To be published to the journal Robotics and Autonomous Systems
in summer 99

# Local Appearance Space for Recognition of Navigation Landmarks

Vincent Colin de Verdière [a] James L. Crowley [a]

[a] *Project PRIMA - Lab. GRAVIR - IMAG, INRIA Rhône-Alpes*
*655, avenue de l'Europe, 38330 - Montbonnot Saint Martin, FRANCE*

**Abstract**

This paper presents a technique for visual recognition in which the appearance of objects is represented by families of surfaces in a local appearance space. An orthogonal family of local appearance descriptors is obtained by applying principal components analysis to small image windows. These descriptors define the axes of a local appearance space. Each local neighborhood of an image projects to a point in this space. By projecting the set of all neighborhoods of a certain size which compose an image we obtain a discrete sampling of a surface. Projecting neighborhoods from images taken at different viewing positions gives a family of surfaces which represent the possible local appearances from those viewing directions. In this manner we compose a representation in which an object or a landmark can be identified by directly addressing into the local appearance space.

Visual landmarks (as well as objects) may be recognized by projecting windows from newly acquired images into the descriptors space and associating them nearby surfaces. To reduce memory requirements, we propose an efficient tree-search algorithm for the association of points to surfaces. This algorithm directly produces a list of surfaces sorted by distance from the point which represents the observed neighborhood.

Our results show that in many common situations, a single window is sufficient to obtain the correct recognition. Furthermore, the confidence of a recognition is easily estimated from the distance from the point to the surface, and the uniqueness determined by number of surfaces which lie near the point. Robust recognition is easily obtained by reinforcing matching using multiple windows and their mutual spatial coherence. In this paper we present experimental results in the use of this technique for view-point invariant object recognition. We then consider the extension of the technique to visual landmark recognition.

*Key words:* local appearance, PCA, recognition, landmark

## 1 Introduction

Visual recognition is the problem of determining the identity and position of a physical object or landmark from a projection to an image. This problem is made difficult in most practical situations because of the need to accommodate changes in illumination, viewpoint, viewing distance, internal camera parameters, and object deformations. Accommodation of such variations remains a major research problem in computer vision.

The classical approach in computer vision to visual recognition is based on the idea that 3D structure constitutes the underlying invariance. Therefore, a reconstruction of 3D structure from multiple views and multiple visual cues can provide a basis for recognition which is invariant to viewing conditions and object deformations. However, the construction of the 3D models and matching of models by back-projection has generally been found to be unstable and noisy in a real environment. Such an approach require algorithms which are unreliable and generally entail a very high computational cost and complexity.

In our research we explore an alternative approach to visual recognition based on the projection of the appearance manifold to a local appearance space. Such a local appearance space can be defined by any orthogonal family of local image descriptors, including the raw pixels. In practice, a minimal set of such local descriptors, in which the variance of the projection of image data is maximized, can be obtained by principal components analysis (PCA) of local neighborhoods. A global approach to principal components analysis, in which the method is applied to an entire image, has been shown to provide remarkable recognition rates under precisely controlled imaging situations. Recognition results for faces[TP91], hands [BJ96][MC97] and for objects under variations in viewpoint and lighting [MN95] provide a demonstration that visual recognition can be based directly on projections of the image signal. However, such global approaches are extremely sensitive to object position, size, orientation and background. Overcoming such sensitivities requires techniques for object segmentation and normalization which are generally ad hoc and unreliable. Global methods are also sensitive to partial occlusions which can not be overcome, even by segmentation. To counter these problems, we have investigated local representations for object appearance.

Any size local neighborhood defines an orthogonal space with one dimension per pixel. We call such a space a local appearance space. An image projects to this space as a surface. A family of images projects as a family of surfaces. However, using a sufficiently large set of pixels to capture image information gives a large sparse appearance space. Principal components analysis makes it possible to determine an optimal subspace for local appearance. Principal components of the covariance of neighborhood projections gives an orthog-

onal space in which the axes are ranked by the variance of the projections. Selecting the first N principal components yields a linear sub-space which simultaneously minimizes the number of dimensions and maximizes the variance of projections. The volume of such a space can be tuned to a particular recognition problem by varying the number of dimensions and the quantification applied to each dimension. Such an approach can provide a powerful tool for building visual recognition systems.

