least fits within the boundaries of the scene object.) The scaling term $Q(\gamma_h)$ counteracts this tendency.

3.4 Run-time modification of masks

At each resolution of the coarse-to-fine search, the current hypothesized locations and scales $\mathbf{z}_{h^*}^{\mathtt{object}}$ are available for each object. We can improve the quality of the similarity measure $E_{h^*}^2(i,j)$ by replacing each minimum SSE mask $\mathbf{z}_{h^*}$ with a new mask in which all pixels are non-occluded, *except for those pixels that would be occluded assuming that the current hypothesized object locations and scales are indeed correct.*

Due to this run-time modification procedure, as the resolution level increases and the hypothesized object locations and scales become more precise, the occluding interactions between the objects will allow the occlusions masks to be continuously improved, becoming better and better approximations to the actual occluded regions $\mathbf{y}^{\mathtt{juice}}$, $\mathbf{y}^{\mathtt{cat}}$, etc. These improved masks are then used in the SSE computations at the next-higher resolution level, and so on. The final goal is a globally consistent scene interpretation, verified at the finest resolution level, in which each object's adaptive mask $\mathbf{z}^{\mathtt{object}}$ has converged to a very good approximation of that object's actual occluded regions $\mathbf{y}^{\mathtt{object}}$.

It should be noted that, because the LM SSE-based searches are performed at very low image resolution, there will sometimes be cases in which one or more model objects will be assigned "incorrect" coarse-resolution occlusion hypotheses $(i^*, j^*; h^*)$, (i.e., when $\mathbf{z}_{h^*}^{\mathtt{object}} = \arg\min_{i,j;h} E_h^2(i,j)$ is not, in fact, a good approximation to $\mathbf{y}^{\mathtt{object}}$.) Therefore if the computed SSE of a hypothesis exceeds an empirical threshold at a certain resolution level, the hypothesis is rejected and the coarse-to-fine verification search is repeated using the next most promising coarse-level hypothesis generated during the initial scene search. This procedure implements a simple form of *backtracking*.

4 Experimental results

In this section, we follow two typical scenes through the interpretation process, in order to better illustrate the algorithm. The set of $L = 6$ model objects used in these examples are those of Fig. 1b. To evaluate the robustness of the adaptive-mask approach, we also performed two sets of experiments. In both experiments, the object database displayed in Fig. 1b was used[7], and both the model templates

[7] The two illustrative examples were actually generated using a set of 13 initial occlusion masks for each model object; these 13-mask sets were simplified approximations to the full 20-mask sets which were used in

