### Eigenhistograms

As seen before, color distributions can be efficiently used as signatures for object recognition in the appearance-based framework. The earliest approach [137] showed the usefulness of color histograms for indexing large object databases independently of object's pose. And, as also explained before, most of the recent approaches focus on illumination color invariance [47], known as color constancy, but although these methods perform better than histogram indexing when color illumination changes, they use color information only where surface color varies and are very sensitive to noise.

All these methods do not consider color distribution changes due to illumination pose, which can induce significant changes in color histograms. Eigenhistograms are introduced in [147, 19] where they consider the problem of non constant illumination and introduce the PCA technique to construct a low dimensional space for representing this effect. Collecting a huge amount of local color histograms of objects under different illumination directions, they assume that these variations span a low dimensional linear subspace of $\Re^n$ (being $n$ the number of colors represented in the histogram) which can be computed using PCA. Principal Component Analysis of a histogram set defines an encoding function that projects each histogram from the set on a low dimensional subspace defined by the $m$ principal eigenvectors $e_i$ of the histogram covariance matrix. In this framework, given that $e_i$ form an orthonormal base for object histograms, they call each $e_i$ an *eigenhistogram*.

### Eigenfaces

A particularly successful application of PCA in the field of computer vision is learning an accurate low dimensional representation of face images. In this case, an interesting fact when the domain space samples are images is that, under a liner representation, the resulting bases can also be viewed as points in the space and consequently are also images. In the case of PCA applied to face images these eigenvectors receive their own name: *eigenfaces* [142]. Figure (2.7) shows the ordered set of the first 100 eigenfaces computed from 500 sample set of faces. Faces are represented in a $96 \times 96 = 9216$ dimensional space, so that, the eigenfaces have also $96 \times 96$ pixels.

Some interesting facts can be observed in the set of eigenfaces in figure (2.7). At least the first two eigenfaces are directions of illumination variation, meaning that the original images have not been completely successful normalized with respect to illumination. This fact is common in this context, and usually, the first eigenfaces reflect only changes in illumination. Other eigenfaces can be also directly associated with particular variations. We can also observe the presence of higher frequencies as we advance in the eigenfaces. The fact PCA is a linear transformation makes it simple to interpret these bases. If we project a face image into the PCA space, each principal component indicates the weight of the corresponding basis: any face image can be written as a weighted sum of these eigenfaces. So, it is clear from this example, that if a proper training set is chosen (a training set that generalizes correctly) the reconstruction and interpretation of new faces would be more acceptable.

There are other several local features that have been used in conjunction with Principal Component Analysis. Usually, PCA is used to reduce the dimensionality of

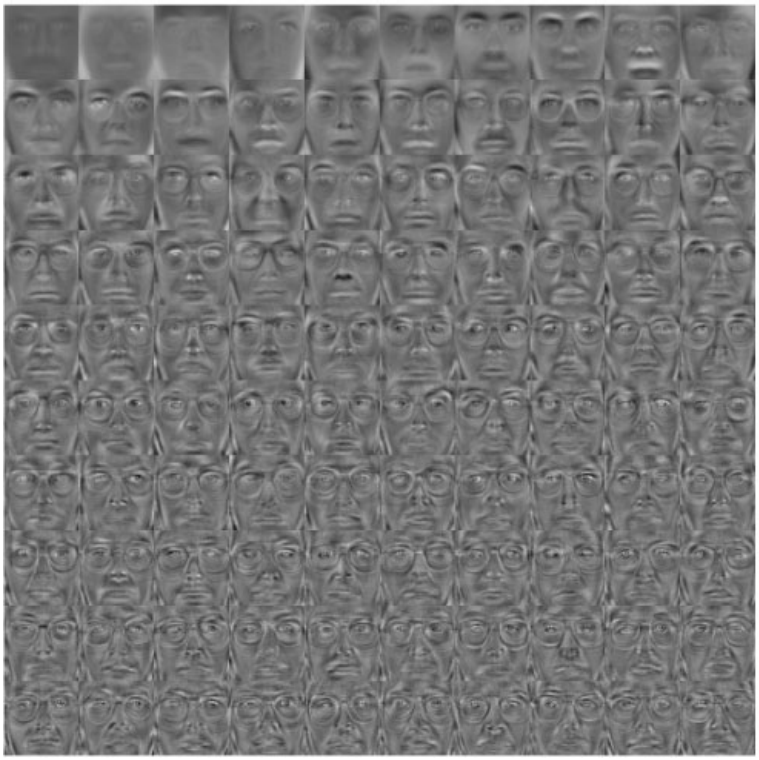


**Figure 2.7:** Typical set of eigenfaces (eigenvectors from an original set of faces). In this particular case, 100 eigenfaces are shown extracted from 500 face images and using a window size of 96 × 96 pixels.

the problem and work with the reduced eigenfeatures using some statistical technique to classify the new projected data.

### 2.1.7 Filters Based on Natural Image Statistics

In the previous section we presented Gaussian derivatives and Gabor filters as two ways to obtain feature vectors. As seen before, these two techniques are simply based on applying a fixed set of filters over an image (or an image patch) and obtain the responses to these filters. After this, we will use these filter responses as our feature vector. As can be seen, this set of filters are described analytically and they are used undistinctively. For example, these filters are used with no distinction even if we want to describe a face image or a chair image.

A first attempt to solve this problem of *uniqueness* is exposed in the previous section where we show the Principal Component Analysis (PCA) as a technique to reduce an original feature space with a given set of eigenvectors. This set of eigenvectors is calculated through a given set of feature vectors that we have previously selected. Then, once we have a given set of eigenvectors, we are able to represent a

new unseen feature vector as the combination of these eigenvectors. This final step can be seen as the result of apply a filter (eigenvector) in the image in order to create the reduced feature vector. As seen, PCA can be viewed as a technique which calculates a set of filters (eigenvectors). These filters come from a set of feature vectors which we considered interesting for our application. Under this point of view, we can formulate one question: what happens if our initial feature vectors are natural images? Here, natural image refers to any image that we can find in the world with no distinction.

Some experiments with natural images (or objects) have been performed and the results are very interesting. We present three techniques that will be used throughout this thesis with their respective applications to natural images in order to find biological/natural based filters.

### PCA with Natural Images

In [110] there is a Principal Component Analysis experiment with natural images. Usually, PCA is applied to a reduced set of objects and the principal components are based on this specific set of images. Given that PCA-based methods require recomputation of the eigenvectors each time a new object is to be recognized, it is natural to ask what the results of PCA would be if one were to take this process to its limit i.e. to perform PCA on a set of arbitrary natural images containing a wide variety of natural and man-made stimuli. When we say that the eigenvectors should be recalculated when a new object is introduced in our object database, we want to reflect that once we have a PCA model for a specific set of objects and we require the inclusion of new objects, it is clear tha PCA has to be recalculated each time we perform some inclusion. An initial work by Hancock et al. [54] sheds some light on this interesting question. They used a neural network to extract the first few principal components of an ensemble of natural images. Random image patches of size $32 \times 32$, $64 \times 64$ and $128 \times 128$ were used as input to the network from a set of 40 natural images. They discovered that regardless of the scale analysis, the eigenvectors obtained were very close approximations of different oriented derivative-of-Gaussian operators.

In [110] they used $32 \times 32$ Gaussian-windowed image patches obtained by scanning across a number of arbitrary images of natural scenes. They extracted the nine first eigenvectors which account for as much as 83% of the input variance. And as observed in their work, the eigenvectors closely approximate different oriented derivative-of-Gaussian operators (see section 2.1.1).

The derivative-of-Gaussian filters can thus be regarded as a set of local natural basis functions, useful for general-purpose object recognition. Part of the rationale for this belief stems from the fact that these basis functions are obtained as result of applying the principle of dimensionality-reduction to arbitrary collections of images containing a plethora of elementary features from natural as well as man-made structures rather than just the training images of particular objects or faces as was the case for the basis functions used in [142, 94]. Further support for this belief comes from the observation that correlation filters generated by principal component expansion maximize signal-to-noise ratio and yield much sharper correlation peaks than

traditional raw image cross-correlation techniques. Indeed, Canny [25] has shown the first and second order Gaussian derivatives to be close to optimal for detecting the elementary features of edges and bars respectively. The Gaussian derivatives filters also allow strategies for rotation normalization in the image plane because they are known to be steerable (as seen in section 2.1.1). Finally, these basis functions are endorsed by some biological studies which show that the different order derivatives of the Gaussian provide an accurate fit to primate cortical receptive field profiles as compared to other mathematical profiles suggested in the literature [159].