The local appearance of an object is captured by a set of images, each of which is decomposed to small overlapping windows. These windows are projected to points in a linear subspace defined by orthogonal local descriptors to form a grid which discretely samples a continuous surface. Observed neighborhoods may be associated to pre-learned images by using the projection as an address into the local appearance space, at a computational cost whose complexity is proportional to the number of dimensions.

The number of dimensions is derived from the number of objects under a sub-linear rule. The memory required to represent a local appearance space can be dramatically reduced by representing the space with a decision tree. While this approach trades computation for memory, the complexity remains linear with the number of dimensions. In this manner, we have been able to build a viewpoint invariant visual recognition system which is invariant to image position and background, and robust to partial occlusions, illumination changes, 2D and 3D camera rotations, and changes in size. This paper reports on experiments to the problem of recognizing visual landmarks for position estimation and navigation. This can be seen as an extension of the work of [MII96] and [AJC97].

The first section of this paper focuses on the definition of the local eigenspace, then the application of this eigenspace to visual recognition. Subsequent chapters report some experimental results.

## 2 Local Eigenspace of Appearance

Visual recognition requires selecting which features will describe the objects to be recognized. Robustness to partial occlusion appears to restrict the choice to local descriptors. A local descriptor can be defined as a scalar value extracted from a small region around a point in an image: a window. A vector of $N$ local descriptors will characterize a window. These descriptors require stability in presence of variations in viewpoint and illumination. It is also desirable that the descriptors be highly discriminant for the patterns to be distinguished and that they be inexpensive to compute.

The computation of a vector of descriptors can be modeled as a projection from the image pixel space to a new space more suitable for recognition. In our technique, this descriptors space is composed of $N$ orthogonal filters. In our experiments, the value $N$ is typically 10, but we will soon perform a systematic variation of this parameter. By using methods based on projection to a linear subspace we inherent the use of an mathematical foundation for analysis of our methods. For example, in a local appearance space, visual recognition is obtained by using the property that close points in the projection space correspond to similar image windows. The confidence in the recognition is given by the value of the distance between points. The uniqueness of a recognition is provided by the number of surfaces which pass near to the point.

The first paragraph of this section presents the computation of the descriptors space using a Principal Components Analysis on image windows. Then, in the next paragraphs, the values of the different parameters of the technique (window size, number $N$ of dimensions and quantification of descriptors) are discussed.

### *2.1 Computing the projection space*

In our approach, we chose to use Principal Components Analysis to compute the descriptors space ($\mathcal{R}$). We wish to stress that the use of PCA is not intrinsic to our approach and that other families of orthogonal operators can be used for such a technique. For example, Rao [RB95] used Gaussian derivatives as local descriptors and Schiele [Sch97] uses Gabor filters and Gaussian derivatives. These analytic descriptors present an important property: it is possible de compute them at any scale or orientation [FA91]. Their drawback is that they are chosen arbitrarily.

Principal Components Analysis is a statistical tool which has some attractive properties in the for data compression and recognition. The Karhunen-Lœve or PCA [Fuk90] changes the space of representation of any data. The new space, sometimes called an eigenspace, is obtained by a linear transformation of the initial space. Its dimensions are defined by pairwise orthogonal eigenvectors. These vectors can be sorted by their eigenvalues which are equal to the variance of the data along the dimensions. Selecting the first few dominant eigenvectors leads to an optimal approximation of the data in the least-square-error sense. This subspace can therefore be used to represent the data as it is an optimally compressed space.