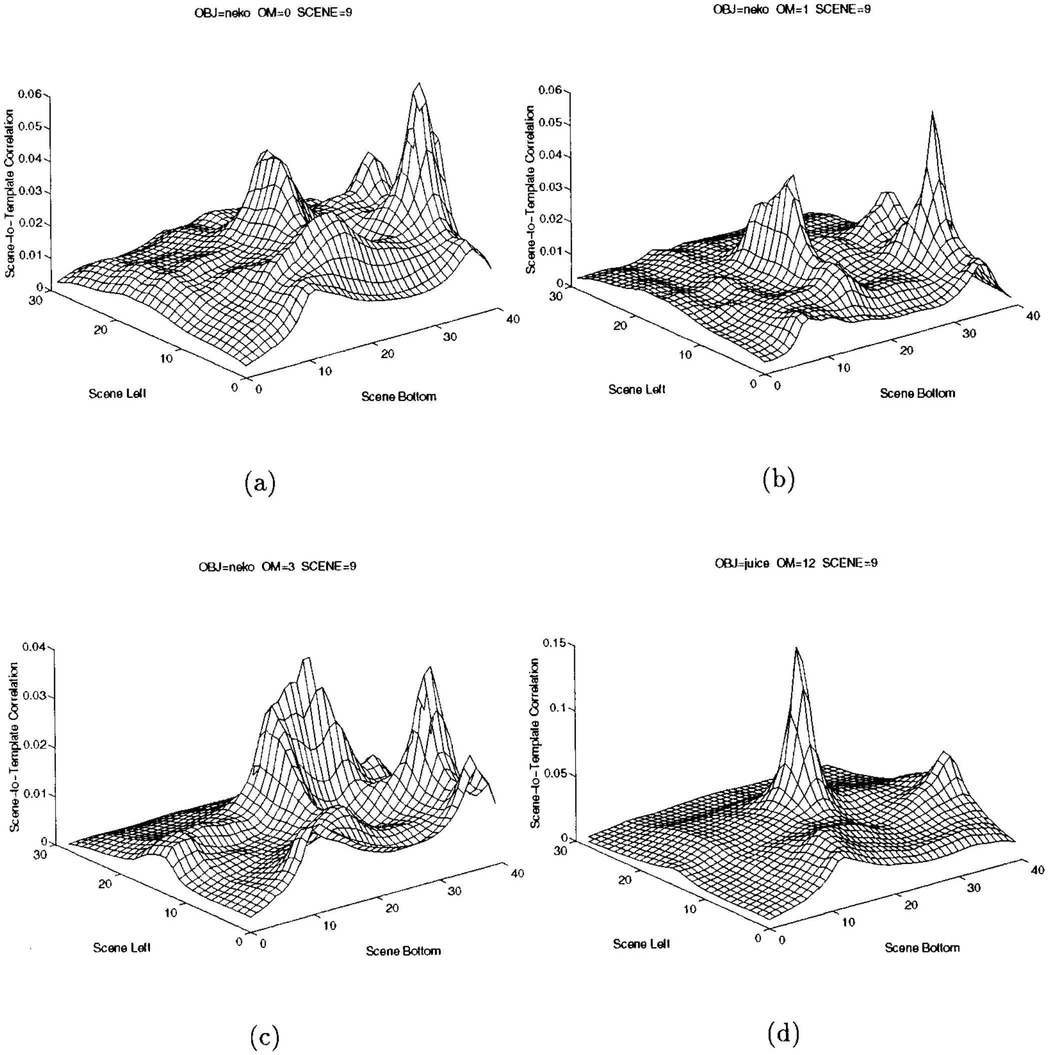


Fig. 5a–d. SSE maps (inverted for clarity) from coarse resolution version of Fig. 1a. **a** Using mask $\mathbf{z}_1^{\mathtt{cat}}$ (no occlusion.) **b** Using mask $\mathbf{z}_6^{\mathtt{cat}}$ (left half occluded.) **c** Using mask $\mathbf{z}_7^{\mathtt{cat}}$ (top half occluded.) **d** Using mask $\mathbf{z}_{13}^{\mathtt{juice}}$ (bottom right occluded)

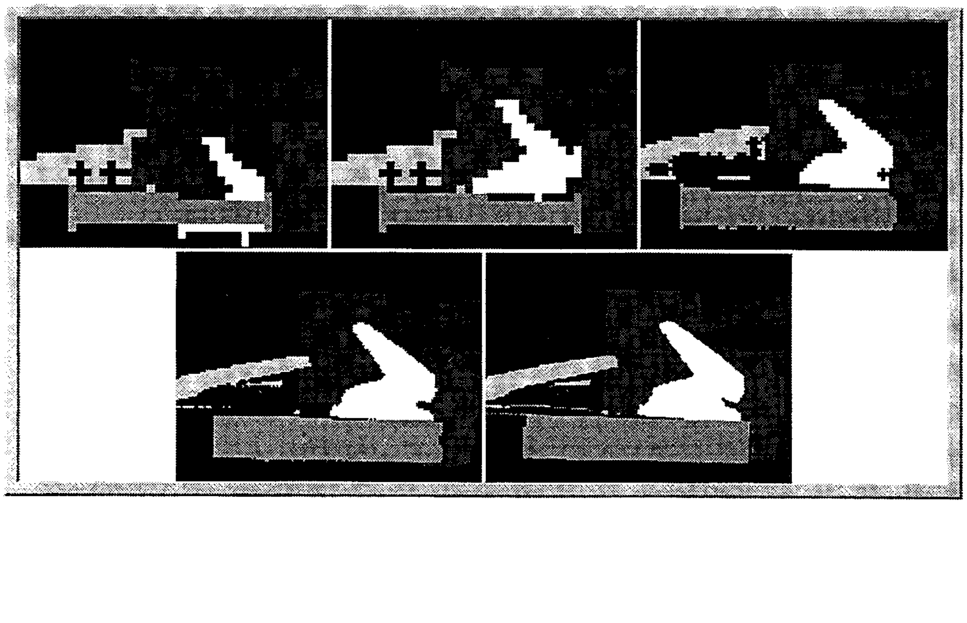


Fig. 6. Coarse-to-fine evolution of object hypotheses

and test scenes were RGB color images (displayed here in grey-scale format.)