### ICA with Natural Images

Independent Component Analysis (ICA) first irrupted as a solution to the problem of Blind Source Separation (BSS) and many of its practical applications are found within this field. It is not the scope of this section to explore the different applications of ICA since it will be explained later in this thesis (see section 3.6.1). But without doubt, the most influencial application of ICA has been as a feature extraction technique. These results, though more theoretical than practical, have proven to be a major impulse for spreading ICA. The main interest of the features extracted by ICA is found on the close relationship between this technique and the representation principle known as *sparse coding*. Such coding of a given dataset should satisfy that only a small number of basis vectors are activated at the same time, or equivalently, most components of the coded data are zero, or close to zero, and only a few are significantly nonzero. In a neural network interpretation this means that, if each basis vector corresponds to a single neuron and its coefficient the corresponding activation, then we have that a given neuron is rarely activated in the network. It is said that such data has a sparse distribution. Notice that this distribution should have a strong peak in the value zero and heavy tails, so sparseness can be equated to supergaussianity.

Though sparse coding has practical utility in signal processing such as data compression and denoising, it was first developed as a model for image representation in the primary visual cortex of mammals (V1) [5, 41, 100]. In the late nighties, a highly successful experiment with natural images finally related sparse coding with $V1$ response through independent component analysis [10, 145]. This experiment consisted of randomly extracting patches of fixed size from natural images and using this data for ICA estimation. Results showed that the obtained basis filters have the three principal properties of simple cells in V1: they are localized, oriented and bandpass. In [17] we can find an experiment where they reproduce this result by extracting 13000 patches from natural images. In this experiment, the patches were normalized on mean and variance and had the mean (an approximately flat image) substracted. Dimensionality was reduced to 144 using PCA in a preprocessing stage to remove noise. Results of estimating ICA on this dataset are shown in figure (2.8). Similar experiments to this of natural image patches have been performed on color and stereo images [60], video data [144], audio data [9] and hyperspectral data [102].

Traditionally, the sparse coding principle was associated to the information theoretic concept of redundancy reduction for compressive coding. On the other side, it has for long being defended that the coding strategy followed by the early visual system is adapted to the input statistics through combined evolution and neural learning.

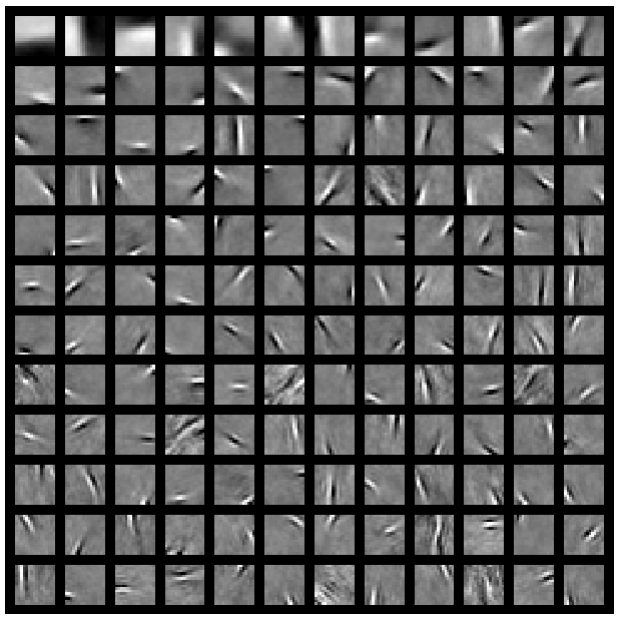


**Figure 2.8:** The ICA basis vectors extracted from natural image patches. As seen, they are oriented, localized and bandpass.

The emphasis of this learning process was biased towards the assumption that the cortex seeks to represent highly redundant sensory data efficiently, so sparse coding seemed a natural solution to the problem. In a recent work Barlow [6] has strongly contributed to redirect the understanding of the mechanisms linking perception and redundancy, arguing that importance should be shifted from economy and efficiency to be set upon its contribution for building an accurate estimation of a probabilistic model for the environment. From this perspective, a compressed representation of sensory experience is doubtfully useful for the brain due to the unreliability of the estimates it can produce. Yet a representation with as little redundancy as possible is desirable since it allows easy access to the event probabilities and statistical dependencies among the responses. This change of emphasis favours the following reasoning: ICA provides a response similar to that present in the visual cortex not thanks to its sparse coding nature but thanks to the fact it reduces statistical dependencies providing a framework where probability estimation is greatly simplified.

There is a growing interest in using ICA as a basis to extract statistical structures of natural images. As example, there is a work of Hyvarinen et al. [63] where they try to propose a model that predicts properties of biological visual systems. They propose a unifying framework based on the concept of spatiotemporal activity "bubbles". For them, a bubble means an activation of simple cells (linear filters) that is contiguous both in space and in time. Another study of Karklin et at. [69] tries to discover higher-order structure of natural images. That is, they try to understand and discover how higher-order properties of images might be represented or what the higher-order

properties might be. They propose a statistical model for higher-order structure that learns a basis on the variance regularities in natural images. This higher-order, non-orthogonal basis describes how, for a particular visual image patch, the coefficient variances deviate from the default assumption of independence. This view offers a novel description of higher-order image structure and provides a way to learn sparse distributed representations of abstract image properties such as object location, scale, and surface texture.

### NMF with Natural Images

Information processing capabilities of embedded systems presently lack the robustness and rich complexity found in biological systems. Endowing artificial systems with the ability to adapt to changing conditions requires algorithms that can rapidly learn from examples. Non negative Matrix Factorization (NMF) appeared in this context where there was the need of discovering the salient visual and auditory cues that describe objects uniquely. Lee and Seung [78] presented an unsupervised learning technique that discovers features in data without external interaction. Non negative matrix factorization (NMF) is the method that they presented which using nonnegative matrix factorization they were able to automatically learn the different parts of objects. They show that a parts-based representation of data is crucial for robust object recognition. The method was tested using a robot constructed to perform simple sensorimotor tasks.

Non-negative Matrix Factorization is used throughout this thesis and it will be explained later (see section 3.3.1). Here, we only present one of the positive aspects that we can exploit from this technique, its ability to discover the most salient visual features. It is argued that the main reason to consider positive descriptions is that they are found in biological neural networks. Since in computer vision one common trait is to deal with positive data descriptions (image pixels), it seems plausible to use also positive factorizations of data instead of negative representations (PCA or ICA). Furthermore, NMF is able to learn parts as features by modeling positive coactivation in the inputs. Of course, such a parts-based representation is valuable because it is invariant to perturbations or occlusions that affect localized regions of the input space.

The online learning algorithm described in [78] allows a robot to rapidly adapt and correct mistakes when given supervisory feedback by a 'teacher'. But there are many situations in which it is impossible for any teacher to be present. In these situations, would it still be impossible for a system to adapt based upon raw sensory stimuli without being told the appropiate thing to do? This is generically quite a difficult problem, but there are some well-stablished algorithms that allow the system to continue to adapt even without a teacher. Generally, these unsupervised learning algorithms attempt to extract common sets of features present in the input data. The system can then use these features to reconstruct structured patterns from corrupted input data. This invariance of the system to noise allows for more robust processing in performing recognition tasks.

NMF originated in the context of unsupervised feature extraction. This technique decomposes images into their representative parts. In the initial studies of NMF [78, 79] we can find a detailed analysis of the reconstructed objects modeled using this

technique with respect to PCA and VQ (Vector Quantization). The interesting result of this analysis is that NMF finds a decomposition into localized parts even though no prior information about the topology of the images was built into the input data. Since many of the parts are used to form the reconstruction, NMF has the combinatorial expressiveness of a distributed representation. But because the nonzero elements of its matrices are all positive, NMF allows only additive combinations. So unlike PCA, no cancellations can occur. Thus, NMF learns a parts-based representation of the data that is sparse as well as distributed. In section 3.3.1 we will find a detailed description of NMF.

