Combining greyvalue invariants with local constraints for object recognition

C. Schmid R. Mohr

GRAVIR
655 avenue de l'Europe
38330 MONTBONNOT SAINT-MARTIN
e-mail : Cordelia.Schmid@imag.fr

Abstract

This paper addresses the problem of recognizing objects in large image databases. The method is based on local characteristics which are invariant to similarity transformations in the image. These characteristics are computed at automatically detected keypoints using the greyvalue signal. The method therefore works on images such as paintings for which geometry based recognition fails. Due to the locality of the method, images can be recognized being given part of an image and in the presence of occlusions. Applying a voting algorithm and semi-local constraints makes the method robust to noise, scene clutter and small perspective deformations. Experiments show an efficient recognition for different types of images. The approach has been validated on an image database containing 1020 images, some of them being very similar by structure, texture or shape.

1 Introduction

Recognition and matching are considered as a major problem in computer vision. We want to address the ambitious goal of identifying and locating objects under the following conditions: 1)partial visibility 2)different viewing angles 3)complex scenes, complex background 4)hundreds or thousands of potential reference shapes.

Furthermore, recognition should also be fast. Methods which realize such a recognition have potential applications ranging from consulting image databases to identifying complex objects in real environments, and even visual servoing for a robot arm manipulator.

There exists two approaches to object recognition in the literature. One of them uses geometric features of an object. The other one relies on the luminance signature of an object, that is on its appearance. The interested reader is referred to the extended version of this paper for a state of the art about existing recognition methods [7].

Appearance based systems are capable to model any kind of object and to differentiate between objects of the same geometrical shape. However, previous methods proposed in the literature are global and therefore do not work if the object is only partially visible or contained in a complex scene. Additionally, these methods are not invariant to any kind of image transformation (except [4] who uses steerable filters). This paper presents a new appearance based approach which overcomes these drawbacks. The proposed method is local and invariant to similarity transformations in the image. It uses a vector of differential greyvalue invariants computed at automatically detected keypoints (see figure 1). A robust recognition algorithm based on voting and semi-local constraints uses the proposed characterization and obtains a high recognition rate. This algorithm is resistant to miss-detection and noise. Furthermore, indexing via a multi-dimensional hash-table makes fast recognition possible.

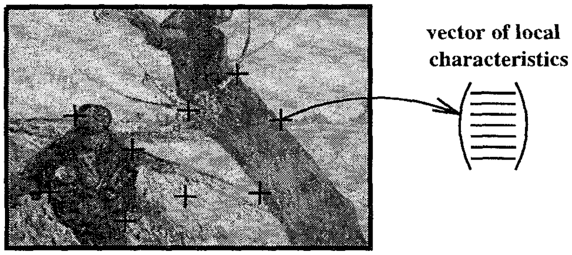

Figure 1: Representation of an image.

Our approach is an important contribution to object recognition. It makes recognition possible for objects in situations which could not be dealt with before. We can identify and locate objects in case of partial visibility, image transformations and complex scenes. In case of 3D objects we are not only capable to retrieve the corresponding object correctly, but also its pose. The success of our approach is based on the combination of differential invariants computed at keypoints with a robust voting algorithm and semi-local constraints. It has been shown that these invariants can be implemented with a sufficiently small filter size to capture local discriminant greylevel information. Moreover, the multi-scale approach makes our method robust to scale changes up to a factor 2

1063-6919/96 $5.00 © 1996 IEEE

which has never been reported in the literature.

2 Keypoint detector

A wide variety of detectors for keypoints exists in literature. Keypoints can be detected using contours or directly from the greyvalue signal. The main advantage of detectors operating on the greyvalue images is that their performance is not dependent on the success or failure of a prior contour extraction step.

For object recognition as well as matching it is important that the detector is repeatable, that is results have to be invariant to image transformations. An imprecision in the location of keypoints leads to different characterizations which can no longer be used for recognition purposes. In a previous work [1], different detectors have been compared in the presence of image rotation, scale change, light changes and image noise. This work has shown a poor stability of existing methods and best results for the Harris detector. A stabilized implementation of this detector has been used in the present work. The stabilization has been obtained by using Gaussian derivative filters. A recursive implementation of these filters guarantees fast detection.

