Visual Recognition Using Local Appearance

Vincent Colin de Verdière and James L. Crowley

Project PRIMA - Lab. GRAVIR - IMAG
INRIA Rhône-Alpes
655, avenue de l'Europe
38330 - Montbonnot Saint Martin, France
{Vincent.Colin-de-Verdiere, Jim.Crowley}@imag.fr

Abstract. This paper presents a technique for visual recognition in which physical objects are 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 image neighborhoods. The principal components with the largest variance are used to define a space for describing local appearance. The projection of the set of all neighborhoods from an image gives a discrete sampling of a surface in this space. Projecting neighborhoods from images taken at different viewing positions gives a family of surfaces which represent the possible local appearances from those viewing directions.
Recognition is achieved by projecting windows from newly acquired images into the local appearance space and associating them to nearby surfaces. An efficient search tree data structure is used to associate projected points to surfaces. Our results show that in many common situations, a single window is sufficient to obtain the correct recognition. Robust recognition is easily obtained by reinforcing matching using multiple windows and their mutual spatial coherence.

1 Introduction

Visual recognition is the problem of determining the identity of a physical object from an arbitrary viewpoint under arbitrary lighting conditions. Changes in the appearance of the object under variations in lighting and viewpoint make this a very difficult problem in computer vision. The classic approach is based on the idea that the underlying 3D structure is invariant to viewpoint and lighting. Thus recovery of the 3D structure should permit the use of techniques such as back-projection and features matching [6] to match the observed structure to a data base of objects.

While such approaches can be made to work in a controlled laboratory environment, they are known to generate several problems in a real environment. For example, the detection of commonly used image features, such as key-points, edges and lines, tends to be unreliable. Photo-optic effects and image noise contribute to creating feature drop-outs and spurious features. Correcting or compensating for such errors leads to the use of ad-hoc techniques with high computational complexities. In addition, such techniques are generally fragile, and their failure conditions are difficult to predict.

An alternative to 3D reconstruction is to remain in the 2-D image space and to work with measurements of the object appearance. A variety of techniques are possible with this approach. A pure information theory model suggests making an inner product (or sum of squared difference) with the very large set of possible appearances masks for each object [12]. This approach can be studied formally and it can be shown to lead to the smallest possible probability of error. However, except for highly constrained problems, this approach is generally rejected as requiring an impossible number of correlation masks. Sirovitch and Kirby [21] showed that in the case of face recognition, principal components analysis (PCA) of the set of images could be used to generate an orthogonal space in which inner products could be reduced to a neighborhood search in a relatively small number of dimensions. Turk and Pentland [22] refined and popularized this approach for face recognition, greatly enhancing the acceptance of principal components analysis as a vision technique. Murase and Nayar [14] extended this idea by expressing the set of appearances of objects as a trajectory in a PCA space. Black and Jepson[2] demonstrated that the appearance of a hand making a gesture could also be expressed and matched as a trajectory in a PCA space.

All of the above techniques suffer from several important difficulties. It is often noted that projection to a PCA space is sensitive to partial occlusions, although recently, a new projection technique has been developed which operates under occlusion [1]. More fundamentally, the projection of an image to a PCA space depends on the precise position of the object in the image, on the intensity and shape of back-ground regions, and on the intensity and color of illumination. Thus, these techniques require that the object be detected, segmented and normalized in intensity, size and position. Such segmentation and normalization is a very difficult problem, and perhaps unsolvable in the general case.

The technique described in this paper began as a search for a method to describe and match the appearance of objects independent of illumination intensity, background, image position, and occlusions. Our approach is to minimize the effects of background and occlusion by working with as small a neighborhood as possible. Our approach overcomes the problem of image position by mapping the locally connected structures in the image space into surfaces in a space of local appearances. Robustness to changes in illumination intensity are obtained by energy normalization during the projection from image to the appearance space. The result is a technique which reliably produces object hypotheses from a large data base of objects when presented with a very small number of very local neighborhoods from a newly acquired image.

