A New Kernel Method for Object Recognition: Spin Glass-Markov Random Fields

BARBARA CAPUTO

Doctoral Thesis
Stockholm, Sweden 2004
Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologie doktorsexamen i Kungl Tekniska högskolan, Valhallavägen 79, Stockholm.

© Barbara Caputo, June 2004

Tryck: Universitetsservice US AB
Recognizing objects through vision is an important part of our lives: we recognize people when we talk to them, we recognize our cup on the breakfast table, our car in a parking lot, and so on. While this task is performed with great accuracy and apparently little effort by humans, it is still unclear how this performance is achieved. Creating computer methods for automatic object recognition gives rise to challenging theoretical problems such as how to model the visual appearance of the objects or categories we want to recognize, so that the resulting algorithm will perform robustly in realistic scenarios; to this end, how to use effectively multiple cues (such as shape, color, textural properties and many others), so that the algorithm uses the best subset of cues in the most effective manner; how to use specific features and/or specific strategies for different classes.

The present work is devoted to the above issues. We propose to model the visual appearance of objects and visual categories via probability density functions. The model is developed on the basis of concepts and results obtained in three different research areas: computer vision, machine learning and statistical physics of spin glasses. It consists of a fully connected Markov random field with energy function derived from results of statistical physics of spin glasses. Markov random fields and spin glass energy functions are combined together via nonlinear kernel functions; we call the model Spin Glass-Markov Random Fields. Full connectivity enables to take into account the global appearance of the object, and its specific local characteristics at the same time, resulting in robustness to noise, occlusions and cluttered background. Because of properties of some classes of spin glass-like energy functions, our model allows to use easily and effectively multiple cues, and to employ class specific strategies. We show with theoretical analysis and experiments that this new model is competitive with state-of-the-art algorithms for object recognition.
Acknowledgements

I want to thank all the people who contributed to this thesis in one way or another. The first two years of my doctoral studies were spent at the University of Erlangen-Nuremberg. I want to thank Prof. H. Niemann, Dietrich Paulus, Joachim Hornegger, Paul Baggenstoss, Ulrike Ahlrichs, Uwe Ohler, Sahla Bouattour, Gyury Dorko and all the present and former members of the Pattern Recognition group in Erlangen, for many fruitful discussions and the friendly atmosphere. I would like to thank Bernt Schiele for his advice, and for hosting me in his group at ETH Zurich during my research visits. A significant amount of the work reported in this thesis was influenced by discussions with him and Bastian Leibe.

I am grateful to Bernhard Schoelkopf for introducing me to the joy of kernel methods, and to Tony Bell for hospitality at the Salk Institute and unforgettable discussions on independent component analysis. A very special thank to the group of statistical physics of disordered systems and neural networks of the University of Rome “La Sapienza”, particularly to Daniel Amit, Miguel Angel Virasoro, Giorgio Parisi, Enzo Marinari, Paolo Del Giudice and Maurizio Mattia. They first have made me discover the beauty of spin glass theory, and then they have made sure I didn’t forget it along the way. I also want to thank Giovanni Ettore Gigante for his wise advices and friendship. I spent the last year of my PhD thesis as visiting fellow at the Smith Kettlewell Eye Research Institute in San Francisco. I gratefully thank Alan Yuille and James Coughlan for many stimulating discussions; San Francisco would have not been as much fun as it was without the friendship of Barbara Rosario.

I finally concluded my PhD thesis at NADA, KTH in Stockholm. A thousand times thank to my advisor Stefan Arnborg, to Jan-Olof Eklundh and Henrik Christensen, without whom this thesis would not exist. Many thanks to Vanka, Eric, Josephine, Peter, Gareth, Ola and all the members of CVAP for moral support and great parties (both much needed!).

Whoever has ever earned a PhD knows that without the love, friendship, support and patience (patience above all) of our beloved, it would have been impossible. Carissimi Alessandra, Amelia, Andrea, Armando, Costantino, Laura, Lorenzo, Micaela, Quick and Sahla, grazie.

This work is dedicated to Emilio Caputo, Miguel Angel Virasoro and Nicolino Barra, the best guides I have ever had.