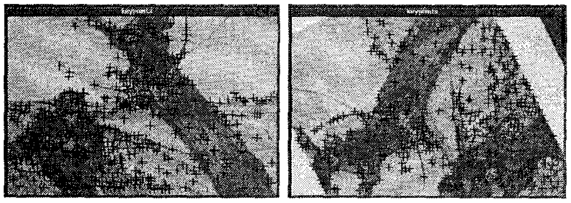

Figure 2: Keypoints detected on the same scene under rotation. The image rotation between the left image and the right image is 155 degrees.

Figure 2 shows keypoints detected on the same scene under rotation. It shows the repeatability of the obtained detector. Notice that not all points are repeated. However, it is sufficient if at least 50% of points are detected in two images and if these points are precise that is at the same location of the scene. On the two images shown in this figure, the number of keypoints is 553 and 554. The average number of keypoints detected on the images of the database is around 150. However, this number depends on the image, and it ranges from 20 to 800.

3 Multi-scaled differential greyvalue invariants

Differential invariants have been studied theoretically by Kœnderink [3] and Romeny and al. [5]. Interested readers are referred to these works. They have made explicit mathematical results first derived by Hilbert and they have proposed a stable implementation for the computation of differentials. This is crucial especially if third order differentials are used. For this purpose, we have used the rotationally invariant Gaussian function to smooth the image.

3.1 Complete set of differential invariants

We propose to use a complete set of invariants under the group $GL(2)$ of rigid displacements in the image to characterize the signal. This set will be denoted $\vec{\mathcal{V}}$. The first part of this vector $\vec{\mathcal{V}}$ contains the complete and irreducible set of differential invariants up to 2nd order (cf. equation 1). The formulation of this first part of the vector is given in tensorial invariant manifest notation, so-called Einstein notation. Notice that the first component of $\vec{\mathcal{V}}$ represents the average luminance, the second component the square of the gradient magnitude and the fourth the Laplacian.

$$\vec{\mathcal{V}}[0..4] = \begin{bmatrix} L \\ L_i L_i \\ L_i L_{ij} L_j \\ L_{ii} \\ L_{ij} L_{ji} \end{bmatrix} \qquad (1)$$

The L_i are the convolution of the derivatives of the Gaussian function with the luminance function L. It is possible to compute them for different sizes σ of the Gaussian.

The second part of the vector $\vec{\mathcal{V}}$ contains a complete set of invariants of third order. Equation 2 presents this set.

$$\vec{\mathcal{V}}[5..8] = \begin{bmatrix} \varepsilon_{ij}(L_{jkl} L_i L_k L_l - L_{jkk} L_i L_l L_l) \\ L_{iij} L_j L_k L_k - L_{ijk} L_i L_j L_k \\ -\varepsilon_{ij} L_{jkl} L_i L_k L_l \\ L_{ijk} L_i L_j L_k \end{bmatrix} \qquad (2)$$

with ε_{ij} the 2D antisymmetric Epsilon tensor defined by $\varepsilon_{12} = -\varepsilon_{21} = 1$ and $\varepsilon_{11} = \varepsilon_{22} = 0$.

3.2 Multi-scale approach

The vector $\vec{\mathcal{V}}$, presented in section 3.1, makes object recognition or matching possible in the presence of any rigid displacement. To be resistant to scale changes, that is to similarity transformations, the vector of invariants has to be calculated at several scales. For a function f, a scale change α can be described by a simple change of variables, $f(x) = g(u)$ where $g(u) = g(u(x)) = g(\alpha x)$. We then obtain :

$$f^{(n)}(x) = \alpha^n g^{(n)}(u) \qquad (3)$$

where $f^{(n)}(x)$ represents the nth derivative of f.

$\frac{[f^{(n)}(x)]^{\frac{k}{n}}}{f^{(k)}(x)}$ is a theoretical invariant to scale change. However, in case of a discrete representation of the function, as for an image, the previous equation 3 is rewritten as :

$$\int_{-\infty}^{+\infty} I_1(\vec{x}) G_{i_1 \dots i_n}(\vec{x}, \sigma) d\vec{x} = \alpha^n \int_{-\infty}^{+\infty} I_2(\vec{u}) G_{i_1 \dots i_2}(\vec{u}, \sigma\alpha) d\vec{u} \quad (4)$$

where $G_{i_1 \dots i_2}$ are the derivatives of the Gaussian.

