



**Universitat
Autònoma
de Barcelona**

Statistical Local Appearance Models for Object Recognition

A dissertation submitted by **David Guil-
lamet Monfuleda** at Universitat Autònoma de
Barcelona to fulfil the degree of **Doctor en In-
formàtica**.

Bellaterra, December 11, 2003

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



This document was typeset by the author using L^AT_EX 2_ε.

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

Copyright © 2003 by David Guillaumet Monfuleda. All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the author.

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 decaïgut 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 tranquil·litat 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 meua 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'agradaria 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 meua família més directe pel seu suport i la seva gran voluntat per suportar-me durant la realització d'aquesta tesi. En especial a la meua 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 èxit 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 complexos, 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 complexos.

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'altas 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 introduït 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 extreïdes de punts rellevants dels objectes i té en compte les relacions espacials 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, complexos 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 maintains 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 deter-

mine 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

Agraiments	i
Resum	iii
Abstract	v
1 Introduction	1
1.1 Motivation	3
1.2 Background	4
1.3 Thesis Statement and Contributions	6
1.4 Outline of the Thesis	7
2 Precedents	9
2.1 Local Features	14
2.1.1 Gaussian Derivatives	15
2.1.2 Gabor Filters	20
2.1.3 Color Features	21
2.1.4 Contour Features	33
2.1.5 Local Invariants	34
2.1.6 Eigenfeatures	36
2.1.7 Filters Based on Natural Image Statistics	38
2.2 Location and Geometry	44
2.2.1 Location of Local Features	44
2.2.2 Geometry of Local Features	48
2.3 Local Object Recognition Techniques	50
2.3.1 Feature Based Schemes	50
2.3.2 Higher-order Schemes	53
2.3.3 Parts-based Schemes	57
2.4 Conclusions	62
3 Object Recognition Using Local Features	65
3.1 Example with Local Features	69
3.2 Data Representation using PCA	73
3.3 Data Representation using NMF	74
3.3.1 Non-negative Matrix Factorization	75

3.3.2	Euclidean Distance Algorithm	76
3.3.3	Divergence Algorithm	77
3.3.4	Graphical Example of NMF	78
3.3.5	Projection of Unseen Vectors	81
3.4	Data Representation using WNMf	81
3.4.1	Weighted Non-negative Matrix Factorization	82
3.4.2	Weight Matrix Selection	84
3.4.3	Comparison of NMF versus WNMf	85
3.5	Empirical Analysis of Linear Transformations using Local Features	88
3.5.1	Classification using Reconstruction Distances	90
3.5.2	Classification using a Parametric Model	101
3.5.3	Classification using a Nonparametric Model	104
3.6	Estimation of the Original Probability Density Function	122
3.6.1	Independent Component Analysis	124
3.6.2	Class Conditional Independent Component Analysis	130
3.6.3	Application of ICA/CC-ICA	134
3.7	Conclusions	140
4	Higher-Order Dependencies	143
4.1	Introduction	143
4.2	Methodology	145
4.2.1	Joint Distribution of k -tuples	145
4.2.2	Density Factorization with ICA	146
4.3	Experimental Results	150
4.3.1	From Histograms to Mixture Models	150
4.3.2	Appearance + Color Models	160
4.3.3	Incorporating Local Geometry	161
4.3.4	Invariance to Partial Occlusion	162
4.3.5	Tuple Selection	165
4.3.6	Modeling Higher-Order Dependencies	170
4.4	Conclusions	174
5	Concluding Remarks	177
5.1	Conclusions	177
5.2	Future Work	181
A	Principal Component Analysis	183
A.1	PCA by Maximizing Variance	183
A.2	Properties of PCA	185
A.3	Gaussian Interpretation of PCA	188
A.4	Applications of PCA	190
B	Experiments using WNMf	193
B.1	Outperformance of WNMf with respect to NMF	193
B.2	Evolution Curves of NMF and WNMf	203
C	Object Image Databases	207

