

Combining Local Recognition Methods for Better Image Recognition

Bart Lamiroy* Patrick Gros† Sylvaine Picard

INRIA

CNRS

DGA

MOVI – GRAVIR‡ – IMAG – INRIA RHÔNE-ALPES

ZIRST – 655, Av. de l'Europe

Montbonnot – FRANCE

Abstract

In this paper we propose a comprehensive framework that allows existing local appearance methods to collaborate in order to overcome their mutual drawbacks. Our approach tends to use the best suited local descriptors for a recognition task, and is capable of combining evidence of different methods in the case where no clearly superior type of descriptor exists. We achieve this collaboration by locally matching geometric configurations and let each match contribute to the computation of the apparent motion between a model image and the unknown query image. We show in this paper that, if we have a set of local methods conforming to a small set of conditions, they can share information about evidence of objects in a scene. This shared evidence results in recognition performances that lie beyond the capacities of any of the currently used individual methods.

Introduction and Motivation

Many scientific contributions in object and image recognition techniques are related to matching of primitives (see for instance [1, 8] for general tools with nice applications). However, they are not adapted for large sets of images, since they require a sequential image by image comparisons. Indexing and appearance based recognition methods for use with large sets of images were therefore introduced by initially approaching image recognition in a global way. They have been proven to be very efficient ([10] *e.g.*), but unsuited for use in a cluttered environment. Therefore, a large number of local, appearance based recognition methods have been developed [7, 11, 13]. They can be roughly divided in two main approaches. The one kind consists of calculating grey-level descriptors that capture the local signature of the image signal [11, 13, 14]. They usually get excellent recognition rates on complex textured images, but have difficulties encoding the general object geometry. The other kind of approach clearly takes its distance from the image signal by using more geometric information as a basis for the local descriptors [5, 7]. They present interesting recognition results, but tend to be less performant than the previously presented techniques.

* Bart Lamiroy is now with ISA at LORIA, Nancy-France. Bart.Lamiroy@loria.fr

† Patrick Gros is now with VISTA at IRISA, Rennes-France. Patrick.Gros@irisa.fr

‡ GRAVIR is a joint research programme between CNRS, INPG and UJF.

All local methods rely on a global consensus for finding the right answer. It can be a simple voting algorithm (sometimes quite similar to an extended Hough transform, or a more complex, improved approach, based on probabilities [12]). Methods based on statistics and probabilities proved to be of a limited impact if the descriptors used are very discriminant [9].

This paper considers a way of bringing together the two major types of local approaches mentioned above, taking advantage of either methods' strengths, avoiding their weaknesses and possibly be even more efficient in domains that the individual models can't address. The basic idea of this paper is founded on two main observations.

1. The first category of methods, relying in purely signal based descriptors, severely lacks a geometrical structure on order to evolve to a more generic recognition paradigm. This is mainly due to the fact that they simply match local signal patterns without taking into account more abstract shape patterns.
2. The other category of methods, using geometry, obtains a good recognition rate on structured images where geometry encodes the implicit morphology of the scene [3]. However, they strongly depend on a valid image segmentation and extraction of significant geometric cues, difficult to obtain with noisy or textured images.

As an example of what we want to obtain with our approach, Figure 4 shows images to be matched in a context of many car engines as reference images. Notice the specular highlights and the occlusions. The previously enumerated individual techniques fail on these kinds of images, but we show that a combination of them will be successful. This paper provides a quite simple framework for engineering efficient combinations of existing techniques.

The outline of this paper is as follows. First we shall briefly recall the underlying notions related to geometric coherence, and detail how this can lead to a cooperation between methods. Next we shall show that this cooperation naturally leads to the creation of new types of local descriptors. Finally, before concluding, we develop a complete example of recognition with two existing methods.

1 Combining Methods

In this section we describe how we obtain the framework allowing different methods to share information. It is based on existing work by Gros *et al.* [5, 7]. We shall briefly outline its principle, before explaining how we adapt it to be used by other methods.

1.1 Geometric Coherence

Exposing the whole matching and recognition techniques in [5, 7] is of no use. One of the ideas is to achieve recognition in two distinct phases. The first being a rough matching, based on image descriptors, knowing that the set of obtained correspondences contains a high percentage of mismatches and incoherent noise. The second phase consists of selecting the correct matches. This selection is obtained by considering the apparent motion defined by each pair of matched local configurations. These configurations contain enough DOF to compute a movement belonging to reasonable family of transforms. In our

case, based on the quasi-invariant approach [2], the considered possible motions belong to the family of similarity transforms.