Barbara Caputo
“When a subject is highly controversial, one cannot hope to tell the truth. One can only show how one came to hold whatever opinion one does hold. One can only give one’s audience the chance of drawing their own conclusions as they observe the limitations, the prejudices, the idiosyncrasies of the speaker [...] but there may perhaps be some truth mixed up with them; it is for you to seek out this truth and to decide whether any part of it is worth keeping. If not, you will of course throw the whole of it into the waste-paper basket and forget all about it.”

(Virginia Woolf, *A room of one’s own*)

If we would know what it is that we are doing, we wouldn’t call it research, would we?
# Contents

## 1 Introduction
1.1 Contribution of this Work ........................................... 3  
1.2 Outline ............................................................... 5

## 2 A Few Landmarks
2.1 State of the Art ....................................................... 10  
2.1.1 Geometry-based Object Recognition Systems .................. 10  
2.1.2 Appearance-based Object Recognition Systems ............... 13  
2.1.3 Scene Recognition Systems .................................... 15  
2.2 The General Framework .............................................. 16  
2.2.1 Appearance-based Methods: the General Formulation ....... 16  
2.2.2 Appearance-based Methods: the Probabilistic Approach .... 16  
2.3 Markov Random Fields ................................................. 18  
2.4 Spin Glasses and Associative Memories ............................ 20  
2.4.1 The General Spin Glass Model .................................. 20  
2.4.2 A Particular Spin Glass Model .................................. 23  
2.5 Kernel Methods ....................................................... 25

## 3 Spin Glass-Markov Random Fields
3.1 Introduction ........................................................... 30  
3.2 Problem Statement .................................................... 30  
3.3 Kernel Associative Memories ....................................... 31  
3.4 Choice of Kernels ..................................................... 33  
3.5 Choice of Prototypes ................................................. 35  
3.5.1 The Naive Ansatz .................................................. 35  
3.5.2 The ICA Ansatz .................................................... 36  
3.6 A Spin Glass-Markov Random Fields ............................... 37  
3.7 The Algorithm: Learning the Kernel Parameter .................. 38  
3.8 Related Methods ...................................................... 40  
3.8.1 Kernel Parzen Windows .......................................... 40  
3.8.2 Support Vector Machines ....................................... 41
3.8.3 Synergetic Pattern Recognition ............................................ 43
3.8.4 The FRAME Model .......................................................... 44
3.9 Experiments ................................................................. 45
  3.9.1 Columbia Database Experiments ...................................... 46
  3.9.2 NELSON Database Experiments ...................................... 52
  3.9.3 Discussion ............................................................... 53
3.10 Summary ................................................................. 53

4 Robustness of SG-MRF ......................................................... 55
  4.1 Introduction ............................................................... 56
  4.2 Robustness of Spin Glass-Markov Random Fields: A Statistical Mech-
        anics View ............................................................ 57
  4.3 Robustness of Spin Glass-Markov Random Fields: a Kernel View 58
  4.4 Robustness of SG-MRFs: Experiments ................................ 60
    4.4.1 Robustness to Noise ............................................... 60
    4.4.2 Robustness to Occlusion .......................................... 65
    4.4.3 Robustness to Decreasing Training Set ........................ 71
    4.4.4 Robustness to Heterogeneous Background ..................... 72
  4.5 Summary ............................................................... 74

5 Ultrametric SG-MRF .......................................................... 75
  5.1 Introduction ............................................................ 76
  5.2 Ultrametric Spin Glass-Markov Random Field ......................... 76
    5.2.1 The Ultrametric Energy ......................................... 76
    5.2.2 Ultrametric Bayes Classifier .................................. 80
    5.2.3 Learning The Kernel Parameters .............................. 81
  5.3 Hierarchical Appearance-based Object Recognition ................. 82
    5.3.1 Problem Statement ............................................. 82
    5.3.2 The Ultrametric Approach ..................................... 83
    5.3.3 Experiments ..................................................... 85
  5.4 Combining Shape and Color Information for Object Recognition. 87
    5.4.1 Problem Statement ............................................. 87
    5.4.2 The Ultrametric Approach ..................................... 88
    5.4.3 Experiments ..................................................... 89
  5.5 A Probabilistic Model of a Scene .................................... 91
    5.5.1 Contextual Information versus Heterogeneous Background 92
    5.5.2 The Ultrametric Organization of Contextual Information 94
    5.5.3 Experiments ..................................................... 96
  5.6 Summary ............................................................. 98