It is relatively easy to apply this technique to local appearance expressed in image neighborhoods. A data vector is an $M \times M$ image window. An energy

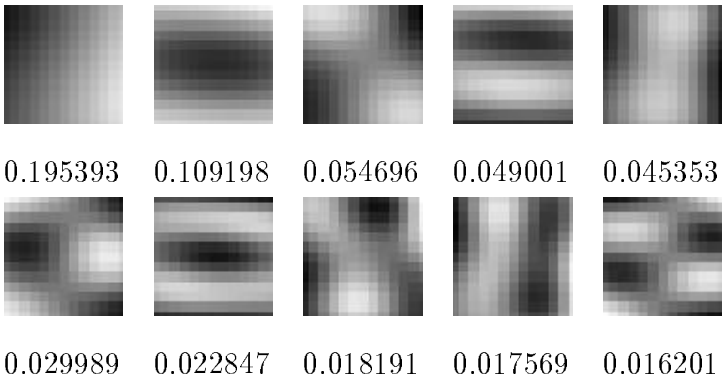


Fig. 1. Descriptors space and eigenvalues

normalization on the windows provides robustness to luminosity variations [Sch97]. A physical object is represented by images from its different appearances. The appearances of an object include several scales and several viewpoints. The set of data is composed by all its overlapping windows from all objects images. Applying the PCA produces a descriptor space $\mathcal{R}$. Distances in these two spaces are equivalent. This leads to the fact that two points close in the PCA space correspond to close windows in the original space. Distances in the subspace provide an accurate approximation of distances in the initial space. The figure 1 presents the 10 first eigenvectors and their associated eigenvalues computed on $13 \times 13$ windows. These vectors were obtained using all possible windows from image dataset extracted from the Columbia database [NNM96].

Objects are likely to be observed under different scales. Robustness to the scale factor requires a multi-scale approach. As experiments show a 10% stability to scale of $9 \times 9$ windows, images need to be derived onto pyramids with a 20% scale change between its successive levels. Two approaches are possible: learning images at every scales or searching unknown images within all scales. While the first approach requires memory, the second requires cost in computation. On the whole, a trade-off can be obtained depending on the application. Currently, our system is only based on the first approach.

### *2.2 Parameters of our approach*

The computation of the descriptor space depends on 3 main parameters: the image windows size, the number of dimensions and the quantification in each dimension.

- Image Windows Size: The objective is to use very small windows so as to be very robust to changes in background and occlusion. Experiments have already shown $9 \times 9$ provides enough information to achieve recognition.

Further experiments (section 4) show the impact a windows size on recognition.

- Number of dimensions: The higher the number of dimensions, the greater the discriminatory power the descriptor vector will have. However, memory requirements are expressed as the number of quanta in each dimension to the power of the number of dimensions. Thus increasing the number of dimensions dramatically increases the amount of memory required. This fact leads to a trade-off between the number of dimensions and the discriminatory power between object classes and appearance variations.
- Quantification of dimensions: The quantification of dimensions generates imprecision in distances calculations. Specifically, signal processing theory shows that quantification acts as Gaussianly distributed random noise. The variance of this noise is related to the logarithm of the number of bits. Another aspect is that it is not relevant to use a quantification which is more accurate than the data noise itself. The descriptors have a certain stability to small changes and the separability of different descriptors should be maintained by the quantification.

The choice and the study of the local descriptors is a key point to our recognition system as they the foundation of our technique. The next session shows the use of these descriptors to objects recognition.

## 3 Application to Objects Recognition

After the choice of local descriptors, some questions remain: where do we apply them on the images ? How can we store these vectors so as to have a fast access and search time ? One window cannot give more information than itself and so we must be able to use multiple windows and spatial constraints criteria between them to achieve high recognition [CC98].

### *3.1 The Local Appearance Grid*

Two main approaches are possible: selecting the windows according to a criteria or selecting windows from a predefined grid.

For the first approach. the selection criteria is also called a detector. Schmid and Mohr [SM96] employ the Harris Detector [HS88] to select some interest points for indexing. Ohba and Ikeuchi [OI96] select windows with a local measure of trackability and a global measure of similarity. Rao and Ballard [RB95] used both approaches: selecting discriminant windows and selecting windows on an object centered circular grid.