Each pair of matches defining a local movement, all matches can be represented as points in the parameter space of the considered family of transforms. A clustering method or Hough transform is then used to select the predominant apparent motion between the two images and guarantees the geometric coherence of the retained matches. In this context, each of the initial matches is considered as a vote in a parameter space. This search for the most coherent global movement is referred to as the *geometric coherence* paradigm.

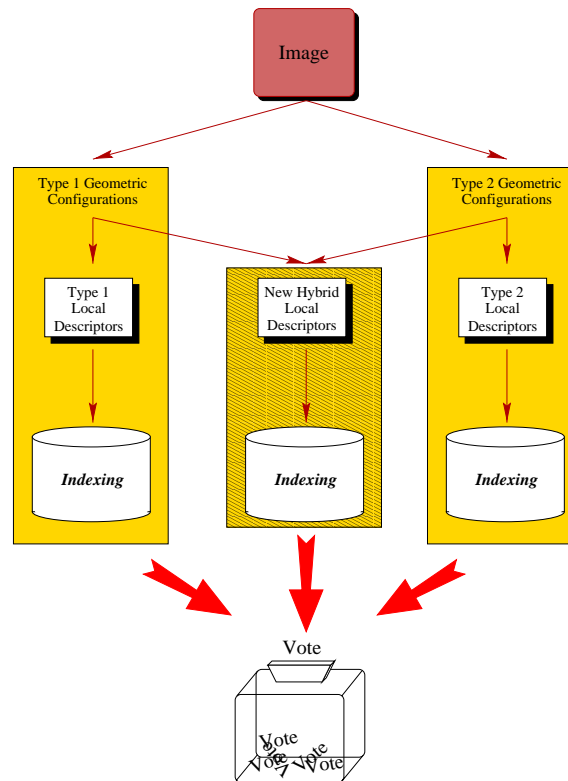


Figure 1: Collaborating Local Methods: different methods (in yellow) compute descriptors on local configurations, extracted from an image. The descriptors are matched with recorded values of known images. These matches give rise to geometric coherence votes in a shared vote space. It is also possible to compute new, combined descriptors (greyed column in the middle).

1.2 Using a Shared Vote Space

In order to combine different kinds of descriptors, *e.g.* angles from [5], grey-level invariants from [14] and local histograms from [13], we are not aiming to construct a super-descriptor combining all approaches. This would defeat our goal of having recognition

methods profit from each other strengths and overcome their mutual weaknesses (and would probably be impossible to realize anyway).

However, we shall try to have all methods conform to the principle of geometric coherence. We then can easily adapt the recognition algorithms to make them collaborate. In what follows we make a clean distinction between configurations or supports on the one side, and descriptors on the other side. The former refer to the physical image pixel set (to which we can attach a geometric interpretation : point, segment, region) on which the latter are computed (angles, length-ratios, intensity derivatives, *etc.*).

Collaboration between different methods is represented in Figure 1 and consists in two main phases (reading the figure top-to-bottom):

1. (in yellow, on the left and right hand sides) different local methods extract the supports for their descriptors from the image to recognize. By matching their descriptors, they find candidate matches between the known images and the unknown image (generally through an indexing scheme).
2. (bottom) instead of using their proper selection schemes for filtering the correct matches, all methods use the geometric information of the matched supports to impose a geometric coherence by voting in a common Hough-transform voting space. The match candidates having contributed to the best fitting apparent motion will be selected as final matches.

There is an interesting side effect to our approach, since we can have *new geometric configurations* emerging from the existing descriptor supports (greyed, middle column in Figure 1). This aims to combine the descriptive power of all approaches to obtain matches in cases where they individually fail. We can create new descriptors from a combination of existing configurations, which again express votes in the same vote space. This particular extension will be described more in detail in section 3.

1.3 Recovering Apparent Motion

It is the notion of “*configuration*” that allows the computation of a local apparent motion between two images when a match exists. In order to make methods collaborate and express their votes in the same space, we need to formalize this concept as well as the way of defining the local apparent motion.

