Universitat
Autònoma
de Barcelona

Statistical Local Appearance Models for Object Recognition

A dissertation submitted by **David Guillamet Monfulleda** at Universitat Autònoma de Barcelona to fulfil the degree of **Doctor en Informàtica**.

Bellaterra, December 11, 2003

Director: **Dr. Jordi Vitrià i Marca**
Universitat Autònoma de Barcelona
Dept. Informàtica & Computer Vision Center

Centre de Visió
per Computador

This document was typeset by the author using LaTeX 2_ε.

The research described in this book was carried out at the Computer Vision Center, Universitat Autònoma de Barcelona.

ISBN 84–932156–8–6

Printed by Ediciones Gráficas Rey, S.L.

A la meva família

Nunca consideres el estudio como una obligación, sino como una oportunidad para penetrar en el bello y maravilloso mundo del saber.

Albert Einstein

No puede usted resolver un problema? Pues bien, póngase a investigar su situación actual y sus antecedentes! Cuando haya investigado cabalmente el problema, sabrá cómo resolverlo. Toda conclusión se saca después de una investigación, y no antes. Únicamente un tonto se devana los sesos, sólo o unido a un grupo, para encontrar una solución o elaborar una idea sin efectuar ninguna investigación. Debe subrayarse que esto no conducirá en absoluto a ninguna solución eficaz ni a ninguna idea provechosa.

Contra el culto a los libros (mayo de 1930).

Agraïments

Aquesta tesi ha estat realitzada gràcies a una beca de la Generalitat de Catalunya. Gràcies a la qual he pogut realitzar la tesi sense problemes temporals i he pogut gaudir de dues magnífiques estades a l'estranger (Zurich i Boston) que han servit com a complement a tota la meva recerca.

En primer lloc voldria agrair la paciència, el suport i el recolzament que el meu director de tesi, el Dr. Jordi Vitrià, ha tingut amb mi durant aquests últims 4 anys. Gràcies a la seva confiança, la seva disponibilitat, els seus coneixements i consells generosos aquesta tesi és el que és en aquest moment. I gràcies a ell, no he decaigut en el llarg i dur camí de la recerca.

En segon lloc voldria agrair al Dr. Juan José Villanueva que, com a director del Centre de Visió per Computador, m'ha brindat l'oportunitat d'investigar en aquest entorn privilegiat.

Molts dels treballs i experiments que es presenten en aquesta tesi son producte de dues estades de recerca. En aquest sentit, vull agrair al Dr. Bernt Schiele del grup Perceptual Computing and Computer Vision (PCCV) de la Eidgenössische Technische Hochschule (ETH) de Zürich l'oportunitat de pasar 3 mesos en el seu laboratori. De la mateixa manera, vull agrair al Dr. Baback Moghaddam de Mitsubishi Electric Research Laboratories (MERL) de Boston l'oportunitat de realitzar un internship de 3 mesos amb un dels seus projectes. Gràcies a aquesta última estada, he pogut descobrir quina és la sensació de treballar en una de les empreses més importants del món de la recerca i aprendre que la qualitat de recerca no depèn del país d'origen, sinó de la persona que la porta a terme.

L'algorisme Class-Conditional Independent Component Analysis (CC-ICA) i el desenvolupament del mètode Weighted Non-negative Matrix Factorization (WNMF) presentats en aquesta tesi han sorgit de l'estreta relació mantinguda durant aquests anys de recerca amb el Dr. Marco Bressan, un gran professional que m'ha ajudat en tot moment en tots els meus dubtes existencials referents a la recerca. A més, com a gran company de despatx que va ser durant un període d'un any, vull agrair-li les converses que hem tingut no referents a la recerca i que m'han ajudat a entendre altres perspectives de la vida desconegudes per mi.