CONTENTS

ix

C.1 Pharmaceutical Products	208
C.2 Corel Database	208
C.3 AR Face Database	215
C.4 MNIST Digit Database	215

D Publications	227
-----------------------	------------

Bibliography	231
---------------------	------------

List of Tables

3.1	Experimental comparison between global and local techniques when recognizing a set of 100 pharmaceutical products. These results show the effectiveness of local approaches versus global ones. Notation: NNL2- Nearest Neighbor using L_2 metric distance.	73
3.2	Ordered list of data classes according to the mean error using the euclidean distance in the original space of local color histograms. Here, all vectors of a given class are compared with the other ones of the same class. The first class (Trees) is the most compact class and the last one (Sky) is the most dispersed one. This experiment has been done using 1000 local color histograms per class randomly selected from the whole database.	91
3.3	Ordered list of data classes according to the sum of the eigenvalues of the corresponding covariance matrices of each data class. The first class of this table denotes a data class with low eigenvalues indicating a compact class. This ordered list is the same as the one presented in table (3.2).	92
3.4	Reconstruction error using PCA and NMF. Second and third columns show the mean error of the retroprojected histograms using PCA and NMF respectively. It can be seen that PCA can always represent the original data better than NMF in global terms but the fourth column shows the number of local color histograms that are better represented with PCA. This table is sorted according to the number reflected in the fourth column, so that, the grass class contains a large number of local vectors that are better represented using PCA than NMF and the sky class contains more vectors that are better represented using NMF than PCA.	93
3.5	Confusion matrix obtained using the PCA technique in a 60 dimensional subspace and using the euclidean distance of the retroprojected vectors.	94
3.6	Confusion matrix obtained using the NMF technique in a 60 dimensional subspace and using the euclidean distance of the retroprojected vectors.	94
3.7	This table shows the α_{CLASS} for each data class with respect to the reduction technique used (PCA or NMF). In bold face we can appreciate which technique performs better for all data classes.	95

3.8	Confusion Matrix for the combined PCA and NMF technique in $60D$ using the euclidean distance of the retroprojected vectors. Grass, Ice, Sand and Sky models are represented with a NMF model and the rest of classes are represented using the PCA technique.	95
3.9	This matrix reflects all the testing vectors that are bad classified using PCA. When an original data vector is classified to belong to another data class, the matrix reflects this error with its corresponding value.	96
3.10	This matrix reflects all the testing vectors that are bad classified using NMF. When an original data vector is classified to belong to another data class, the matrix reflects this error with its corresponding value.	96
3.11	Confusion Matrix for the combined hierarchical PCA and NMF technique in $60D$ using the euclidean distance of the retroprojected vectors.	97
3.12	(a) and (b) are two imaginary situations of two models X and Y with their corresponding reconstruction distances. (c) Reflects the classification results considering the minimal reconstruction distances of both models X and Y of (a) and (b). Each local region is classified without taking into account its neighborhood.	99
3.13	Classification results considering the most frequent model appeared in the neighborhood of each local region. These results are based on table (3.12). In this example, there is one ambiguous region labeled as ? that has 3 local regions with model X and 3 with model Y . In this particular case of having the same number of models in the neighborhood, we can classify the local region as the model that has the minimal distance by summing its reconstructed distances. Model X sums $0.4 + 0.2 + 0.3 = 0.9$ and model Y sums $0.5 + 0.8 + 0.5 = 1.8$, so, this ambiguous region is labeled as model X	100
3.14	(a) and (b) are the mean of the neighborhood reconstruction distances of each local region. These two tables are the result of considering the neighborhood of each model of table (3.12). (c) Shows the final classification results considering the tables of (a) and (b).	100
3.15	Experimental results using the whole database of 50864 testing vectors. It can be seen that the best configuration is achieved using a PCA + NMF mixed representation in conjunction with both neighborhood techniques obtaining a recognition rate of 69.21%. The particular case of using a hierarchical representation in conjunction with the neighborhood mean technique produces an error because we obtain the result of 55.6% (the same as using the PCA). This is due because in the hierarchical representation, all classes are represented at some level of the tree by the PCA model and this model leads to represent vectors with lower reconstruction distances. So, if we combine the reconstruction distances obtained by PCA using the neighborhood mean technique, this leads to obtain the same results as using the PCA technique.	100
3.16	Confusion matrix obtained using the PCA technique in a 60 dimensional subspace. This is the probabilistic approach of PCA.	101
3.17	Confusion matrix obtained using the NMF technique in a 60 dimensional subspace. This is the probabilistic approach of NMF.	102