In a sense this technique trades memory for computational cost. Once the data base of appearances has been constructed by an off-line process, the identity of objects in a newly acquired image of a scene can be determined at a very small computational cost. Further more, the technique is composed of linear transformations and distance matching techniques, all of which are formally analyzable. Thus, no ad-hoc tricks are required. The operating conditions and probability of error can be analyzed and predicted.

This approach is similar in spirit to the work of Rao and Ballard [17] who used very large vectors of local measurements as features for pattern recognition. We have taken inspiration from Schiele [19] who recognizes objects in scenes using multi-dimensional statistics of local operators and from Schmid and Mohr [20] who uses a local jet of viewpoint invariant descriptors to detect and match key-points in a view-invariant manner. Another close approach to this problem is the system SEEMORE [13] developed by B.W. Mel which efficiently recognized objects by using a large set of image features.

With our approach, local appearance is described using the appearance of small (9 by 9) neighborhoods. This size was selected for our initial experiments as the smallest size for which the local spatial frequency is not perturbed by the sinc function which is inherent in neighborhood selection. The set of all possible 9 by 9 neighborhoods have an 81 by 81 covariance matrix with 81 principal components. Any (9 by 9) neighborhood can be exactly represented by a projection into this 81 dimensional space. For real images, the variance (principal values) for most of these dimensions is extremely small. Thus it has proved possible to use a much smaller number of principal components to define an orthogonal subspace for expressing local appearance.

The experiments described below use an orthogonal space defined by the first principal components (10) from a very large collection of 9 by 9 neighborhoods. An image is modeled in this space as a surface. A collection of images from different viewpoints and different lighting conditions is represented as a discrete sampling of a manifold in this 10 dimensional space. The number of dimensions of the manifold structure is determined by the number of dimensions variation of appearance.

The recognition criterion for local appearance is the distance between an observed window and the surface points of the manifold. Thus such a data-base structure can potentially be searched by table lookup. In reality, the appearance space in which the manifolds are expressed is both very large and very sparse. Thus for a fixed memory cost, a very large gain in the number of objects (manifolds) and in the variations of appearance (dimensions in the manifold structure) can be obtained by using an appropriate search algorithm.

In many of our experiments, a single 9 by 9 neighborhood proves sufficient to recall the correct objects, the correct viewpoint and the correct position on the objects surface. However, many ambiguous neighborhoods, such as uniform regions and high-contrast also exist. Thus we obtain a high robustness by using spatial coherence criteria between the windows.

The problem of recognition with local features may be divided into three independent sub-problems. The first sub-problem consists in choosing a set of local descriptors. The experiments described here are based on PCA of small neighborhoods, but we believe that other orthogonal families of neighborhood descriptors such as Gabor filters could also be used. The second sub-problem is to define where to apply the descriptors: on selected interest points or on a predefined grid. The third sub-problem is to define a data structure to store

and search these descriptors. The following sections address each of these subproblems.

2 Local Eigenspaces of Appearance

The first part in recognition is to choose which local descriptors of images are to be used. A local descriptor is a scalar value computed from a small region around a point in an image: a window. A window can be characterized by a vector of N local descriptors. These descriptors should have some certain required properties such as stability to change of viewpoint and illumination, stability under partial occlusion. It is also desirable that the descriptors be highly discriminate for the patterns to be distinguished and that they be inexpensive to compute.

Formally, 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 descriptor space is composed of N orthogonal filters (in the experiments below, N equals 10). This space provides the desirable property that an image window which is a vector in a $M \times M$ space is represented by a single vector in the descriptor space with $N \ll M$. 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 first paragraph of this section presents the computation of the descriptor 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.

Computing the projection space: In our approach, we chose to use a Principal Components Analysis to compute the descriptors space (P). We wish to stress that the use of PCA is not intrinsic to our approach. Other families of orthogonal operators can be used for such a technique. For example, Rao [17] used Gaussian derivatives as local descriptors. Schiele [18] uses Gabor filters and Gaussian derivatives. These analytic descriptors present an important property: it is possible de compute them at any scale or orientation [7]. Their drawback is that they are chosen arbitrary.