4.1 Illustrative example 1

In this example, the scene of Fig. 1a was used. The first step is to recursively downsample the scene three times, yielding the four scenes at different resolutions, as shown in Fig. 3, in addition to the original scene[8]. Next, the coarse resolution search is performed. For each of the $L=6$ model objects and M initial occlusion masks, an SSE-based scan of the scene was performed. Figure 5 shows examples of the scene-to-model SSE varying across the image for various occlusion masks $\mathbf{z}_h$. For purposes of clarity, these plots are inverted such that peaks correspond to regions of low SSE, and therefore high similarity.

Figure 5a shows $E_1^2(i,j)$ using the mask $\mathbf{z}_1^{\mathtt{cat}}$ corresponding to the first mask in Fig. 2 with no occlusion. The minimum SSE occurs at the correct location of the `cat` in the lower right corner of the image (see Fig. 1a.) In Fig. 5b, the mask $\mathbf{z}_6^{\mathtt{cat}}$ (corresponding to the sixth mask in Fig. 2 with the left half of the `cat` occluded) was used to compute $E_6^2(i,j)$, and again the minimum occurred in the lower right corner of the image, except that the peak is sharper and higher than in Fig. 5a. This is to be expected, since the left side of the `cat` is, in fact, occluded in the scene, so $\mathbf{z}_6^{\mathtt{cat}}$ should yield a better `cat` detector than $\mathbf{z}_1^{\mathtt{cat}}$. In Fig. 5c, the mask $\mathbf{z}_7^{\mathtt{cat}}$ (corresponding to the seventh mask in Fig. 2 with the top half occluded) was used, and we see that this grossly incorrect occlusion hypothesis results in an incorrect point of minimum SSE. Finally, Fig. 5d shows the sharp similarity peak that results from using the thirteenth guessed mask $\mathbf{z}_{13}^{\mathtt{juice}}$ associated with the `juice` object (not shown.) This guess corresponds to the bottom right corner of the `juice` being occluded (which happens to be the case for this scene.) Consequently, the `juice` is correctly located near the center of the scene. These plots show that, when a particular occlusion mask $\mathbf{z}_h$ is "correct" (i.e., it closely approximates the actual occlusion $\mathbf{y}$), then the resulting scene-to-model SSE becomes a reliable indicator of object location. Note, however, that incorrect local minima may still exist, which may be incorrectly interpreted as likely object locations, hence the need for the verification stage and the possibility of backtracking.

Following this series of LM coarse searches, the iterative verification stage is performed. For each of the L model objects, the minimizing location hypothesis $(i^*, j^*; h^*) = \arg\min\limits_{(i,j;h)} \mathscr{C}(i,j,h)$ is selected. These L object hypotheses provide the starting point for the coarse-to-fine staged search discussed in Sect. 3.2. Figure 6 shows a graphical representation of these object hypotheses (in terms of their adaptive masks) as the search progresses to full resolution.

Note that in this scene five of the six initial coarse-level object hypotheses were approximately correct. However, the location of the `stapler3` object (displayed as white in Fig. 6) was grossly in error. Consequently, at the first iteration of the verification stage, the objective function $\mathscr{C}(i,j,h)$ of the `stapler3` object hypothesis exceeded a threshold. The search backtracked, and the incorrect hypothesis was replaced with the second-best (and correct) candidate. As the verification stage proceeded through finer resolution levels, the residual ambiguities in object location and scale were resolved during the perturbation procedure. At termination, the algorithm had converged to the correct hypothesis. Figure 7a shows the final result of the scene interpration by displaying the adaptive masks in their final configuration. In addition, Fig. 8 focuses on the evolving configuration of the `cat`'s adaptive mask through the iterations.

4.2 Illustrative example 2

For the second example, Fig. 7b shows the final interpretation result super-imposed over the original scene. In this example, the verification stage performed a significant amount of backtracking prior to achieving a final convergence to the correct interpretation (unlike the first example, in which only a single backtracking event occurred on the first iteration.)

Figure 9 shows a schematic of the backtracking sequence for this example, in which the adaptive mask of the `cat` is superimposed over the scene at each iteration. A total of ten iterations were required before convergence was achieved.