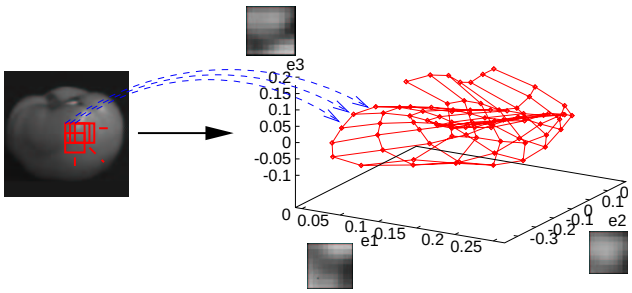


Fig. 2. Representation of an image as a surface in a 3D subspace of $\mathcal{R}$

Techniques based on detectors provide good results, but there is a major drawback on the stability of the detectors. This stability cannot be guaranteed when the viewpoint or the luminosity changes. Two images of the same scene will not necessarily generate corresponding interest points, and therefore matching them becomes difficult. Our approach starts on the opposite viewpoint: at first, no selection or detection of *good* windows is effected. Then, an analysis of the occurrences of the windows inside the database leads to the suppression of non discriminant windows.

In our technique, the training set is composed of all the windows from all the images. A window is projected into a point of the descriptors space $\mathcal{R}$. Images are divided into a grid of $M \times M$ windows. The figure 2 presents an image decomposed in overlapping windows and its corresponding surface in a 3 dimensional subspace of $\mathcal{R}$. There is a one pixel shift between neighboring windows. Then, an image is represented by a surface in the space $\mathcal{R}$. This surface is stored as a discrete grid. As the overlap between two successive windows is very large, unless high discontinuities in the image space, two neighboring windows in an image have close projections in $\mathcal{R}$. Therefore, a continuity property can be defined which holds between most points of the grid which represents a surface on $\mathcal{R}$. A physical object is represented by a set of images and therefore, in the database, by a set of surfaces. This image representation provides the recognition as the projection of an unknown window can be directly searched in $\mathcal{R}$.

### *3.2 Indexing and Searching descriptors*

As shown in the previous paragraphs, our recognition technique needs to store and search $N$ dimensional descriptor vectors. A naive structure would be to store vectors in images (i.e. 2D arrays). This structure provides easy access to proximity criteria during a search but requires an impossible exhaustive search. Therefore, a tree search structure was chosen.

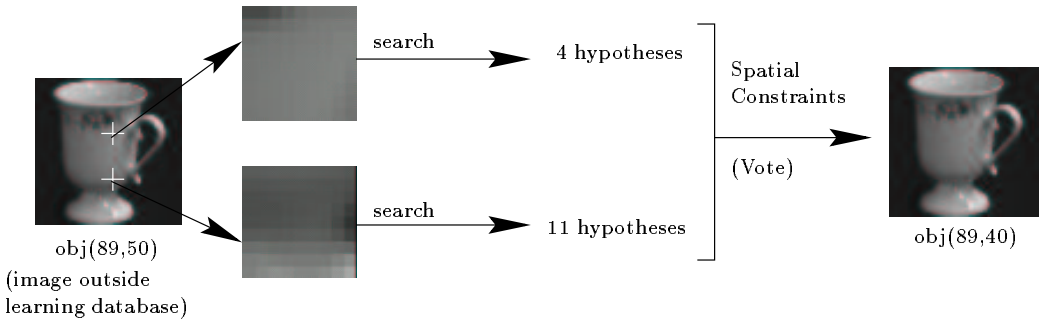


Fig. 3. Example of a two windows search

Each $N$ dimensions are, successively, decomposed into 4 clusters. During the database learning, dimensions are decomposed according to the need to separate the data. This structure provides a very easy addition of elements, an a-posteriori characterization of the discriminant power of vectors in the database (the tree gives directly an histogram of probability of vectors) and an effective searching technique. As vectors are not exact keys, the search consists in finding all vectors in the tree which are inside a sphere centered on the searched vector. Therefore, an exploration of several branches of the tree is required but can be done effectively by approximating distances at each search level to reduce the number of explored branches. Experiments show that this structure is very efficient and that the mean number of searched leaves in a search is around 30 on an images database extracted from the Columbia database [NNM96].

### *3.3 Searching multiple vectors*

As a unique window does not contain enough information for a full recognition, it is possible to increase the recognition by searching multiple windows. A classic technique is to use a voting algorithm based on the idea of k-Nearest-Neighbor rule. Each window provides a list of object hypotheses. An hypothesis is a vote for an object. The object that obtains the largest number of vote is recognized.

