Efficient Matching with Invariant Local Descriptors

Roger Mohr, Patrick Gros, and Cordelia Schmid

Imag – Inria
655 avenue de l'Europe
F-38330 Montbonnot
first_name.last_name@imag.fr

Abstract. We are addressing the problem of matching images of scene or of objects when a large collection of reference objects is considered. The paper addresses also the issue of dealing with illumination change and camera position changes. Our approach is firstly based on the use of invariants. Invariants have to be computed locally so that the resulting values will not affected by partial occlusion or accidental highlights. Invariants proved to be a very discriminant piece of information and stored in a hash table they allow efficient indexing of visual shape. Final recognition can be performed using simply a robust voting technique or can be improved using Bayesian decision.

1 Introduction

This paper addresses the problem of matching an image to a large set of reference images. The query image is a new (partial) image of an object imaged in the database, and it might be taken from a different viewing angle or under different illumination conditions.

1.1 Related work

Existing approaches in the literature are of two types: those that use geometric features of an object; and those that rely on the luminance properties.

Geometric approaches model objects by 3D properties such as lines, vertices and ellipses and try to extract these features in order to recognise the objects. General surveys on such model-based object recognition systems are presented in [4, 9]. These methods generally comprise three components : matching, pose computation, and verification. The key contribution of several recognition systems has been a method of cutting down the complexity of matching. For example tree search is used in [6] and recursive evaluation of hypotheses in [1].

The novelty of indexing is that the feature correspondence and search of the model database are replaced by a look-up table mechanism [15, 25]. The

* This work was supported by Inria Action incitative and by the Région Rhône-Alpes

major difficulty of these geometry based approaches is that they use human-made models or require CAD-like representations. These representations are not available for objects such as trees or paintings; in the case of "geometric" objects the CAD-like features used are difficult to extract from the image [8]. An alternative approach is to use the luminance information of an object. The idea is not to impose what has to be seen in the image (points, lines ...) but rather to use what is really seen in the image to characterise an object. The first idea was to use colour histograms [31]. Several authors have improved the performance of the original colour histogram matching technique by introducing measures which are less sensitive to illumination changes [12,19,20,30]. Instead of using color, greyvalue descriptors can also be used for histograms [26].

An alternative idea was to to use the reference image itself for the correspondance, but in order to reduce the size of space, these images were projected on the principal eigenspaces. This approach was first used in [32] for face recognition and then in [18] for general objects. A different reduction is proposed in [33] who learns features which best describe the image. It is also possible to compute local greyvalue descriptors at points of a global grid. The descriptors are either steerable filters [23] or Gabor filters [35]. In the case of partial visibility grid placement gets difficult, as the grid cannot be centred.

1.2 The proposed approach

The approach proposed here holds in four keywords: *invariant local* signatures used for *indexing* the set of potentiel matches; the matching process can be made *robust* by exploiting the redondance in the images.

Almost all of the existing luminance approaches are global and therefore have difficulty in dealing with partial visibility and extraneous features. On the other hand, geometric methods have difficulties in describing "non-geometric" objects and they have problems differentiating between many objects. Local computation of image information is necessary when dealing with partial visibility; photometric information is necessary when dealing with a large number of similar objects and the luminance information is very discrimant, as it known by all the people doing matching by correlation. The approach described here uses local greyvalue or color features computed at interest points as displayed in figure 1.

The invariant characteristics used in this work are based on differential greyvalue invariants [14,24] and they are extended to color in section 4. This ensures invariance under the group of displacements within an image, and it is easily extended toward an affine transformation of luminance in section 2.2. A multi-scale approach [17,34] makes this characterisation robust to scale changes, that is to similarity transformations.

A voting algorithm makes retrieval robust to miss-matches as well as outliers. Outliers are caused by miss-detection of feature points and extraneous features. Semi-local constraints reduce drastically the number of miss-matches.

Such an approach allows the handling of partial visibility and transformations such as image rotation and scaling. Experiments have been conducted on a set

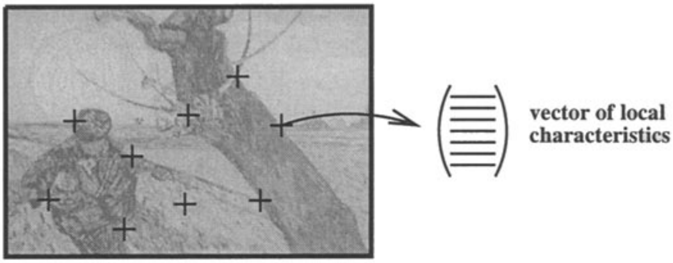


Fig. 1. Representation of an image.

of more than a thousand images, some of them very similar in shape or texture. The high recognition rate is the result of careful design in which robustness to outliers and tolerance to image noise were considered at each step. The use of a priori information of the distribution of these invariant characteristic signatures allows to improve the overall results, and this will be introduced and discussed in section 5.

2 Computing local grey level invariant

Local image signature could be computed around every pixel. As this is too space consuming, many authors compute them on a sparse grid. Our experiences showed instability when the local signatures are not computed at the right location, so the invariants were computed at interest point locations. These points are extracted using the Harris detector [13] which proved to be the most stable one.

2.1 The local jet invariant

The local jet: The image in a neighbourhood of a point can be described by the set of its derivatives. Their stable computation is achieved by convolution with Gaussian derivatives [17, 34]. This set of derivatives has been named "local jet" by Kœnderink [14] and defined as follows:

Let I be an image and σ a given scale. The "local jet" of order N at a point $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2})$ is defined by

$$J^N[I](\mathbf{x},\sigma) = \{L_{i_1 \ldots i_n}(\mathbf{x},\sigma) \mid (\mathbf{x},\sigma) \in I \times I\!R^+ \, ; n = 0, \ldots, N\}$$

in which $L_{i_1 \ldots i_n}(\mathbf{x},\sigma)$ is the convolution of image I with the Gaussian derivatives $G_{i_1 \ldots i_n}(\mathbf{x},\sigma)$ and $i_k \in \{x_1, x_2\}$.