3.18	Confusion matrix obtained using the WNMf technique in a 60 dimensional subspace. This is the probabilistic approach of WNMf.	102
3.19	α_{class} defined in expression (3.35) per each technique. Bold face indicates the best technique of each class.	102
3.20	Confusion Matrix for the combined classifier (PCA, NMF and WNMf) in 60D using a probabilistic approach.	103
3.21	Confusion Matrix for the optimal hierarchical classifier (PCA, NMF and WNMf) in 60D using a probabilistic approach.	103
3.22	Recognition rates using the MNIST database with the PCA and NMF techniques using several metric distances (L_1 , L_2 , Cos and EMD). We also used a $k = 1$ and $k = 5$ nearest neighbor classifiers.	108
3.23	Results with occlusions in a 20D subspace.	111
3.24	Results with occlusions in a 50D subspace.	112
3.25	Facial expression results. This table reflects how both techniques can deal with facial expressions. Note that scream expression (AR 04) is hard to recognize.	116
3.26	Illumination results. This table reflects how both techniques manage illumination changes in faces. Note that in this case, NMF obtains the best classification results when the number of dimensions starts to be considerable (100 or 150).	117
3.27	Occlusion results when sunglasses are present. Note that in this case, NMF only is better when we use a high dimensional feature space and no lighting conditions are considered. When lighting conditions are considered, Bayesian approach obtains the best recognition rates. . . .	117
3.28	Occlusion results when a scarf is present. In this case, FaceIt obtains the best recognition results. And we have to note that NMF is better than the Bayesian approach in this situation. Again, NMF is always better than PCA.	118
3.29	Expression results when we consider a previous gender classifier. This table must be compared with table (3.25) where we can appreciate some improvements.	119
3.30	Illumination results when we consider a previous gender classifier. This table must be compared with table (3.26). In this particular case, AR 07 is specially improved in low dimensional spaces.	121
3.31	Occlusion results when we consider sunglasses and a previous gender classifier. This table must be compared with table (3.27). We can see that in this particular case, recognition rates are not really improved.	121
3.32	Occlusion results when we consider a scarf and a previous gender classifier. This table must be compared with table (3.28). In this case, recognition rates present a general improvement.	122
3.33	Recognition percentages in the first five places for the six compared classification methods in the object recognition experiment.	139

