Computer Vision and Image Understanding **78,** 99–118 (2000)
doi:10.1006/cviu.1999.0830, available online at http://www.idealibrary.com on IDEAL

Robust Recognition Using Eigenimages

Aleš Leonardis

Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25, 1001 Ljubljana, Slovenia
E-mail: alesl@fri.uni-lj.si

and

Horst Bischof

Pattern Recognition and Image Processing Group, Vienna University of Technology,
Favoritenstr. 9/1832, A-1040 Vienna, Austria
E-mail: bis@prip.tuwien.ac.at

Received February 12, 1999; accepted November 5, 1999

The basic limitations of the standard appearance-based matching methods using eigenimages are nonrobust estimation of coefficients and inability to cope with problems related to outliers, occlusions, and varying background. In this paper we present a new approach which successfully solves these problems. The major novelty of our approach lies in the way the coefficients of the eigenimages are determined. Instead of computing the coefficients by a projection of the data onto the eigenimages, we extract them by a robust hypothesize-and-test paradigm using subsets of image points. Competing hypotheses are then subject to a selection procedure based on the Minimum Description Length principle. The approach enables us not only to reject outliers and to deal with occlusions but also to simultaneously use multiple classes of eigenimages. © 2000 Academic Press

Key Words: appearance-based matching, principal component analysis, robust estimation, occlusion, minimum description length.

1. INTRODUCTION AND MOTIVATION

The appearance-based approaches to vision problems have recently received renewed attention in the vision community due to their ability to deal with combined effects of shape, reflectance properties, pose in the scene, and illumination conditions [16, 19]. Besides, the appearance-based representations can be acquired through an automatic learning phase, which is not the case with traditional shape representations. The approach has led to a variety of successful applications, e.g., illumination planning [20], visual positioning and

1077-3142/00 $35.00

tracking of robot manipulators [21], visual inspection [34], "image spotting" [18], and human face recognition [3, 32].

As stressed by its proponents, the major advantage of the approach is that both learning and recognition are performed using just two-dimensional brightness images without any low- or mid-level processing. However, there still remain various problems to be overcome since the technique rests on direct appearance-based matching [19]. The most severe limitation of the method in its standard form is that it cannot handle problems related to occlusion, outliers, and varying background. In other words, the standard approach is not robust, where the term *robustness* refers to the fact that the results remain stable in the presence of various types of noise and can tolerate a certain portion of outliers [11, 29]. One way to characterize the robustness is through the concept of a *breakdown point*, which is determined by the smallest portion of outliers in the data set at which the estimation procedure can produce an arbitrarily wrong estimate. For example, in the standard approach even a single erroneous data point (having an arbitrary value) can cause an arbitrary wrong result, meaning that the breakdown point is 0%.[1]

Different approaches have been proposed in the literature to estimate the coefficients of the eigenspace projections more reliably. Pentland suggested the use of modular eigenspaces [25] to alleviate the problem of occlusion. Ohba and Ikeuchi [24] proposed the eigenwindow method to be able to recognize partially occluded objects. The methods based on "eigenwindows" partially alleviate the problems related to occlusion but do not solve them entirely because the same limitations hold for each of the eigenwindows. Besides, due to local windows, these methods lack the *global* aspect and usually require further processing.

To eliminate the effects of varying background Murase and Nayar [18] introduced the search-window, which is the AND area of the object regions of all images in the training image set. This was further extended to an adaptive mask concept by Edwards and Murase [6]. However, the assumption on which the method has been developed is rather restrictive; namely, a target object can only be occluded by one or more of the other $M - 1$ target objects, rather than occluded by some unknown entity or perturbed by a different background.

On the other hand, Black and Jepson [5] proposed to use a conventional robust M-estimator for calculating the coefficients; i.e., they replaced the standard quadratic error norm with a robust error norm. Their main focus was to show that appearance-based methods can be used for tracking. The approach of Black and Jepson is less robust than our resampling approach. In addition, only the dominant structure is identified. If more than one object is present in the scene, the same procedure has to be reapplied to the outliers. In contrast, our method has a mechanism that can simultaneously deal with multiple objects. Rao [27] introduced a robust hierarchical form of the MDL-based Kalman filter estimators that can tolerate significant occlusion and clutter. The limitations of this approach are similar to those of the approach of Black and Jepson. Again, the critical steps are the initialization and simultaneous recovery of occluding objects.

In this paper we present a novel approach, which extends our previous work [13], that successfully solves the problems related to occlusion, cluttered background, and outliers. The major novelty of our approach lies in the way the coefficients of the eigenimages are determined. Instead of computing the coefficients by a projection of the data onto the

[1] Even when the outlier cannot attain an arbitrary value, which is usually the case in practical applications where the values come from a bounded interval, the bias, although finite, can completely destroy the estimate [10].

eigenimages, we apply random sampling and robust estimation to generate hypotheses for the model coefficients. Competing hypotheses are then subjected to a selection procedure based on the Minimum Description Length (MDL) principle. The approach enables us not only to reject outliers and to deal with occlusions but also to simultaneously use multiple classes of eigenimages.

Our robust approach extends the domain of applicability of the appearance-based methods so that they can be used in more complex scenes, i.e., scenes that contain occlusion, background clutter, and outliers. While this is a significant step forward in the area of appearance-based recognition, some problems pertinent to appearance-based matching still remain. In particular, the objects in the input images should match in scale those that are modeled in the eigenspace. Also, the translation and plane-rotation invariance is achieved by initiating the hypotheses exhaustively at regularly spaced points and at various orientations. We have recently proposed a multiresolution approach that can efficiently cope with these problems [4]; however, its description is outside the scope of this paper.

The paper is organized as follows: We first review the basic concepts of the traditional appearance-based matching methods and point out their main limitations. In Section 3 we outline our proposed approach and detail its basic components. In Section 4 we present the results on complex image data using the standard image database from Columbia University [22]. We conclude with a discussion and outline the work in progress.

2. APPEARANCE-BASED MATCHING

The appearance-based methods consist of two stages. In the first, off-line (training) stage a set of images (templates), i.e., training samples, is obtained. These images usually encompass the appearance of a single object under different orientations [34] or different illumination directions [20] or multiple instances of a class of objects, e.g., faces [32]. In many cases, images not in the training set can be interpolated from the training views [26, 33]. The sets of images are usually highly correlated. Thus, they can efficiently be compressed using principal component analysis (PCA) [1], resulting in a low-dimensional eigenspace.

In the second, on-line (recognition) stage, given an input image, the recognition system projects parts of the input image (i.e., subimages of the same size as training images) to the eigenspace. In the absence of specific cues, e.g., when motion can be used to presegment the image, the process is sequentially applied to the entire image. The recovered coefficients indicate the particular instance of an object and/or its position, illumination, etc.