També m'agradaria agrair la paciència de tots els companys amb els que he compartit despatx durant aquest llarg camí de 4 anys. En primer lloc, agrair als primers integrants de despatx: Juanma, Xevi, Botey, Ramon i Poal. En especial atenció al Juanma per la multitud de converses que hem tingut en relació a la recerca i a d'altres aspectes de la vida. En ell he trobat un bon company amb el que he disfrutat grans moments. També vull fer especial menció als meus dos últims companys de despatx: la Anna i el Poal. En particular, vull agrair a la Anna la seva paciència respecte els comentaris bursàtils esporàdics que feien perillar la tranquilitat de l'ambient del despatx. També voldria agrair al meu company Poal els moments que hem viscut junts, en especial menció als moments de festa donat que sempre hi ha coses positives a recordar.

Gràcies a tots aquells que formen part del CVC. Per començar, moltes gràcies a tots aquells que fan funcionar el CVC: [mjose, pilar, montse, mcmerino, pedro]@cvc.uab.es. També vull donar les gràcies als doctors del CVC que m'han ajudat en alguns dels petits dubtes que he tingut durant tots aquests anys. A la resta de gent de la unitat, també moltes gràcies. Voldria fer especial menció al David Masip, gran company de recerca, d'activitats físiques (SAF) i de converses bursàtils amb el qual he passat grans moments d'alegria i/o pessimisme en les nostres inversions.

També voldria agrair als meus amics la seva presència en la meva vida. Primer de tot, donar les gràcies al David i Edu la seva companyia i molts dels moments que hem viscut junts durant aquest temps. També m'agradaira mencionar l'amistat sorgida amb la Isa, la Lupe, Greig i Luz durant aquest últim any. Finalment, donar les gràcies a la Alicia pel recolzament que m'ha donat en aquesta última etapa de la tesi.

Voldria també donar les gràcies a la meva família més directe pel seu suport i la seva gran voluntat per soportar-me durant la realització d'aquesta tesi. En especial a la meva mare, per haver-me facilitat l'oportunitat d'estudiar, per haver confiat en mi, per haver-me estimat com ho ha fet i sense la qual no hauria arribat on sóc actualment. També agrair la presència de les meves dues germanes: Mercè i Gemma en els moments difícils de la realització d'aquesta tesi.

Resum

Durant els últims anys, hi ha hagut un interès creixent en les tècniques de reconeixement d'objectes basades en imatges, on cadascuna de les quals es correspon a una aparença particular de l'objecte. Aquestes tècniques que únicament utilitzen informació de les imatges són anomenades tècniques basades en l'aparença i l'interès sorgit per aquestes tècniques és degut al seu éxit alhora de reconèixer objectes. Els primers mètodes basats en l'aparença es recolzaven únicament en models globals. Tot i que els mètodes globals han estat utilitzats satisfactòriament en un conjunt molt ampli d'aplicacions basades en la visió per computador (per exemple, reconeixement de cares, posicionament de robots, etc), encara hi ha alguns problemes que no es poden tractar fàcilment. Les oclusions parcials, canvis excessius en la il·luminació, fons complexes, canvis en l'escala i diferents punts de vista i orientacions dels objectes encara són un gran problema si s'han de tractar des d'un punt de vista global. En aquest punt és quan els mètodes basats en l'aparença local van sorgir amb l'objectiu primordial de reduir l'efecte d'alguns d'aquests problemes i proporcionar una representació molt més rica per ser utilitzada en entorns encara més complexes.

Usualment, els mètodes basats en l'aparença local utilitzen descriptors d'alta dimensionalitat alhora de descriure regions locals dels objectes. Llavors, el problema de la maledicció de la dimensionalitat (*curse of dimensionality*) pot sorgir i la classificació dels objectes pot empitjorar. En aquest sentit, un exemple típic per alleujar la maledicció de la dimensionalitat és la utilització de tècniques basades en la reducció de la dimensionalitat. D'entre les possibles tècniques per reduir la dimensionalitat, es poden utilitzar les transformacions lineals de dades. Bàsicament, ens podem beneficiar de les transformacions lineals de dades si la projecció millora o manté la mateixa informació de l'espai d'alta dimensió original i produeix classificadors fiables. Llavors, el principal objectiu és la modelització de patrons d'estructures presents als espais d'altes dimensions en espais de baixes dimensions.