It should be noted that biological neural networks may use similar types of constraints to achieve analogous representations. The firing rates of neurons cannot be negative, and the strengths of synapses do not generally change sign [78]. These onesided constraints could possibly be important in developing the sparsely, distributed coding of sensory input that give rise to robust biological information processing. In figure (2.9) we can see an example where the obtained NMF bases are sparse and distributed.

**Figure 2.9:** 64 NMF bases computed from a training set of 7000 frontal view face images. Part of the bases correspond to intuitive parts of faces.

Up to now, we have explained different ways to obtain feature vectors which can be used to describe an object. Depending on the captured image of an object and assuming that we are able to know a priori what kind of information we want to use, we will select some of the explained techniques (or other techniques similar to the ones explained here) to extract feature vectors. We present figure (2.10) to summarize the above explained techniques which can be used to extract local information. Figure (2.10) presents a toy car and some of the previous explained local features which can be used to represent local image patches. For example, if we have a set of image patches which we know that are invariant to illumination, we can use color histograms as seen in the figure. But since we have an object with a tremendous number of contours, we can use contour features as also seen in the figure. In contrast to this, we can obtain the responses to a fixed set of filters and, as example, we can use the Gabor filters or the Gaussian derivative filters. Instead of this, we can also calculate a set of filters that depend on the data. As a graphical example, having the set of image patches shown in figure (2.10), we obtained a set of filters using PCA, ICA and NMF that are also shown in the figure.

## 2.2 Location and Geometry

In the previous section we have described *what* we want to extract from in images in order to be able to identify a subset of objects that we previously know a priori. Information extracted from objects is coded into a feature vector that takes into account different natural properties of a local patch of an object. Assuming that we know which are the features that can be used to extract information from image pixels we should we able to represent from *where* we have extracted this information in order to determine the presence and location of an object in an image. Here is where we show different computational strategies to obtain reliable or possible locations of "important" patches of image objects. For us, the term *where* is a mixture of the location and topology of the local features which are extracted from images. Then, we can use this information to extract local features as described in the previous section.

### 2.2.1 Location of Local Features

Previous section shows a different set of local features that are usually used in computer vision. But, the problem now is where in the image we have to extract this set of local features? There are two common approaches to solve this problem: (i) Use a fixed grid of keypoints placed above objects; (ii) Calculate some relevant locations in our objects and extract the local information only from this set of relevant points (keypoints).

#### Local Points from a Fixed Grid

In [129] it is shown that while vertical (intracolumnar) connections develop at the beginning of the third trimester of gestation in humans, long horizontal connections within layers do not appear until shortly before birth, and they continue developing

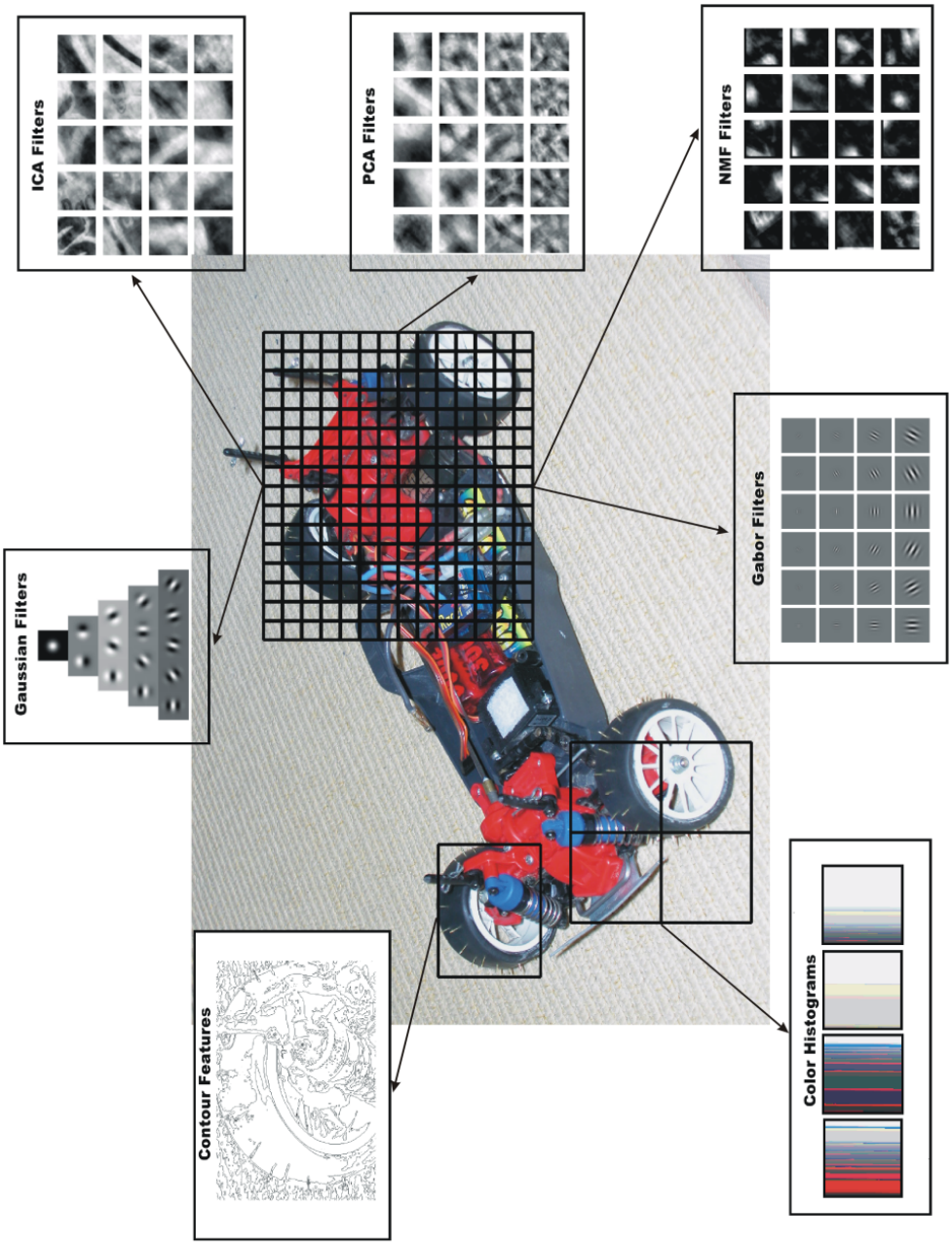



**Figure 2.10:** Graphical representation of a set of local features which can be used to extract local information from image patches of an object.

until the first postnatal year [22]. For this reason it is not unlikely that newborns and young infants process and store segmented objects as a matrix of local features without much interaction among such features. This can be modeled by a grid of hypercolumns covering the portion of the visual field segmented as the object. It seems that this is the biological reason that motivates to use a cell representation instead of working with the relevant regions of objects.

From the computational point of view, a fixed cell representation is better since we already know a priori the number of points to be used for our task. This simplifies the problem and can be used for faster applications. But depending on the nature of our objects, a fixed cell representation can be enough or not. For example, if we have objects in our database that are completely homogeneous, i.e. a tomato, where are the interesting points of a tomato? Inside or in the border of the tomato? This is a difficult question and in this particular case, a fixed cell representation would be the best solution because there are not relevant regions that can describe homogeneous objects. In figure (2.11) we can see a homogeneous object (a tomato) where its relevant points are difficult to obtain because any of the point detectors will fail to localize relevant points. It is clear that a fixed cell representation, for this particular case, would be a good solution since all the regions of the tomato would be considered.

**Figure 2.11:** Typical example of an object that we can find in an object database. This is a homogeneous object and there are not relevant points in the object surface.