We now introduce the notation. Let $\mathbf{y} = [y_1, \ldots, y_m]^{\mathrm{T}} \in \mathbb{R}^m$ be an individual template, and let $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ be a set of templates; throughout the paper a simple vector notation is used since the extension to 2-D is straightforward. To simplify the notation we assume $\mathcal{Y}$ to be normalized, having zero mean. Let $\mathbf{Q}$ be the covariance matrix of the vectors in $\mathcal{Y}$; we denote the eigenvectors of $\mathbf{Q}$ by $\mathbf{e}_i$, and the corresponding eigenvalues by λ_i. We assume that the number of templates n is much smaller than the number of elements m in each template; thus an efficient algorithm based on SVD can be used to calculate the first n eigenvectors [19]. Since the eigenvectors form an orthogonal basis system, $\langle \mathbf{e}_i, \mathbf{e}_j \rangle = 1$ when $i = j$ and 0 otherwise, where $\langle\,\rangle$ stands for a scalar product. We assume that the eigenvectors are in descending order with respect to the corresponding eigenvalues λ_i. Then, depending on the correlation among the templates in $\mathcal{Y}$, only p, $p < n$, eigenvectors are needed to represent the $\mathbf{y}_i$ to a sufficient degree of accuracy as a linear combination of eigenvectors $\mathbf{e}_i$,

$$\tilde{\mathbf{y}} = \sum_{i=1}^{p} a_i(\mathbf{y})\mathbf{e}_i. \tag{1}$$

The error we make by this approximation is $\sum_{i=p+1}^{n} \lambda_i$ and can be calculated by $\|\mathbf{y}\|^2 - \sum_{i=1}^{p} a_i^2$ [17]. We call the space spanned by the first p eigenvectors the *eigenspace*.

To recover the parameters a_i during the matching stage, a data vector $\mathbf{x}$ is projected onto the eigenspace,

$$a_i(\mathbf{x}) = \langle \mathbf{x}, \mathbf{e}_i \rangle = \sum_{j=1}^{m} x_j e_{i,j} \quad 1 \le i \le p. \tag{2}$$

$\mathbf{a}(\mathbf{x}) = [a_1(\mathbf{x}), \ldots, a_p(\mathbf{x})]^{\mathrm{T}}$ is the point in the eigenspace obtained by projecting $\mathbf{x}$ onto the eigenspace. Let us call the $a_i(\mathbf{x})$ coefficients of $\mathbf{x}$. The reconstructed data vector $\tilde{\mathbf{x}}$ can be written as

$$\tilde{\mathbf{x}} = \sum_{i=1}^{p} a_i(\mathbf{x})\mathbf{e}_i. \tag{3}$$

It is well known that PCA is among all linear transformations the one which is optimal with respect to the reconstruction error $\|\mathbf{x} - \tilde{\mathbf{x}}\|^2$.

2.1. Weaknesses of Standard Appearance-Based Matching

In this section we analyze some of the basic limitations of the standard appearance-based matching methods and illustrate the effect of occlusion with an example. Namely, the way the coefficients a_i are calculated poses a serious problem in the case of outliers and occlusions.

Suppose that $\hat{\mathbf{x}} = [x_1, \ldots, x_r, 0, \ldots, 0]^{\mathrm{T}}$ is obtained by setting the last $m - r$ components of $\mathbf{x}$ to zero; a similar analysis holds when some of the components of $\mathbf{x}$ are set to some other values, which, for example, happens in the case of occlusion by another object. Then

$$\hat{a}_i = \hat{\mathbf{x}}^{\mathrm{T}}\mathbf{e}_i = \sum_{j=1}^{r} x_j e_{i,j}. \tag{4}$$

The error we make in calculating a_i is

$$(a_i(\mathbf{x}) - \hat{a}_i(\hat{\mathbf{x}})) = \sum_{j=r+1}^{m} x_j e_{i,j}. \tag{5}$$

It follows that the reconstruction error is

$$\left\| \sum_{i=1}^{p} \left(\sum_{j=r+1}^{m} x_j e_{i,j} \right) \mathbf{e}_i \right\|^2. \tag{6}$$

Due to the nonrobustness of linear processing, this error affects the whole vector $\tilde{\mathbf{x}}$. Figure 1 depicts the effect of occlusion on the reconstructed image. A similar analysis holds

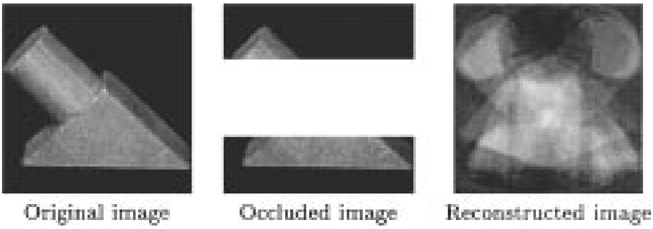


FIG. 1. Demonstration of the effect of occlusion using the standard approach for calculating the coefficients a_i.

for the case of outliers (occlusions are just a special case of spatially coherent outliers). We can show that the coefficient error can get arbitrarily large by just changing a single component of **x**, which proves that the method is nonrobust with a breakdown point of 0%.

Since in calculating the eigenimages there is no distinction made between the object and the background (which is usually assumed to be black), the effect of a varying background is, in the recognition phase, similar to that of occlusion. Therefore, to obtain the correct result in the case of the standard method, the object of interest should be first segmented from the background and then augmented with the original background. Our robust approach does not require this segmentation step and can thus cope with objects that appear on various backgrounds.

The problems that we have discussed arise because the complete set of data **x** is required to calculate a_i in a least square fashion (Eq. (2)). Therefore, the method is sensitive to partial occlusions, to data containing noise and outliers, and to changing backgrounds.

In the next section we explain our new approach which has been designed to overcome precisely these types of problems.

3. OUR APPROACH

The major novelty of our approach lies in the way the coefficients of the eigenimages are determined. Instead of computing the coefficients by a projection of the data onto the eigenimages (which is equivalent to determining the coefficients in a least-squares manner), we achieve robustness by employing *subsampling*. This is the very principle of high breakdown point estimation such as Least Median of Squares [29] and RANSAC [7]. In particular, our approach consists of a twofold robust procedure: We determine the coefficients of the eigenspace projection by a robust hypothesize-and-test paradigm using only *subsets* of image points. Each hypothesis (based on a random selection of points) is generated by the robust solution of a set of linear equations (similarly to α-trimmed estimators [29]). Competing hypotheses are then selected according to the Minimum Description Length principle.

In the following we detail the steps of our algorithm. For clarity of the presentation, we assume that we are dealing with a single eigenspace at a specific location in the image. At the end of the section we describe how this approach can be extended to multiple eigenspaces for different objects.

3.1. Generating Hypotheses

Let us first start with a simple observation. If we take into account all eigenvectors, i.e., $p = n$, and if there is no noise in the data x_{r_i}, then in order to calculate the coefficients a_i

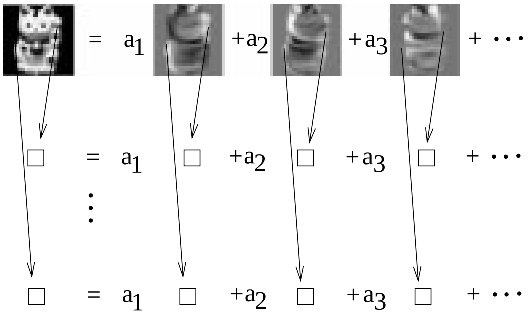


FIG. 2. Illustration of using linear equations to calculate the coefficients of eigenimages.