PCA has certain properties which make it convenient to use as a starting point. The main advantage of these vectors is that are derived directly from the image database. The PCA descriptors provide a basis for which the image data will project with a maximum variance for a minimum number of dimensions. Selecting the first eigenvectors is optimal in the least-square-error sense for representing the data. However this does not mean that this choice is also optimal for recognition.

Using the Karhunen-Lœve or PCA [8], it is possible to compute an orthogonal space from a list of data vectors. This space is composed of as many dimensions as the original data, but it is easy to derive from it an optimal subspace on the reconstruction viewpoint. In this new space, distances have the same meaning

as in the image space. This leads to the fact that two points close in PCA space correspond to close windows in the original space. Distances in the subspace provide an accurate approximation of distances in the original space. In our case, vectors are $M \times M$ image windows.

A physical object is represented by images from its different appearances. Thetraining set is composed by all overlapping windows from all images of the objetcs toe recognized. The application of PCA on this set provides M^2 eigenvectors e_i with their associated eigenvalues λ_i. The eigenvectors e_i form an orthogonal set. For a complete and accurate representation of the training set, all M^2 eigenvectors should be maintained. However, the amount of information (defined as the ability to discriminate) captured by the projection onto an eigenvector is related to the variance (the eigenvalue) of the projection of all neighborhoods on that eigenvector. Thus it is possible to use the eigenvalues to sort and select the dominant eigenvectors to compose the descriptors space P.

Distances in P are an approximation of correlations in the image space. Correlation techniques [12] are efficient for recognition of a small number of features, such as in the case of tracking an object. The projection into the eigenspace an association between an observed window and a prelearned window to be determined by direct addressing. In a sense, this is correlation without search. This the approach is easily adapted to real time.

As an example, a 400 image database extracted from the Columbia database [15] provides a space P for windows of size 9×9. The first 10 eigenvectors were selected to define this projection space P. Figure 1 presents the 10 first eigenvectors and their associated eigenvalues.

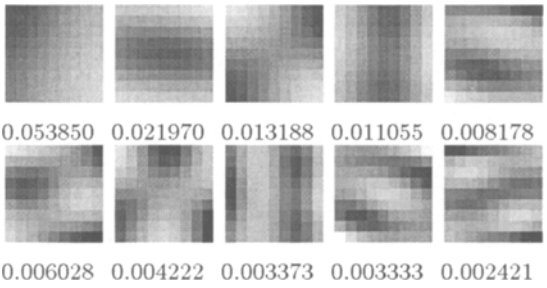

Fig. 1. Descriptors space and eigenvalues

In the space P, an image window W is associated to a single point A and close windows are associated to close points. A is obtained by projecting W into P, i.e. by computing a dot product between the window W and each eigenvectors e_i. In order to increase the stability to luminosity changes, a normalized dot

product is used:

$$A_i =< W, e_i >= \frac{\sum_{j=1}^{M^2} W_j . e_{i,j}}{\sqrt{\sum_{i=j}^{M^2} W_j^2} \sqrt{\sum_{i=j}^{M^2} e_j^2}}$$

Division by the energy of windows provides stability of A against the changes of the intensity of the light. The choice of this energy normalization rather then a mean and variance normalization or a max-min normalization [3] is motivated by Schiele's study [18]. A better invariance against light change can also be obtained by the use of color as in [9,4].

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

- Image Window Size: The objective is to use very small windows so as to be robust to changes in background and occlusion. Previous work shows that 9×9 is the minimum size size stable enough to image noise. Good recognition results have confirmed the validity of this windows size. Systematic exploration of possible window sizes remains to be done. We note that 9 by 9 neighborhoods extracted from a pyramid can be used to capture the multi-scale structure of appearance.
- 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.
  Using one dimension, it is possible to separate at best k different vectors. The parameter k depends on the stability of the descriptors to small changes in the images. Experimentally, it is possible to evaluate a distance threshold th which define the maximum distance between two vectors corresponding to the same window. Vectors values are in the interval $[0; 1]$, so we have $k = \frac{1}{th}$. For N dimensions and L_∞ distance, the maximum number of different descriptors vectors is $k_N = k^N = \frac{1}{th^N}$. For $th = 0.1$ and $N = 10$, it leads to $k_N = 10$ billions which is a lot more than the number currently stored cells. This calculus gives an idea of the storage capacity of the data structure. However, in the data itself, all the possible values are not used: the distribution of the data on the vectors is not regular (see Fig.2 on the distribution of vectors on the 3 first dimensions). The irregular repartition encourages to increase the number of dimensions. However, memory requirements limits this number and leads to a trade-off.
- Quantification of dimensions: The quantification of dimensions results create imprecision in distance. Specifically, signal processing theory shows that