Local descriptors are usually based on values that invariant under a given family of transforms (*e.g.* angles based on connected line segments [7], or second order gradient moments [13, 14]). By definition, the number of degrees of freedom obtained from a match of local configurations is sufficient to compute a local transform.

In our case, a similarity transform is fully determined by four parameters (t_x, t_y, α, σ); two for translation, t_x and t_y , one for rotation α and one for scale σ . We therefore need at least 4 DOF in order to compute a local transform between configurations, which is the case as soon as they consist of at least two points. Unfortunately, several methods [13, 14] are based on simple points in the image, thus lacking the required number of DOF. We propose, in that case, to calculate local, variant (as opposed to *invariant*) measures increasing the number of DOF. For instance, the rotation angle α can be retrieved by locally measuring the *variant* gradient orientation value. The gradient orientation varies with rotation and thus introduces the needed supplementary DOF. An excellent way of dealing with scale is described in [4].

2 Collaborating Methods

We now have a complete framework covering all local recognition methods. The fact that they can collaborate by expressing votes in a shared vote space allows us to develop a system that naturally selects the most appropriate descriptors for recognition. This section shows how.

In our case the shared vote space is the four dimensional similarity transform Hough space. All methods express hypotheses of plausible apparent motions, and the accumulation of coherent votes allows us to select the correct matches. Since this is not easily represented graphically, we use Figure 2, which represents a 2D voting space, for sake of clarity. Figure 2 represent two cases of voting behaviour that can occur. It shows the voting result two methods working on identical images.

On the left hand side, the first method finds a correct cluster, but the second method does not really succeed in finding a clear one, and ends up determining a wrong accumulation point. Combining the votes of both methods results in a clear density cluster, allowing to discard the solution found by the second method. On the right hand side, nei-

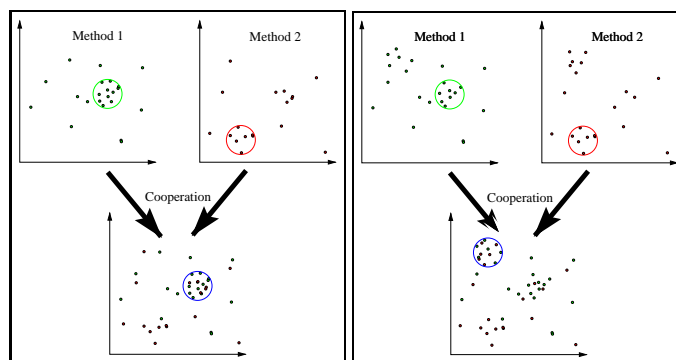


Figure 2: Collaborating methods sharing votes. *Left*, natural selection of best method. *Right*, correction of erroneous recognition.

ther method initially finds the correct transformation parameters. However, collaborating, they create the emergence of a new density cluster. This is exactly what happens in our four dimensional case.

The following setup shows an example of this. We dispose of 12 model images, and we present 120 query images to this image base, 10 instances for each model on average. We first present the subset of query images corresponding to the first model image, then the subset corresponding to the second model image, *etc*. By representing the recognition results on a graph and by putting query numbers on the x -axis and the model responses on the y -axis, we should observe (for a perfect recognition algorithm) a step graph. The first 2 graphs of Figure 3 show the obtained recognition results for geometric invariants based on either Z or Y formed segment configurations [7]. The third graph shows the recognition results for configurations based on two segments and a point called SSP (SSP configurations will be explained in section 4.4). The fourth graph represents the recognition results of the three methods collaborating.

The first remark we can make is that global recognition is far better than that of the individual methods. A more interesting result is the behavior of the three individual meth-

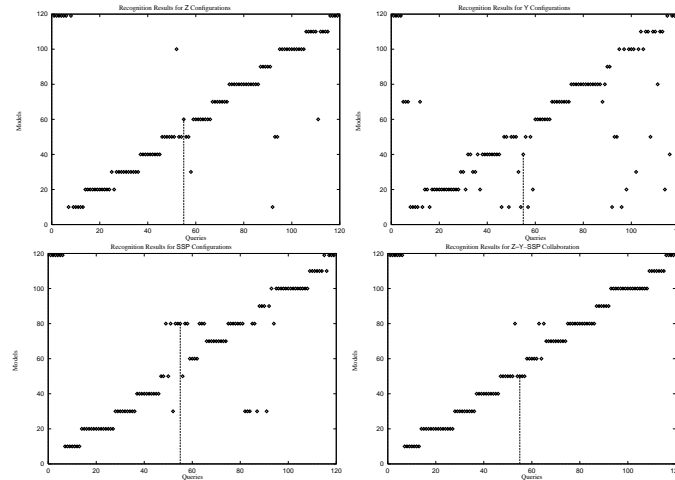


Figure 3: Collaboration between Z, Y and Segment-Segment-Point (SSP) configurations.