(Eq. (2)) we need only n points $\mathbf{r} = (r_1, \ldots, r_n)$. Namely, the coefficients a_i can simply be determined by solving the following system of linear equations (see Fig. 2):

$$x_{r_i} = \sum_{j=1}^{n} a_j(\mathbf{x}) e_{j,r_i} \quad 1 \leq i \leq n. \tag{7}$$

However, if we approximate each template only by a linear combination of a subset of eigenimages, i.e., $p < n$, and there is also noise present in the data, then Eq. (7) can no longer be used, but rather we have to solve an over-constrained system of equations in the least-squares sense using k data points ($p < k \leq m$). In most cases, $k \ll m$, since the number of pixels is usually three orders of magnitude larger than the number of eigenimages. Thus we seek the solution vector $\mathbf{a}$ which minimizes

$$E(\mathbf{r}) = \sum_{i=1}^{k} \left(x_{r_i} - \sum_{j=1}^{p} a_j(\mathbf{x}) e_{j,r_i} \right)^2 . \tag{8}$$

Of course, the minimization of Eq. (8) can only produce correct values for coefficient vector $\mathbf{a}$, if the set of points r_i does not contain outliers, i.e, not only extreme noisy points but also points belonging to different backgrounds or some other templates due to occlusion. Therefore, the solution has to be sought in a robust manner.

The following robust procedure has been applied to solve Eq. (8). Starting from the randomly selected k points $r_1, \ldots, r_k$, we seek the solution vector $\mathbf{a} \in \mathbb{R}^p$ which minimizes Eq. (8) in a least-squares manner. Then, based on the error distribution of the set of points, we keep reducing their number by a factor of α (i.e., those points with the largest error) and solve Eq. (8) again with this reduced set of points.

In order to analyze the robustness of this procedure with respect to the parameters and noise conditions, we have performed several Monte Carlo studies (see Appendix A). It turns out that our robust procedure can tolerate up to 50% outliers. To increase the quality of the estimated coefficients we can exploit the fact that the coefficients that represent a point in the eigenspace are not arbitrarily spread in the eigenspace but are either discrete points close to the training samples in the eigenspace or points on a parametric manifold [19]. Therefore, we determine from the estimated coefficients the closest point in the eigenspace or on the parametric manifold, respectively, which gives us the coefficients of the closest training image (or an interpolation between several of these coefficients). For the search we use the exhaustive search procedure implemented in SLAM [23].

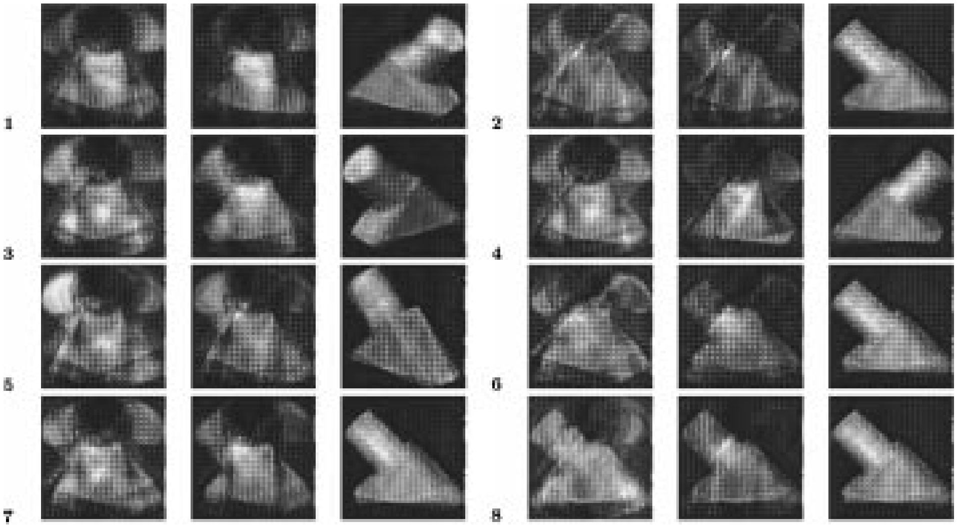


FIG. 3. Some hypotheses generated by the robust method for the occluded image in Fig. 1b with 15 eigenimages; for each hypothesis (1–8), from left to right: reconstructed image based on the initial set of points, reconstruction after reduction of 25% of points with the largest residual error, and the reconstructed image based on the parameters of the closest point on the parametric manifold.

The obtained coefficients a_i are then used to create a hypothesis $\tilde{\mathbf{x}} = \sum_{i=1}^{p} a_i \mathbf{e}_i$, which is evaluated both from the point of view of the error $\boldsymbol{\xi} = (\mathbf{x} - \tilde{\mathbf{x}})$ and from the number of compatible points. For good matches (i.e., objects from the training set) we expect an error of $\frac{1}{m} \sum_{i=p+1}^{n} \lambda_i$ on the average (see Section 2). Therefore we can set the error margin for the compatible points to $\Theta = \frac{2}{m} \sum_{i=p+1}^{n} \lambda_i$ (the factor of 2 is used because the λ_i are calculated from the training set only, and we deal also with objects not included in the training set). A hypothesis is acceptable if the cardinality of the set of compatible points is above the acceptance threshold. However, unlike with RANSAC [7], this condition can really be kept minimal since the selection procedure (cf. Section 3.2) will reject the false positives. The accepted hypothesis is characterized by the coefficient vector $\mathbf{a}$, the error vector $\boldsymbol{\xi}$, and the domain of the compatible points $D = \{ j \mid \xi_j^2 < \Theta\}$, $s = |D|$.

Figure 3 depicts some of the generated hypotheses for the occluded image in Fig. 1. One can see that four out of eight hypotheses are close to the correct solution. Figure 4

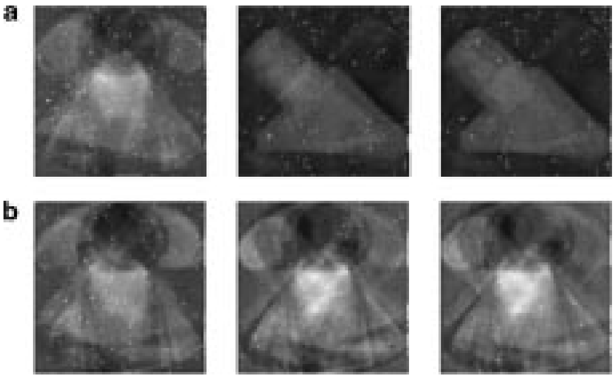


FIG. 4. Points that were used for calculating the coefficient vector (overlaid over the reconstructed image); (a) a good hypothesis, (b) a bad hypothesis.

depicts the points (overlaid over the reconstructed image) that were used to calculating the coefficient vector in the case of a good (a) and bad (b) hypothesis. One can observe that for the bad hypothesis not all the points on the occluded region were eliminated.

However, as depicted in Fig. 3, one cannot expect that every initial randomly chosen set of points will produce a good hypothesis if there is one, despite the robust procedure. Thus, to further increase the robustness of the hypotheses generation step, i.e., increase the probability of detecting a correct hypothesis if there is one, we initiate, as in [2, 7], a number of trials. In Appendix B, we calculate, based on combinatorial arguments, the number of hypotheses that we need to generate for a given noise level.

3.2. Selection

The set of hypotheses, described by the coefficient vectors $\mathbf{a}_i$, the error vectors $\boldsymbol{\xi}_i$, and the domains of the compatible points $D_i = \{j \mid \boldsymbol{\xi}_j^2 < \Theta\}$, $s_i = |D_i|$, which has been generated is usually highly redundant. Thus, the selection procedure has to select a subset of “good” hypotheses and reject the superfluous ones. The core of the problem is how to define optimality criteria for a set of hypotheses. Intuitively, this reduction in the complexity of a representation coincides with a general notion of simplicity [12]. The simplicity principle has been formalized in the framework of information theory. Shannon [30] revealed the relation between probability theory and the shortest encoding. Further formalization of this principle in information theory led to the principle of *Minimum Description Length* (MDL) [28]. A derivation (see [14, 15]) leads to the minimization of an objective function encompassing the information about the competing hypotheses [14].