Following the coarse search, the initial hypothesized locations of the `stapler2`, `glue box`, and `juice` objects were correct (within the spatial ambiguities associated with the initial coarse resolution.) However, both the first- and second-best `cat` hypotheses, in terms of minimized $\mathscr{C}(i,j,h)$, were incorrect: the `cat` was initially hypothesized to be present in the upper right corner of the scene. This incorrect hypothesis survived until the finest resolution level before being rejected by the error threshold criterion. The search backtracked to the coarsest level, and the next (again incorrect) `cat` hypothesis survived to the second-finest resolution level before being rejected. Finally, the third (correct) `cat` hypothesis is tried, and results in a correct convergence. This is a good example of how the absence of medium and high spatial frequencies at the coarsest level can lead to mistakes, and hence the need for the verification stage.

4.3 Robustness experiment 1

The purpose of the first robustness experiment was to evaluate the performance of the algorithm over a large set of scenes. An image database containing 50 different occluded scenes was generated by arranging the model objects in random occluded configurations, and against cluttered backgrounds. The average number of model objects present per scene was 3.3.

The scene interpretation algorithm was performed on each of the 50 test scenes. The results are summarized in Table 1. A total of 164 instances of the model objects appeared in the 50 scenes; the algorithm correctly identified and located 154 of them, or 94%. In the case of 3 object instances (2%), the algorithm converged to an incorrect object

the two robustness experiments, and which were statistically generated as discussed in Sect. 3.1.

[8] The coarse-resolution versions of the L model templates (e.g., Fig. 4a) and their associated sets M_l of M initial occlusion masks each (e.g., Fig. 4b) are generated off-line.

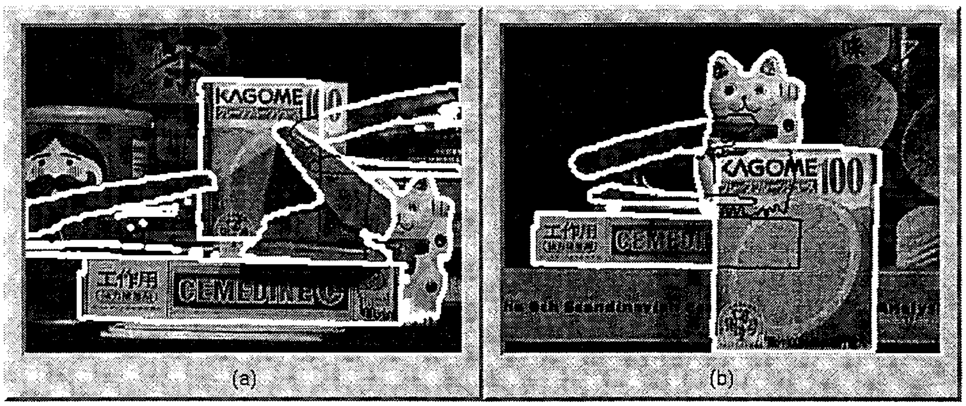


Fig. 7a,b. Scene interpretation following final convergence, with the non-occluded portions of the object hypotheses outlined in white, and the occluded portions outlined in black. **a** Example 1. **b** Example 2