4.1	Recognition rates when we consider the original $9D$ invariant jets with independence assumption and the independent $3D$ ICA projected vectors. First instance is used for training, and 4 new instances are used for testing.	153
4.2	Recognition rates when we consider a mixture of Gaussians. Experiments use 3 independent components that were obtained using an ICA transformation. We present a mixture of Gaussians using 5 and 10 Gaussians.	155
4.3	Recognition rates when we consider a mixture of Gaussians. Experiments use 3 independent components which were obtained using an ICA transformation. We present a mixture of Gaussians using 5 and 10 Gaussians. This experiments contains ordered tuples.	157
4.4	Recognition rates when the number of Gaussians is estimated using the Minimum Description Length (MDL).	159
4.5	Recognition rates when the number of Gaussians is estimated using the Minimum Description Length (MDL) and homogeneous objects contain a significant number of keypoints (at least 50 keypoints per object). The 3 columns in the left show results obtained using the product of probabilities, and the 2 columns in the right show results when a voting scheme is used.	159
4.6	Recognition rates with 100 objects using $k = 1$ tuples and considering all the possible models. All spaces are modeled using a mixture of 10 Gaussians.	160
4.7	Recognition rates with 100 objects using $k = 2$ tuples and considering all the possible models. All spaces are modeled using a mixture of 10 Gaussians.	161
4.8	Recognition rates with 100 objects using $k = 3$ tuples. All spaces are modeled using a mixture of 10 Gaussians.	161
4.9	Recognition rates with 100 objects using $k = 3$ tuples. This table reflects a two-stage classifier: the first for the appearance/color model and the second for the geometry model. First column is an appearance/color model using a mixture of 10 Gaussians and the geometry model used 2 Gaussians. Second column uses a MDL estimator. . . .	162
4.10	Recognition rates with 100 objects using $k = 3$ tuples and various occlusions according to different quadrants (Q_1, Q_2, Q_3 and Q_4) and combinations of two quadrants.	165
4.11	Experiments done using 500 digits per each class as training and 10,000 testing digits. First column of results indicates that each training digit instance is represented using one joint distribution (a total of 5,000 training models) and second column indicates that each class is represented using only one model (10 training models in total). Recognition rates are represented according to the tuple order and tuple selection technique used.	173

List of Figures

2.1	(a) and (b) are two images that seem to contain "regular" faces. (c) and (d) correspond to the rotated images of (a) and (b) and we see that they are not real face images. Here we see the importance of the spatial layout of local features. This example has been extracted from [108].	13
2.2	14 oriented Gaussian derivative basis filters of up to the fourth-order and the initial Gaussian. From top to down: a Gaussian ; 2 first order Gaussian derivatives; 3 second order Gaussian derivatives ; 4 third order Gaussian derivatives and 5 fourth order Gaussian derivatives. Bright regions denote positive magnitude while darker regions denote negative magnitude.	20
2.3	A typical Gabor filter bank. The real part (see equation 2.39) of the filters is shown. This particular filter bank contains 6 different orientations and 4 different scales.	22
2.4	Intensity histogram of a gray image. The horizontal axis correspond to the gray level and the vertical axis reflects its frequency.	24
2.5	Distribution of points corresponding to skin pixels in the RGB space. Such distribution of pixels is usually modeled using color histograms for fast skin-detection purposes (real time applications).	25
2.6	Example of features that can be extracted from an image. (a) Image with an object used to extract its features, (b) Curve segments extracted from the object in (a), (c) L-junctions and parallel curve segments extracted from the object in (a). This graphical example has been extracted from [107].	34
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.	38
2.8	The ICA basis vectors extracted from natural image patches. As seen, they are oriented, localized and bandpass.	41
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.	43
2.10	Graphical representation of a set of local features which can be used to extract local information from image patches of an object.	45

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.	46
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).	48
3.1	A standard object recognition scheme.	67
3.2	(a) A sample object image. (b) Gray values represent all the interesting points. White crosses represent local maxima. (c,d,e,f) Different instances of the same product of (a) taken in different poses and their interesting points.	70
3.3	Given a set of detected keypoints, a circular region around each keypoint is used to extract its local color histogram representation achieving rotational invariance. It can be seen that the non homogeneous regions are used to identify an object.	70
3.4	(a) 3 original basis functions used to make up the set of 10 training vectors used in our experiment. (b) to (f) are the different basis functions obtained after applying the NMF method with different configurations of the parameter r . If we expect to find a small amount of bases ($r = 1, 2$), the obtained bases are combinations of the global behaviours of the sample data. If we only want 3 basis functions ($r = 3$), the obtained bases are exactly the same as the original ones (a). Furthermore, if we want to extract a large amount of basis functions ($r = 4, 5$), the extracted basis functions are combinations of the specific behaviours of the sample data.	79
3.5	10 different samples created by weighting the 3 Gaussian functions defined in figure (3.4.a). As it can be seen, these samples contain mixed behaviours from the original data vectors.	80
3.6	Five different color newspapers each of them containing some characteristic local regions.	86
3.7	NMF (light line) and WNMF (solid line) reconstruction error against iterations on the test data. Using a small number of bases $r = 30, 35, 40$ from (a) to (c), NMF outperforms WNMF. When $r = 45, 50, 55$ from (d) to (f), WNMF outperforms NMF.	87
3.8	WNMF error subtracted to the NMF error against r , the number of bases of the factorization.	88
3.9	Histogram bases obtained using (a) NMF and (b) WNMF. Note that the number of histogram bases containing the white color are reduced using WNMF and that WNMF contains 5 different histogram bases with green tonalities against 2 green tonalities of the classical NMF.	89