The objective function has the following form:

$$F(\mathbf{h}) = \mathbf{h}^{\mathrm{T}}\mathbf{C}\mathbf{h} = \mathbf{h}^{\mathrm{T}} \begin{bmatrix} c_{11} & \dots & c_{1R} \\ \vdots & & \vdots \\ c_{R1} & \dots & c_{RR} \end{bmatrix} \mathbf{h}. \tag{9}$$

Vector $\mathbf{h}^{\mathrm{T}} = [h_1, h_2, \dots, h_R]$ denotes a set of hypotheses, where h_i is a *presence-variable* having the value 1 for the presence and 0 for the absence of the hypothesis i in the resulting description. The diagonal terms of the matrix $\mathbf{C}$ express the cost–benefit value for a particular hypothesis i,

$$c_{ii} = \mathrm{K}_1 s_i - \mathrm{K}_2 \|\boldsymbol{\xi}_i\|_{D_i} - \mathrm{K}_3 N_i, \tag{10}$$

where s_i is the number of compatible points, $\|\boldsymbol{\xi}_i\|_{D_i}$ is the error over the domain D_i, and N_i is the number of coefficients (eigenvectors).

The off-diagonal terms handle the interaction between the overlapping hypotheses,

$$\begin{aligned} c_{ij} &= \frac{-\mathrm{K}_1 |D_i \cap D_j| + \mathrm{K}_2 \xi_{ij}}{2}, \\ \xi_{ij}^2 &= \max\left(\sum_{D_i \cap D_j} \boldsymbol{\xi}_i^2, \sum_{D_i \cap D_j} \boldsymbol{\xi}_j^2 \right), \end{aligned} \tag{11}$$

where D_i denotes the domain of the ith hypothesis and $\sum_{D_i \cap D_j} \boldsymbol{\xi}_i^2$ denotes the sum of squared errors of the ith hypothesis over the intersection of the two domains D_i, D_j.

Equation (9) supports our intuitive thinking that an encoding is efficient if

- the number of pixels that a hypothesis encompasses is large,
- the deviations between the data and the approximation are low,
- while at the same time the number of hypotheses is minimized.

The coefficients K_1, K_2, and K_3, which can be determined automatically [14], adjust the contribution of the individual terms. These parameters are weights which can be determined on a purely information-theoretic basis (in terms of bits) or can be adjusted in order to express the preference for a particular type of description. In general, K_1 is the average number of bits which are needed to encode an image when it is not encoded by the eigenspace, K_2 is related to the average number of bits needed to encode a residual value of the eigenspace approximation, and K_3 is the average cost of encoding a coefficient of the eigenspace. Due to the nature of the problem, i.e., finding the maximum of the objective function, only the relative ratios between the coefficients play a role, e.g., K_2/K_1 and K_3/K_1.

The objective function takes into account the interaction between different hypotheses which may be completely or partially overlapped. However, we consider only the pairwise overlaps in the final solution. From the computational point of view, it is important to notice that the matrix **C** is symmetric, and depending on the overlap, it can be sparse or banded. All these properties of the matrix **C** can be used to reduce the computations needed to calculate the value of $F(\mathbf{h})$.

We have formulated the problem of selection in such a way that its solution corresponds to the global extremum of the objective function. Maximization of the objective function $F(\mathbf{h})$ belongs to the class of combinatorial optimization problems (quadratic Boolean problem). Since the number of possible solutions increases exponentially with the size of the problem, it is usually not tractable to explore them exhaustively. Thus the exact solution has to be sacrificed to obtain a practical one. We are currently using two different methods for optimization. One is a simple greedy algorithm and the other one is Tabu search [8, 31]. In our experiments we did not notice much of a difference between the results produced by the two optimization methods, therefore we report here only the results obtained by the computationally simpler greedy method.

One should note that the selection mechanism can be considerably simplified when we know that there is a single object in the image or multiple nonoverlapping objects. In these cases only the diagonal terms need to be considered. However, for multiple overlapping objects the optimization function has to be used in its full generality (see Fig. 9).

3.3. *Complete Algorithm*

The complete algorithm is outlined in Fig. 5. The left side depicts the training (off-line) stage. The input is a set of training images for each object. The output consists of the eigenimages and the coefficients of training images, or alternatively, of the parametric eigenspaces. The right side of Fig. 5 depicts the recognition (on-line) stage. As input, it receives the output of the training stage (eigenspaces and coefficients for each object) and an image in which instances of training objects are to be recognized. At each location in the image, several hypotheses are generated for each eigenspace. The selection procedure then reasons among different hypotheses, possibly belonging to different objects, and selects those that better explain the data, thereby delivering automatically the number of objects, the

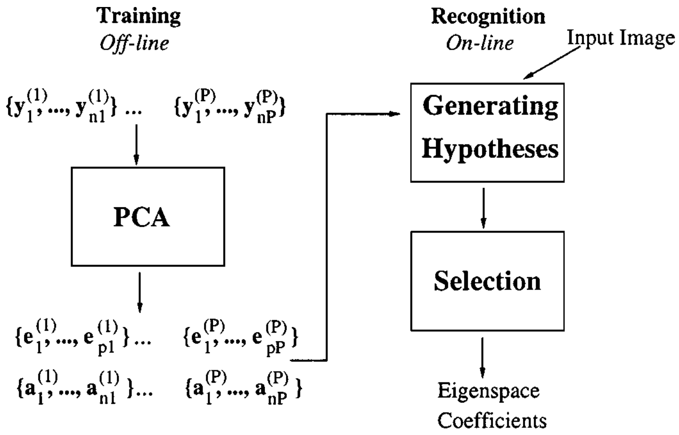

Training
Off-line
$\{y_1^{(1)},\ldots,y_{n1}^{(1)}\}\ \ldots\ \{y_1^{(P)},\ldots,y_{nP}^{(P)}\}$
PCA
$\{e_1^{(1)},\ldots,e_{p1}^{(1)}\}\ldots\ \{e_1^{(P)},\ldots,e_{pP}^{(P)}\}$
$\{a_1^{(1)},\ldots,a_{n1}^{(1)}\}\ldots\ \{a_1^{(P)},\ldots,a_{nP}^{(P)}\}$
Recognition
On-line
Input Image
Generating Hypotheses
Selection
Eigenspace Coefficients

FIG. 5. A schematic diagram outlining the complete robust algorithm.

eigenspaces they belong to, and the coefficients (via the nearest neighbor search performed already at the hypotheses generation step).

4. EXPERIMENTAL RESULTS

In this section we first present several single experiments to demonstrate the utility of our robust method. In the next section we report on extensive testing that we performed to compare the standard method with the robust method. We performed all experiments on the standard set of images (Columbia Object Image Library, COIL-20) [22]. Figure 6 shows some of the 20 objects in the COIL-20. Each object is represented in the database by 72 images obtained by the rotation of the object through 360° in 5° steps (1440 images in total). Each object is represented in a separate eigenspace, and the coefficients of the eigenspace specify the orientation of the object via nearest neighbor search on the parametric manifold.