### Interesting and Relevant Points (keypoints)

By *interest point* we simply mean any point in the image for which the signal changes two-dimensionally. Conventional *corners* such as L - corners, T - junctions and Y - junctions satisfy this, but so do black dots on white backgrounds, the endings of branches and any location with significant 2D texture. As it will be presented later, this thesis uses the idea of interesting points but it is not our scope to find the best point detector. There are several works in the literature that have analyzed different point detectors in order to show which is the best one [3, 16, 57]

There is a recent work developed by Cordelia Schmid et al [124] where they introduce two novel criteria in order to evaluate interest points: *repeatability* and *information content*. Those two criteria directly measure the quality of the feature for tasks like image matching, object recognition and $3D$ reconstruction. They apply to any type of scene, and they do not rely on any specific feature model or high-level interpretation of the feature.

*Repeatability* explicitly compares the geometrical stability of the detected interest

points between different images of a given scene taken under varying viewing conditions. Previous methods have evaluated detectors for individual images only. An interest point is "repeated", if the $3D$ scene point detected in the first image is also accurately detected in the second one. The repeatability rate is the percentage of the total observed points that are detected in both images.

*Information content* is a measure of the distinctiveness of an interest point. Distinctiveness is based on the likelihood of a local greyvalue descriptor computed at the point within the population of all observed interest point descriptors. Descriptors characterize the local shape of the image at the interest points. The entropy of these descriptors measures the information content of a set of interest points.

From these studies, it seems that the best point detector is the Harris detector [55]. The Harris point detector, as other point detectors, is based on a matrix related to the auto-correlation function. The local auto-correlation function measures the local changes of the signal. This measure is obtained by correlating a patch with its neighbouring patches, that is with patches shifted by a small amount in different directions. In the case of an interest point, the auto-correlation function is high for all shift directions.

Given a shift $(\Delta_x, \Delta_y)$ and a point $(x, y)$, the auto-correlation function is defined as:

$$f(x,y) = \sum_{(x_k,y_k)\in W} (I(x_k,y_k) - I(x_k+\Delta_x, y_k+\Delta_y))^2 \tag{2.62}$$

where $(x_k, y_k)$ are the points in the window $W$ centered on $(x, y)$ and $I$ the image function.

If we want to use this function to detect interest points we have to integrate over all shift directions. Integration over discrete shift directions can be avoided by using the auto-correlation matrix. This matrix is derived using a first-order approximation based on the Taylor expansion:

$$I(x_k+\Delta_x, y_k+\Delta_y) \approx I(x_k,y_k) + \begin{pmatrix} I_x(x_k,y_k) & I_y(x_k,y_k) \end{pmatrix} \begin{pmatrix} \Delta_x \\ \Delta_y \end{pmatrix} \tag{2.63}$$

Substituting the above approximation (2.63) into equation (2.62), we obtain:

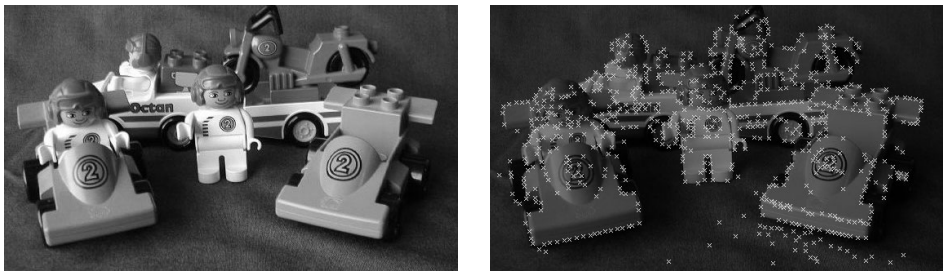
$$\begin{aligned} & f(x,y) = \\ = & \sum_{(x_k,y_k)\in W} \left( \begin{pmatrix} I_x(x_k,y_k) & I_y(x_k,y_k) \end{pmatrix} \begin{pmatrix} \Delta_x \\ \Delta_y \end{pmatrix} \right)^2 \\ = & \begin{pmatrix} \Delta_x & \Delta_y \end{pmatrix} \mathbf{A}(x,y) \begin{pmatrix} \Delta_x \\ \Delta_y \end{pmatrix} \end{aligned} \tag{2.64}$$

being the matrix $A(x, y)$ defined as:

$$\mathbf{A}(x,y) = \begin{bmatrix} \sum_{(x_k,y_k)\in W} (I_x(x_k,y_k))^2 & \sum_{(x_k,y_k)\in W} I_x(x_k,y_k) I_y(x_k,y_k) \\ \sum_{(x_k,y_k)\in W} I_x(x_k,y_k) I_y(x_k,y_k) & \sum_{(x_k,y_k)\in W} (I_y(x_k,y_k))^2 \end{bmatrix} \tag{2.65}$$

This matrix captures the structure of the neighborhood. If this matrix is of rank two, that is both of its eigenvalues are large, an interest point is detected. A matrix

of rank one indicates an edge and a matrix of rank zero a homogeneous region. Harris uses the auto-correlation matrix $A$ and the use of discrete directions and discrete shifts is avoided. Instead of using a simple sum, a Gaussian is used to weight the derivatives inside the window and interest points are detected if the auto-correlation matrix $A$ has two significant eigenvalues.



(a) (b)

**Figure 2.12:** Example where we show the results of using the well-known Harris point detector. Image (a) has been used to extract a relevant set of interesting points (keypoints) that are shown in image (b).

### 2.2.2 Geometry of Local Features

We should consider the possibility of using the geometrical relationship among the local features in order to increase our representation scheme. Representations which do not contain geometrical information are simple and fast. It is argued by advocates of this approach that this representation scheme is biologically plausible because it can reproduce the fast recognition performed by the brain. Some examples of systems exploiting no topology between features are Seemore [87], Multidimensional indexing [24] as well as Geometric hashing [134]. These schemes with no topology among local features suffer from ambiguity in the representation. They may fail to distinguish between the image of a given object and other rearrangements of the features (a jumbled image). They also rely heavily on attributes other than shape since the distinct and relevant shape features can only be extracted by capturing the relationship among two or more local shape features in one way or another. Thus, the spatial layout of the local features is essential in distinguishing shape differences among objects [129]. Both of these characteristics (response to jumbled images and the reliance on non-shape features for recognition) are quite in contrast with the functioning of human visual system [129]. It can be argued, of course, that the spatial relationships can be implicitly encoded in such schemes given an appropiate set of features. For example, a set of features, each encompassing a pairwise topology of more primitive features, may be sufficient for enforcing a global topology. Although this assertion is true in general, it is hard to imagine that there exists a feature set which would be capable of capturing the topological relationships unambiguously, for *all* scenes, considering the astronomical number of possible layouts of objects and pose.

The other category of local feature representations consists of the schemes which

explicitly encode the topological relationships among local features. An old example of such approach is Dynamic Link Architecture [149] and its algorithmic version, labeled graph matching [148], where each node is labeled with a feature, and the spatial relationships are encoded using arcs between the nodes. This approach has shown success in dealing with variations such as size and illumination and distortion. Dynamic Link Architecture is a neural model which only utilizes local interaction among neurons and temporal correlations of the neural activity, thus offering a biologically plausible scheme for object representation and recognition.

When topology between features is considered it appears another point of view with respect to object recognition: A parts-based representation. That is, if we take into account the relationship between some of our features, we can consider that these features can have a new label for us: an object *part*. The definition of an object part is difficult and it depends on the final application. In general, we can define a part of an object as a collection of features that have something in common, the spatial location of them has something interesting for us because they can be distributed closely and we can take advantage of this fact. For example, if we use the color as the main feature to classify our object, an object part could be a subset of the color features that have the same tonality. The last studies seem to show that this intermediate representation (neither local nor global) makes possible recognition of objects with invariance against size, view point, partial occlusion, etc.

Many feature-based schemes operate based on template matching and thus, would require a huge number of examples (or templates) per object in order to cope with variations in size, orientation in depth and plane, occlusion, distortions, texture, surface marking, color and lighting. Clearly this is a poor approach (or not a very convincing one) since even with only a couple of degrees of freedom, it quickly results in combinatorial explosion. Some models demonstrate robustness to some important variations. While some of the feature-based methods may be robust to some variations due to their specific representations, they all share the way they describe the 3D spatial information, and therefore, unanimously suffer from the consequences of such type of representation.