ods *vs.* that of the collaboration of the three. For image number 55, for instance, each individual method gives an erroneous answer (model 6 for the Z configurations, model 4 for the Y configurations and model 8 for the SSP configuration) while the collaborative approach finds the correct answer.

We can therefore conclude that cooperation by vote sharing conveys more information than just the sum of information contained in the results of individual methods.

3 Introducing New Descriptors

As shown in Figure 1, new, hybrid descriptors can be used to enforce cooperation between methods when both participants fail to find a good density cluster in transformation space. The motivation behind this is that, when a local method cannot cope with a given image it is probably due to one of the following reasons: either the extracted configurations are unreliable (loss in precision, repeatability or noisy signal data), either the computed descriptors cease to contain the pertinent information needed for matching. In either case, the computed descriptors are of no use, leaving only the geometric data as a plausible alternative.

By combining the most robust information contained in the configurations of both collaborating methods and by computing new, independent descriptors on them, we extract the remaining, unused information to recover from the failure of the latter. Since the local configurations used by different methods vary greatly we cannot propose a general combination algorithm. We shall give a detailed analysis of a concrete collaboration at the end of this paper, but two general constraints can yet be established.

- Given that each of the intervening configurations contains enough DOF to compute an apparent motion, a combination of them has a sufficient high number of DOF to be able to introduce geometric quasi-invariants.
- If the combined configurations are sufficiently rich, it may be interesting to delib-

erately reduce their number of DOF in exchange for a more robust characterization of the resulting configuration.

These points shall be made clear in the next section, as we will be detailing a concrete example of collaborating methods.

4 Complete Example of Collaboration

This section examines in detail the integration of two local recognition methods. One, purely geometrical, based on quasi-invariants, developed in [7]. An other, developed in [14, 4] based on grey-level luminance invariants. We shall quickly detail both methods and then give a detailed step by step analysis on how cooperation between the methods was established.

4.1 Segments and Quasi-Invariants

The first method we use is based on the indexing on quasi-invariants [7], calculated on configurations formed by connected segments. It has a very good recognition rate on neatly structured images where line segments are abundant. As soon as image line segmentation fails to capture image semantics, the recognition quality of the method rapidly declines. Its major advantage resides in its geometric approach through which it is able to capture the morphology of the scene, and be totally independent on illumination or texture.

4.2 Luminance and Local Jet

The second method comes from [14]. Its configuration consist of interest points in an image, on which a luminance invariant is calculated, inspired by the *local jet* introduced by Koenderink [6]. It obtains excellent results on complex images, especially those with distinct textured zones. The method fails to integrate forms or geometric image structures, and is therefore too rigid for possible further generalization to more flexible image classes. Another inconvenient is its difficulty to cope with 3D texture and specularities.

It is to note that this approach is not inherently invariant to similarity transforms. In order to absorb scale change, the author uses multi-scale descriptors that need to be checked over different support sizes.

4.3 Integrating Votes

The segment based method already conforms to the geometric coherence paradigm. We still need to compute the translation, rotation and scale parameters from one matched pair of interest points in order to compute the apparent motion for the second method. Translation is trivial. We obtain the rotation factor by computing the gradient vector in both points of a matched pair, and by taking the difference between the obtained orientation angles. Scale information is obtained by integrating the multi-scale matching described in [4].

At this point the two considered methods are able to express their knowledge in a same vote space. A first step towards cooperation has been made.