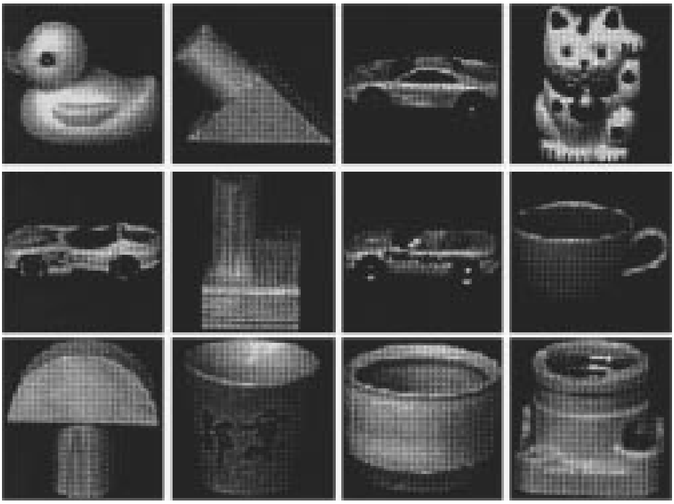


FIG. 6. Some of the test objects used in the experiments.

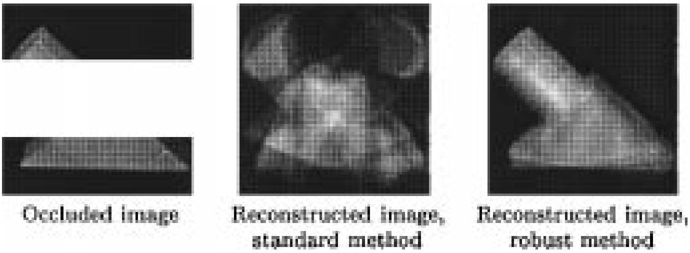

Occluded image
Reconstructed image, standard method
Reconstructed image, robust method

FIG. 7. Demonstration of insensitivity to occlusions using the robust methods for calculating the coefficients a_i.

Unless stated otherwise, all the experiments are performed with the following parameter setting:

Number of eigenimages p	15
Number of initial hypotheses H	10
Number of initial points k	$12p = 180$
Reduction factor α	0.25
K_1	1
K_2	0.1
K_3	50
Compatibility threshold Θ	100

Figure 7 demonstrates that our approach is insensitive to occlusions. One can see that the robust method outperforms the standard method considerably. Note that the blur visible in the reconstruction is the consequence of taking into account only a limited number of eigenimages.

Figure 8 shows several objects on a considerably cluttered background (image size 540 × 256). At every second pixel 10 hypotheses per eigenspace where initiated. All the objects have been correctly recovered by the robust method. The run-time for the robust method on this image is approximately 3.8 h for our nonoptimized MatLab implementation on a Pentium II-450 PC. However, this execution time can be reduced to approximately 10 min with a hierarchical implementation and an intelligent search technique [4].

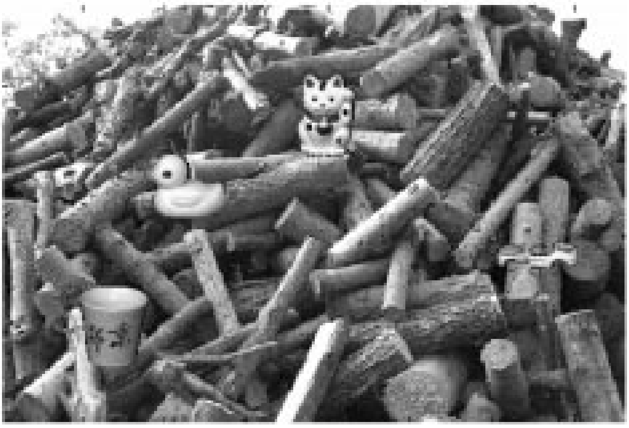


FIG. 8. Test-objects on a cluttered background.

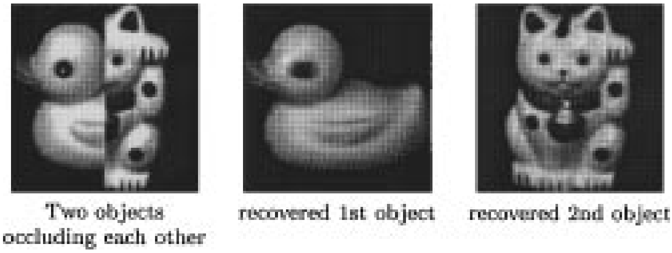


FIG. 9. Two objects occluding each other.

Figure 9 demonstrates that our approach can cope with situations where one object occludes another. One can see that the robust method is able to recover both objects. One should note that in this case the selection mechanism based on the MDL principle delivers automatically that there are two objects present in the scene (i.e., we do not need to specify the expected number of objects in advance).

4.1. Comparison of the Methods

In this section we report on the extensive testings that we performed to compare the standard and the robust method. We performed two types of experiments:

- orientation estimation and
- classification.

In the first experiment, we compared the methods also on the level of estimated coefficients. Namely, a comparison on the level of recognition may be misleading since sometimes the recovered coefficients are fairly poor (especially with the standard method), but still the object is correctly recognized.

We have performed tests with Gaussian, Salt & Pepper, and Replacement noise. Due to the lack of space and the qualitative similarity between the Salt & Pepper and Replacement noise, we only report on the tougher Salt & Pepper noise (5–80%) (Fig. 10). In addition we report on the experiments with occlusions (10–60%), see also Fig. 10.

4.1.1. Orientation estimation. For this set of experiments we constructed the eigenspace of a single object with images from 36 orientations (each 10° apart) and used the remaining views to test the generalization ability of the methods. In addition, we varied the number of eigenvectors and the number of hypotheses. As a performance measure we used the median of the normalized coefficient error. The following plots (Fig. 11) show the typical results of