The point here is that if we consider the set of shapes of the objects in the world, natural as well as human-made, we can see that although quite diverse, there is a high amount of redundancy in the set: there are certain shapes which take part in a large number of objects. An optimal coding scheme would certainly have to take advantage of such enormous redundancy in the input space. By coding all views of all objects, the redundant object parts get implicitly coded over and over again. Such redundant encoding not only leads to the need for an enormously large storage space, but also fails to provide generalization capacity useful for coping with new objects. In the particular case of the view-point invariant recognition of objects in such paradigm relies on information about different views of whole-objects, an unfamiliar object which is seen from only one vantage point, cannot be recognized from other views, even though it is composed of parts which have each been explored from all views previously within various familiar objects.

## 2.3 Local Object Recognition Techniques

As seen, we explained that we are able to extract local information (*what*) from specific locations (*where*) in images. However, it remains to know *how* we can use this information to perform a classification of an object that appears in an image. Here, we refer to classification to the process of assigning a class label to an instance described by a feature vector. There are many well-stablished frameworks for classification, i.e. nearest-neighbor methods, decision trees, neural networks, support vector machines, etc.

It is clear that once a particular set of feature vectors are extracted from objects, we should define a measure of similarity at some point. Even using a complex classification framework, we would require a similarity measure between feature vectors. It is clear that two objects are similar if they contain similar feature vectors of their appearance measures.

This section presents some of the well-known techniques which have appeared in recent years in the computer vision community. Our aim is to present the main features of each technique in order to know and to present the basics of each framework. Throughout this thesis, we will compare our work with some of these techniques presenting the main advantages or drawbacks with respect to them.

We have classified these techniques as (i) Feature based schemes, (ii) Higher-order schemes and (iii) Parts-based schemes. Feature based schemes are those that are simply based on the local features extracted from objects without taking into account the geometrical relationship among themselves. That is, when no information about other features is considered. When a method deals with a feature that also considers the geometrical information provided by the other features in an object, we are in front of a higher-order scheme. Finally, when the higher-order dependencies between a subset of the whole features of an object are topologically grouped and labelled as an object part and obey to some a priori restrictions imposed by us, we are talking about parts-based schemes. Furthermore, each technique is also explained in terms of its classification scheme as we are interested in knowing the internal matching process between an object model and new unseen objects. Then, this information will be used in next chapters in order to describe and compare our technique with the ones presented in the following.

### 2.3.1 Feature Based Schemes

The methods presented here are those that once they have an object in an image, they simply extract a set of local features to obtain an object model without any information about the geometrical relationship among features.

#### Object Recognition using Multidimensional Receptive Field Histograms

Schiele and Crowley [120, 119] developed an object recognition technique that is based on the use of multidimensional receptive field histograms. They observed that the color histogram approach is an attractive method for object recognition because of its simplicity, speed and robustness. Additionally the approach does not rely on the correspondence between the object model and the test image. However, the reliance

of color histograms on object color and light source intensity make it inappropiate for many recognition problems. The idea of histograms has been used as the main skeleton for the work of Schiele and Crowley. They developed a similar technique using local descriptions of an object's shape provided by a vector of linear receptive fields. They take as a reference $2D$ local characteristics instead of color information and they build a histogram that can be viewed as a probability function of an object. In their work they presented several local features to build receptive field histograms. The best results have been achieved considering the first Gaussian derivative in $x$-direction, the first Gaussian derivative in $y$-direction, the Laplacian operator and a rotation invariant Gaussian filter based on a mixture of first and second order Gaussian derivatives (see equation 2.13). Also, they combined these filters analyzing the influence of different histogram resolutions. Also, different histogram measurements are evaluated showing the performances of each scheme.

They achieved robustness under the presence of image plane rotation, the presence of scale changes and image plane rotation simultaneously. Also, they analyze the presence of viewpoint changes and the presence of partial occlusions. Their results reported that the best histogram matching metric is the $\chi^2$ and good results are obtained with the typical histogram intersection measurement. For specific details of this method see references [120, 119].

Classification in this method is based on the use of the k-nearest neighbor technique of histogram representations using several metric distances between histograms.

### SEEMORE: Combining Color, Shape, and Texture Histogramming in a Neurally-Inspired Approach to Visual Object Recognition

In [87] Mel presented a novel approach to computer vision: the integration of multiple cues in an unique classifier based on a biological model. The main inspiration of his work was the conjecture that the early phase of object recognition in the brain is based on a feedforward feature-extraction hierarchy. In order to prove the plausibility of this conjecture in an engineering context, a difficult 3D object recognition domain was developed to challenge a pure feedforward, receptive-field-based recognition model that he called SEEMORE. SEEMORE was based on 102 viewpoint-invariant nonlinear filters which were as a group sensitive to contour, texture, and color cues.

In evaluating the plausibility of the feedforward feature-extraction scenario of SEEMORE [87] as a model for biological vision, Mel considered several desirable properties of a feature-space representation as dictated by the "computational ecology" of natural vision. SEEMORE's visual representation is composed of 102 feature channels that emphasize spatially-localized filter computations, and are collectively sensitive to contour shape, color, and texture cues. SEEMORE's architecture is similar in spirit to the color histogramming approach of Swain and Ballard [137] (see section 2.1.3), but includes spatially-structured features that provide also for shape-based generalization.

SEEMORE's channels fall into in five groups: (1) 23 color channels, each of which corresponds to a small blob of color parameterized by "best" hue and saturation, (2) 11 coarse-scale intensity corner channels parameterized by open angle, (3) 12 "blob" features, parameterized by the shape (round and elongated) and size (small,

medium, and large) of bright and dark intensity blobs, (4) 24 contour shape features, including straight angles, curve segments of varying radius, and parallel and oblique line combinations, and (5) 16 shape/texture-related features based on the outputs of Gabor functions at 5 scales and 8 orientations.

Experiments reveal good recognition performance in a viewpoint- and configuration-invariant 3D object recognition problem with 100 objects, including objects that are rigid, non-rigid, and "statistical" in nature, and photographs of complex scenes. For more information see [87].

Classification in this method is based on the use of the k-nearest neighbor technique of histogram representations using several metric distances between histograms. These histograms are of different nature with respect to the previous explained framework.

**Learning to recognize objects in images: acquiring and using probabilistic models of appearance**

In [107] Pope presented a singular scheme that provides a general point of view for object recognition techniques. He presented a method of modeling the appearance of objects, of acquiring such models from training images, and of using the models to accomplish recognition. In theory, this method can handle complex, real-world objects better than previously known methods (up to 1995). Indeed, in principle, it can be used to recognize any object by its appearance, provided it is given a sufficient range of training images, sufficient storage for model views, and an appropiate repertoire of feature types. Is for these reasons that this method can thus be considered a completely general approach to recognition by appearance.

The main features of his approach are:

- Appearance is described using discrete features of various types, ranging widely in scale, complexity, and specificity. He have adopted a basic, yet versatile repertoire of feature types comprising intensity edge segments and their groupings. The repertoire used contains low-level features such as straight, circular and elliptical segments of intensity edges and higher-level features representing groupings of these, such as junctions, groups of adjacent junctions, pairs of parallel segments, and convex regions.

- An object model represents a probability distribution over possible appearances of the object, assigning high probability to the object's most likely manifestations. Thus, learning an object model from training images amounts to estimating a distribution from a representative sampling of that distribution.

- An object model divides the range of possible appearances into a number of discrete views. Within each view, independent probability distributions characterize the occurrence, position, and intrinsic attributes of individual features.

- A match quality measure provides a principled means of evaluating a match between a model and an image. It combines probabilities that are estimated using distributions recorded by the model. The measure leads naturally to an efficient matching procedure, probabilistic alignment, used to accomplish both

learning and recognition. To identify good matches quickly, the procedure employs constraints based on features' relations, positions, and intrinsic attributes, each commensurate with its uncertainty as recorded by the model.

Finally, this method cannot distinguish objects on the basis of scale; instead it tries to recognize objects regardless of their scaling in the image. The method does, however, report the apparent scale of an object as measured directly in the image; under restrictive assumptions of viewing conditions, objects of different scale may be distinguished on that basis. And changes in perspective distortion of an object's appearance, if they are significant, must be included in the training set as views of the object, as this method does not attempt to infer possible ranges of distortion. Similarly, illumination changes must be covered by the training set if those changes affect features appreciably.