The σ of the Gaussian function determines the quantity of smoothing. This σ also coincides with a definition of scale-space which will be considered later on. σ will be referred to as the *size* of the Gaussian.

The differential invariants: Following the consideration that, even if there are no invariant in images coming from 3D scenes, the similarity group capture up to the first order to variation of shape in an image (see [5]). As translation is already set by selecting point feature, we have to cancel out rotation and scaling in our differential descriptors.

We first consider the case of rotation. Differential invariants were studied theoretically by Kœnderink [14] and Romeny et al.[10]. A complete set of invariants can be computed that locally characterises the signal. The set of invariants used in this work is limited to third order. This set is stacked in a vector, denoted by $\mathcal{V}$. In equation (1) vector $\mathcal{V}$ is given in tensorial notation – the so-called Einstein summation convention. Notice that the first component of $\mathcal{V}$ represents the average luminance, the second component the square of the gradient magnitude and the fourth the Laplacian.

$$\mathcal{V}[0..8] = \begin{bmatrix} L \\ L_i L_i \\ L_i L_{ij} L_j \\ L_{ii} \\ L_{ij} L_{ji} \\ \varepsilon_{ij}(L_{jkl} L_i L_k L_l - L_{jkk} L_i L_l L_l) \\ L_{iij} L_j L_k L_k - L_{ijk} L_i L_j L_k \\ -\varepsilon_{ij} L_{jkl} L_i L_k L_l \\ L_{ijk} L_i L_j L_k \end{bmatrix} \tag{1}$$

with L_i being the elements of the "local jet" and ε_{ij} the 2D antisymmetric Epsilon tensor defined by $\varepsilon_{12} = -\varepsilon_{21} = 1$ and $\varepsilon_{11} = \varepsilon_{22} = 0$.

2.2 Extension to scale and luminane changes

To be insensitive to scale changes the vector of invariants has to be calculated at several scales. A methodology to obtain such a multi-scale representation of a signal has been proposed in [17, 34].

For a function f, a scale change α can be described by a simple change of variables, $f(x) = g(u)$ where $g(u) = g(u(x)) = g(\alpha x)$. For the nth derivatives of f, we obtain $f^{(n)}(x) = \alpha^n g^{(n)}(u)$. Theoretical invariants are then easy to derive, for example $\frac{[f^{(n)}(x)]^{\frac{k}{n}}}{f^{(k)}(x)}$ is such an invariant.

However for such a computation the size of the Gaussian has to be adjusted; this implies a change of the calculation support. As it is impossible to compute invariants at all scales, scale quantisation is necessary for a multi-scale approach. Often a half-octave quantisation is used. The stability of the characterisation has proven this not to be sufficient. Experiments have shown that matching based on invariants is tolerant to a scale change of 20% (see [28]). We have thus chosen a scale quantisation which ensures that the difference between consecutive sizes is less than 20%. As we want it to be insensitive to scale changes up to a factor of 2, the size σ varies between 0.48 and 2.07, its values being: 0.48, 0.58, 0.69, 0.83, 1.00, 1.20, 1.44, 1.73, 2.07.

If the scene is locally planar, and has locally a regular reflection property, and if the camera has a linera response in intensity, then the change of illumination can be expressed locally as a linear fonction (this will be discussed more in details in 4); let I and I' be surface before and after the illumination change, so $I' = aI + b$ for some unkonwn parameter a and b.

It is straightforward to see that the differential invariants are not sensitive to b, except for the first component which is average value which has to be discarded. Then, computing ratio of derivatives allows to directly have components which are invariant to a, and this leads to 7 independant invariants, as we had to cancel out two more parameters in the model.

3 The indexing and retrieval

3.1 The basic voting algorithm

Similarity between vectors has to be estimated using Mahalanobis distance. For this purpose the variance–covariance matrix has been estimated by collecting data when traking interest points on different scenes.

Then a database is constructed containing a set $\{M_k\}$ of models. Each model M_k is defined by the vectors of invariants $\{\mathcal{V}_j\}$ calculated at the interest points of the model images. During the storage process, each vector $\mathcal{V}_j$ is added to the database with a link to the model k for which it has been computed. Formally, the simplest database is a table of couples $(\mathcal{V}_j, k)$.

Recognition consists of finding the model $M_{\hat{k}}$ which corresponds to a given query image I, that is the model which is most similar to this image. For this image a set of vectors $\{\mathcal{V}_l\}$ is computed which corresponds to the extracted interest points. These vectors are then compared to the $\mathcal{V}_j$ of the base by computing: $d_M(\mathcal{V}_l, \mathcal{V}_j) = d_{l,j}\ \forall(l,j)$. If this distance is below a threshold t according the χ^2 distribution, the corresponding model gets a vote.

As in the case of the Hough transform [29], the idea of the voting algorithm is to sum the number of times each model is selected. This sum is stored in the vector $T(k)$. The model that is selected most often is considered to be the best match: the image represents the model $M_{\hat{k}}$ for which $\hat{k} = \arg\ \max_k T(k)$.

Figure 2.a shows an example of a vector $T(k)$ in the form of a histogram. Image 0 is correctly recognized.

Multi-dimensional indexing: Without indexing the complexity of the voting algorithm is of the order of $l \times N$ where l is the number of features in the query image and N the total number of features in the data base. As N is large (about 150,000 in our tests) efficient data structures need to be used.