Our technique is based on voting on a couple (object, position) rather than just the object. This means that we include spatial constraints between hypotheses in the recognition algorithm. Therefore, incoherent hypotheses do not improve an object's score. Most hypotheses will have a single vote and the few remaining will be classified on the number of votes for the recognition. This technique can be interpreted as a form of Hough Transform. The object recognized corresponds to a major cluster in the object/position space.

Figure 3 presents the example of the recognition of an unknown image (outside the database) by searching two different windows. Each search return a few

hypotheses. An hypothesis $H$ is a vector $H = (Obj, View, X, Y, Distance)^t$. By selecting only the hypotheses which are coherent (same object and position), in this example, the correct match is selected.

The use of a single window for recognition has shown good results in simple scenes containing only the searched object. Then, using more windows appears more robust. In complex scenes, the selected window may correspond to an object from the database or from something else. In this case, the use of multiple windows appears intrinsically essential: in fact, it is required that a window from the search object is selected during the search.

## 4 Recognition Experiments

The technique proposed in the previous sections has been successfully evaluated in some experiments. The first paragraph show results on a large object database, then the second paragraph show some results for robot visual navigation.

### *4.1 Experiments using the Columbia database*

An experimental database has been extracted from the Columbia database [NNM96]. Four images of the 100 objects are indexed in our search database (i.e. 400 images). The descriptors space is composed of 10 dimensions which means that one window is defined by a vector of 10 coordinates. Each image is selected with a 20 (and 10) degrees shift in viewing position (angles 0, 20, 40 and 60). The test images are all the remaining images between the 5 and 65 degrees viewpoints (i.e. 1000 images).

Roughly, our technique provides for a unknown window a sorted list of window hypotheses whose distances with the unknown window are below a threshold. Some window patterns occur very often and generate a very high number of hypotheses. In this case, the window is rejected as non discriminant. In the remaining cases, the right object is recognized at a certain rank in the hypotheses list. A low rank enforces intuitively that the window is recognized. But, as the threshold correspond to the experimental stability of the descriptor vectors, any hypothesis of the list might be the correct one. Figure 4 evaluates recognition obtained by a single random window of the test database under changes in viewpoint, using the COIL benchmark database. Local windows of size $9 \times 9$ were used in this experiment. The abscissa axis presents the rank of recognition of the right object. As intuitively predicted, using more viewpoints leads to higher recognition rate. A 70% recognition rate by a single

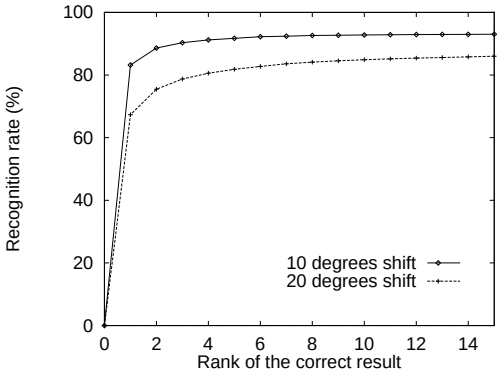


Fig. 4. Recognition Rate by a Single Random Window

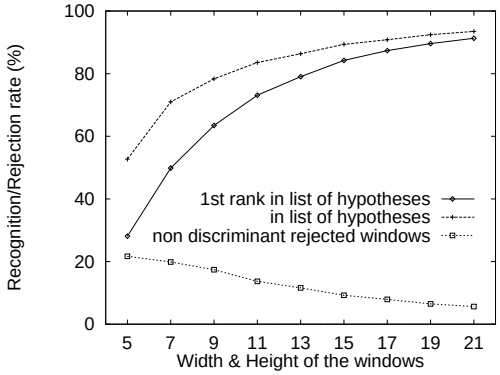


Fig. 5. Evaluation of the window size parameter on recognition

window is very high given that windows were selected randomly. In fact, 10% of the failure case correspond to the rejection of non discriminative windows, therefore the false match rate is very low.