4.4 Hybrid Configurations

Let's now consider what happens if neither of the methods is capable of recognizing a given image. This means that the first method failed to extract valid line configurations, and that the illumination changes or 3D effects were too important for the second. In an attempt to retrieve the lost information, we introduce two new configurations that will allow us to compute quasi-invariants. We hope that those new descriptors will contain sufficient scene information to proceed to a correct recognition. By keeping only the more robust and reproducible parts of the existing configurations where segmentation is concerned, we attempt to reduce influence of noise in the image. Therefore, the line segments of the first method will be reduced to the line they define, discarding the end points. The points of the second method are kept as geometrical entities, without any adjoined grey-level information. By grouping two lines and a point (this is the previously mentioned Segment–Segment–Point (SSP) configuration), or two points and a line, we then obtain configurations containing enough DOF to compute a local apparent motion and 2 independent quasi-invariants. The quasi-invariants are computed from pure geometrical information. We use the distance ratio $\frac{d_1}{d_2}$ and the angle α between both lines as descriptors.

In the general case, there is no obligation to taking geometry based descriptors. As a matter of fact, any invariant information of either of the two intervening methods may be combined. In our case, however, the first method did not contain any luminance or grey-level based invariants, and was based on angles and length ratios, and the second method was not readily extendible to line configurations. The only common factor left over was geometry.

4.5 Experimental Results

The following experience shows a sample of images where individual methods were incapable of performing a correct recognition. Our database of known images consists in 9 views of different car engines. We presented a series of other views of the same engines (taken from a different viewpoint) to our system, and obtained a successful identification of the images.

Figure 4 shows the kind of queries we are capable of treating. The model base is represented on the right hand side, while the correctly identified query images are shown on the left hand side. We observe that we can allow 10° to 20° viewpoint changes without loss of recognition quality.

If we compare the ranks of the first and second model having obtained the highest (and second highest) number of votes corresponding to the queries represented in Figure 4, we observe that the second choice attains on average 65% of the voting score of the first choice, with a standard deviation of 15%. Selection of the best model is therefore without ambiguity in most cases.

Conclusion

In this paper, we have presented a general framework covering a wide range of local recognition models. We have shown that, by introducing the concept of geometric coherence, existing models can share recognition information. Sharing this information en-