In order to speed up the search, the vectors were stored in a variant of k-d trees. Each dimension of the space is considered sequentially. Access to a value in one dimension is made through fixed size 1-dimensional buckets. Corresponding buckets and their neighbours can be directly accessed. Accessing neighbours is necessary to take into account uncertainty. A bucket is extended in the next

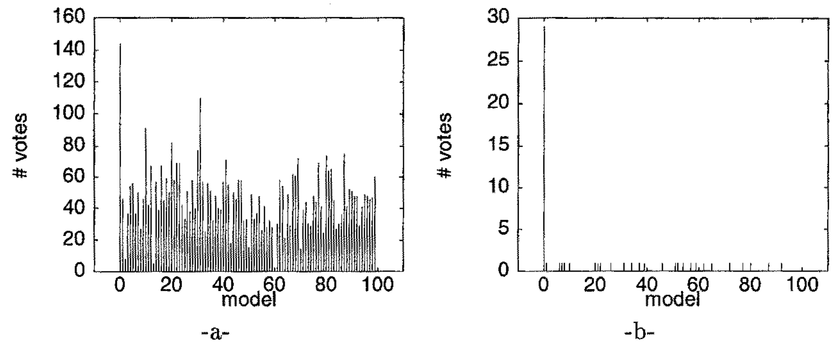


Fig. 2. Result of the voting algorithm : the number of votes are displayed for each model image. Image 0 is recognised correctly. a) direct voting. b) using semi-local constraints

dimension if the number of values stored is above a threshold. Therefore the data structure can be seen as a tree with a depth which is at most the number of dimensions of the stored vectors. The complexity of indexing is of the order of l (number of features of the query image).

This indexing technique leads to a very efficient recognition. The database contains 154030 points. The mean retrieval time for our database containing 1020 objects is less than 5 seconds on a Sparc 10 Station.

3.2 Semi-local constraints

At this stage, no structural has been taken into account for the image, due to our local strategy and we have seen that such a poor strategy provides already acceptable results. Errors occurs mainly when different images share very similar local descriptors, as it is the case for the aerial images. Califano [7] suggested that using longer vectors decreases this probability. Yet the use of higher order derivatives for our invariants is not practical. Another way to decrease the probability of false matches and still stick with our local approach, is to use local shape configuration as shown in figure 3.

For each feature (interest point) in the database, the p closest features in the image are selected. We require that at least 50% of the neighbours match. In order to increase the recognition rate further, a geometric constraint is added: The angles between neighbour points have to be equal, as for example the angles α_1 and α_2 in figure 3. The impact of the geometrical coherence and the semi-local constraints is displayed in figure 2.b. The score of the object to be recognised is now much more distinctive, at the cost of loosing many potential matches.

3.3 Experimental results

Conducted for an image database containing 1020 images, experiments have shown the robustness of the method to image rotation, scale change, small viewpoint variations, illumination changes, partial visibility and extraneous features.

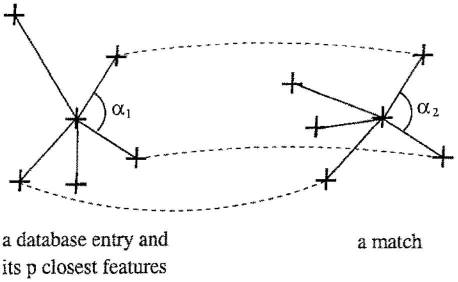


Fig. 3. Semi-local constraints: neighbours of the point have to match and angles have to correspond. Note that not all neighbours have to be matched correctly.

The obtained recognition rate is above 99% for a variety of test images taken under different conditions.

Content of the database: The database includes different kinds of images such as 200 paintings, 100 aerial images and 720 images of 3D objects (see figure 4). 3D objects include the Columbia database. These images are of a wide variety. However, some of the painting images and many of the aerial images are very similar. This leads to ambiguities which the recognition method is capable of dealing with.

In the case of a planar 2D object, an object is represented by one image in the database. This is also the case for nearly planar objects as for aerial images. A 3D object has to be represented by images taken from different viewpoints. Images are stored in the database with 20 degrees viewpoint changes.

Recognition results: Some examples illustrate the conditions under which the method operates correctly. A systematic evaluation for a large number of test images taken under different conditions is then shortily presented.

Firstly in the following three examples are displayed, one for each type of image. For all of them, the image on the right is stored in the database. It is correctly retrieved using any of the images on the left. Figure 5 shows recognition of a painting image in the case of image rotation and scale change. It also shows that correct recognition is possible if only part of an image is given.

In figure 6 an example of an aerial image is displayed. It shows correct retrieval in the case of image rotation and if part of an image is used. In the case of aerial images we also have to deal with a change in viewpoint and extraneous features. Notice that buildings appear differently because viewing angles have changed and cars have moved.

Figure 7 shows recognition of a 3D object. The object has been correctly recognised in the presence of rotation, scale change, change in background and partial visibility. In addition, there is a change of 10 degrees of viewpoint position between the two observations. Notice that the image of the object has not only been recognised correctly, but that the closest stored view has also been retrieved.