La primera part d'aquesta tesi utilitza primordialment histogrames color, un descriptor local que ens proveeix d'una bona font d'informació relacionada amb les variacions fotomètriques de les regions locals dels objectes. Llavors, aquests descriptors d'alta dimensionalitat es projecten en espais de baixes dimensions tot utilitzant diverses tècniques. L'anàlisi de components principals (PCA), la factorització de matrius amb valors no-negatius (NMF) i la versió ponderada del NMF són 3 transformacions lineals que s'han introduit en aquesta tesi per reduir la dimensionalitat de les dades i proporcionar espais de baixa dimensionalitat que siguin fiables i mantinguin les estructures de l'espai original. Una vegada s'han explicat, les 3 tècniques lineals són àmpliament comparades segons els nivells de classificació tot utilitzant una gran diversitat de bases de dades. També es presenta un primer intent per unir aquestes tècniques en un únic marc de treball i els resultats són molt interessants i prometedors. Un altre objectiu d'aquesta tesi és determinar quan i quina transformació lineal s'ha d'utilitzar tot tenint en compte les dades amb que estem treballant. Finalment, s'introdueix l'anàlisi de components independents (ICA) per modelitzar funcions de densitat de probabilitats tant a espais originals d'alta dimensionalitat com la seva extensió en subespais creats amb el PCA. L'anàlisi de components independents és una tècnica lineal d'extracció de característiques que busca minimitzar les dependències d'alt ordre. Quan les seves assumpcions es compleixen, es poden obtenir característiques estadísticament independents a partir de les mesures originals. En aquest sentit, el ICA s'adapta al problema de reconeixement estadístic de patrons de dades d'alta dimensionalitat. Això s'aconsegueix utilitzant representacions condicionals a la classe i un esquema de decisió de Bayes adaptat específicament. Degut a l'assumpció d'independència aquest esquema resulta en una modificació del classificador ingenu de Bayes.

El principal inconvenient de les transformacions lineals de dades esmentades anteriorment és que no consideren cap tipus de relació entre les característiques locals. Conseqüentment, es presenta un mètode per reconèixer objectes tridimensionals a partir d'imatges d'escenes desordenades, tot utilitzant un únic model après d'una imatge de l'objecte. Aquest mètode es basa directament en les característiques visuals locals extretes de punts rellevants dels objectes i té en compte les relacions espaials entre elles. Aquest nou esquema redueix l'ambigüitat de les representacions anteriors. De fet, es presenta una nova metodologia general per obtenir estimacions fiables de distribucions conjuntes de vectors de característiques locals de múltiples punts rellevants dels objectes. Per fer-ho, definim el concepte de k-tuples per poder representar l'aparença local de l'objecte a k punts diferents i al mateix moment les dependències estadístiques entre ells. En aquest sentit, el nostre mètode s'adapta a entorns desordenats, complexes i reals demostrant una gran habilitat per detectar objectes en aquests escenaris amb resultats molt prometedors.

Abstract

During the last few years, there has been a growing interest in object recognition techniques directly based on images, each corresponding to a particular appearance of the object. These techniques which use only information of images are called appearance based models and the interest in such techniques is due to its success in recognizing objects. Earlier appearance-based approaches were focused on the use of holistic approaches. In spite of the fact that global representations have been successfully used in a broad set of computer vision applications (i.e. face recognition, robot positioning, etc), there are still some problems that can not be easily solved. Partial object occlusions, severe lighting changes, complex backgrounds, object scale changes and different viewpoints or orientations of objects are still a problem if they should be faced under a holistic perspective. Then, local appearance approaches emerged as they reduce the effect of some of these problems and provide a richer representation to be used in more complex environments.