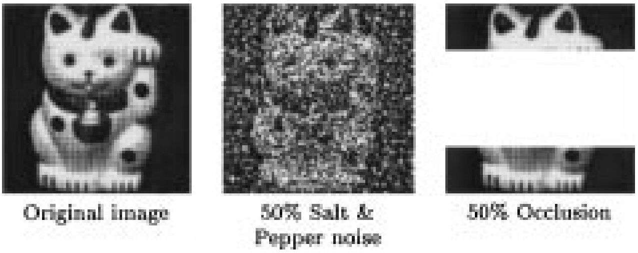
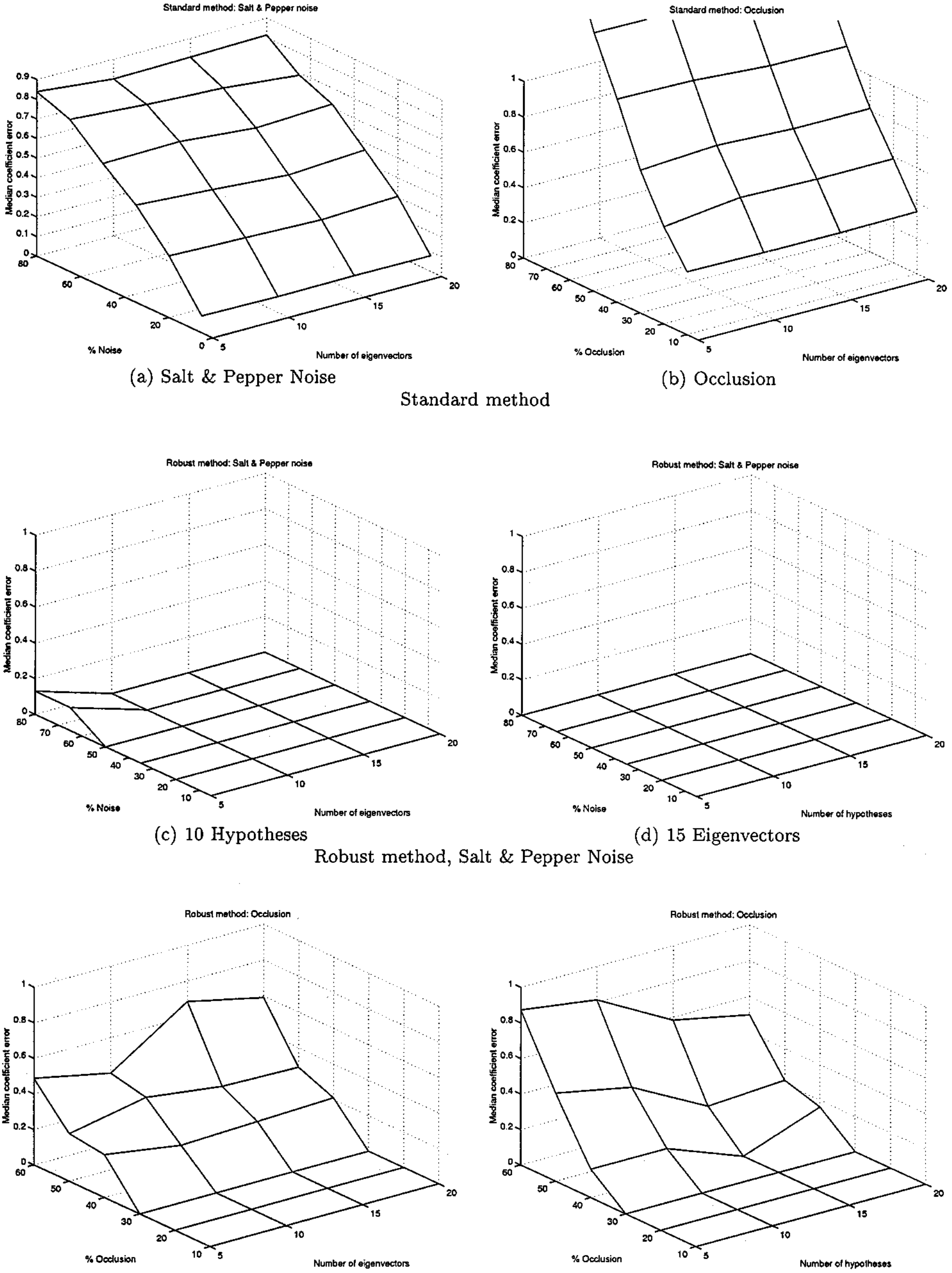


FIG. 10. Test image subjected to Salt & Pepper noise and occlusion.



(a) Salt & Pepper Noise

(b) Occlusion

Standard method

(c) 10 Hypotheses

(d) 15 Eigenvectors

Robust method, Salt & Pepper Noise

(e) 10 Hypotheses

(f) 15 Eigenvectors

Robust method, Occlusion

FIG. 11. Orientation estimation results.

the standard and the robust method obtained for the test set of one object under the various noise conditions.

Figures 11c and 11d show that the robust method is insensitive to the Salt & Pepper noise up to 80%. For the standard method one observes a linear increase of the error with the level of noise, Fig. 11a.

TABLE 1
Summary of Orientation Estimation Experiments (Median of the Absolute Orientation Error in Degrees)

	Salt & Pepper [%]						Occlusions [%]					
Method	5	20	35	50	65	80	10	20	30	40	50	60
Standard	0	0	0	5	10	10	0	0	5	95	90	90
Robust	0	0	0	0	0	0	0	0	0	0	5	40

Figures 11e and 11f show the results for occluded objects, which again demonstrate the superiority of the robust method. One can see that the robust method is insensitive to occlusions of up to 30%; then its performance begins to deteriorate.The standard method is very sensitive to occlusions; i.e., even for 10% occlusions the errors are already quite high, Fig. 11b.

We also performed experiments with Gaussian noise (not depicted here). As expected, the standard method (which is the optimal method for the Gaussian noise) produced best results. However, the results produced by the robust method were almost as good (only a slight increase in error).

The experiments have shown that the robust method outperforms the standard method considerably when dealing with outliers and occlusion. Table 1 summarizes the results for several objects. To generate the table, we used 15 eigenimages and generated 10 hypotheses. In this case, the error measure is the median of the absolute orientation error given in degrees.Comparing this table to the plots, one can observe that the recovered coefficients might be fairly poor (especially with the standard method), but still the pose is correctly estimated due to the high dimensionality and sparseness of the points in the eigenspace.

4.1.2. Classification. For this set of experiments we used all 20 objects. As a training set we used, similarly to the previous experiment, images from 36 orientations of each object. For the standard method, we calculated the universal eigenspace [19] with all training images, and when the object was correctly recognized, we used the object's eigenspace to determine the orientation. For the robust method, we used only the objects' eigenspaces, and in the hypothesis generation stage, we generated for each object eight hypotheses. For all eigenspaces, 15 eigenvectors were used. As for the error measure, we used the classification accuracy, and for those objects which have been correctly recognized, we also calculated the median absolute error in the orientation.

Table 2 shows the results for 50% Salt and Pepper noise, and Table 3 shows the results for 50% occlusions.

TABLE 2
Classification Results on Images with 50% Salt & Pepper Noise

Method	Recognition rate	Median absolute orientation error
Standard	46%	50°
Robust	75%	5°

TABLE 3
Classification Results on Images with 50% Occlusion

Method	Recognition rate	Median absolute orientation error
Standard	12%	50°
Robust	66%	5°

These results clearly indicate the superiority of the robust method over the standard method. The higher error rates for the occlusion can be explained by the fact that certain objects are already completely occluded in some orientations. In Figs. 12 and 13 we have depicted those objects that caused the highest error, in either orientation estimation or classification.

In summary, these experiments demonstrate that our robust method can tolerate considerable amount of noise, can cope with occluded multiple objects on varying background and is therefore much wider applicable than the standard method.

5. DISCUSSION AND CONCLUSIONS

In this paper we have presented a novel robust approach which enables the appearance-based matching techniques to successfully cope with outliers, cluttered background, and occlusions. The robust approach exploits several techniques, e.g., robust estimation and the hypothesize-and-test paradigm, which combined together in a general framework achieve the goal. We have presented an experimental comparison of the robust method and the standard one on a *standard database* of 1440 images. We identified the “breaking points” of different methods and demonstrated the superior performance of our robust method. We have shown experimentally (by Monte Carlo simulation) the influence of the adjustable parameters on the breakdown point (Appendix A). A general conclusion drawn from these experiments is as follows: The robust method can tolerate much higher levels of noise than the standard parametric eigenspace method under reasonable computational cost. In terms of speed the standard method is approximately seven times faster than the robust method. However, in the case when we have “well-behaved” noise with low variance in the images, we can show that the robust method is approximately 20 times faster than the standard method. This is because, with well-behaved noise, we do not need to explore many hypotheses and we do not need to perform the selection and back-projection of the coefficients. Also, the number of initial points (see Eq. (8)) can be significantly reduced. What should be emphasized is that in this case the robust method is independent of the size of the matching image.