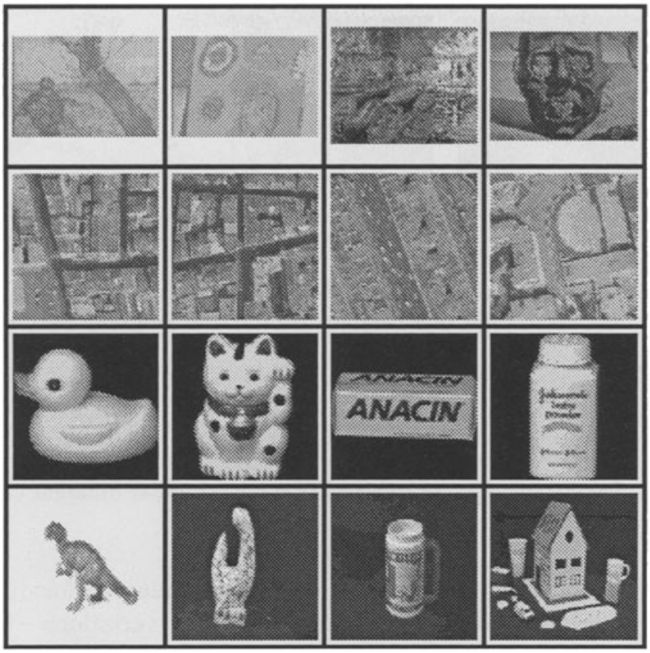


Fig. 4. Some images of the database. The database contains more 1020 images.

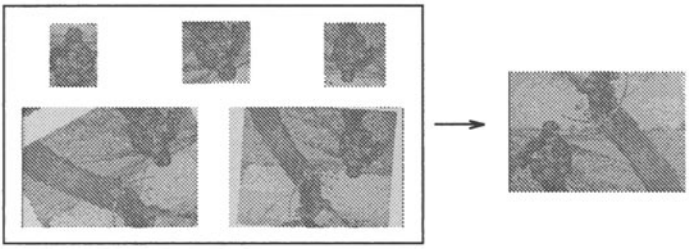


Fig. 5. The image on the right is correctly retrieved using any of the images on the left. Images are rotated, scaled and only part of the image is given.

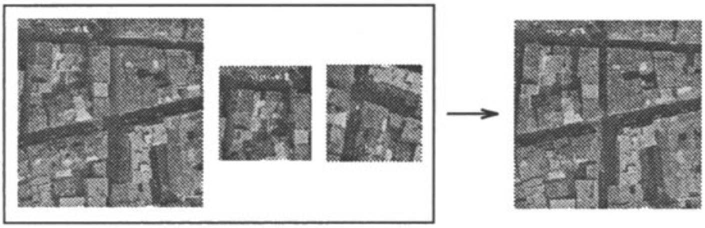


Fig. 6. The image on the right is correctly retrieved using any of the images on the left. Images are seen from a different viewpoint (courtesy of Istar).

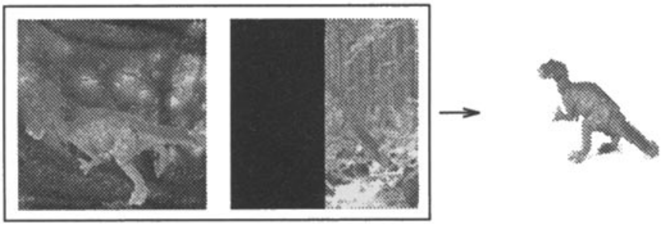


Fig. 7. The image on the right is correctly retrieved using any of the images on the left. The 3D object is in front of a complex background and only partially visible.

Rotation	Scaling	Viewpoint change aerial	Viewpoint change 3D	30% partial visibility	20% partial visibility	10% partial visibility
100 %	100 %	99 %	99.8 %	100 %	95 %	90.5 %

Table 1. Percentage of good retrieval (firt guess) under different conditions.

Systematic evaluation of retrieval The method is evaluated for different transformations – image rotation, scale change, viewpoint variations – as well as for partial visibility. It has to be noticed that all the changes mentionned does not come from artificial motion in the images, but from new view taken each time. One particular case is the partial visibility: subimages were taken from the intial image in the case of painting, and from a new view for the case of aerial images. Average correct retrieval are summarized in Table 1; a retrieval is considered as correct if the *first* guess for the retrieved image is the correct one. For a more detailed evaluation the reader is refered to [28].

4 The color case

As shown in the previous section, the direct use of the grey-level information in the images provide very discriminant and powerful local descriptors. It is thus natural to wonder whether the use of color would not provide even better results.

4.1 The local jet in color

The first problem when using color images is to choose a representation system for color information. In the present work, the RGB representation has been chosen as the most convenient: It is directly available on most image devices and it allows to find simple models of color variation (see next section). In this representation, a color image is composed of three monochromatic images, corresponding to each of the three R, G or B channels.

From such an image, it is possible to compute a grey-level image, using a well established formula [22]: $G = 0.2125R + 0.7154G + 0.0721B$. Interest points can be detected from this image, using the Harris detector, as it is done with

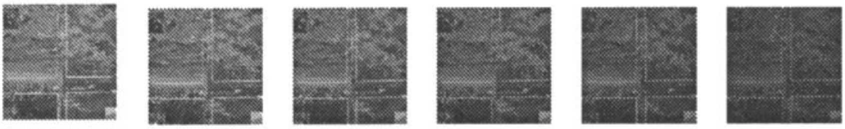


Fig. 8. 6 images with light intensity variation (numbered from 1 to 6).

Model	Median and maximum of errors between two images				
	Im. 1 et 2	Im. 1 et 3	Im. 1 et 4	Im. 1 et 5	Im. 1 et 6
M1	13.1909 44.2041	38.3275 94.3769	60.1664 141.128	83.1204 186.786	107.75 247.348
M2	5.568 33.643	5.67147 34.8631	5.80588 31.5682	5.88565 33.1802	6.04758 39.7918
M3	5.56175 33.6486	5.56269 33.9358	5.65976 32.1776	5.73528 33.223	5.51266 33.3612
M4	5.55503 33.5326	5.62512 34.5064	5.74105 31.6982	5.78087 34.1667	5.82957 39.7582
M5	5.54143 33.5563	5.53818 33.5722	5.62793 32.1895	5.68712 33.8315	5.45086 32.9997

Table 2. Evaluation of the different models.

ordinary grey level images. Given these points, it is then possible to compute a local jet for each of the three original R, G and B images of the color image. This gives a total of $30 = 3 \times 10$ signal characteristics for each interest point.