A second experiment evaluates recognition as a function of the window size on figure 5. Only 40 objects were used for this particular experiment. The figure presents two recognition curves which increase with the size of windows and a rejection curve. This last curve show the rate of windows which were dismissed as non discriminant. As predicted, the larger the window are, the greater the recognition rate. However, windows of larger size do not appear to be local features and will be more sensitive to partial occlusion. The larger windows are also far more sensitive to background variations. Therefore, using a window size around $10 \times 10$ appears is a good trade-off: searching multiple windows will compensate for the loss in recognition. Other experiments using windows of size $9 \times 9$ have shown that using multiple windows reject most bad matches and objects are recognized at a rate of 90% when only 10% of objects are visible.

### *4.2 Results on a robot visual navigation*

The robustness of the object recognition technique lead us to consider using it for mobile robot navigation. In fact, a classic navigation technique in controlled environment is to position artificial landmarks that the robot can detect and track in order to follow its path. It is possible to navigate without artificial landmarks. Matsumo [MII96] and Jones [AJC97] succeeded a visual based navigation by a "View Sequenced Route Representation". This technique requires other sensors and an intelligent control system to counter positioning failure due to possible occlusion. Transversal deviation or partial occlusion are likely to happen in case of obstacle avoidance for example when the robot needs to modify its predefined path. Finding a visual landmark will

Fig. 6. Visual tour of the robotic hall

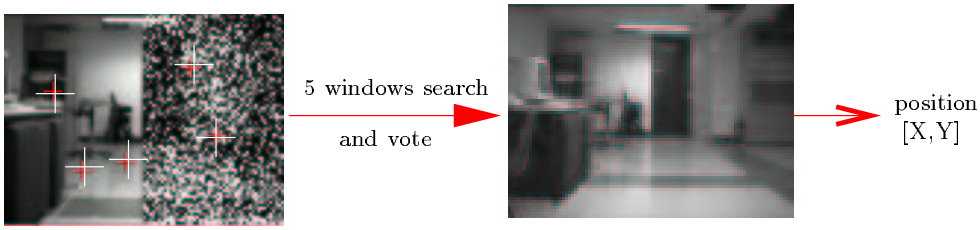


Fig. 7. Example of successful search using 5 windows in a half occluded image

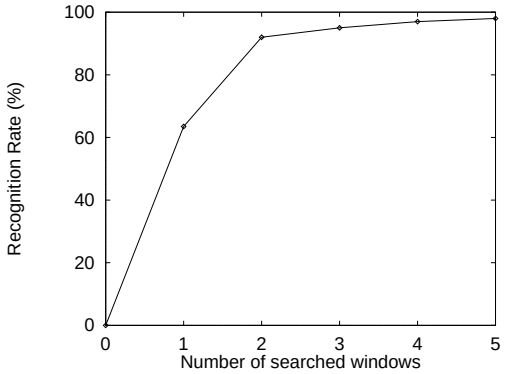


Fig. 8. Using multiple searches increases the recognition rate

sometime fail. Using our local features, we provide with a tool to estimate its position which is robust occlusion and translation. Position estimation is based on scene recognition by our object recognition system.

A visual path (Fig 6 of our lab is defined by a sequence of images [AJC97]. The goal of our system is to detect in a newly acquired image which database image is the closest and therefore get an approximate of the absolute position of the robot. A more accurate position can be obtain by using other techniques as Pourraz in [PC98]. Figure 7 show an example of a partially occluded image which was recognized by searching 5 windows. Half of the image is replaced by random noise. Searching windows centered on the crosses of the image, the vote algorithm selects the correct result. More, generally, Figure 8 illustrates the quality of recognition as a function of the number of used windows. Windows generating too many hypotheses were removed from the test because non discriminant. The set of test images consists in images taken at intermediate viewpoints of the database images.

These results show a near perfect recognition rate by using multiple windows

on non occluded images. 4 windows provides 97% recognition and for the 3% remaining, the correct answer was found at the second or third rank.

## 5 Conclusions

Our experiments indicate that using local appearance surfaces to represent objects is an efficient and robust method of visual recognition. The primary drawback in this representation is its memory requirement. In fact, no a-priori selection of discriminant window is required. An a-posteriori suppression of non-discriminant windows without any loss in recognition has been tested on $9 \times 9$ windows and results in more than 10% reduction of the data structure. The evaluation of the structure redundancy and discriminant power is one of our current research activities.