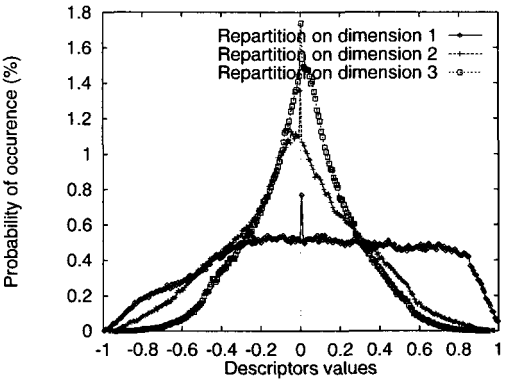

Fig. 2. Histograms of distribution of descriptors on the 3 first dimensions

quantification acts as Gaussianly distributed random noise. The variance of this noise is related to the logarithm of the number of bits.
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.

3 The Local Appearance Grid

After the choice of local descriptors, the next question is where to apply the discriptors to images. Two main approaches are possible: selecting the windows according to a criteria or selecting windows from a predefined grid.

The first approach is equivalent to a detector for interest points (or keyponts). Schmid and Mohr [20] employ the Harris Detector [10] to select interest points for indexing. Ohba and Ikeuchi [16] select windows with a local measure of trackability and a global measure of similarity. Rao and Ballard [17] used both approaches: selecting discriminant windows and selecting windows on an object centered circular grid.

Techniques based on detectors provided good results, but there is a major drawback on the stability of the detectors. This stability is hard to guarantee when the viewpoint or the luminosity changes. Therefore, our approach consists of first selecting and storing all possible windows. Subsequently, windows which are not discriminant are removed from the training test. Discriminability is determined by the number of nearby projections in the descriptor space.

Appearance Grid. 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 space P. Images are divided into a grid of 9×9 windows.

Figure 3 presents an image decomposed in non overlapping windows and the corresponding surface in a 3 dimensional subspace of P. In practice, the images

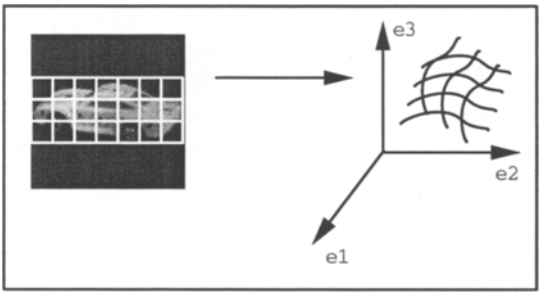

Fig. 3. Representation of an image as a surface in a 3D subspace of P (not real data)

are decomposed in overlapping windows with a shift of one pixel between two neighboring windows. Then, an image is represented by a surface in the space P. This surface is stored as a discrete grid. As the overlap between two successive windows is very large (89%), except in rare cases, two neighboring windows in an image generally have close projections in P. Therefore, a continuity property can be defined which holds between most points of the grid which represents a surface on P. A physical object is represented by a set of images and therefore, in the database, by a set of surfaces.

Recognition can be obtained by using this image representation. An unknown window is projected onto the space P into a vector v. Recognition is achieved by searching vectors in P which are close to v. Formally, the goal is to find the set S of all vectors w from the training set which are inside a ball of radius R: $S_{solutions} = \{w/dist(w, v) < R\}$. R is a threshold defining the maximum distance between two similar vectors and then image windows.