Usually, local appearance methods use high dimensional descriptors to describe local regions of objects. Then, the curse of dimensionality problem appears and object classification degrades. A typical example to alleviate the curse of dimensionality problem is to use techniques based on dimensionality reduction. Among possible reduction techniques, one could use linear data transformations. We can benefit from linear data transformations if the projection improves or mantains the same information of the high dimensional space and produces reliable classifiers. Then, the main goal is to model low dimensional pattern structures present in high dimensional data.

The first part of this thesis is mainly focused on the use of color histograms, a local descriptor which provides a good source of information directly related to the photometric variations of local image regions. Then, these high dimensional descriptors are projected to low dimensional spaces using several techniques. Principal Component Analysis (PCA), Non-negative Matrix Factorization (NMF) and a weighted version of NMF, the Weighted Non-negative Matrix Factorization (WNMF) are 3 linear transformations of data which have been introduced in this thesis to reduce dimensionality and provide reliable low dimensional spaces. Once introduced, these three linear techniques are widely compared in terms of performances using several databases. Also, a first attempt to merge these techniques in an unified framework is shown and results seem to be very promising. Another goal of this thesis is to determine when and which linear transformation might be used depending on the data we are dealing with. To this end, we introduce Independent Component Analysis (ICA) to model probability density functions in the original high dimensional spaces as well as its extension to model subspaces obtained using PCA. ICA is a linear feature extraction technique that aims to minimize higher-order dependencies in the extracted features. When its assumptions are met, statistically independent features can be obtained from the original measurements. We adapt ICA to the particular problem of statistical pattern recognition of high dimensional data. This is done by means of class-conditional representations and a specifically adapted Bayesian decision scheme. Due to the independence assumption this scheme results in a modification of the naive Bayes classifier.

The main disadvantage of the previous linear data transformations is that they do not take into account the relationship among local features. Consequently, we present a method of recognizing three-dimensional objects in intensity images of cluttered scenes, using a model learned from one single image of the object. This method is directly based on local visual features extracted from relevant keypoints of objects and takes into account the relationship between them. Then, this new scheme reduces the ambiguity of previous representations. In fact, we describe a general methodology for obtaining a reliable estimation of the joint distribution of local feature vectors at multiple salient points (keypoints). We define the concept of k-tuple in order to represent the local appearance of the object at k different points as well as the statistical dependencies among them. Our method is adapted to real, complex and cluttered environments and we present some results of object detection in these scenarios with promising results.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

There are a number of computer vision applications which involve, one way or another, the recognition of objects in images. In common robot applications, a robot searches for known objects in images continuously acquired by a camera. In common image retrieval applications, a collection of images is searched for views of a specific object or images showing objects from an object category, for instance a specific actor, members of a sports team or images of flowers. In object-based video annotation, video frames are automatically labeled with symbolic descriptions which may be connected to the presence of certain objects in the sequence. Using the symbolic descriptions, efficient archival and retrieval of images or sequences of interest is possible. Recognition may be also useful for image catalogue searching/browsing common in art, trademark and other commercial applications. Recently, web-based systems have been developed, searching on the internet for images showing desired objects.

As seen, object detection and recognition is a pervasive activity in our lives. We are constantly looking for, detecting and recognizing objects: people, streets, buildings, tables, chairs, desks, sofas, beds, automobiles, etc. Yet it remains a mystery how we perceive objects so accurately and with so little apparent effort. Is for this reason that during the last few years, there has been a growing interest in object recognition directly based on images, each of them containing an object or a set of objects. So that, object recognition can be defined as the task of determining whether (and/or which) objects appear in a collection/sequence of images. But in order to recognize an object one must know, at least, something about the object. Thus, one must have an object representation that describes what are the differences of this object with respect to the other ones. In this context, detection or recognition of objects has resulted in a very difficult task and in over 30 years of research, progress has been rather limited. The central problem is how to deal with a huge amount of variation in visual appearance. That is, how we can obtain a universal representation of an object that is able to cope with both the variation within the object and with the diversity of visual imagery that exists in the world. It is not the scope of this dissertation to propose a universal object representation scheme. However, this thesis shows how a computer system can acquire a possible and feasible (in terms of computational costs) model of an object's appearance directly from intensity images only using local descriptors, and then use that model to recognize the object in other scenes.