Classification in this method is based on the estimation of probability density functions. A bayes formulation is derived in order to perform object classification.

### 2.3.2 Higher-order Schemes

When an object is represented by a set of features and their geometrical relationships, we consider that the method contains higher-order information. Here we show a set of techniques that make use of this higher-order information about features.

#### Local Greyvalue Invariants for Image Retrieval

Schmid and Mohr [123] presented an extended work about local greyvalue invariants for image retrieval where they addressed the problem of retrieving images from large image databases. From a general point of view, the method is based on local greyvalue invariants which are computed at automatically detected interest points. After this, a voting algorithm and semi-local constraints make retrieval possible. Indexing allows for efficient retrieval from a database of more than 1000 images (year 1997). Experimental results show correct retrieval in the case of partial visibility, similarity transformations, extraneous features, and small perspective deformations.

This scheme is based on a set of extracted keypoints. They are extracted using the well known harris detector [55] (see section 2.2). Using this set of keypoints as a reference, they extracted a feature vector on each keypoint based on derivatives which locally describe an image. This feature vector consist of the local jets presented in section 2.1.5 because they are invariant to $2D$ image rotations of rigid displacements in the image. After this, these invariants are then inserted into a multi-scale framework in order to deal with scale changes (see section 2.1.5 or [55] for more information).

Once all the whole set of invariant vectors are obtained they evaluate whether two invariant vectors are similar using the Mahalanobis distance (see equation 2.50). This distance is adequated to this feature comparison task because it is a standard method to model the uncertainties in the components as random variables with Gaussian distribution. Since the different components of the invariant vectors have different magnitudes, this distance takes into account the different magnitude as well as the covariance matrix of the components. Finally, once they know how to compare invariant features, they applied a voting algorithm in order to select a model from the

database.

But one of the interesting things of this work is the inclusion of semi-local constraints to the model in order to reduce possible ambiguities because a given feature might vote for several models. Having a large number of models or many very similar ones raises the probability that a feature will vote for several models. In their work, they propose the use of local shape configurations. For each feature (interest point) in the database, the $p$ closest features in the image are selected. Requiring that all $p$ closest neighbours are matched correctly would mean that there is no miss-detection of points. Thus, they impose that at least 50% of the neighbours match. And in order to increase the recognition rate further, a geometric constraint is added to the model. This constraint is based on the angle between neighbour points. Applying this set of conditions to the recognition scheme, they are able to detect objects robustly.

Their experiments have been conducted for an image database containing 1020 images. They have shown the robustness of the method to image rotation, scale changes, small viewpoint variations, partial visibility and extraneous features. The obtained recognition rate is above 99% for a variety of test images taken under different conditions. We should mention that this technique has been used as a basis for a lot of applications in computer vision since it combines the use of local invariants, a voting algorithm and multi-dimensional indexing for image retrieval obtaining fast queries in the database. For more information see [123].

Classification in this method is based on the use of the k-nearest neighbor technique of local invariant features extracted from objects. The distance measure between invariant features is the Mahalanobis distance.

### Spectral Correspondence for Point Pattern Matching

Carcassoni and Hancock [26] investigate the correspondence matching of point-sets using spectral graph analysis. Since objects can be represented by a set of interesting points, this method can also be considered as a relevant one to take into account in computer vision. The main philosophy of this method is to make use of the spectral graph theory to solve the matching of two point-sets. Spectral graph theory is a term applied to a family of techniques that aim to characterise the global structure properties of graphs using the eigenvalues and eigenvectors of either the adjacency matrix or the closely related Laplacian matrix [30].

Spectral graph theory, as explained in [26], is not very used in computer vision because although elegant, spectral graph representations are notoriously susceptible to the effect of structural error. That is, spectral graph theory can furnish very efficient methods for characterising exact relational structures, but soon breaks down when there are spurious nodes and edges in the graphs under study. Carcassoni and Hancock consider how spectral methods can be rendered robust correspondence matching with points-sets which contain significant noise and contamination. To do this they make use of the framework provided by EM (Expectation Maximization) algorithm.

The most interesting part of their work is the real world data experiments. They have matched images from a gesture sequence of a moving hand. The feature points in these moving hands are points of maximum curvature on the outline of the hand.

They showed that using the probabilistic method based on the Expectation - Maximization (EM) method, all the correspondence matches were correct. Finally, they have extended these experiments to a longer motion sequence in which a hand is clenched to form a fist. They showed that the dual-step EM algorithm performs very good in contrast to the other spectral methods. For more information see [26].

Classification in this method is based on the estimation of probability density functions. They derived a formulation based on Expectation - Maximization (EM) in order to work with spectral graphs and probabilities at the same time.

**Face Recognition by Elastic Bunch Graph Matching**

In [158] Wiskott, Fellous, Kruger and Malsburg presented a system for recognizing human faces from single images out of a large database containing one image per person. The task was difficult because of image variation in terms of position, size, expression, and pose. The system collapses most of this variance by extracting concise face descriptions in the form of *image graphs*. In these, fiducial points on the face (eyes, mouth, etc.) are described by sets of wavelet components (jets). Image graph extraction is based on the *bunch graph*, which is constructed from a small set of sample image graphs. Recognition is based on a straightforward comparison of image graphs.

Their system has an important core of structure which reflects the fact that the images of coherent objects tend to translate, scale, rotate, and deform in the image plane. Their basic object representation was labeled graph; edges are labeled with distance information and nodes were labeled with wavelet responses locally bundled in jets. Stored model graphs can be matched to new images to generate image graphs, which can then be incorporated into a gallery and become model graphs. They used wavelets because they are robust to moderate lighting changes and small shifts and deformations. Model graphs can easily be translated, scaled, oriented, or deformed during the matching process, thus compensating for a large part of the variance of the images.

This general structure is useful for handling any kind of coherent object and may be sufficient for discriminating between structurally different object types. However, for in-class discrimination of objects, of which face recognition is an example, it is necessary to have information specific to the structure common to all objects in the class. This is crucial for the extraction of those structural traits from the image which are important for discrimination ("to know where to look and what to pay attention to"). In their system, class-specific information has the form of *bunch graphs*, one for each pose, which are stacks of a moderate number (70 in their experiments) of different faces, jet-sampled in an appropiate set of fiducial points (places over eyes, mouth, contour, etc.). Bunch graphs are treated as combinatorial entities in which, for each fiducial point, a jet from a different sample face can be selected, thus creating a highly adaptable model. This model is matched to new facial images to reliably find the fiducial points in the image. Jets at these points and their relative positions are extracted and are combined into an image graph, a representation of the face which has no remaining variation due to size, position.

Experimental results demonstrated that this system performs very good since it is able to cope with face rotations and some facial expressions. For more information

about this method see [158].

Classification in this method is based on the use of the k-nearest neighbor technique between bunch graphs. They define a distance measure that takes into account the whole set of jets of each bunch graph in order to find similar bunch graphs stored in the training stage.

### Shape Matching and Object Recognition Using Shape Contexts

In [12] Belongie, Malik and Puzicha presented a novel approach to measure similarity between shapes and they exploited it for object recognition. In their framework, the measurement of similarity is preceded by (1) solving correspondences between points on the two shapes, (2) using the correspondences to estimate an aligning transform. In order to solve the correspondence problem, they attached a descriptor, the *shape context*, to each point. The shape context at a reference point captures the distribution of the remaining points relative to it, thus offering a globally discriminative characterization. Corresponding points on two similar shapes will have similar shape contexts, enabling to solve for correspondences as an optimal assignment problem. Given the point correspondences, they estimate the transformation that best aligns the two shapes; regularized thin-plate splines provide a flexible class of transformation maps for this purpose. The dissimilarity between two shapes is computed as a sum of matching errors between corresponding points, together with a term measuring the magnitude of the aligning transform. They treat recognition in a nearest-neighbor classification framework as the problem of finding the stored prototype shape that is maximally similar to that in the image. They presented results for silhouettes, trademarks, handwritten digits and the COIL dataset.