It is interesting to note that the basic steps of the proposed algorithm are the same as in ExSel++ [31], which deals with robust extraction of *analytical* parametric functions from

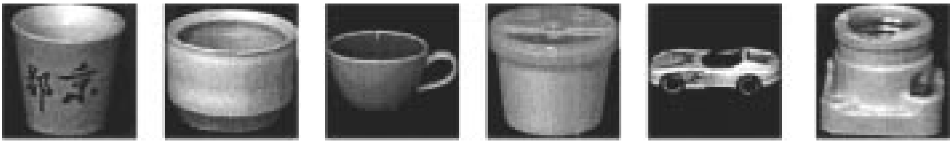


FIG. 12. Gross errors in the orientation determination are mainly caused by rotation-symmetric objects.

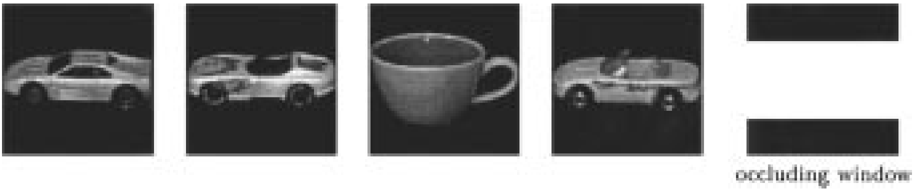


FIG. 13. Gross errors in the classification are mainly caused by objects that are hardly visible under occlusion.

various types of data. Therefore, the method described in this paper can also be seen as an extension of ExSel++ to *learnable classes* of parametric models.

The applications of the proposed method are numerous. Basically everything that can be performed with the classical appearance-based methods can also be achieved within the framework of our approach, only more robustly and on more complex scenes.

The proposed robust approach is a step forward; however, some problems still remain. In particular, the method is sensitive to scale, and the application of the method on a large image is computationally very demanding. In [4], we have recently demonstrated how the robust methods can be applied to convolved and subsampled images yielding the same values of the coefficients. This enables an efficient multi-resolution approach, where the values of the coefficients can directly be propagated through the scales.This property is used to extend our robust method to the problem of scaled and translated images.

APPENDIX A

Robust Fitting

The goal of this Appendix is to test the robustness of the solution of Eq. (A.1) and to determine a suitable range of adjustable parameters. Starting from k points $r_1 \ldots r_k$ we seek the solution vector $\mathbf{a} \in \mathbb{R}^p$ which minimizes

$$E(\mathbf{r}) = \sum_{i=1}^{k} \left(x_{r_i} - \sum_{j=1}^{p} a_j(\mathbf{x}) e_{j,r_i} \right)^2 . \tag{A.1}$$

Based on the error distribution of the set of points, we keep reducing their number by a factor of α until all points are either within the compatibility threshold Θ, or the number of points is smaller than ω.

Since it is hard to derive an analytic expression for this highly nonlinear problem, we tested the robustness using a simulated Monte Carlo approach. The procedure was as follows: We generated eigenimages from a set of test images, and p eigenimages were used for projection. The true coefficients were determined by projection of the test images on the eigenspace. Then for various parameter settings and different levels of replacement noise (i.e., a point r_i is selected at random, and its value is replaced by a uniform random number between 0 and 255), the coefficients were determined and the distance between the recovered and the true coefficients was plotted. Figure A1 shows a plot where each point is an average over 30 trails for the parameter setting $p = 16$, $k = 12p$, $\alpha = 0.25$, $\omega = 3p$. One can see that more than 50% of noise can be tolerated by our method. The reason for this is that the noise amplitude deviations are limited to $[0, \ldots, 255]$, which is the case in images quantized

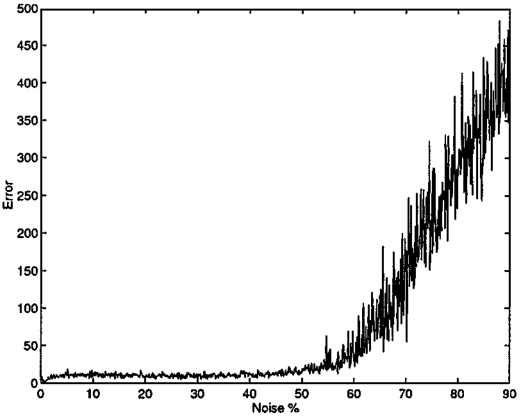


FIG. A1. Monte Carlo simulation showing the robustness of solving the equations.

to eight bits. Monte Carlo simulations have been performed with the following ranges of parameter values: $k \in \{4p \ldots 20p\}$; $\alpha \in [0.1, 0.6]$; $\omega \in \{2p \ldots 5p\}$.

The robustness behavior of solving the equations is similar to Fig. A1 for a wide range of parameter values, $k > 7p$, $0 < \alpha < 0.5$, $2p < \omega < 4p$. Since these parameters also influence the computational complexity of the algorithm, the parameters $k = 12p$, $\alpha = 0.25$, $\omega = 4p$ are a good compromise between robustness and computational complexity.

APPENDIX B

How Many Hypotheses?

In this Appendix we derive how many hypotheses need to be generated in order to guarantee at least one correct estimate with a certain probability. The arguments used here are very similar to those of Grimson [9]. Having determined the amount of noise that can be tolerated by solving the equations (denoted by τ) we can now calculate the number of hypotheses H that need to be generated for a given noise level ζ in order to find at least one good hypothesis with probability η. The derivation is straightforward: When we generate

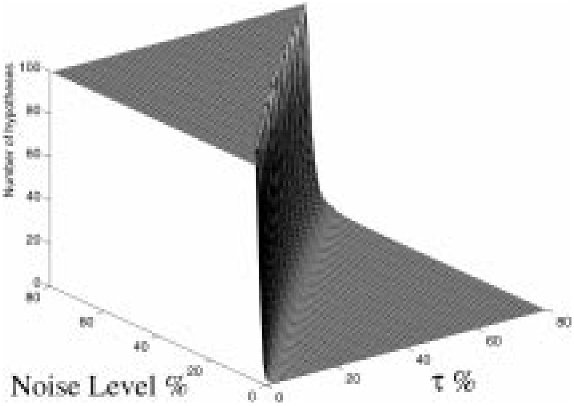


FIG. B2. Number of necessary hypotheses H versus noise and noise tolerance τ of robustly solving the equations; values larger than 100 are set to 100.