Equation 4 shows that the size of the Gaussian, that is the calculation support, has to be adjusted. As it is impossible to compute invariants at all scales, scale quantization is necessary for a multi-scale approach. Often a half-octave quantization is used. The stability of the characterization has proven this not to be sufficient. Experiments have shown that matching based on invariants resists to scale change of 20% (see [6]). We have thus chosen a scale quantization which ensures that the difference between consecutive sizes is less than 20%. As we want to be resistant to scale changes up to a factor of 2, the size σ varies between 0.48 and 2.07, being chosen at values: 0.48, 0.58, 0.69, 0.83, 1, 1.2, 1.44, 1.73, 2.07.

4 Recognition Algorithm

In this section the recognition algorithm is presented. It is based on the computation of the similarity between two invariant vectors.

4.1 Mahalanobis distance

To recognize an image, it is necessary to decide if two invariants are similar. We propose to model the uncertainties in the components of $\vec{\mathcal{V}}$ as random variables with Gaussian distribution and use the Mahalanobis distance to compare invariant vectors. This distance takes into account the different magnitude as well as the covariance matrix Λ of the components:

$$d_M(\vec{b},\vec{a}) = \sqrt{(\vec{b}-\vec{a})^T \Lambda^{-1} (\vec{b}-\vec{a})}$$

By thresholding this distance, it is possible to decide statistically if two invariants are similar as the square of the Mahalanobis distance is a random variable with a χ^2 distribution. As the covariance matrix is a real symmetric (semi) definite positive matrix, it can be decomposed as follows:

$$\Lambda^{-1} = P^T D P = P^T \sqrt{D}\sqrt{D} P$$

where P is orthogonal and D is diagonal. We then have:

$$d_M(\vec{a},\vec{b})^2 = d_E(\sqrt{D}P\vec{a}, \sqrt{D}P\vec{b})^2$$

To compute the Mahalanobis distance between two invariants vectors is thus equivalent to transform these vectors by multiplying them by the matrix $\sqrt{D}P$ and to compute the Euclidean distance between the two transformed vectors.

4.2 Voting algorithm

Recognition consists of finding the model $M_{\hat{k}}$ which corresponds to a given image I. That is the model which is "closest" (i.e. the most similar) to this image.

As for the Hough transform the idea of the voting algorithm is to sum the number of times each model is selected. Thus, each time a model M_k is selected, a voting table T is updated such that the value $T(k)$ is incremented by one. Note that a point can select each model but only once. The model that is selected most often is considered to be the best match: the image represents the model $M_{\hat{k}}$ for which

$$\hat{k} = \underset{k}{\mathrm{argmax}}\, T(k)$$

Figure 3 shows an example of a voting table in form of a histogram. There are 100 objects, object 0 is correctly recognized. However, some of the other objects have obtained almost equivalent scores.

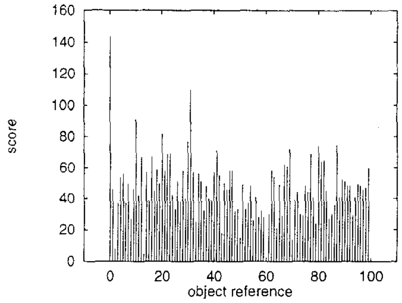

Figure 3: Result of the voting algorithm.

4.3 Indexing into a multi-dimensional hash table

The complexity of the voting algorithm can be controlled by organizing the database as a hash table. Given a vector $\vec{\mathcal{V}}$, it is possible to define a neighbourhood in which all plausible candidates have to lie. An indexing technique allows an implementation by ordering the vectors in a multi-dimensional table. Each level of this multi-dimensional hash table indexes one component of a characterization vector. Thus the hash table realizes a partition of the Euclidean space. The problems associated to such a multi-dimensional hash table are the granularity of this partitioning, and the dimensionality of the hash table. We have observed that a high dimensionality with a coarse granularity at each level is better than a low dimensionality with a fine granularity at each level. This can be easily explained by the fact that high dimensionality allows to better spatially differentiate points. Using high dimensionality, a regular partition is very memory consuming. In our implementation we thus stop further partitioning when a cell contains at most a given number of points.

This indexing technique leads to a very efficient recognition. The database contains 154030 points. However, the mean retrieval recognition time for our database containing 1020 objects on a Sparc 10 Station, is less than 5 seconds. Performance could be further improved by parallelization, as each characterization vector is searched for separately.