3.10 Optimal hierarchical representation in order to classify the original 10 data classes. Given a node in the tree and having a local color histogram to classify between the right or the left leaves, we obtain the reconstruction distance of this vector and we choose the leaf that contains the lower reconstruction distance. Under the name of each data class, we have the name of the technique used to represent the data. 97

3.11 An image of the database that is divided into $10 \times 10 = 100$ local regions and its corresponding labeled regions. 98

3.12 Optimal hierarchical representation that classifies our original 10 data classes. Given a node in the tree and having a local color histogram to classify between the right or the left leaves, we obtain the reconstruction distance of this vector and we choose the leaf that contains the minimal distance. Under the name of each data class, we have the name of the technique used to represent the data. 103

3.13 Different set of bases obtained using PCA and NMF in two dimensional subspaces, 20 and 50. We can appreciate the parts-based representation of NMF and the holistic representation of PCA. 107

3.14 2 different bases obtained with NMF in the 50 dimensional subspace described in figure (3.13.d). As seen, these two bases share some spatial pixels. 108

3.15 Recognition rates using the MNIST database with PCA and NMF techniques. This graphical representation of results corresponds to table (3.22). We analyzed all subspace dimensions of 20, 25, 30, 35, 40, 45, 50 and 100. (a) is the graphical result of using a $k = 1$ nearest neighbor classifier, (b) is the graphical result of using a $k = 5$ nearest neighbor classifier. 109

3.16 Reconstruction examples with occlusions with PCA and NMF in two different dimensional subspaces. Original digit is the first column, the second column reflects the occlusion applied to each digit, third column reflects the PCA reconstruction and last column the NMF reconstruction. 111

3.17 Recognition rates with different levels occlusions using two different subspaces (20 and 50 dimensions) and using two k -nn nearest neighbor classifiers ($k = 1$ and $k = 5$). 113

3.18 Conditions of an individual of the AR face database: (1) neutral, (2) smile, (3) anger, (4) scream, (5) left light on, (6) right light on, (7) both lights on, (8) sunglasses, (9) sunglasses/left light, (10) sunglasses/right light, (11) scarf, (12) scarf/left light, (13) scarf/right light 115

3.19 Bases obtained by both techniques, PCA and NMF. 115

3.20 Gender classification results when we classify males (a) and females (b) in a 50 dimensional feature subspace. (c) and (d) are the results in a 100 dimensional subspace and (e) and (f) in a 150 dimensional subspace. 120