In our experiments of visual robot navigation, scale was not used for the recognition and was responsible for some of the failure. Some structure are stable on robot path as a corridor for example. In this case, the position information might be accessible only at a specific scale. Other experiments showed that using analytic descriptors such as Gaussian derivatives or Gabor filters provide similar recognition results to PCA filters. Therefore it is possible to use their scalability property to generate a multiscale representation of the scenes. Detecting objects on one of the level of the representation will provide the robot localization.

## References


[AJC97] C.S. Andersen, S.D. Jones, and J.L. Crowley. Appearance Based Processes for Visual Navigation. In *5th International Symposium on Intelligent Robotic Systems, SIRS'97*, pages 227–236, Royal Institute of Technology, Stockholm, Sweden, July 1997.

[BJ96] M.J. Black and A.D. Jepson. EigenTracking: Robust Matching and Tracking of Articulated Objects using a View-Based Representation". In Springer Verlag, editor, *ECCV'96, Fourth European Conference on Computer Vision*, pages 329–342, 1996.

[CC98] V. Colin de Verdière and J.L. Crowley. Visual Recognition using Local Appearance. In *Fifth European Conference on Computer Vision, ECCV'98*, volume 1, pages 640–654, Freiburg, Germany, June 1998.

[FA91] W.T. Freeman and E.H. Adelson. The Design and Use of Steerable Filters. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 13(9):891–906, september 1991.

[Fuk90] K. Fukunaga. *Statistical Pattern Recognition*, chapter Feature Extraction and Linear Mapping for Signal Representation. Academic Press, School of Electrical Engineering, West Lafayette, Indiana, 1990.

[HS88] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. 4th Alvey Vision Conference*, 1988.

[LZ97] A. Lux and B. Zoppis. An Experimental Multi-language Environment for the Development of Intelligent Robot Systems. In *5th International Symposium on Intelligent Robotic Systems, SIRS'97*, pages 169–174, 1997. more informations at `http://www-prima.imag.fr/Ravi/`.

[MC97] J. Martin and J.L. Crowley. An Appearance–Based Approach to Gesture–Recognition. In Alberto Del Bimbo, editor, *International Conference on Image Analysis and Processing*, number 1311 in Lecture Notes in Computer Science, Florence, Italia, Sept. 17–19, 1997. Spriger Verlag.

[MII96] Y. Matsumoto, M. Inaba, and H. Inoue. Visual Navigation using View-Sequenced Route Representation. In *International Conference on Robotics and Automation*, volume 1, pages 83–88. IEEE, 1996.

[MN95] H. Murase and S. K. Nayar. Visual Learning and Recognition of 3D Objects from Appearance. *International Journal of Computer Vision*, 14:5–24, 1995.

[NNM96] S. A. Nene, S. K. Nayar, and H. Murase. Columbia Object Image Library (COIL-100). Technical report, Columbia University, New York, February 1996.

[OI96] K. Ohba and K. Ikeuchi. Recognition of the Multi Specularity Objects for Bin-picking Task. *IROS 96*, 3:1440–1448, 1996.

[PC98] F. Pourraz and J.L. Crowley. Continuity Properties of the Appearance Manifold for Mobile Robots Position Estimation. In *Symposium for Intelligent Robotics Systems, SIRS'98*, 1998.

[RB95] R. P. N. Rao and D. H. Ballard. Object Indexing using an Iconic Sparse Distributed Memory. In *ICCV'95 Fifth International Conference on Computer Vision*, pages 24–31, 1995.

[Sch97] B. Schiele. *Object Recognition using Multidimensional Receptive Field Histograms*. PhD thesis, I.N.P. Grenoble, France, 1997. English translation.

[SM96] C. Schmid and R. Mohr. Combining Grayvalue Invariants with Local Constraints for Object Recognition. In *International Conference on Computer Vision and Pattern Recognition*, 1996.

[TP91] M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

Experiments presented in this article were programmed using the RAVI multi-language environment [LZ97].