4.4 Semi-local constraints

In the presence of noise, a given feature might vote for several models. Having a large number of models or many very similar ones raises the probability that a feature will vote for several models. Califano [2] as well as Rao [4] have suggested that using longer vectors $\vec{\mathcal{V}}$ decreases this probability. Yet it is not practicable to increase the order of derivation of our invariants. Adding invariants computed at different scales, as has been proposed by Rao , would make recognition in a multi-scale context impossible.

A way to decrease the probability of false matches is to use global features. However, global characteristics are sensitive to scene clutters and occlusions. We thus propose to use local shape configurations. For each feature F in the database, the p closest features in the image are selected and we require that at least 50% of the neighbours match.

In order to further increase the recognition rate, a simple geometric constraint has been added. This constraint is based on a direction calculated locally from the signal: the angle α of the gradient. As we suppose the transformation can be locally approximated by a similarity, the difference between these angles has to be locally consistent.

An example for using the geometrical coherence and the semi-local constraints is displayed in figure 4. It gives the score if constraints are applied to the example in figure 3. The score of the object to be recognized is now quite distinctive.

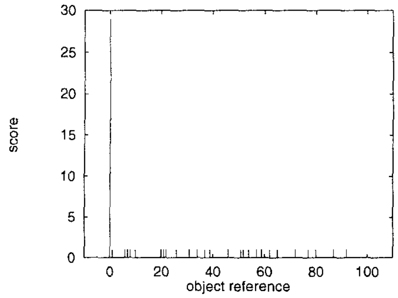

Figure 4: Result of applying geometric and semi-local constraints

Experiments (see section 5) have shown the importance of these constraints. Each constraint decreases the number of ambiguities. A consequence of using the geometric constraint is that the threshold t used to select a model has less importance.

These constraints permit to select only discriminant points of an image. They decrease the number of false matches and reduces the overall number of matches. There are from 18 to 75 points matched during the recognition process depending on the image. This small number of matched points compared to the number of detected keypoints illustrates the rejection of non discriminant points and explains why an images stored in the database with only 20 points can be correctly retrieved even if images of the database are stored with 600 points.

5 Experimental Results

Experiments have been conducted for an image database containing more than 1000 images. They have shown the robustness of the method to image rotation, scale change, partial visibility and scene clutter. Moreover, an object can be recognized in complex scenes and in the presence of occlusion.

5.1 Content of the database

The database used for the experiments presented in the following contains 1020 images. This includes 200 paintings images, 100 aerial images and 720 images of 3D objects. These images are of a wide variety. However, when we consider the paintings images or the aerial images , we observe that some images are very similar. This leads to ambiguities which the recognition method is capable of dealing with. Notice also the small size of the details in case of aerial images.

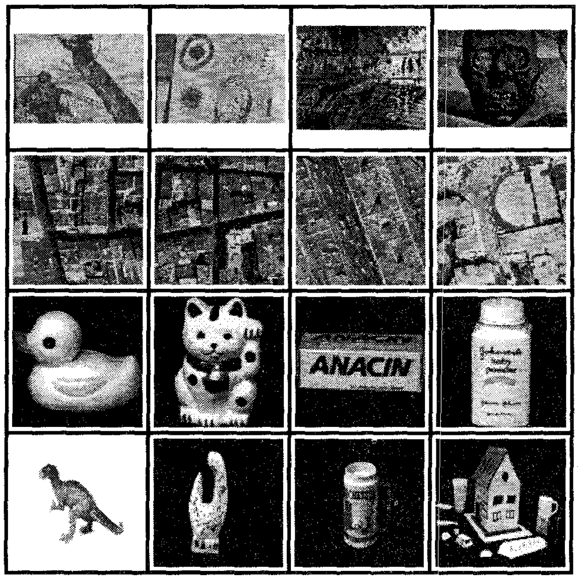

Figure 5: Some images of the database. The database contains more than 1000 images.

In case of 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 which corresponds to paraperspective projection. In order to recognize a 3D object from any viewpoint, it has to be represented by several images which are stored in the database. Each image corresponds to a different aspect of the object and is in the following referred to by model image. The required number of model images depends on the complexity of the object. To obtain a set of model images, equally spaced views of an object were taken (images uniformly distributed over a circle).

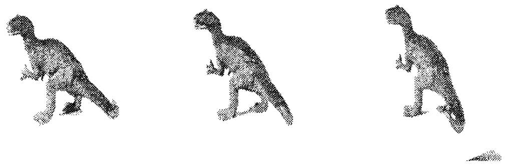

Figure 6: Some model images of the "Dinosaur" contained in the database.