The earlier recognition systems were mainly focused on a holistic approach. In spite of the fact that global representations have been successfully used in different computer vision applications (i.e. face recognition, robot positioning, etc.), there are some problems that can not be easily solved under this framework. Such problems are partial object occlusions, severe lighting changes, complex backgrounds, object scale changes and different viewpoints or orientations of objects. In contrast to this, local representations have been recently proposed to solve these problems or to reduce the effect of some of them (i.e. occlusions are difficult to deal with global representations but occlusion effects are reduced when dealing with a local scheme). The idea behind local approaches is that "local" models of objects allow a richer representation and, consequently, the object model can be used in more complex situations.

Regarding natural vision systems, it seems clear that our brain works using local stimuli when detects "things" in the world through visual images coming from our eyes [87]. Then, the conjunction of all local perceived stimuli seems to create a final label (i.e. the name of an object) that helps us to understand what we are looking at. What it not seems clear is how the brain works and, usually, when one tries to obtain a computational approach of this scheme, the method fails because it is not possible to cope with all the possible world situations.

The decision whether a given object appears in the image is made by comparing image measurements with an object representation which is called the *object model*. And to represent an object, a recognition system may exploit various quantities dependent on object properties such as its shape, surface reflectance, etc. Thus, the object model consists of some descriptions of object properties and/or assumptions about the scene parameters which influence the variation in object appearance in different views. So that, given an object model and measurements from a test image, recognition of an object is achieved if some measurements from the test image can be explained as originating from the learned object.

In our work, a model of object appearance is derived from a diversity of measurements which are directly related to the local appearance of the object. We present different local appearance descriptors but we are mainly focused on the positive ones. In particular, we present comparative studies between traditional methods (Principal Component Analysis, Independent Component Analysis) and more recent methods (Non-negative Matrix Factorization) when extracting and representing local information of objects in frameworks such as recognition of faces, handwritten digits, pharmaceutical products and natural scenes.

The huge amount of local information which can be extracted from scenes and objects usually generates high dimensional spaces. High dimensional spaces tend to produce a common problem known as *curse of dimensionality* [11]. The expression *curse of dimensionality* is due to Bellman and in statistics it relates to the fact that the convergence of any estimator to the true value of a smooth function defined on a space of high dimension is very slow. In terms of object recognition and detection, this means that, a priori, we need an enormous amount of observations of an object to obtain a good estimate of a function that identifies this object unambiguously. And one of the important strategies to overcome the curse of dimensionality of such high dimensional representations, is the reduction of the input space. Then, we present a set of methods which are based on a linear transform of the original data that also involve a reduction of the original data dimensionality.

1.1 Motivation

The thesis proposes the use of different statistical techniques for the representation and recognition of objects. Under the assumption that the nature of the object motivates the use of a different technique and that we do not know any universal technique to model objects, we present a set of statistical methods that are compared in laboratory cases and real scenes. Our main goal is to present various alternatives and allow the reader the chance to select the best one for her/his problem. As exposed before and as it will be treated throughout the thesis, we propose a local appearance scheme in order to obtain a flexible model of an object which is able to deal with problems such as occlusions, slight variations in the point of view, different lighting conditions, etc. There are several local techniques to be used in object recognition/classification but, at the end, what we have is a set of feature vectors that are high-dimensional. For instance, the number of pixels can be used as a feature vector and this implies an internal high-dimensional representation. In general, adopting a local framework may offer the following advantages:

- **Stability**: Small changes in an object should produce small changes in an object model. A little modification of the object would affect to the whole description of the object if we consider a global representation. Thus, a local scheme reduces this effect.
- **Efficiency**: A local scheme is based on a portion of the input object, so that, its representation is more efficient than obtaining a global descriptor.

but, also, the following disadvantages:

- **Uniqueness**: Usually, a global approach obtains a unique representation of an object in contrast to a local representation that could generate a set of local descriptors with the same local information. This is due to the fact that objects are composed of similar parts. But, this problem can be solved by using high-level information about the object or building a model that takes only the relevant information about objects.
- **Speed/Complexity/Storage**: A global approach is, in general, very fast because it is based on the whole object and usually, we only have to compare two feature vectors (the vector corresponding to the analyzed object with respect to one feature vector stored during the training stage of another object). But,

a local representation leads to have a large amount of feature vectors of one object that should be compared with other amounts of feature vectors corresponding to other objects in the database. It is clear that using a local scheme, the complexity of the object model is increased deriving in a complex method to match two different representations. Furthermore, the storage required for such local representations is also increased with respect to a global scheme since more information is needed.

A good trade-off between these criteria lead most researchers to several different techniques. But, what we have to take in mind is that a model based on local features should be described by a combination of local features, each pertaining to a specific region of an object. Local features can be computed with relative efficiency as each one is based on a limited portion of the object. Moreover, a description composed of local features can be relatively stable as only some of the features need be affected by any small change in appearance. And, of particular importance for object recognition, partial occlusion of an object will only partly affect recovery of a local feature description. This last statement is the one that motivates that this thesis is completely based on local descriptors.

As exposed before, a local representation leads to high-dimensional and complex representations where we need to estimate a high-level function which classifies objects. With a local representation in our hands, what we try to pursue with this thesis is:

- **Simplicity** in our representation. We believe that complex representations lead to non understandable models or highly complex estimators that should be avoided.
- **Low complexity**. We present a set of local techniques which are based on reducing dimensionality in order to overcome the curse of dimensionality problem. This is a first step to keep the same information and reduce time and storage.
- **Higher-order dependencies**. Dependencies among local features can also be considered in order to analyze possible higher-order dependencies. So that, we also analyze one possible scheme for such a representation.

1.2 Background

Object recognition is neither a new problem nor one that we can consider solved. Indeed, several studies working on this specific field of computer vision have been presented during the last years. Even having a set of real world environments analyzed with complex techniques which are able to deal with specific objects under a huge variety of natural conditions, it is always possible to find a particular position of the object in a specific background where the method fails in its attempt to detect it. That is, object recognition is difficult and still will be in the future. The reason is easy: we are trying to extract some meaningful concept (for example, the name of an object) from a collection of pixels. These set of pixels can contain an object that we previously learned and, of course, it will be embedded in scenes, combined and recombined in a highly variable manner. Would it be possible to find this universal object representation scheme? We do not know this answer and it is not the scope of this dissertation to find a universal representation, but we believe that objects can not be described uniquely.

Object appearance has a large range of variation. It varies with changes in viewpoint, in lighting, and, if the object is flexible, with changes in shape. Of course, since we are working with images, the object viewed through a camera suffers from an optical distortion and pixel quantization. Also, when recognizing a class of similar objects, we have to use discriminant features to distinguish between them. Moreover, objects are embedded in scenes, combined and recombined with other objects in a highly variable manner. Such problems should be addressed and the possible solution will surely have implications throughout the system.