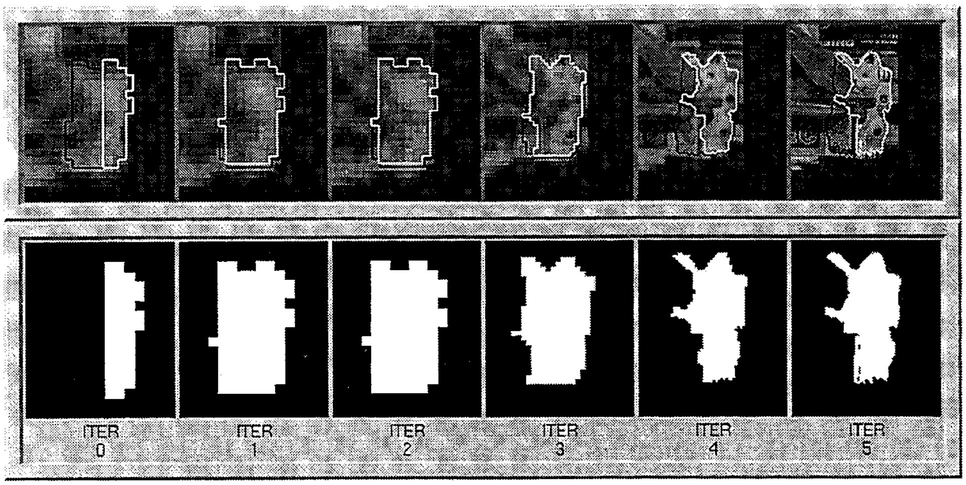


Fig. 8. History of the `cat` hypothesis, with the black and white boundaries indicating the regions of the `cat` hypothesized to be occluded and not-occluded, respectively

hypothesis. In an additional 7 instances (4%), the algorithm rejected all object hypotheses, despite the fact that the object was present in the scene.

In an actual application, the error thresholds would be highly optimized for the particular database of objects. However, these results were obtained with very little threshold optimization. We suspect that, in a real application, performance could be improved further by conducting a thorough study of the optimal threshold settings for this particular object database.

4.4 Robustness experiment 2

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

4.4.1 Variations in scale

In the first part of the experiment, a simple occluded scene was constructed and tested 21 times. For each scene, the objects were moved further from the camera in order to independently investigate algorithm performance subject to scale changes only. Figures 10a and b show the two extremes of scale. In each of the 21 test scenes, the algorithm determined the presence and correct location and scale of the `cat` and `glue box` objects.

4.4.2 Variations in 3D rotation

In the second portion of the experiment, we investigated the robustness of the adaptive-mask approach to variations in viewing direction via 3D object rotations. We were interested in how much 3D rotation could be present before scene interpretation failures occurred.

The same simple configuration from the first portion of the experiment was used to construct ten different scenes in which the 3D rotation of the `cat` varied from −25° to +20° (with respect to the `cat`'s template image.) Figure 10c and d shows the two extremes of this range of rotation. As the `cat` was rotated through these angles, the algorithm performed successfully seven times, and failed three times, with the failures occuring at angles of −25°, −15°, and +20°.

4.4.3 Variations in illumination

In the third portion of the experiment, the tolerance of the algorithm to changes in scene illumination was investigated. Fifteen test scenes were constructed, in which scene illumination was varied over several degrees of freedom by turning on and off various overhead lights, and adjusting the position and intensity of various spotlights. Figure 10e and f shows two such scenes.

The adaptive-mask approach was applied to each of the 15 scenes. In 5 of the scenes, recognition failed, as the algorithm reported that both the `cat` and the `glue box` objects