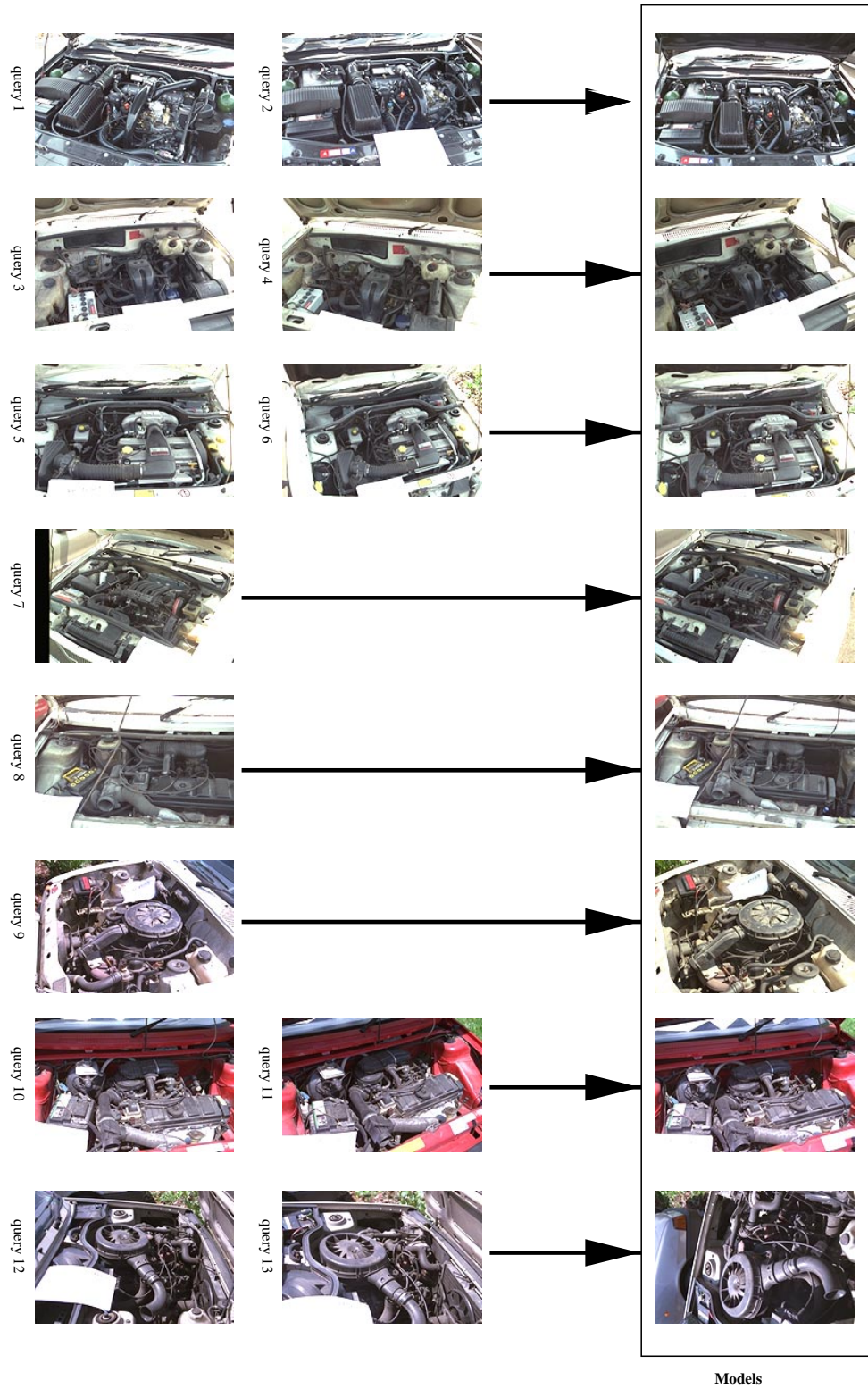


Figure 4: Successful queries. The corresponding models are represented on the right.

hances the chances of recognition beyond those of the individual methods. Furthermore we have shown that this collaboration can give rise to new local descriptors allowing further enhancement of recognition results. We have shown the validity of our approach of a series of very difficult images consisting of views of car engines.

Extensions of this work will most certainly integrate more signal based hybrid descriptors. The hybrid descriptors presented in this paper were purely geometric, but luminance based invariants would probably render our approach even more robust and flexible.

References

- [1] N. Ayache and O.D. Faugeras. HYPER: a new approach for the recognition and positioning of 2D objects. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8(1):44–54, 1986.
- [2] T.O. Binford and T.S. Levitt. Quasi-invariants: Theory and exploitation. In *Proceedings of DARPA Image Understanding Workshop*, pp. 819–829, 1993.
- [3] S. Carlsson. Combinatorial geometry for shape representation and indexing. In J. Ponce, A. Zisserman, and M. Hebert, eds., *Proc. of the ECCV'96 Intl. Workshop on Object Representation in Computer Vision, Cambridge, England*, Lecture Notes in Computer Science, pp. 53–78. Springer-Verlag, April 1996.
- [4] Y. Dufournaud, C. Schmid, and R. Horaud. Matching images with different resolutions. In *Proc. of the Conf. on Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, USA*, June 2000.
- [5] P. Gros, O. Bournez, and E. Boyer. Using local planar geometric invariants to match and model images of line segments. *Computer Vision and Image Understanding*, 69(2):135–155, 1998.
- [6] J.J. Koenderink and A.J. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375, 1987.
- [7] B. Lamiroy and P. Gros. Rapid object indexing and recognition using enhanced geometric hashing. In *Proc. of the 4th Eur. Conf. on Computer Vision, Cambridge, England*, volume 1, pp. 59–70, April 1996.
- [8] D.G. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, Norwell, Massachusetts, 1985.
- [9] R. Mohr, S. Picard, and C. Schmid. Bayesian decision versus voting for image retrieval. In *Proc. of the 7th Intl. Conf. on Computer Analysis of Images and Patterns, Kiel, Germany*, pp. 376–383, 1997.
- [10] H. Murase and S.K. Nayar. Visual learning and recognition of 3D objects from appearance. *International Journal of Computer Vision*, 14:5–24, 1995.
- [11] R.C. Nelson and A. Selinger. A cubist approach to object recognition. In *Proc. of the 6th Intl. Conf. on Computer Vision, Bombay, India*, pp. 614–621, 1998.
- [12] C.F. Olson. Probabilistic indexing for object recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17(5):518–522, May 1995.
- [13] B. Schiele and J.L. Crowley. Object recognition using multidimensional receptive field histograms. In *Proc. of the 4th Eur. Conf. on Computer Vision, Cambridge, England*, pp. 610–619, 1996.
- [14] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(5):530–534, May 1997.