Grid constraints for filtering hypotheses. A search for vectors within a ball centered on a vector v projected from a window W usually gives several possible matches. Several techniques are possible to select the correct match. The first idea is to sort the set of solutions with the distance to v. Then, the first element x of this set correspond to the window X the closest to V. Nevertheless, the set of solutions is often too large to be completely relevant and needs some treatment. The first problem is very large sets of solutions from some very low discriminancy windows: for example, a window of constant gray value. Two approaches are possible to solve this problem:

- During the learning of the training data, it is possible to detect that some projections occurs very often and should be suppressed from the database as they will nether produce a correct recognition. This can also be done during the search phase: if a search has given too many results, the database can be updated by suppressing these irrelevant vectors.
- If the refinement of the database has not been done during the construction, it is also possible to abort a search when too many possible answers are detected.

Another problem is that neighboring windows usually project into neighboring points in P. Therefore, if a window w_i is recognized, its neighboring window w_{i+1} is likely to be recognized as well. For example, on Fig.4, on a two dimensional subspace of P, the points A, B, C and D are inside the X centered ball but corresponds to neighboring image windows. So, only the closest C should be selected as an answer.

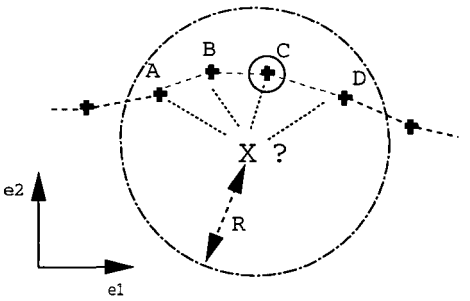

Fig. 4. redundancy of solutions in a 2D subspace of P

In this case, there are two solutions also:

- It is possible to remove some redundancy from the database during the learning phase. After adding all the windows from an image, it is possible to reproject and search all the windows with a low distance threshold. When a neighbor window is recognized, the searched window can be suppressed. The redundant window is marked so as not to be also suppressed from the database.
- It is also possible after the search to detect which projections are similar and to select the best one as the answer.

The latter technique is implemented and gives acceptable results. The drawback is some computation time for the search. The first solution should improve the technique. This logical representation needs some data structure to be usable for searching.

4 Recognizing and Searching descriptors

As shown in the previous sections, our recogntion technique needs to store and search N dimensional descriptor vectors. A naive structure would be a 2D array storing each vectors according to their positions on images. This structure provides easy access to proximity criteria during a search but requires an impossible exhaustive search. Therefore, a tree search structure is necessary.

Hierarchical search tree. The data structure used for our technique is a search tree. Each 10 dimensions are decomposed into 4 parts successively. During the database training, new dimensions are divided according to the need to separate the data. On the example of Fig.5, L_i represents short lists of stored vectors.

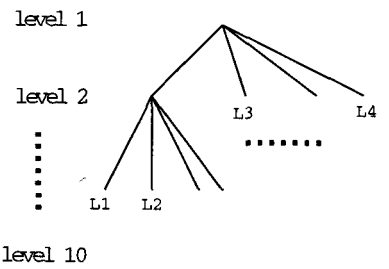

Fig. 5. Indexation Storage Tree

L_1 and L_2 have been classified up to the 3^{rd} level of the tree while, for L_3 and L_4, the 2^{nd} level was sufficient. This structure provides a very easy addition of elements: just find the right leaf and add the new vector to the list. If the list length is larger than a predefined threshold, the list is separated into 4 new lists of the next level. Large lists on the last level of the structure are likely to happen for very frequent and low discriminancy windows. But, these lists are useless to store as they will not provide recognition and therefore can be suppressed. A very nice property of this structure is, that, it is very easy to detect when adding a new vector that this vector is redundant or not discriminant. So, suppressing irrelevant vectors can be done very easily. In fact, a substitution of vector lists by their length converts the structure into a tree of the probability of each vectors.

Searching the tree. During the search phase, it is necessary to find all vectors w in the tree which are inside a sphere centered on the searched vector v. The radius R depends on the stability of the descriptors (see Sect.2). A classic technique would be to just look for the vectors in the leaf of the tree corresponding to v. This technique would miss some solutions. There is no guarantee that the search ball is included in a leaf. As a result, the search needs to explore some neighboring branches of the tree depending on the radius R. At each level of the tree, it is possible to compute the minimum distance between the searched vector and each branches of the tree. Branches with distances higher than the search threshold can be removed from the current search.