4.2 Illumination Models

Obtaining image descriptors invariant to illumination variations implies that an illumination model is known first. Two sources of variations can be distinguished: The first one is a color or intensity variation of the light source, that will be called an internal variation, and the secondone is due to a variation of position or orientation of the source and will be called an external variation.

Models for Internal Variations Several models of internal variations have been compared. They describe how the color vector $\mathbf{p} = (r, g, b)$ of an image pixel is transformed in $\mathbf{p}' = (r', g', b')$ at each image pixel when the light source internally changes ($\mathbf{T}$ is a translation vector, $\mathbf{D}$ a diagonal matrix and $\mathbf{M}$ a 3×3 matrix): M1: $\mathbf{p}'=\mathbf{p}$ M2: $\mathbf{p}'=\mathbf{D}\mathbf{p}$ M3: $\mathbf{p}'=\mathbf{D}\mathbf{p}+\mathbf{T}$ M4: $\mathbf{p}'=\mathbf{M}\mathbf{p}$ M5: $\mathbf{p}'=\mathbf{M}\mathbf{p}+\mathbf{T}$

The performance of each model was first evaluated on real images, and then we tried to determine which ofthe estimated parameters were really significant. *Model Evaluation.* To evaluate the different models we took 6 images representing a same scene. Between the shots the light intensity was the only variation (see fig. 8).

In order to compute the different model parameters saturated pixels were removed, and a least median square method based on the SVD decomposition was used. For each pair of images the median and the maximum of errors between the first image and the second image corrected by the model were computed. Table 2 presents the results.

According to these results modele M5 appears to have the best quality / complexity ratio. Six additional parameters are needed to obtain slightly better results. The diagonal model without translation is good when the images are not too different.

Test of Estimated Parameters. More parameters usually provide better results, but these additional parameters may estimate noise rather than the model itself. To check if it is the case, the method proposed by Florou was used [11].

This method is based on a statistical test. The noise is assumed to be centered and for each parameter a confidence interval is computed for a confidence level of 95%. The radius of this interval is: $R(p_i) = \sqrt{\chi^2(95\%, m)}\sqrt{\frac{f}{n}}\sqrt{\sigma_i^2}$, where $\chi^2(95\%, m)$ is the value of χ^2 distribution for the given confidence level, m is the number of estimated parameters, f is the sum of the errors, n is the number of pixels and σ_i^2 is the variance of the estimated parameter. If $0 \in [p_i - R(p_i), p_i + R(p_i)]$, 0 appears to be an estimation as good as p_i for the i-th parameter. In such a case the estimation p_i is considered not to be significant.

The significance of model M5 parameters has been tested with images of different sizes. The following table provides the minimal size for which each parameter of model M7 has been estimated significantly.

parameter	a_{11}	a_{12}	a_{13}	a_{21}	a_{22}	a_{23}	a_{31}	a_{32}	a_{33}	t_r	t_g	t_b
size	5	60	125	60	20	25	90	90	30	30	15	15

It appears that a significant estimation of all the parameters, and especially of the non diagonal terms, requires large sub-images. On the other hand the parameters present in the model M5 are significantly estimated even with small sub-images.

Models for External Variations In the same fashion as Finlayson proposed the diagonal model for internal variations, he proposes a very simple model for external variations [2]: This model states that each color vector is multiplied by a scale factor. The bad news is that this factor depends of the relative position and orientation of the light source and the illuminated point, and thus varies from pixel to pixel. The full model contains a scale factor for each pixel.

4.3 Local color invariants

There are two main ways to exploit the models presented in the previous section.

The first one consists in deriving a normalization scheme: each image is transformed in order to be independent from the illumination conditions. According to the illumination model taken into account, different normalization techniques are possible. An example is presented on Fig. 11.

To stay in the global frame of the present paper, we focus the presentation on local color invariants. The basic information used to compute these invariants are the 30 components of the color local jet. The deal is to combine these components in such a way that the result is invariant under illumination and/or geometric variations.

Scale invariance. As with grey level images, scale invariance may be obtained using a multi-scale approach : the local jet is computed for several values of σ. Invariants are computed at these different scales, and their comparison provides the scale factor between points of different images.

Rotational Invariance. Rotation is defined by a single parameter: Its angle. Thus there should exist 29 independent invariants: Each channel provides 9 of them, two other ones can be chosen among the 3 following ones:

$$R_xG_x+R_yG_y \qquad G_xB_x+G_yB_y \qquad B_xR_x+B_yR_y$$

Illumination and Rotational Invariance. The model M3 was chosen as the reference model in the latter. To compute illumination invariants from the set of rotational invariants just presented 6 parameters have to be eliminated. The translational parameters t_r, t_g, and t_b are eliminated by suppressing the 3 invariants R, G, and B. The 3 diagonal terms are eliminated by dividing each invariant by the correct power of the gradients and by suppressing the 3 gradients from the invariant list. For example one of the invariants presented before become:

$$\frac{R_xG_x + R_yG_y}{(R_xR_x + R_yR_y)^{1/2}(G_xG_x + G_yG_y)^{1/2}}$$

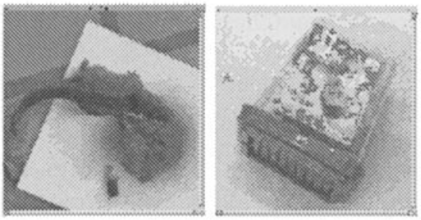


Fig. 9. An image of each of the rotational test sequences.