In any case, a system based on the recognition of objects should be composed of two different parts: (i) Representation/Learning and (ii) Classification/Recognition. The first stage, named representation, also includes the learning of an object. At some point, representation is synonymous of learning and they are complementary. How we learn an object? We learn an object by extracting relevant information from images which contain the objects we want to learn in order to build an object model. This final object model should be able to generalize its knowledge to images which contain unseen instances of the same learned objects. Recognition or classification of an object is performed by finding a match between a set of measurements of the image and an object model stored during the learning process. However, in order to decide upon what information shall we use to represent and classify objects, we have to take in mind: (i) what objects we want to identify and (ii) how we can do it. Several decisions should be made at this point:

- **What** local information should be represented? Since we have objects in images and images are collections of pixels, there is a large amount of information to be considered as relevant or discriminant for object recognition. For example, we can use color information [137, 62, 61], texture information [86, 130, 98, 109], contour features [107], etc. And there are some methods that use a collection of different features in order to find the most discriminant ones and to be able to model the maximum amount of objects [87].

- From **where** we have to extract the local information and where are the objects? Once we know we have an object in an image, one question appears in our mind: all the parts of the object are relevant or maybe we only have to take into account some ”discriminant” ones for our learning/representation scheme? Different approaches have been presented, some of them based on a fixed grid of points [120] from where we extract local information and other ones based on some ”relevant” points of the object (named keypoints [55]) obtained using some pixel statistics (i.e. from zones of maximum curvature).

- What **technique** should be used in order to learn the information extracted from images? There is not enough to extract a lot of local information from arbitrary regions of objects, we also need a technique which is capable to learn all this information in a reasonable computer time with the possibility to recognize the learned objects in an efficient way. This technique should be able to provide the object entity present in an image and its spatial location.

At the end, when we already know what local information should be extracted from specific locations in an image it remains to know how we should computationally represent this information. Usually, information is stored as a collection of feature vectors, that is, a set of n-dimensional feature vectors which encode the information of specific points of an object. Thus, an object is composed of several feature vectors extracted from relevant regions of the object (we would have to think about what is relevant for us). There are several methods and local descriptors to be used at this point. However, we usually obtain a high-dimensional representation. This representation is high-dimensional because we deal with a great amount of information. As explained before, this possible high dimensional space can produce the well-known problem of *curse of dimensionality* when we try to work with all this information. There are several methods that group some feature vectors under the assumption that they belong to the same object part. Thus, the complexity of our initial representation is considerably reduced. Here, we define the word *part* as a particular region of an object which contains several local feature vectors with similar appearances. Then, recognition of an object derives, for example, to estimate a probability density function. In order to estimate reliable probability density functions from a set of feature vectors, researchers conducted their experiments to find efficient techniques to do this task. Possible techniques are histogram representations [120, 87], ICA based schemes [20, 91, 64, 19], etc. All these techniques are valid but they have produced object recognition systems that are still severely limited in the variety of objects and/or viewing situations with which they can cope. Usually, these limitations are, in considerable part, due to a reliance on models that contain too little information. Thus, it is clear that introducing more information one may obtain an improved representation but this means a higher-dimensional space to deal with.

Furthermore, once we have a set of feature vectors corresponding to some relevant regions/points of objects, one may consider the relationships among these local and relevant points. It is clear that a purely based local approach without information about the spatial layout of the neighbor local points is severely limited because produces ambiguity in the representation. So that, given a local point in an object, its neighborhood can be very important to reduce the level of ambiguity. This is the main objective of the higher-order representations because they try to take into account the spatial configuration of the local points of an object.