As an example, our descriptor space is constituted of 10 dimensions which values are between 0 and 255. A high number of vectors need to be stored (more than 300.000 currently in our database). Each dimension is divided into 4 branches: 0-63, 64-127, 128-191 and 192-255. The radius R is 30 which is lower that half the width of a tree cell. This leads to, in the worst case, in a full tree,

1024 cells will need to be evaluated out of a one million cells tree. Experimentally, the numbers of cells is nether over 200 and its mean is around 30 cells.

Overall, 70 to 80 % of images windows provides a direct recognition of the object. This result was achieved without any selections of the windows in the unknown images (images not used in the construction of the database).

5 Searching multiple vectors

The previous sections described the search of a single window within the database. A single window cannot guarantee a good recognition especially if randomly selected. Another approach to the recognition problem is to search multiple vectors. A classic technique is to use a voting algorithm based on the idea of k-Nearest-Neighbor rule [5]. Each search provides a list of object hypotheses. An hypothesis generates a vote for an object. The object that obtains the maximum number of votes is recognized. The first paragraph presents another approach which focused on spatial constraints between hypotheses of the same object.

Another aspect is that as the descriptors are not invariant to scale or rotation, a multiscale and multirotation search is required.

Spatial Constraints. If two different windows from an image are searched in the database, there relative position is well known. So, the results of such search are only coherent if they have an equivalent relative position. This spatial criteria can be used to reject many false recognition. More precisely, in the simple case where no 2D rotation or scale change are accepted, two windows w_1 and w_2 from an image are searched. $\delta_p = w_2 - w_1$ is the displacement vector between the two windows. Two solutions s_1 and s_2 are compatible with δ_p if $dist((s_2 - s_1), \delta_p) < th$ with th a predefined threshold. This constraint allow to select the right solution in a smarter way than the basic voting technique.

The technique consists in defining an hypothesis as being an object and its position in the image. A new hypothesis for an object generates a vote for a previous hypothesis only it is coherent. At the end, only the coherence sets of hypotheses remained. The larger one gives the recognition. In this case, two incoherent hypotheses of the same object won't disturb the recognition algorithm.

Generalization over scale and 2D orientation. The descriptors which have been chosen for the recognition are not invariant to rotation or scale. Recognizing robustly in case of scale or rotation depends on the stability of the descriptors to these images transformations. Recognition beyond scale change may be obtain by two approaches. The first approach is a multiscale technique: images of object at different scale are learned by the search database. A searched image will have a sufficiently close scale to one of the images used to create the database and thus will be recognized. The other approach consists in storing a single scale into the database but to compute the descriptors at several scales. One of the computed scale will match the database scale. Using simultaneously both techniques is also possible. For orientation, the same two approaches can be used.

If multiple objects are in the image, the first answers will hopefully correspond to the objects.

All theses approaches are linked to the possibility of computing the descriptors from different scales and orientations. Our descriptors based on the PCA need to compute the transformation on the image and then project into the descriptor space whereas other analytic descriptors can be computed directly at different scales or orientations. As a conclusion, robustness to scale and orientation is accessible but requires a additional computational or memory consumption.

6 Recognition Experiments

The technique presented in the previous sections have been evaluated in preliminary recognition experiments. The first experimental database is extracted from the Colombia database [15]. Four images of 49 first objects are used in our search database. Each image is selected with a 20 degrees shift in viewing position (angle 0, 20, 40 and 60). The test images are all the remaining images between the 5 and 65 degree viewpoints.

A 20 degree change in viewing angle is an arbitrary choice which provides good recognition results. To improve the technique, the image viewpoints should be selected for their appearance rather than for any arbitrary rule. Some objects are very stable along some viewpoints change.