3.21	Recognition rates according to the method used. Solid line indicates the recognition rates obtained without using any gender information and the dashed line indicates the recognition rates when considering gender information. The best method according to the whole set of face situations is the Non-negative Matrix Factorization in a 150 dimensional space using the correlation distance as a metric obtaining a recognition rate of 66.74%.	123
3.22	Full cream milk box with predominant rose tones (f-milk) and semi skimmed milk box with predominant green tones (ss-milk).	136
3.23	Shows the value of the 19 th independent component for the representative histograms of the full milk image (continuous line) and for the representative histograms of the semi skimmed milk image (line with dots). (b) Shows how this component is the only one activated upon the appearance of a certain color distribution in the full milk image.	137
3.24	(a,b) Probability maps for component 19, activated by the color distribution of the human figure, and component 21 activated by the color distribution of the "calcio" tag. (c,d) Thresholded probability map shown over original image.	138
4.1	Graphical models: (a) fully-connected graph denoting no independence assumptions (b) the ICA-factorized model with pair-wise only dependencies.	147
4.2	Two graphical models. (a) Undirected graphical model corresponding to the graphical model of figure (4.1.b). (b) A directed version of the graphical model in (a) where we can use conditional probabilities and the arrow indicates the direction of causation.	148
4.3	System diagram for k -tuple density factorization using ICA and Gaussian mixture models.	149
4.4	Synthetic "cluttered" scene and a detection example. (a) The synthetic test image of 20 objects from COIL; (b) The rotated and occluded version of (a); (c) The likelihood map for detecting "piggy bank" in (b). The white dots are the interest points and high-likelihood points are highlighted.	151
4.5	The average detection rank of 20 objects of figure (4.4.a) using different Independent Components ($m=1,2,\dots, 9$) versus original $n = 9$ dimensional jets (shown as the rightmost bar). Dataset used in this particular example: COIL 20 objects.	152
4.6	Three 2 dimensional spaces representing the tuple space of one object.	154
4.7	Three 2 dimensional spaces representing one of the objects of the database where keypoints of tuples are ordered. We also show how a mixture of Gaussians can be adapted to our tuple distributions.	156
4.8	Two objects of our database. These two objects are similar in terms of local appearance and their respective tuple spaces are shown in figure (4.9).	157

4.9	Three joint distributions of tuples of two objects corresponding to each $m = 3$ independent components. One of the objects is modeled using a mixture of Gaussians (circles) that is also seen in figure (4.7). Crosses represent tuples of the non-modeled object. We see a high level of intersection between these two tuple spaces.	158
4.10	Likelihood maps (c), (d) and (e) of different models for the input image (b) when using object (a) as the search target. Note: We are only showing the first 20 objects of COIL-100.	163
4.11	Experiment with occlusions. Likelihood maps (c), (d) and (e) of different models for the input image (b) when using object (a) as the search target. Note: We are only showing the first 20 objects of COIL-100.	164
4.12	Two objects (a) and (b) with similar colors but differing in shape used to train our models. First column (c) contains 4 testing images where the two learned objects are present under different occlusions and poses. Columns (d) and (e) show the detection maps for objects (a) and (b), respectively.	166
4.13	(a) Satellite image of the US Pentagon building (prior to 9/11/01). (b) extracted building region used for learning. (c) a new test image of the same region taken after 9/11/01 under different natural conditions and with the damaged portion of the building missing (removed after site cleanup). (d) the highest probability target area of the test image given our local appearance model of (b). (Note: All images have been rescaled for display purposes.)	167
4.14	Given a tuple $(\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3)$, we obtain the middle point using the spatial location of each keypoint. When the three distances between each keypoint and the middle point (R_1, R_2 and R_3) are lower than a predefined radial value (R_{thr}), this tuple is considered for learning.	168
4.15	Detection maps corresponding to different radial thresholds (from $R_{thr} = 25$ to $R_{thr} = 45$ pixels).	169
4.16	Toy example used to test our technique in cluttered scenes. Having a complex scene as shown in (b), we have extracted a toy object as shown in (a). Note that this toy object appears in (b) three times. One instance is the same instance as in (a), another instance of the object appears occluded and there is another instance where the object has a completely different pose. Interesting points of this scene are reflected in (c). Size of toy object is 100×140 pixels.	170
4.17	Likelihood maps obtained when we try to search for the object defined in figure (4.16.a) when using a radial tuple threshold of $R_{thr} = 25$ pixels. As seen, the color and appearance model can detect all the three instances of the learned object under different object poses.	171
4.18	Different likelihood maps that have been thresholded in order to visually perceive what parts of the scene are detected. These likelihood maps have been obtained according to the radial tuple threshold (R_{thr}) considered in each particular case.	172
A.1	Typical (a) eigenspectrum, (b) energy distribution.	187