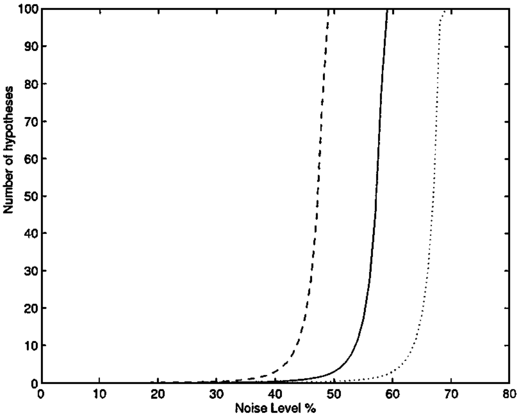


FIG. B3. Number of necessary hypotheses H versus noise. The noise tolerance τ of robustly solving the equations was 0.4 (dashed), 0.5 (full), 0.6 (dotted); values larger than 100 are set to 100.

one hypothesis, the probability of finding a good one is:

$$\rho(\text{good hypo}) = \rho(\text{from } k \text{ points at most } k\tau \text{ are outliers}),$$

where $\rho(a)$ denotes the probability of event a. Therefore,

$$\rho(\text{good hypo}) = \sum_{i=0}^{\lfloor k\tau \rfloor} \binom{k}{i} \zeta^i (1-\zeta)^{k-i}.$$

Now we generate H hypotheses to satisfy the following inequality ρ(at least one good hypo) $> \eta$

$$1 - \left(1 - \sum_{i=0}^{\lfloor k\tau \rfloor} \binom{k}{i} \zeta^i (1-\zeta)^{k-i}\right)^H > \eta,$$

which gives us the required number of hypotheses:

$$H > \frac{\log(1-\eta)}{\log\left(1 - \sum_{i=0}^{\lfloor k\tau \rfloor} \binom{k}{i} \zeta^i (1-\zeta)^{k-i}\right)}.$$

Figures B2 and B3 demonstrate the behavior of H graphically, when the number of equations is $k = 150$. One can clearly see that as long as the amount of noise is within the range of the noise tolerance of solving the equations (τ) we need only a few hypotheses; however, as soon as we have approximately 5% more noise than can be tolerated by solving the equations, we would need to generate a very large number of hypotheses to guarantee to find at least one good hypothesis with probability η. Therefore, we can conclude that only a few (<5) hypotheses need to be explored when the noise level is within the required bounds, a fact which has also been demonstrated by the experimental results.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their constructive and detailed comments, which helped considerably in improving the quality of the paper. This work was supported by a grant from the Austrian National

Fonds zur Förderung der wissenschaftlichen Forschung (S7002MAT). A.L. also acknowledges support from the Ministry of Science and Technology of the Republic of the Slovenia (Projects J2-0414, J2-8829).

REFERENCES

1. T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Wiley, New York, 1958.
2. A. Bab-Hadiashar and D. Suter, Optic flow calculation using robust statistics, in *Proc. CVPR'97, 1997*, pp. 988–993.
3. D. Beymer and T. Poggio, Face recognition from one example view, in *Proceedings of 5th ICCV'95*, pp. 500–507, IEEE Computer Society Press, 1995.
4. H. Bischof and A. Leonardis, Robust recognition of scaled eigenimages through a hierarchical approach, in *Proc. of CVPR'98*, pp. 664–670, IEEE Computer Society Press, 1998.
5. M. Black and A. Jepson, Eigentracking: Robust matching and tracking of articulated objects using a view-based representation, *Int. J. Comput. Vision* **26**(1), 1998, 63–84.
6. J. Edwards and H. Murase, Appearance matching of occluded objects using coarse-to-fine adaptive masks, in *Proc. CVPR'97*, 1997, pp. 533–539.
7. M. A. Fischler and R. C. Bolles, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, *Commu. ACM* **24**(6), 1981, 381–395.
8. F. Glover and M. Laguna, Tabu search, in *Modern Heuristic Techniques for Combinatorial Problems* (C. R. Reeves, Ed.), pp. 70–150, Blackwell, 1993.
9. W. E. L. Grimson, The combinatorics of heuristic search termination for object recognition in cluttered environments, *IEEE Trans. Pattern Anal. Mach. Intell.* **13**, 1991, 920–935.
10. F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics—The Approach Based on Influence Functions*, Wiley, New York, 1986.
11. P. J. Huber, *Robust Statistics*, Wiley, New York, 1981.
12. Y. G. Leclerc, Constructing simple stable descriptions for image partitioning, *Int. J. Comput. Vision* **3**, 1989, 73–102.
13. A. Leonardis and H. Bischof, Dealing with occlusions in the eigenspace approach, in *Proc. of CVPR'96*, pp. 453–458, IEEE Comput. Soc. Press, 1996.
14. A. Leonardis, A. Gupta, and R. Bajcsy, Segmentation of range images as the search for geometric parametric models, *Int. J. Computer Vision* **14**(3), 1995, 253–277.
15. A. Leonardis and H. Bischof, An efficient MDL-based construction of RBF networks, *Neural Networks* **11**(5), 1998, 963–973.
16. B. W. Mel, SEEMORE: Combining color, shape, and texture histograming in a neurally inspired approach to visual object recognition, *Neural Comput.* **9**(4), 1997, 777–804.
17. B. Moghaddam and A. Pentland, Probabilistic visual learning for object representation, *IEEE Trans. Pattern Anal. Mech. Intell.* **19**(7), 1997, 696–710.
18. H. Murase and S. K. Nayar, Image spotting of 3D objects using parametric eigenspace representation, in *The 9th Scandinavian Conference on Image Analysis* (G. Borgefors, Ed.), Vol. 1, pp. 323–332, Uppsala, 1995.
19. H. Murase and S. K. Nayar, Visual learning and recognition of 3-D objects from appearance, *Int. J. Comput. Vision* **14**, 1995, 5–24.
20. H. Murase and S. K. Nayar, Illumination planning for object recognition using parametric eigenspaces, *IEEE Trans. Pattern Anal. Mach. Intell.* **16**(12), 1994, 1219–1227.
21. S. K. Nayar, H. Murase, and S. A. Nene, Learning, positioning, and tracking visual appearance, In *IEEE International Conference on Robotics and Automation, San Diego, May 1994*.
22. S. A. Nene, S. K. Nayar, and H. Murase, *Columbia Object Image Library (COIL-20)*, Technical Report CUCS-005-96, Columbia University, New York, 1996.
23. S. A. Nene, S. K. Nayar, and H. Murase, *SLAM: Software Library for Appearance Matching*, Technical Report CUCS-019-94, Department of Computer Science, Columbia University, New York, September 1994.
24. K. Ohba and K. Ikeuchi, Detectability, uniqueness, and reliability of eigen windows for stable verification of partially occluded objects, *IEEE Trans. Pattern Anal. Mach. Intell.* **9**, 1997, 1043–1047.

25. A. Pentland, B. Moghaddam, and T. Straner, *View-Based and Modular Eigenspaces for Face Recognition*, Technical Report 245, MIT Media Laboratory, 1994.

26. T. Poggio and S. Edelman, A network that learns to recognize three-dimensional objects, *Nature* **343**, 1990, 263–266.

27. R. Rao, Dynamic appearance-based recognition, in *CVPR'97*, pp. 540–546, IEEE Comput. Society, 1997.

28. J. Rissanen, A universal prior for the integers and estimation by minimum description length, *Ann. Statist.* **11**(2), 1983, 416–431.

29. P. J. Rousseuw and A. M. Leroy, *Robust Regression and Outlier Detection*, Wiley, New York, 1987.

30. C. Shannon, A mathematical theory of communication, *Bell Systems Tech. J.* **27**, 1948, 379–423.

31. M. Stricker and A. Leonardis, ExSel++: A general framework to extract parametric models, In *6th CAIP'95*, (V. Hlavač and R. Šara, Eds.), Lecture Notes in Computer Science, No. 970, pp. 90–97, Prague, 1995, Springer-Verlag.

32. M. Turk and A. Pentland, Eigenfaces for recognition, *J. Cognitive Neurosci.* **3**(1), 1991, 71–86.

33. S. Ullman, *High-Level Vision*, MIT Press, Cambridge, MA, 1996.

34. S. Yoshimura and T. Kanade, Fast template matching based on the normalized correlation by using multiresolution eigenimages, in *Proceedings of IROS'94, 1994*, pp. 2086–2093.