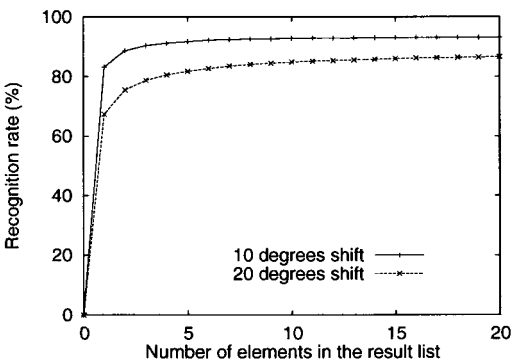

Fig. 6. Recognition Rate by a Single Random Window

A first experiment study the recognition on the Columbia database with a single random window of the test images. Two subsets of the Columbia database were tested (Fig.6). In the first case, there is a 10 degrees shift between viewpoints of objects. The second one uses a 20 degrees shift. The recognition process returns a sorted list of the k best objects. k is represented on the abscissa axis. $k = 2$ means the correct object is recognized in first or second position of the result list. Recognition does not achieve 100% with a single window because some

windows are not discriminant at all (for example, window with a constant gray level). A better interpretation of this result is to say that 70% of image windows are usable for direct recognition and 90% for multiple windows recognition.

To increase the recognition rate, our technique requires the use of multiple windows. An experiment (Fig.7) studies the recognition using two random windows separated from 5 and 10 pixels. The first graph (already shown on the previous figure) is reproduced for comparison. Spatial coherence criteria are used to join searches from the two windows. Recognition does not achieve 100% because no window selection was made and some non discriminant windows were searched. It is possible to increase the recognition by increasing the number of searched windows. The spatial coherence criteria guarantee that the more window are used, the more likely the right window is recognized. Recognizing from a few windows can be considered as an initial stage, to be followed, if necessary, by recognition from a region or an entire image.

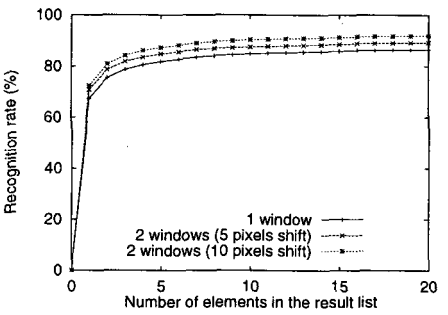

Fig.7. Recognition Rate by a Two Random Window.

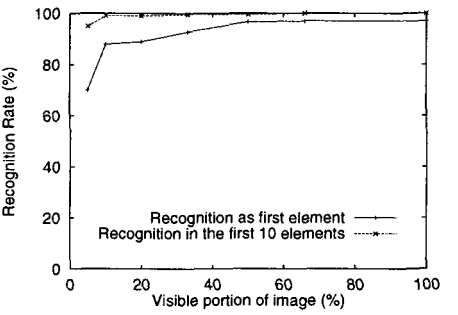

Fig.8. Recognition Rate depending on objects occlusion.

The next experiment evaluates recognition from an image region. This means that part of images are occluded to the recognition process. The test images were occluded for 95% to 0%. Figure 8 shows that recognition is very high as a first element and perfect in the first 10 results.

In the previous experiments, scale was supposed to be stable, but in the real world, object size will change depending on the distance with the camera. It is possible to generate a multiscale database by using images of different scales. An object database of five objects with scale shift of 20% between two images has been computed to test this idea. Test images of different scales (every 5%) were successfully used for recognition. 80% of windows provide a correct recognition. Other windows correspond to non discriminant windows or non scale stable windows.

Overall, experiments show that this technique provides reliable recognition of objects. Robustness to scale and orientation variations is accessible by using multiple images of objects. The memory storage requirements is nevertheless increased and the suppression of non discriminant windows becomes necessary.

7 Conclusions

Recognition of objects using the projection of an entire image to an orthogonal feature space poses serious problems. The point in a feature space to which the entire image is projected is perturbed by changes in background, in illumination, in the image position of objects and occlusions. It is possible to overcome these problems by applying such a method to local neighborhoods together with proper normalization. In this case, an image projects to a surface in a local appearance space. Changes in viewpoint and lighting deform this surface in a continuous manner, creating a family of surfaces which represent the family of appearances of the object. If the basis functions and number of dimensions of this feature space are appropriately chosen, the space will be sufficiently sparse so as permit discrimination of a large number of families of surfaces.