For the "Dinausor" object, (see figure 6), 18 images at 20 degrees increments in pose are sufficient to build the dinosaur's model. For less complex objects, such as the "Abstract Hand" (second image from the left in the fourth row of figure 5), 9 views proved to be sufficient to set up a model base.

5.2 Recognition results

In this section, recognition results are presented. Figure 7 shows the recognition of a painting image in the presence of image rotation and scale change. Moreover, this figure shows that correct recognition is possible if only part of an image is given.

In figure 8 an example of an aerial image is displayed. It shows correct retrieval in case of image rotation and if part of an image is used. However, in case of aerial images we also have to deal with a change in viewpoint, that is a perspective deformation and scene clutter. Notice that buildings appear differently due to changed viewing angles and cars have moved. This example shows that correct retrieval is still possible if only one view has been stored in the base.

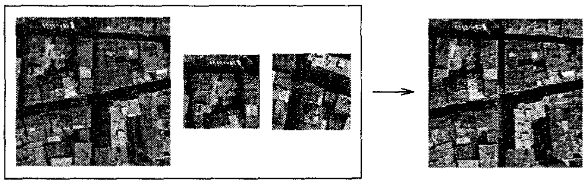

Figure 8: The image on the right is correctly retrieved using any of the images on the left (courtesy of Istar).

Figure 9 shows recognition of a 3D object. The object has been correctly recognized in the presence of rotation, scale change, change in background and occlusion. Notice that the object has not only been recognized correctly, but that the corresponding pose has also been retrieved.

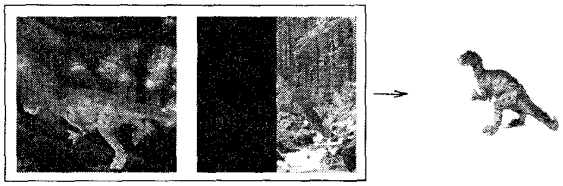

Figure 9: The image on the right is correctly retrieved using any of the image on the left.

The mean recognition rate is of 99.7% for 1180 recognition tests using whole images. Notice that none of the test images is stored in the base. When we consider the different kinds of images separately, we obtain a recognition rate of 100% for the painting images taken under different rotations and scales, 99% for the aerial images taken from different viewpoints (images which are not recognized correctly correspond to an harbor and contain only water) and 99.86% for the 3D objects taken under different viewing angles. We will now show the robustness of the method for different kinds of image transformations, variations of viewpoints for 3D objects and if only part of an image is given.

Invariance to image rotation To test invariance to image rotation, we have taken images of different paintings by rotating the camera around its optical axis. This is possible via a special mechanism of our lenses. Figure 10 shows some rotated images for one of the paintings, the "Sanja" painting. The right most image is the one contained in the base. The recognition rate obtained is 100% for 40 different rotations equally distributed over a circle. This experiment shows that the characterization is completely invariant to image rotation. It is thus not necessary to store in the database more than one image for different rotations.

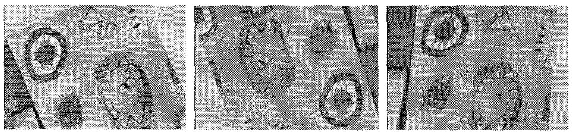

Figure 10: Image rotations of the "Sanja" painting. The right most image is the one stored in the database. All other images have been correctly recognized.

Robustness to scale change To test robustness to scale changes, we have used a zoom lens to take several images of an object with different scale factors. Figure 11 shows some scaled images for one of the paintings, the "Vangogh" painting. Using a multi-scale approach, the recognition rate attains a score of 100% up to a scale change of 2.2.

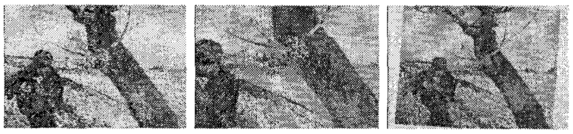

Figure 11: Scale changes of the "sower" painting. The right most image is the one stored in the database. All other images have been correctly recognized using a multi-scale approach.

Robustness to viewpoint variations for 3D objects To test the robustness to a change of the viewing angle, test images are taken at different angles than the images stored in the base. For each 3D object, test images have been taken at 20 degrees difference in viewing angle. The viewing angles of the test images lie in the middle of two images stored in the base which are spaced by 20 degrees. The recognition rate obtained is of 99,86%.