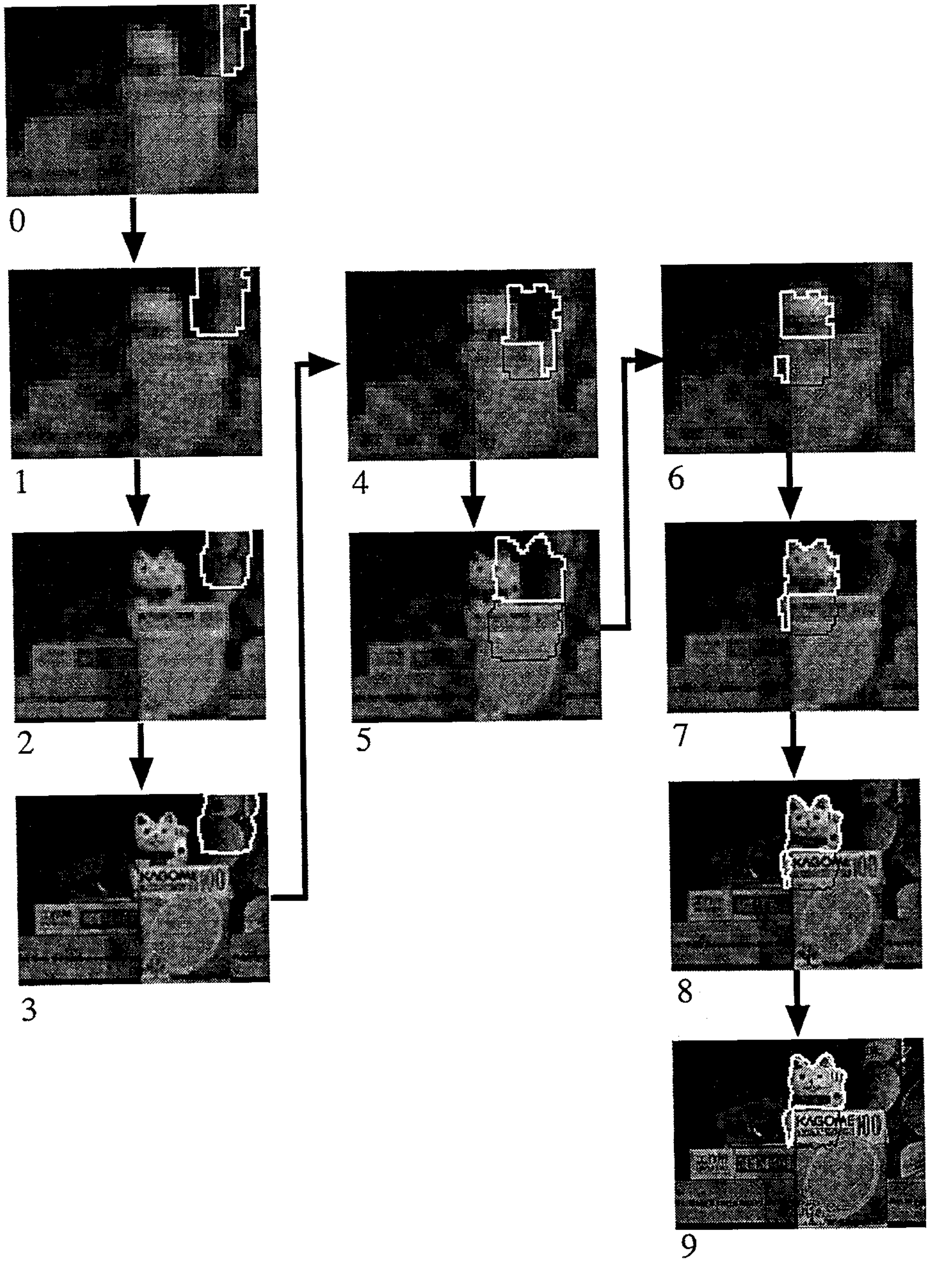


Fig. 9. Illustrative history of cat-backtracking in Example 2

were absent from the scene. In the other 10 scenes, including both scenes presented in Fig. 10e and f, recognition was successful.

It is difficult to show quantitative results from this experiment. All that can be reported is that the algorithm performed reasonably well in the presence of moderate illumination changes, such as those shown in Fig. 10e and f, but failed when the illumination changes became more extreme.

5 Discussion

This paper presents an investigation into the extension of the appearance-matching approach to deal with occluded objects. The core of this extension is the use of an adaptive mask that takes advantage of inter-object occlusion interactions in order to eliminate occluded regions from the scene-to-template SSE computations. Without a priori knowledge of the scene, an initial set of possible occlusion masks must be generated, based on the statistics of the model object geometries. The resulting coarse-resolution squared-error computations yield object location and scale hypotheses. These hypotheses are then verified using a coarse-to-fine staged search.

During this search, the masks are adaptively improved as new information (from higher spatial frequencies and from global occlusion interactions) is taken into account. In addition, a mask may undergo a drastic adaption in the event that an object hypothesis is rejected and the search backtracks.

Most of the strengths of the adaptive-mask approach stem directly from its reliance on scene-to-template squared error. The most compelling strength, shared by all appearance-matching approaches, is the lack of dependence upon object complexity in terms of either shape or surface pat-

Table 1. Robustness results from 50-scene database

	Num. Instances				Percentages		
Object Name	Total	Success	Mislocated	Missing	Success	Mislocated	Missing
glue box	28	28	0	0	100%	0%	0%
stapler1	17	15	1	1	88%	6%	6%
cat	40	34	2	4	85%	5%	10%
juice	43	43	0	0	100%	0%	0%
stapler3	15	14	0	1	93%	0%	7%
stapler2	21	20	0	1	95%	0%	5%
TOTAL	164	154	3	7	94%	2%	4%