The experiments presented here demonstrate that an object can be recognized using a representation in which appearances are represented by surfaces in a local appearance space. This approach has been tested using 9 by 9 image neighborhoods, and a 10 dimensional space using images from the Colombia image data base.

The association of a small neighborhood of an image with a point on a surface of local appearances can be structured as an addressing operation (a table lookup). However, when the space is very large (and very sparse) it is more convenient to use a tree-structure scheme for addressing. In either case, the recognition of an object is possible in a large variety of viewing conditions, provided that those conditions (or similar conditions) have been represented in the space.

The technique is based on the storage a huge number of descriptors vectors in a search tree. An extension to this technique is to reduce effectively the number of vectors by suppressing redundant or non-discriminating vectors. It also appears possible to use filters such as Gaussian derivatives or Gabor functions to provide robustness to orientation and scale. However, this robustness has an effective cost in discrimination: many windows will project to the same point and therefore more dimensions become necessary. Our initial experiments indicate that this method provides a robust and computationally efficient for recognizing physical objects under varying conditions of viewpoint and illumination.

References

1. H. Bischof A. Leonardis and R. Ebensberger. Robust recognition using eigimages. Technical Report PRIP-TR-47, Pattern Recogntion and Image Processing group - Vienna University of Technology, june 1997.
2. 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.
3. P. Bobet. *Tête stéréoscopique, Réflexes oculaires et Vision.* PhD thesis, INPG France, 1995.

4. V. Colin de Verdière. Reconnaissance d'objets par leurs statistiques de couleurs. Master's thesis, projet PRIMA, I.N.P. Grenoble, France, June 1996. only in french (available at `http://pandora.imag.fr/~colin/projetDEA.html`).
5. R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*, chapter 4. J. Wiley and sons, 1973.
6. O. Faugeras. *Three-Dimensional Computer Vision : A Geometric Viewpoint.* The MIT Press, 1993.
7. 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.
8. K. Fukunaga. *Statistical Pattern Recognition*, chapter Feature Extraction and Linear Mapping for Signal Representation. Academic Press, School of Electrical Engineering, West Lafayette, Indiana, 1990.
9. B. V. Funt and G. D. Finlayson. Color constant color indexing. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 17(5):522–529, 1995.
10. C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. 4th Alvey Vision Conference*, 1988.
11. 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://pandora.imag.fr/Prima/Ravi/`.
12. J. Martin and J.L. Crowley. Comparison of correlation techniques. In U. Rembold et al., editor, *Intelligent Autonomous Systems - IAS-4*, pages 86–93, Karlsruhe, Germany, March 27–30 1995.
13. B.W. Mel. Seemore: A view-based approach to 3-d object recognition using multiple visual cues. *Advances in neural information processing systems*, 8:865–871, 1996.
14. H. Murase and S. K. Nayar. Visual learning and recognition of 3d objects from appearance. *International Journal of Computer Vision*, 14:5–24, 1995.
15. S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (coil-100). Technical report, Columbia University, New York, february 1996.
16. K. Ohba and K. Ikeuchi. Recognition of the multi specularity objects for bin-picking task. *IROS 96*, 3:1440–1448, 1996.
17. 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.
18. B. Schiele. *Object Recognition using Multidimensional Receptive Field Histograms.* PhD thesis, I.N.P. Grenoble, France, 1997. English translation.
19. B. Schiele and J. L. Crowley. Object recognition using multidimensional receptive field histograms. In *ECCV'96, Fourth European Conference on Computer Vision, Volume I*, pages 610–619, 14–16 April 1996.
20. C. Schmid and R. Mohr. Combining grayvalue invariants with local constraints for object recognition. In *International Conference on Computer Vision and Pattern Recognition*, 1996.
21. I. Sirovich and M. Kirby. Low-dimensional procedure for the caracterization of human faces. *J. Opt. Soc Am. A*, 4(3):519–524, March 1987.
22. 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 [11].