B.1	Sample image regions and their corresponding class labels (cloud, grass, ice, leave, rocky mountains, sand, sky, snow mountains, tree and water) extracted from the corel image database.	194
B.2	Evolution of the reconstruction error of NMF (thin line) and WNMF (thick line) techniques through dimension 20 to 100 when we consider 2 data classes (Clouds, Grass). Both training and testing results are presented independently. In the specific case of the Grass class, dimension interval goes from 10 to 90.	195
B.3	Evolution of the reconstruction error of NMF (thin line) and WNMF (thick line) techniques through dimension 20 to 100 when we consider 3 data classes (Ice, Leaves and Rocky Mountains).	196
B.4	Evolution of the reconstruction error of NMF (thin line) and WNMF (thick line) techniques through dimension 20 to 100 when we consider 2 data classes (Sand, Sky). Both training and testing results are presented independently.	197
B.5	Evolution of the reconstruction error of NMF (thin line) and WNMF (thick line) techniques through dimension 20 to 100 when we consider 3 data classes (Snow Mountains, Tree and Water). Both training and testing results are presented independently.	198
B.6	Graphical representation of the obtained NMF and WNMF bases of Grass and Leaves classes in a $r = 40$ dimensional subspace (NMF and WNMF).	199
B.7	Graphical representation of the similarity matrices of bases \mathbf{W} of grass and leaves classes in a $r = 40$ dimensional subspace. These similarity matrices are obtained using the bases shown in figure (B.6). White zones correspond to high similarity bases and black zones to non similar ones.	200
B.8	Evaluation of the similarity matrices of both NMF and WNMF techniques through 20 to 100 dimensions in order to determine when the WNMF starts to obtain a less redundant set of bases. We show classes Clouds, Grass, Ice, Leaves, Rock Mountains and Sand. These results are nearly the same as the ones obtained in figures (B.2,B.3,B.4,B.5).	201
B.9	Evaluation of the similarity matrices of both NMF and WNMF techniques through 20 to 100 dimensions in order to determine when the WNMF starts to obtain a less redundant set of bases. We show classes Sky, Snow Mountains, Tree and Water. These results are nearly the same as the ones obtained in figure (B.2,B.3,B.4,B.5).	202
B.10	Evolution of the error curves of NMF (solid line) and WNMF (dotted line) in different subspaces (from $r = 6$ to $r = 34$ in steps of 4 units).	204
B.11	Evolution of the error curves of NMF (solid line) and WNMF (dotted line) in different subspaces (from $r = 38$ to $r = 58$ in steps of 4 units).	205
C.1	A set of 100 pharmaceutical products used in the first experiments. As seen, this set of pharmaceutical products are very conflictive since they are very similar in color.	209
C.2	A first set of 100 pharmaceutical products.	210

C.3	A second set of 100 pharmaceutical products.	211
C.4	A third set of 100 pharmaceutical products.	212
C.5	A fourth set of 100 pharmaceutical products.	213
C.6	Corel image database.	214
C.7	AR database. Individuals from AR01 to AR16.	216
C.8	AR database. Individuals from AR17 to AR32.	217
C.9	AR database. Individuals from AR33 to AR48.	218
C.10	AR database. Individuals from AR49 to AR64.	219
C.11	AR database. Individuals from AR65 to AR80.	220
C.12	AR database. Individuals from AR81 to AR96.	221
C.13	AR database. Individuals from AR97 to AR112.	222
C.14	AR database. Individuals from AR113 to AR117.	223
C.15	MNIST digit database (digits from 0 to 1).	223
C.16	MNIST digit database (digits from 2 to 3).	224
C.17	MNIST digit database (digits from 4 to 5).	224
C.18	MNIST digit database (digits from 6 to 7).	225
C.19	MNIST digit database (digits from 8 to 9).	225

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 view-point, 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 performance 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.