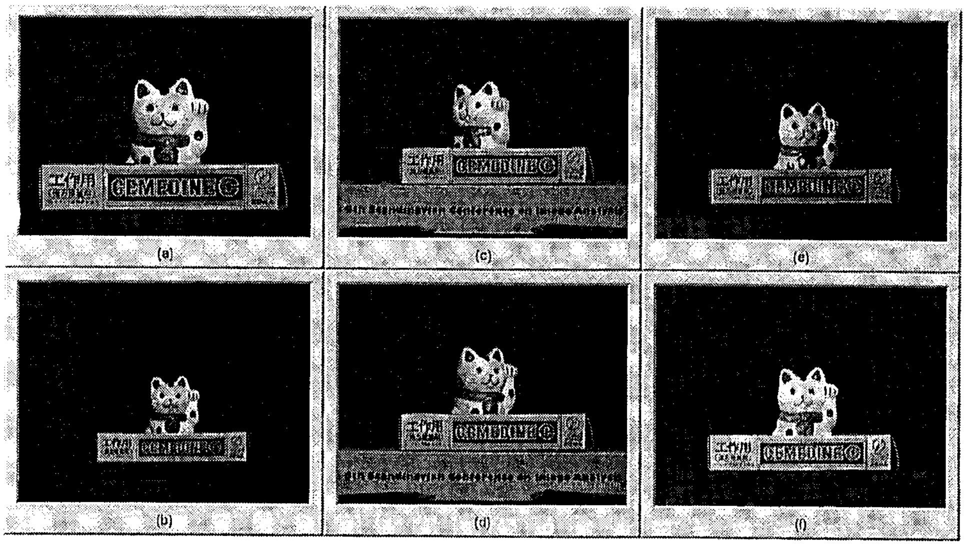


Fig. 10a–f. Test scenes. **a** and **b** Extremes of scale variation. **c** and **d** Extremes of 3D rotation variation. **e** and **f** Extremes of illumination variation, for which successful scene interpretation was achieved

terns. As a result, rigid objects of arbitrary complexity can be handled by this technique, which is a necessary condition for most systems designed to operate outside of a vision laboratory.

The appearance-matching core of this approach also yields a good degree of robustness to image noise and to reasonable amounts of illumination variation, 3D object rotation, and scale variations (e.g., corruption of a small set of scene pixels will have only a small effect on system performance, unlike the situation with many geometric-based methods.) Furthermore, the use of 2D model templates for object representation allows the use of a "teach-by-showing" methodology to build object model databases.

In this paper, the image-spotting problem was restricted to a single canonical viewing direction for each model object in order to study the occlusion issue independently. The next logical step will be to investigate the extension of the adaptive-mask concept to the previous image-spotting work of Murase and Nayar (1995b), in which objects may be recognized over a range of 3D rotations.

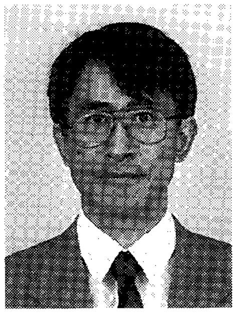
We are also interested in improving upon the SSE (L_2) similarity metric, which is by no means optimal for real images (although it is the easiest to compute.) Recently proposed alternatives to L_2 correlation appear to improve performance for some classes of image (see Boninsegna and Rossi 1994 and Brunelli and Messelodi 1995.) We would like to investigate these and other similarity metrics, such as color histograms, texture measurements, higher order statistical properties, etc., and combinations thereof. By doing so, we hope to extend the domain of appearance-matching techniques into the realm of non-rigid objects.

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

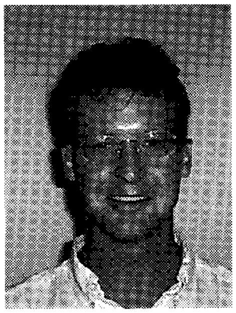
References

1. Anisimov VA, Gorsky ND (1993) Fast hierarchical matching of an arbitrarily oriented template. Pattern Recogn Lett 14: 95–101
2. Ansari N, Delp E (1990) Partial Shape Recognition: A Landmark-Based Approach. IEEE Trans Pattern Anal Mach Intell 12 (5): 470–483
3. Ben-Arie J, Rao KR (1993) On the Recognition of Occluded Shapes and Generic Faces Using Multiple-Template Expansion Matching. In: Proceedings Int. Conf. Computer Vision and Pattern Recognition, New York, N.Y., 15–17 June 1993, pp 214–219
4. Ben-Arie J, Rao R (1994) Optimal template matching by nonorthogonal image expansion using restoration. Mach Vision Appl 7: 69–81
5. Boninsegna M, Rossi M (1994) Similarity measures in computer vision. Pattern Recogn Lett 15: 1255–1260
6. Brunelli R, Messelodi S (1995) Robust Estimation of Correlation with Applications to Computer Vision. Pattern Recogn 28 (6): 833–841
7. Brunelli R, Poggio T (1993) Face Recognition: Features versus Templates. IEEE Trans Pattern Anal Mach Intell 15 (10): 1042–1052
8. Burt PJ (1988) Smart Sensing within a Pyramid Vision Machine. Proceedings of the IEEE 76 (8): 1006–1015