They treat an object as a point set and they assume that the shape of an object is essentially captured by a finite subset of its points. But, in contrast to other approaches, a shape is represented by a discrete set of points samples from the internal or external contours on the object but they do not use key-points such as maxima of curvature or inflection points. After this, for each point on a shape, they want to find the best matching point on a second shape and they introduce a novel descriptor, the shape context, that plays an important role. The shape context is the fact to consider the set of vectors originating from a point to all other sample points on a shape. These vectors express the configuration of the entire shape relative to the reference point. Thus, if an object contains $n$ points, they create a shape histogram using the remaining $n-1$ points of the shape. This histogram is defined for all $n$ points of a shape and is defined to be the shape context of a point. Once two objects are defined using shape histograms, they analyze the correspondence between two objects using the $\chi^2$ test statistic as a distance measure between shape histograms. And once the correspondence is solved, they try to know which is the transformation that affects to one object in order to obtain the other one.

They presented results on the MNIST dataset of handwritten digits. In the experiments, they used 100 points samples from the Canny edges to represent each digit. They claim that their method is the best one obtained using the MNIST dataset with an error rate of 0.63%. Another experiment considers the COIL-20 database in order to perform 3D object recognition and they obtain very favorable recognition results.

They also tested the method with the MPEG-7 shape silhouette database and again, they obtain the best recognition results of about 76.51%. Finally, they tested the method with trademarks since the shape is the only source of information of trademarks. And results are also very good in this particular case. For more information see [12].

Classification in this method is based on the use of the k-nearest neighbor technique between bunch graphs. They define a distance measure that takes into account the whole set of jets of each bunch graph in order to find similar bunch graphs stored in the training stage.

### 2.3.3 Parts-based Schemes

Recently, and following human visual processing results, it has been proposed an specific class of geometrical scheme for coding local features [143]: a parts-based scheme. This scheme is present when a subset of topologically connected features of an object is grouped to form an entity with a particular set of behaviours. This is what is called object part. Here we show a set of techniques that divide an object into parts.

#### Object Detection using the Statistics of Parts

Schneiderman and Kanade [127, 125, 126] have presented a statistical method for 3D object detection. The method decomposes the 3D geometry of each object into a small number of viewpoints. For each viewpoint, they construct a decision rule that determines if the object is present at that specific orientation. Each decision rule uses the statistics of both object appearance and "non-object" visual appearance. They represent each set of statistics using a product of histograms. And each histogram represents the joint statistics of a subset of wavelet coefficients and their position on the object. Their approach is to use many such histograms representing a wide variety of visual attributes. Using this method, they have developed an algorithm that can reliably detect faces that vary from frontal view to full profile view and an algorithm that can reliably detect cars over a wide range of viewpoints.

To build a trainable object detector for detecting faces and cars at any size, location and pose, they propose to use a detector with multiple classifiers, each spanning a different range of orientation. With this, they are able to cope with variation in object orientation. Each of these classifiers determines whether the object is present at a specified size within a fixed-size image window. And to find the object at any location and size, these classifiers scan the image exhaustively.

One characteristic trait of this method is that it is a parts-based approach because the input variables are grouped into sets, where the relationships within each set are more accurately modeled than those across sets. They labeled to each such set as a *part*. Parts of a face can be the eyes, nose, and mouth and can be considered as *parts* and modeled separately. But, of course, it should be mentioned that this *part* does not need to have a natural meaning to us (such as a nose or an eye), and could be defined as a group of pixels, or transform variables, that satisfy certain mathematical properties. In this work and assuming this definition of *part*, they mention that these

*parts* do not have to be composed from disjoint groups of variables. They say that a variable can be reused in multiple *parts*.

In this work, the final classifier uses a variety of *parts* to embody various combinations of locality in space, frequency, and orientation. Some *parts* represent small regions over high frequencies, other *parts* represent large regions over low frequencies, and still other *parts* are specialized in horizontal and vertical information. And these choices are designed to capture common statistical dependencies in appearance of an object. This is the reason that they chose a wavelet representation because it allows to directly design *parts* with these locality properties.

Their method worked very good using cars and faces as examples to classify. At the end, since the classifier uses *parts* across the full extent of the object, they analyzed which *parts* or areas tended to be most influencial. In this sense, we can always think whether the eyes, nose, and mouth regions are really the most important areas to detect faces? Their results in this aspect show that no particular region in a face seemed to be consistently more influential than the others, and the regions of particular positive influences were not sharply localized but tend to be spread out. Occluded areas usually contributed a negative influence. Also, characteristics that were uncommon in the training set, such as the mottled beard on some man gave negative response.

Classification in this method is based on the estimation of probability density functions through the use of histogram representations.

### Visual features of intermediate complexity and their use in classification

In [143] Ullman assumes that the human visual system analyzes shapes and objects in a series of stages in which stimulus features of increasing complexity are extracted and analyzed. The first stages use simple local features, and the image is subsequently represented in terms of larger and more complex features. These include features of intermediate complexity and partial object views. The nature and use of these higher-order representations remains an open question in the study of visual processing by the primate cortex.

In this work, Ullman shows that intermediate complexity (IC) features are optimal for the basic visual task of classification. Moderately complex features are more informative for classification than very simple or very complex ones, and so they emerge naturally by the simple coding principle of information maximization with respect to a class of images.

The work presented by Ullman is motivated by a fundamental question in the study of visual processing and is the problem of "feature selection". That is, which features of an image are extracted and represented by the visual cortex? Several brain areas are involved in visual object processing, and different features are represented at different stages. It seems that in the earliest processing stages, which involve the retina, lateral geniculate nucleus (LGN) and primary visual cortex (V1), the image is represented by simple local features such as center-surround receptive fields and oriented lines and edges. This encoding can arise from the computational principles of decorrelation and redundancy reduction or from faithful reconstruction of the input using sparse encoding. After this early processing, moderately complex features are represented

in areas V4 and the adjacent region TEO, and finally, partial or complete object views are represented in anterior regions of inferotemporal (IT) cortex. Based on this biological motivations, Ullman shows by computational analysis and simulations, that features of intermediate complexity (IC) and partial object views are optimal for visual object classification.

In this section, we have previously seen that complex objects are represented in terms of simpler elements such as wavelets, Gabor basis functions or Gaussian derivative functions. Furthermore, it has also been proposed that the visual cortex can be represented using such representations. Ullman states that these functions are universal in the sense that they are equally applicable to all natural images and he proposes that the visual system encodes features of intermediate complexity that are class-specific, that is, selected for encoding images within a class of related images. He states the these features are used after the encoding of simple features in V1 but before the encoding of complex object views in anterior IT cortex, and they are specifically selected to support visual classification - one of the basic tasks of visual perception.

They experimented with face and car images in order to extract a correct fragment size that contributes to obtain the best performances. All the experiments demonstrated the superiority of intermediate size fragments in contrast to other size fragments and this superiority can be explained as the interplay of two factors: specificity and relative frequency. This can be explained as follows: a large face fragment can provide reliable indication of the presence of a face in an image, although the likelihood of encountering such a fragment in a novel face image is low. Consequently, the information carried by such a fragment with respect to the class is limited. A smaller fragment has a higher likelihood of appearing in different face images, but the likelihood of its presence in non-face images is also higher. For more information see [143].

This method is based on probability density functions in order to find the intermediate-size of fragments that best describes objects. They used a likelihood estimator to find the correct size of object part.

### A Probabilistic Approach to Object Recognition Using Local Photometry and Global Geometry

In [23] Burl, Weber and Perona presented a new framework to detect parts in objects. The basic assumption of their work is that many object classes of the world, including human faces, can be modeled as a set of characteristic parts arranged in a variable spatial configuration.

In their work, they are interested in object classes in which instances from the class can be modeled as a set of characteristic parts in a deformable spatial configuration. They propose the example of considering human faces, which consist of two eyes, a nose, and mouth. These parts appear in an arrangement that depends on an individual's facial geometry, expression, and pose, as well as the viewpoint of the observer. Here, they do not offer a precise definition of what constitutes an object part, but they refer to any feature of the object that can be reliably detected and localized using only the local image information. Hence, a part may be defined through

a variety of visual cues such as distinctive brightness or orientation pattern, texture, color, motion, or symmetry. Also, parts may also be defined at multiple scales. A coarse resolution view of the head can be considered a "part" as can a fine resolution view of an eye corner. The parts may be object-specific (eyes, nose, mouth) or generic (blobs, corners, textures).