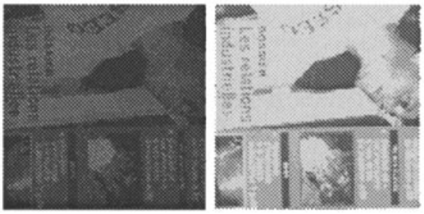
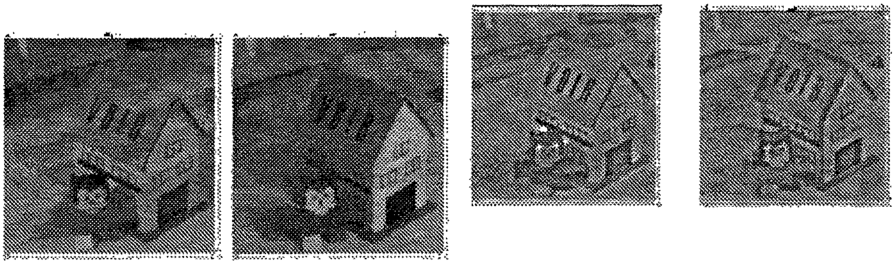


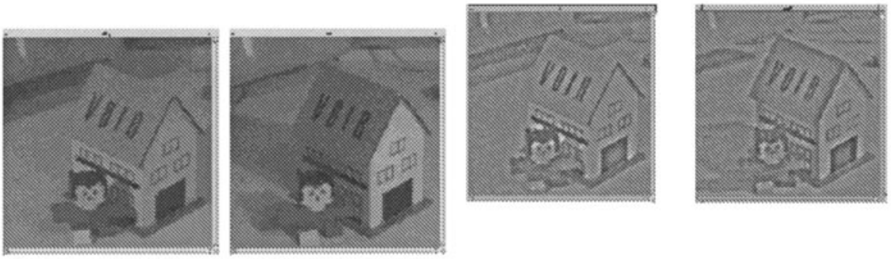
Fig. 10. 2 images of a sequence with a variation of light intensity.

Results with a Scene Rotation. The rotational invariants were tested using two sequences of 30 images taken 6 degrees apart (see Fig. 9). In both cases the axis of rotation is the optical axis of the camera. The first image of each sequence was matched with all the other image of the sequence using the rotational invariants, providing more than 99% of correct matches.

Results for Light Internal Variations. A sequence of 10 images representing a journal cover was used (see Fig.10). Between the shots only the light intensity



nor-



ional

ts for

ches.

which

Results with external light variations. In this last test a sequence of 7 images was used (see Fig. 11). Between the shots the light source moved around the scene. The first image was matched to the other ones using two techniques. First the invariants for rotations and illumination variations were used directly, although they are clearly out of their domain of validity. Second a normalization technique was used (normalized images are chown on the right of Fig. 11), followed by the use of the same invariants . The rate of correct matches is 71% without normalization and 80% with normalization. The effect of normalization is twofold: There were more points detected, and these points were detected in a more repeatable way, and the percentage of correct matches is higher.

5 Extension using a probabilistic model

The basic indexing algorithm described in section 3 assumes that all vectors are equally probable, therefore according equal importance to all matches. However, this is not the case: We observed that the invariants collected in our grey-level image data base are far from evenly distributed. An rough histrogram was used to estimate the distribution, as the large number of data was not sufficient for making an analytical estimation in such a still larger space.

5.1 The basic Bayesian model

In a study on the distribution of the measure in some "receptive fields", Schiele [27] derived a Bayesian model which allowed the distribution of the measures to be taken into account. Here we take a similar approach, deriving a model which uses the a priori knowledge of the distribution of the invariants in the matching process. But, as this work is concerned with matches or potential matches, we have to derive a more sophisticated model in order to take into account the matching process.

Let Q be the query image and R be a database image being considered as a match candidate. It is assumed that Q and R will have a large number of features in common if they match. Q has $n(Q)$ interest points, and R has $n(R)$ such points. From the $n(Q)$ features of Q, $\{m_i\}, i \in I$ is the set of the features which are matched with features in R. Each m_i is a feature vector of Q that has a matching feature vector f_k appearing in R.

We want to evaluate $P(R|\{m_i\})$. Using Bayes formula we get

$$P(R|\{m_i\}) = \frac{P(\{m_i\}|R).P(R)}{P(\{m_i\})} \tag{2}$$

Assuming that the individual matches are independent, this translates to

$$P(R|\{m_i\}) = \frac{\prod_{i \in I} P(m_i|R)P(R)}{\prod_{i \in I} P(m_i)} \tag{3}$$

$P(m_i)$ is the probability that the i-th feature of Q has one match with $n(R)$ random features.

This approach considers only the effect of matches, but the fact that many features fail to match must also be considered. To incorporate this let $\{\overline{m}_j\}, j \in J$ be the set of of features that failed to be matched. Formula (3) can thus be extended:

$$P(R|\{m_i\}, \{\overline{m}_j\}) = \frac{\prod_{i \in I} P(m_i|R) \prod_{j \in J} P(\overline{m}_j|R)P(R)}{\prod_{i \in I} P(m_i) \prod_{j \in J} P(\overline{m}_j)} \tag{4}$$

5.2 Posterior Probability of Retrieved Images

Matching also occurs randomly, inducing false matches, and this was not taken into account in the previous discussion. If we assume that Q is a subimage of R under some new viewing condition, the k-th feature of Q might be a feature of R with the previously defined probability α. It could be also a feature that occured due to some random process with the density probability of R.