6 Moving Beyond .............................................................. 101
  6.1 Introduction ............................................................ 102
  6.2 Kernel Class Specific Classifier .................................... 103
    6.2.1 Problem Statement ............................................ 104
Chapter 1

Introduction

Humans and animals extract rich and detailed information about the environment through vision. The visual process is extremely complex: it includes for instance the analysis of color, shape and texture of the visual pattern under consideration. Moreover, visual information is used for recognition, locomotion and manipulation. In this thesis, we will focus the attention on the task of visual object recognition. With the expression “object recognition” we refer to three different possible tasks: object identification, object classification and object discrimination. Object identification consists of determining to which object a presented view belongs. Recognizing our coat in a closet containing several other items is an example of object identification. Object classification consists of attributing views to object categories: examples of object categories are pens, shoes, cars, glasses, and so on. Finally, object discrimination consists of determining whether a presented view belongs or not to a specific object or category. This thesis will deal with the tasks of object identification and classification.

Object recognition is an important part of our lives. We recognize objects in all our everyday activities: we recognize people when we talk to them, we recognize our cup on the breakfast table, our car in a parking lot, and so on. While this task is performed with great accuracy and apparently little effort by humans, it is still unclear how this performance is achieved. This has challenged the computer vision research community to build artificial systems able to reproduce the human performance. After 30 years of intensive research, the challenge is still open.

An automated system for object recognition would have many uses, for instance:

- **Artificial mobile systems**
  The development of automated systems has dramatically improved the quality of our lives in the last 50 years. Cars, computers, washing machines (just to cite few examples) are habitual elements of our daily experiences. All these automated systems are still completely supervised: their action must be guided thoroughly by a user. Automated object recognition systems could be incorporated in many of these devices; this would open the possibility for
CHAPTER 1. INTRODUCTION

them to be used in a semi- (or total-) unsupervised manner. This would be extremely beneficial for executing tasks in environments of potential danger for humans, and so on.

- **Database search**
The explosion of desktop publishing, Internet usage and multimedia computing gives people access to a dramatically high quantity of digital images, and this number is rising continuously. However, the possibility to use these collections is limited by a lack of effective retrieval methods. Currently, the strategy to find a specific image in such a collection consists in searching using text-based captions and low-level image features such as color and texture. Automatic object recognition could be used to extract more information from these images and help to label and categorize them automatically.

Creating computer methods for automatic object recognition gives rise to challenging theoretical problems: given a set of observations, relative to a set of objects or visual categories,

- **How should we model the visual appearance of the objects or categories we want to recognize?** Objects vary in visual appearance: for example, an object’s orientation and distance from the camera affects its appearance. For visual categories variations are even stronger: for example, cars vary in size, shape, coloring and in small details such as tires and the headlights. A good algorithm for object recognition should be able to generalize from the given set of observations;

- **How can we perform robust recognition?** Objects are located in different environments: they can be partially occluded by other objects in the scene. The presence of other objects can be misleading for the recognition of a specific one. Objects’ appearance changes with respect to lighting conditions. For all these reasons, a good algorithm for object recognition should be robust with respect to noise, occlusion, cluttered background and light changes;

- **How can we use effectively multiple cues?** Objects can be described in terms of many different features, such as shape, color, textural properties and many others. These features can be combined together in an unique feature vector, or to each type of information can correspond a different feature vector. Then, the collection of these features will be used as different cues in the recognition step. A good algorithm for object recognition should make it possible to use as many cues as possible but not more than it is needed, and it should use them in the most effective manner;

- **How can we use class specific strategies?** Different objects and categories can have different distinctive features. For example, a distinctive feature for cups is the handle, while for bananas is the yellow color; and so on. Ideally, an object recognition system should use specific features and/or specific strategies for different classes.
1.1. CONTRIBUTION OF THIS WORK

The present work is devoted to the above issues. We propose to model the visual appearance of objects and visual categories via probability density functions. We propose a new graphical statistical model and we study theoretically and experimentally its performance for object identification and classification. We explore its robustness to noise, occlusion and cluttered background. The proposed statistical model allows to use easily and effectively multiple cues, and to employ class specific strategies. We call this new statistical model Spin Glass-Markov Random Field (SG-MRF).