1.3 Thesis Statement and Contributions

The research described in this dissertation has been divided into two parts: (i) representation of local feature vectors using different statistical methods to allow reliable recognition strategies and (ii) derivation of a framework based on higher-order dependencies of feature vectors to improve the former approach:

- In part I of the thesis, we propose different statistical techniques which are mainly based on reducing the dimensionality of a set of feature vectors to perform object recognition and classification. Then, the *curse of dimensionality* is alleviated. Typical problems of computer vision such as face recognition, handwritten digit recognition, recognition of pharmaceutical products, recognition of patches of natural objects are solved using these techniques. We use Principal Component Analysis (PCA), Non-negative Matrix Factorization (NMF) and Weighted Non-negative Matrix Factorization (WNMF). Performances of these techniques are compared under different kinds of classification frameworks. Classification is performed using different approaches as using (i) reconstruction distances, (ii) parametric models and (iii) non-parametric models. PCA, NMF and WNMF are based on reducing the dimensionality of a problem. They obtain a new subspace which is able to reproduce the original feature space but using less dimensions. So that, we want to determine when and what technique should be used depending on the data we are dealing with. Finally, we introduce Independent Component Analysis (ICA) in order to model probability density functions in the original feature spaces as well as its extension to subspaces obtained using PCA. Object recognition and classification consists of determining the entity of an object given a set of possible objects. Since ICA is not adapted to compare different classes of data, the approach has been extended to work with different classes. We called this extension Class-Conditional Independent Component Analysis (CC-ICA).

- In part II of the thesis, we propose a novel object recognition model that integrates, in addition to basic visual feature frequencies, several higher order characteristics of the objects, including conditional probabilities of visual feature co-occurrences and feature neighborhood arrangements. The combinatorial explosion problem that can arise from this model is solved by using statistical factorization techniques that greatly simplify feature joint probability density estimation. In this framework, we take advantage of the Independent Component Analysis (ICA) to obtain factored joint distributions of image features and reduce the complexity of the problem. Several evaluations using real-world scenes and objects have been performed. Results are satisfactory and provide a robust framework for object detection and recognition mainly adapted to cluttered scenes.

1.4 Outline of the Thesis

In the following, the content of each chapter is summarized.

In chapter 2, we summarize several references which have been source of inspiration for different aspects of the thesis. In particular, we enumerate a set of local descriptors which can be extracted directly from an image. These local descriptors provide a good source of information directly related to the photometric variations of local image regions. We may point out the popularity of these local descriptors in the context of object representation. Once we know *what* local information should be extracted from objects, we introduce the problem of deciding from *where* they should be extracted to obtain reliable descriptions. Then, we survey different object recognition methods in order to know *how* can we merge local descriptors and their spatial arrangement in an optimal manner.

Chapter 3 is devoted to the evaluation of several statistical techniques. We provide a simple derivation of the statistical technique of Principal Component Analysis (PCA), a technique that has dominated the appearance-based approach to vision. Then, we introduce another scheme, the Non-negative Matrix Factorization (NMF) which is similar to the PCA but assuming non-negative data. Finally, we propose a new statistical technique called Weighted Non-negative Matrix Factorization (WNMF) to overcome some of the encountered difficulties of the traditional NMF. These three techniques are widely compared in terms of performances using several databases. Since they are also based on reducing the data dimensionality, the problem of *curse of dimensionality* is inherently solved. At the end, the Independent Component Analysis (ICA) is introduced in order to obtain factored probability density functions. We adapted ICA to perform with problems where we have a set of object classes, is what we called Class-Conditional Independent Component Analysis (CC-ICA). A final example showing its perfomance is provided.

The previous chapter 3 focuses on methods which make use of local descriptors without considering any spatial arrangement of them. In chapter 4, we propose a novel object recognition model that integrates, in addition to basic visual feature frequencies, several higher order characteristics of the objects. The combinatorial explosion problem that can arise from this model is solved by using statistical factorization techniques that greatly simplify feature joint probability density estimation. In this framework, we take advantage of the Class-Conditional Independent Component Analysis (CC-ICA) to obtain factored joint distributions of image features and reduce the complexity of the problem. Several evaluations using real-world scenes and objects have been performed. Results are satisfactory and provide a robust framework for object detection and recognition mainly adapted to cluttered scenes.

Chapter 5 summarizes the main contributions of this thesis and discusses possible avenues of future research.