Let p_R^k be the probability of the k-th random feature vector of Q to appear in image R. p_R^k is estimated in the particular image R. The probability that it might miss all the $n(R)$ features in R is therefore $(1-p_R^k)^{n(R)}$. Thus the corresponding probability of matching one of these features is $1-(1-p_R^i)^{n(R)}$. Combining this two events which are exclusive, the likelihood of match (a correct or a false one) becomes:

$$P(m_i|R) = \alpha + (1-\alpha)(1-(1-p_R^i)^{n(R)}) \tag{5}$$

Similarly the probability of occurence of the j-th feature vector of Q will be p_B^k (B for data *b*ase) and the a priori probability of match with $n(R)$ feature vector is then

$$P(m_j) = \beta + (1-\beta)(1-(1-p_B^i)^{n(R)})$$

Substituting this into equation 3 results in:

$$P(R|\{m_i\}) = \frac{\prod_{i\in I}(\alpha + (1-\alpha)(1-(1-p_R^i)^{n(R)}))}{\prod_{i\in I}(\beta + (1-\beta)(1-(1-p_B^i)^{n(R)}))} P(R) \tag{6}$$

The probability for the non matched point is handled in a similar way.

5.3 Experiments

The experiments reported here are based on querying aerial images only as these are the more difficult ones to process: They are similar in texture and shape, as roofs of old houses look very similar seen from the sky. The query images were taken from an airplane from a position different from the reference images (about 20 degrees change in the almost vertical viewing direction). Altitude was the same, so the scaling effect can be neglegted. However due to the change of viewing direction, the images differ: some facades are visible, illumination has changed, etc.

Experiments: The value for α was set to 0.5 and 0.35. The value for β was set to $\frac{\alpha}{1020}$ as the data base has 1020 images.

The experiments were conducted in the following way: for each query image, a subimage is extracted which represents $x\%$ of the initial image as it is taken as the query. x ranges from 100 to 9. The rank of the right answer was measured as the output. As the standard deviation of such answer is high for small window, random selection of such windows were multiplied in order to get significant mean values.

Four different matching strategies were investigated: direct voting and use of the semi-local constraints, and on the top of the strategies, the use of the Bayesian models. Fig. 12 displays the behavior of the different strategies. The abcissa represents the size of the subwindows considered for the request (percentage of the image surface). The ordinate shows the mean rank of the correctly matching image.

The four curves displayed correspond, from top to bottom, to the simple voting on invariant, the Bayesian model with $\alpha = 0.5$, the Bayesian model with $\alpha = 0.35$, and the use of semi local context.

The results show that there is an clear advantage to using the Bayesian decision rule for the case of simple voting. The behavior of the Bayesian decision rule is not too much affected by the value of α. This gain is more limited for smaller windows where the number of matched features decreases largely.

The use of semi-local constraints is much more discriminant and provide the best result, and introduction of the Bayesian model almost does not improve

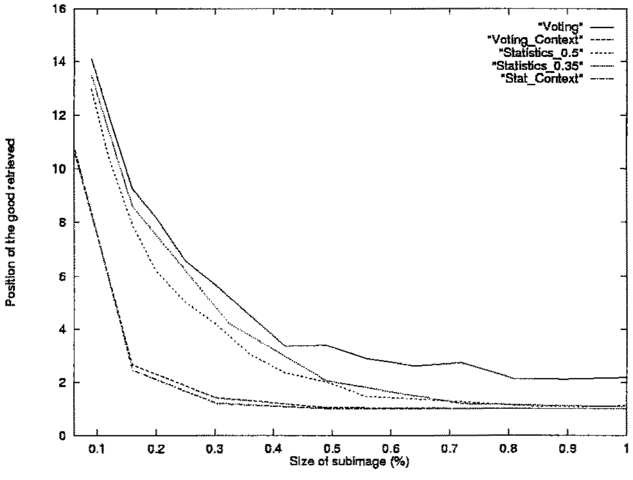


Fig. 12. Behavior of voting *vs* posterior probability

its quality This can be forcasted from the analysis of formula (5): when p_R^i is very small, equation (5) simplifies to $P(m_i|R) = \alpha$, i.e. to the case when false matches are not considered. This is exactly what the use of semi-local constraint does.

6 Discussion, conclusion

This paper has clearly illustrated how invariant signatures are powerfull index for retrieving images. Invariance is however not always a solution and two approximations were used here in order to be able to compute such invariant signatures: firstly limited camera motion for which the similarity group is sufficient for modeling the geometric image changes, secondly a simplification of the illumination model in order to be able to compute significant brightness invariants.

The signatures are discriminant enough that a simple voting strategy allows to find the right image from a fragment taken under different conditions. However this strategy in largely improved by the introduction of geometric constraints, the so-called semi-local constraints, or by introducing a Bayesian decision criterion.

This approach has been widely experimented under different conditions: illumination changes, change of viewpoints, zooming, etc. and the results are impressive, even when using only the gray level values. Extensive tests with color image has right now not been performed; the corresponding data base is under construction and can be reached at `www.inrialpes.fr/movi/pub/Images/index.html`.

The main limitation we observed during experiments were the large changes in the invariants caused by strong shadows on details, for instance, the shadow around a car in an aerial image. For such purpose robust descriptors should be developped like for instance what Zabih developped for stereomatching [36]. A preliminary step in this direction can be found in [16].

Such a technique allows to consider for instance object modeling by using a large collection of images instead artificial models. But the major application is probably in image data base. Even if the experiments conducted here were not realistic with respect of data base size, the kind of difficult images we were processing, in particular the similarity in texture and representation (see the second row in Fig. 4) allows us to forcast it applicabitity to large set of images. The key issue we forsee when dealing with 10^6 images is the indexing problem. With such a size, the data structure we are using has to be stored on disks, and right now the structure has too many links for allowing efficient use on secondary memory. No solution for such uncertain index in such large space has been provided yet. Right now solutions mainly focussed on searching for the nearest neighbors (see for instance [21] and [3]).