In Section 1.1 we describe the main features of SG-MRF, and we discuss the main contributions of this work. We conclude the Chapter with a short outline of the thesis.

1.1 Contribution of this Work

The design of an algorithm for object recognition must take into account several steps: the expression “object recognition” implies that we have a knowledge of the objects we wish to recognize. Then we have to deal with the problem of how to represent the information that characterizes the objects under consideration (feature extraction). This will define how the objects description, or models, will be stored as a database. Thus, the recognition step will consist in the comparison of image data with the model database. The algorithm will typically require the estimation or learning of the model parameters from training data; the final stage will consist in choosing an appropriate strategy so to use this information for recognition purposes. Once the recognition problem has been solved, a subsequent task can be to determine 3-D position and orientation of the considered object. This problem is known in literature as pose estimation; in this thesis it will not be considered.

Most of the research work on object recognition has focused the attention on the feature extraction step, trying to build representations effective for large collections of objects, that permits to recognize them under different viewpoints and lighting conditions, in different environments, occluded and (or) from noisy images. In all these approaches (for a review on the state of the art in object recognition we refer the reader to Chapter 2, Section 2.1), the classification step is performed using methods developed by the machine learning community.

In this thesis we tackle the object recognition problem focusing the attention on how the extracted information is combined together. We develop a new probabilistic model and a new probabilistic classifier, and we show with theoretical analysis and experiments that this new classifier improves performance of a given representation, with respect to other state-of-the-art methods, commonly employed in the research community. The algorithm is developed on the basis of concepts and results obtained in three different research areas: computer vision, machine learning and statistical physics of spin glasses.

The main features of our probabilistic model are:
A fully connected Markov random field is used for estimating the probability distribution of the model objects. Full connectivity enables to take into account the global appearance of the object, and its specific local characteristics at the same time. Moreover, full connectivity makes it possible to define a neighborhood system for 3D objects in spite of pose variations (for a detailed discussion on this point we refer the reader to (Li, 1995) and Chapter 2, Section 2.3). Defining a neighborhood system for Markov random field modeling of the appearance of a 3D object was an open problem which made unfeasible using Markov random fields for this task (while many successful examples can be found for 2D object recognition, see (Li, 1995) and reference therein).

The energy function, that characterizes the Markov random field to be used, is derived from results of statistical physics of spin glasses. Thus we can benefit of the theoretical knowledge developed by the physics community on those class of energies. Moreover, as they have a parametric form, we can learn the optimal parameters for each model object. This is equivalent, in a fully connected Markov random field, to achieve a globality sum of the significant localities.

Markov random fields and spin glass energy functions can actually be combined together via nonlinear kernel functions. Kernel functions, and the wide class of algorithms that use kernel functions and thus are described as kernel methods (for instance support vector machines, kernel principal component analysis and many others (Schölkopf et al, 2002)), have become increasingly popular within the machine learning community in the last years. Several papers have shown the potential usefulness of these algorithms for object recognition (we refer the reader to Chapter 2, Section 2.2, for a review on this topic). An open challenge for kernel algorithms is the choice of the kernel type (note that, once the kernel type is fixed, it is always possible to select the kernel parameters during the training stage, for instance via cross-validation); the type of kernel chosen determines the metric space where the data are mapped, and consequently the algorithm’s performance (Schölkopf et al, 2002). Till today, how to choose a kernel type for a given task is a lively research area; in practice, for the vast majority of kernel algorithms (and particularly for those used in object recognition) this choice is largely heuristic. The algorithm presented here has theoretical limitations regarding the kernel type which can be used, thus it eliminates the element of heuristic.