They presented a mathematical framework that starts considering a set of candidate locations for object parts and a probabilistic approach tries to find the best combination. They considered independent part positions, that is, there is not intersection between the object parts in the image. Also, they also considered the case of jointly distributed part positions. To test their method, they chose the problem of detecting faces from frontal views. A grayscale image sequence of 400 frames was acquired from a person performing head movements and facial expressions in front of a cluttered background. The images were $320 \times 240$ pixels in size, while the face occupied a region of approximately 40 pixels in height. Their face model was comprised of five parts, namely eyes, nose tip and mouth corners.

The optimal detector for object classes was derived for the case of independent part positions (not overlapping regions). When the part positions are jointly distributed the optimal detector was too complicated to evaluate, but it can be approximated using a winner-take-all simplification. In both cases, the detector was composed of two terms: the first term measures how well the hypothesized parts in the image match the actual model parts, while the second term measures how well the hypothesized spatial arrangement matches the ideal model arrangement. The resulting method combines the part match with shape and is invariant to translation, rotation, and scaling.

This method is based on creating a probability approach which tries to find a set of locations for object parts that best describe objects.

### Unsupervised Learning of Models for Recognition

In [152, 153] Weber, Welling and Perona presented an extension to the framework explained before. They also worked with the problem of object parts but under another point of view. Here, they presented a method to learn object class models from unlabeled and unsegmented cluttered scenes for the purpose of visual object recognition. They focus on a particular type of model where objects are represented as flexible constellations of rigid parts (features). After this, the variability within a class is represented by a joint probability density function (pdf) on the shape of the constellation and the output of part detectors. In a first stage, the method automatically identifies distinctive parts in the training set by applying a clustering algorithm to patterns selected by an interest operator. It then learns the statistical shape model using expectation maximization. And the method achieves very good classification results on human faces and rear views of cars.

This work is very interesting because they are motivated in the problem of recognizing members of object classes with no previous knowledge about the members. Here, they define an object class as a collection of objects which share characteristic features or parts that are visually similar and occur in similar spatial configurations. They presented three problems in this work which they also tried to solve. The prob-

lems are: (i) *Segmentation or registration of training images*: Which objects are to be recognized and where do they appear in the training images? (ii) *Part Selection*: Which object parts are distinctive and stable? (iii) *Estimation of model parameters*: What are the parameters of the global geometry or *shape* and of the appearance of the individual parts that best describe the training data?

They tested the performance on two data sets: images of rear views of cars and images of human faces. For each of the two object classes they took 200 images showing a target object at an arbitrary location in cluttered background. They also took 200 images of background scenes from the same environment, excluding the target object. No images were discarded by hand prior to the experiments. The face images were taken indoors as well as outdoors and contained 30 different people (male and female). The car images were taking on public streets and parking lots where they photographed vehicles of different sizes, colors and types, such as sedans, sport utility vehicles, and pick-up trucks. In the case of faces, discrimination of images containing the desired object vs. background images exceeds 90% correct with simple models composed of 4 parts. Performance on cars is 87% correct.

The main goal of this work was to demonstrate that it is feasible to learn object models directly from unsegmented cluttered images, and to provide ideas on how one may do so. For more information about this article see [152, 153].

This method is based on creating a probability approach which tries to find a set of locations for object parts that best describe objects. Furthermore, this approach tries to learn what are the most distinctive and stable parts from objects without previous knowledge.

### Minimizing Binding Errors Using Learned Conjunctive Features

In [88] Mel and Fiser have studied some of the design trade-offs governing visual representations based on spatially invariant conjunctive features detectors, with an emphasis on the susceptibility of such systems to false-positive recognition errors (Malsburg's classical binding problem). They derive an analytical model that makes explicit how recognition performance is affected by the number of objects that must be distinguished, the number of features included in the representation, the complexity of individual objects, and the clutter load, that is, the amount of visual material in the field of view in which multiple objects must be simultaneously recognized, independent of pose, and without explicit segmentation. Using the domain of text to model object recognition in cluttered scenes, they show that with corrections for the nonuniform probability and nonindependence of text features, the analytical model achieves good fits to measured recognition rates in simulations involving a wide range of clutter loads, word sizes, and feature counts. They then introduced a greedy algorithm for feature learning, derived from the analytical model, which grows a representation by choosing those conjunctive features that are most likely to distinguish objects from the cluttered background in which they are embedded. They show that the representations produced by this algorithm are compact, decorrelated, and heavily weighted toward features of low conjunctive order. Their results provide a more quantitative basis for understanding when spatially invariant conjunctive features can support unambiguous perception in multiobject scenes, and lead to several insights

regarding the properties of visual representations optimized for specific recognition tasks.

In [88] we can find a model for object recognition based on text in order to present a greedy algorithm to reduce hallucinated objects (in this case, words). The studies of this work were carried out in text world because it contains many of the complexities of vision in general: target objects are numerous (more than 40.000 words in the database), are highly nonuniform in their prior probabilities, are constructed from a set of relatively few underlying parts (26 letters), and individually can contain from 1 to 20 parts. Furthermore, the parts from which words are built are highly nonuniform in their relative frequencies and contain strong statistical dependencies.

They tested two databases: (i) a word database that contains 44.414 entries, representing all lowercase punctuation-free words and their relative frequencies found in 5 million words in the Wall Street Journal (WSJ); (ii) an english text database that consisted of approximately 1 million words drawn from a variety of sources [88]. Their work is based on $n$-grams (defined as a binary detector that responds when a specific spatial configuration of $n$ letters is found anywhere (one or more times) in the input array).

This works presents a lot of interesting issues that are usually involved in actual object representation techniques and for more information see [88].

This approach is based on a probabilistic scheme which tries to choose the conjunctive features that best distinguish objects from cluttered backgrounds.

## 2.4 Conclusions

In this introductory chapter we have introduced one of the most common problems of computer vision: The representation of object information (*what* and *where*) and using this information to detect and recognize objects.

Firstly, we introduced the problem of object recognition and we realized that it is not an easy task. Since the art of identify objects in images is very complex, researchers tried to solve this problem proposing different frameworks. Appearance-based models seem to be a good alternative for current approaches. The main motivation behind the use of appearance-based models is that it seems that the human visual system works directly with retinal stimulations. Analogously, these retinal stimulations can be seen as the response to certain filters which are directly related to the appearance of objects (gaussian derivatives, etc). Thus, we organized this chapter in three main parts: (i) A section which is related to *what* local information is extracted from images, (ii) A section which is related to *where* this local information is taken from, (iii) A section which is related to *how* the *what* and *where* information has been historically used in order to detect and recognize objects in images. This last part consists of a survey of well-known techniques where we describe each technique and we explain the main issues related to object classification.

As observed in this chapter, we exposed a typical framework for object recognition. It has been shown that appearance-based methods served as a first mechanism to decide upon the main strategy of object recognition and classification. Then, we described why local feature based approaches are usually selected to perform classi-

fication of objects. It is stated that global representations have several drawbacks. They perform badly in front of scenes with occlusions, severe lighting conditions, changes in viewpoint, etc. As advantages, we should consider its simplicity and fast recognition performances. They could be adapted to partially solve all this set of drawbacks, but they would lose its simplicity and speed. Local approaches are presented in order to inherently solve some of these drawbacks. However, it turns out that they are complex and they require a lot of computational resources and time.

Chapter 3 introduces a set of statistical techniques which are based on local representations. Without considering the spatial relationship among local features, we will use different statistical representations in order to perform object recognition and classification. In conjunction with well-known techniques, we present new schemes for object recognition which are specially adapted to positive local representations. Then, we perform several comparisons among all this set of statistical methods.

Chapter 4 introduces a new scheme for representing local features as well as their spatial arrangement in an object. We create a model that takes into account *what* information is used and from *where* is taken from in one single model. In doing so, we use a parametric approach which models *what* and *where* information providing a robust framework for object recognition and classification as the computational costs of such a scheme are acceptable.