References

1. N. Ayache and O.D. Faugeras. HYPER: a new approach for the recognition and positioning of 2D objects. *PAMI*, 8(1):44–54, 1986.
2. K. Barnard, G. Finlayson, and B. Funt. Colour constancy for scenes with varying illumination. In *ECCV*, pages 3–15, July 1996.
3. S. Berchtold, D.A. Keim, and H.P. Kriegel. The X-tree: An index structure for high-dimensional data. In *Proceedings of the 22nd VLDB Conference, Mumbai (Bombay), India*, pages 28–39. the Very Large Database Endowment, 1996.
4. P.J. Besl and R.C. Jain. Three-dimensional object recognition. ACM *Computing Surveys*, 17(1), 1985.
5. T.O. Binford and T.S. Levitt. Quasi-invariants: Theory and exploitation. In *Proceedings of* DARPA *Image Understanding Workshop*, pages 819–829, 1993.
6. R.C. Bolles and R. Horaud. 3DPO : A three-dimensional Part Orientation system. *IJRR*, 5(3):3–26, 1986.
7. A. Califano and R. Mohan. Multidimensional indexing for recognizing visual shapes. *PAMI*, 16(4):373–392, April 1994.
8. J.L. Chen and G.C. Stockman. Matching curved 3D object models to 2D images. In A.C. Kak and K. Ikeuchi, editors, *Proceedings of the Second CAD-Based Vision Workshop*, pages 210–218, Los Alamitos, California, February 1994. IEEE Computer Society Press.
9. R.T. Chin, H. Smith, and S.C. Fralik. Three-dimensional object recognition. *ACM Computing Surveys*, 17(1):75–145, 1986.
10. L.M.T. Florack, B. ter Haar Romeny, J.J Koenderink, and M.A. Viergever. General intensity transformation and differential invariants. *Journal of Mathematical Imaging and Vision*, 4(2):171–187, 1994.
11. G. Florou and R. Mohr. What accuracy for 3D measurements with cameras? In *ICPR*, volume I, pages 354–358, 1996.
12. B.V. Funt and G.D. Finlayson. Color constant color indexing. *PAMI*, 17(5):522–529, 1995.
13. C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, 1988.
14. J.J. Koenderink and A.J. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375, 1987.
15. Y. Lamdan and H.J. Wolfson. Geometric hashing: a general and efficient model-based recognition scheme. In *ICCV*, pages 238–249, 1988.

16. Z.D. Lan and R. Mohr. Non-parametric invariants and application to matching. Technical Report 3246, INRIA, September 1997.
17. T. Lindeberg. *Scale-Space Theory in Computer Vision.* Kluwer Academic Publishers, 1994.
18. H. Murase and S.K. Nayar. Visual learning and recognition of 3D objects from appearance. *IJCV*, 14:5–24, 1995.
19. K. Nagao. Recognizing 3D objects using photometric invariant. In *ICCV*, pages 480–487, 1995.
20. S.K. Nayar and R.M. Bolle. Computing reflectance ratios from an image. *Pattern Recognition*, 26(10):1529–1542, 1993.
21. S.A. Nene and S.K. Nayar. A simple algorithm for nearest neighbor search in high dimensions. *PAMI*, 19(9):989–1003, 1997.
22. C.A. Poynton. Frequently asked questions about color, 1997.
23. R.P.N. Rao and D.H. Ballard. Object indexing using an iconic sparse distributed memory. In *ICCV*, pages 24–31, 1995.
24. B.M Romeny, L.M.J. Florack, A.H. Salden, and M.A. Viergever. Higher order differential structure of images. *Image and Vision Computing*, 12(6):317–325, 1994.
25. C.A. Rothwell. *Object Recognition Through Invariant Indexing.* Oxford Science Publication, 1995.
26. B. Schiele and J.L. Crowley. Object recognition using multidimensional receptive field histograms. In *ECCV*, pages 610–619, 1996.
27. B. Schiele and J..L. Crowley. Probabilistic object recognition using multidimensional receptive field histogram. In *ICPR*, pages 50–54, 1996.
28. C. Schmid. *Appariement d'images par invariants locaux de niveaux de gris.* Thèse de doctorat, Institut National Polytechnique de Grenoble, GRAVIR – IMAG – INRIA Rhône–Alpes, July 1996. ftp.imag.fr/pub/MOVI/theses/schmid.ps.
29. S.D. Shapiro. Feature space transforms for curve detection. *Pattern Recognition*, 10(3):129–143, 1978.
30. D. Slater and G. Healey. The illumination-invariant recognition of 3D objects using color invariants. *PAMI*, 18(2):206–210, 1996.
31. M.J. Swain and D.H. Ballard. Color indexing. *IJCV*, 7(1):11–32, 1991.
32. M. Turk and A. Pentland. Face recognition using eigenfaces. In *CVPR*, pages 586–591, 1991.
33. P. Viola. Feature-based recognition of objects. In *Proceedings of the AAAI Fall Symposium Series: Machine Learning in Computer Vision: What, Why, and How?, Raleigh, North Carolina, USA*, 1993.
34. A.P. Witkin. Scale-space filtering. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence, Karlsruhe, Germany*, pages 1019–1023, 1983.
35. X. Wu and B. Bhanu. Gabor wavelets for 3D object recognition. In *ICCV*, pages 537–542, 1995.
36. R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondance. In *ECCV*, pages 151–158, 1994.