Robustness in case of part of an image To test robustness being given part of an image, about 100

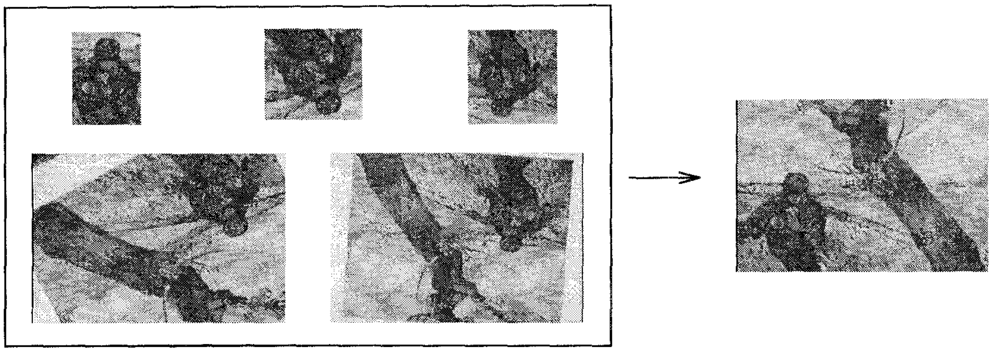

Figure 7: The image on the right is correctly retrieved using any of the images on the left.

parts have been extracted for painting images, see figure 12. These parts are sufficient to recognize the corresponding paintings correctly. Considering the size of our database, this can be explained only by the fact that points are very discriminant.

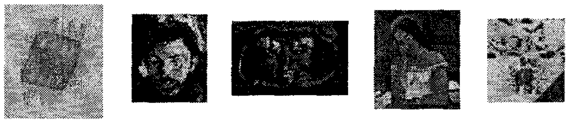

Figure 12: Parts of paintings.

One possible application of the recognition of a part is to find a small object in a big image, for example to find a hotel on a city map. Experiments have shown good results for this task.

6 Conclusion

This paper has shown that the differential greylevel invariants introduced by Kœnderink efficiently characterize points. These invariants describe the image locally. They also present the advantage of being continuous. Used in a multi-scale approach, they allow to derive a method robust to scale changes and hence provide a description of points robust to the group of image similarity transformations. As automatically detected keypoints are characteristics of patterns, invariants calculated at keypoints can be used for indexing 2D greylevel patterns. A voting algorithm in a multi-dimensional hash table permits then to retrieve images. However, blindly voting on individual invariants is not sufficient to guarantee the correctness of the answer in database indexing. It is then crucial to introduce a semi-local coherence between these identifications. This increases the recognition rate. By adding a measure of geometrical coherence a recognition rate of at least 99% is then attained. Experiments were conducted on a database containing 1020 images: paintings, aerial parts of cities and 3D objects. It has been shown that even small parts of images can be recognized. This is due to the fact that the proposed characterization is very discriminant. This has been shown by the small number of points used for recognition.

However, this method is limited by the robustness of the keypoint detector to scale changes. The next step is therefore to propose a detector robust to such changes. Another possible extension of our work is to include affine deformations. The problem is not only to determine affine invariants, but mainly to provide an algorithmic framework within which to allow affine variations.

7 Acknowledgements

Cordelia Schmid was supported by the European Community within the framework of the Human Capital and Mobility program.

References

[1] C. Bauckhage and C. Schmid. Evaluation of keypoint detectors. Technical report, 1996. To appear.

[2] A. Califano and R. Mohan. Multidimensional indexing for recognizing visual shapes. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 16(4): 373–392, April 1994.

[3] J. J. Koenderink and A. J. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375, 1987.

[4] R.P.N. Rao and D.H. Ballard. Object indexing using an iconic sparse distributed memory. In *Proceedings of the 5th International Conference on Computer Vision, Cambridge, Massachusetts, USA*, pages 24–31, 1995.

[5] B.M. ter Haar Romeny. *Geometry-Driven Diffusion in Computer Vision.* Kluwer Academic Publishers, 1994.

[6] C. Schmid and R. Mohr. Matching by local invariants. Technical report, INRIA, August 1995.

[7] C. Schmid and R. Mohr. Object recognition using local characterization and semi-local constraints. Extended version to CVPR'96, ftp://ftp.imag.fr/pub/MOVI/publications/ Schmid_cvpr96.ps.gz.