9. Chaudhury S, Acharyya A, Subramanian S, Parthasarathy G (1990) Recognition of Occluded Objects with Heuristic Search. Pattern Recogn 23 (6): 617–635
10. Fukunaga K (1990) Introduction to Statistical Pattern Recognition. Academic Press, Boston, Mass
11. Han M-H, Jang D (1990) The Use of Maximum Curvature Points for the Recognition of Partially Occluded Objects. Pattern Recogn 23 (1): 21–33
12. Liu Z-Q, Caelli TM (1988) Multiobject Pattern Recogn and Detection in Noisy Backgrounds Using a Hierarchical Approach. Comput Vision Graphics Image Process 44: 296–306
13. Margalit A, Rosenfeld A (1990) Using Probabilistic Domain Knowledge to Reduce the Expected Computational Cost of Template Matching. Comput Vision Graphics Image Process 51: 219–534
14. Murase H, Nayar S (1995a) Visual Learning and Recognition of 3D Objects from Appearance. Int J Comput Vision 14 (1): 5–24
15. Murase H, Nayar S (1995b) Image Spotting of 3D Objects using Parametric Eigenspace Representation. In: The 9th Scandinavian Conference on Image Analysis, Uppsala, Sweden, June 1995, pp 323–332
16. Ohba K, Ikeuchi K (1996) Recognition of the Multi Specularity Objects for Bin-Picking Task. In: 1996 IEEE/RJS Int. Conf. on Intelligent Robots and Systems, Osaka, Japan, 4–8 November 1996, pp 1440–1447
17. Ray KS, Majumder DD (1991) Recognition and positioning of partially occluded 3D objects. Pattern Recogn Letter 12: 93–108
18. Rosenfeld A, Vanderbrug GJ (1977) Coarse-Fine Template Matching. IEEE Trans Syst Man Cybern 2: 104–107
19. Salari E, Balaji S (1991) Recognition of Partially Occluded Objects Using B-Spline Representation. Pattern Recogn 24 (7): 653–660
20. Sista S, Bouman C, Allebach J (1995) Fast Image Search Using a Multiscale Stochastic Model. In: Proc. 1995 Int. Conference on Image Processing, Washington, DC, October 1995, Vol 2, pp 225–228
21. Turk MA, Pentland AP (1991) Face Recognition Using Eigenfaces. In: Proc. of IEEE Conf. on Comput Vision and Pattern Recognition, Maui, Hawaii, June 1991, pp 586–591
22. Wiles CR, Forshaw MRB (1993) Recognition of volcanoes on Venus using correlation methods. Image Vision Comput 11 (4): 188–196



Hiroshi Murase received the B.E., M.E., and Ph.D. degrees in electrical engineering from the Unversity of Nagoya, Japan. From 1980 to the present he has been engaged in pattern recognition research at Nippon Telegraph and Telephone Corporation (NTT). From 1992 to 1993 he was a visiting research scientist at Columbia University, New york. He was awarded an IEICEJ Shinohara Award in 1986, and a Telecom System Award in 1992, and the IEEE CVPR best paper award in 1994, and the IEE ICRA best video award in 1996. His research interests include computer vision, video analysis, character recognition, and mul timedia recognition. He is a member of the IEEE, IEICEJ, IPSJ.



Jeff L. Edwards was born in Dallas, Texas, USA on Feb. 3, 1970. He received BSME and MSME degrees from Purdue University in 1992 and 1993. After working in industry, he returned to Purdue and received his MS in Computer Engineering in 1996, and subsequently took a research position at NTT Basic Research Labs in Atsugi-shi, Japan. He is now working for a software start-up company called Electric Planet Interactive in Palo Alto, California.