Although the model we developed has been intended to be used for object recognition, it is a new probabilistic model and a new probabilistic classifier, that can be used for any pattern recognition application. Thus, the research work presented in this thesis represents a contribution for two research fields: the field of computer vision and the field of machine learning.
**1.2. OUTLINE**

**Contribution to the Field of Computer Vision.** This thesis introduces a novel probabilistic appearance-based object recognition system (Caputo *et al*, 2001) that achieves very good recognition performances on a variety of 3D shapes, ranging from objects like cups, cars and planes, to lizards and snakes. We report results of experiments on several databases and we evaluate performance with increasing number of items in the database, with decreasing number of views per object in the training set, in presence of noise, occlusion and background changes (Caputo *et al*, 2002a).

We investigate the capability of the new probabilistic method to generalize and thus recognize visual categories. To this purpose, we tested successfully the system on objects never previously seen or modeled. We report results of experiments on several databases and we benchmark with state-of-the-art categorization algorithms.

We then develop and implement an extension of the proposed probabilistic method that presents a hierarchical structure. We apply this new algorithm to hierarchical object recognition (Caputo *et al*, 2002f), to appearance-based object recognition combining together shape and color information (Caputo *et al*, 2002e; Caputo *et al*, 2002g) and to the recognition of objects in different scenes. All these applications show promising results.

Although we used a particular representation in almost all the experiments we performed, the probabilistic method we developed can be used with any kind of global features.

**Contribution to the Field of Machine Learning.** This thesis presents a new class of graphical models that are inspired by results of statistical physics of disordered systems. We study theoretically and experimentally the properties of this model and its connections with other probabilistic methods presented in machine learning and statistical mechanics literature.

Then we develop a new probabilistic classifier that permits to use different features, specific to each class, for different classes of patterns (Caputo *et al*, 2002b). This result extends and generalizes recent work on class specific classifiers (Baggenstoss *et al*, 2000). The major advantage of our extension is that it allows to use this new family of classifiers for vision applications, a prohibitive task before.

**1.2 Outline**

The rest of this thesis is organized as follows: Chapter 2 presents the state of the art in object recognition and scene modeling, and review some theoretical background. Chapter 3 describes the derivation of the new probabilistic model and reports extensive experiments that show the effectiveness of the new algorithm with respect to existing approaches. Chapter 4 reports theoretical analysis and extensive experiments that show the robustness of the new model for appearance-based
object recognition. Chapter 5 extends the novel probabilistic model and shows with theoretical analysis and experiments how it can be applied for hierarchical object recognition, for combining together shape and color information for appearance-based object recognition, and in a probabilistic framework for recognizing objects in a scene, that makes use of the contextual information given by the scene as well as of the appearance of objects. We report experiments that prove the concept and shows the effectiveness of our approach. Chapter 6 explores some possible directions for future research. With respect to computer vision, we present research work on the recognition of visual categories; with respect to machine learning, we present a new kernel-based classifier that allows to use different features for different object classes to be recognized; with respect to statistical mechanics of spin glasses, we present research work on the properties of a new spin glass system, generated by our model. The thesis concludes with a summary discussion and possible future direction of research.

List of Publications Most of the work presented in this thesis has previously appeared in the following publications:


Other articles by the author that are not directly related to the topic of the thesis are:


Chapter 2

A Few Landmarks

This Chapter sets the scene for the rest of the thesis. It reviews the major results obtained in object recognition up to now, and introduces some theoretical background which will be of central importance in the following chapters. Section 2.1 discusses the state of the art in object recognition and scene modeling. The technical part of the Chapter starts with the mathematical formulation of probabilistic appearance-based object recognition (Section 2.2). Section 2.3 introduces Markov Random Fields (MRF), and Section 2.4 gives a concise description of some ideas and theoretical results of Spin Glass (SG) theory. Finally, Section 2.5 introduces briefly kernel functions and kernel methods.
2.1 State of the Art

Object recognition is one of the most researched areas of computer vision, with applications in many fields. Most methods developed so far can be categorized as geometry-based or appearance-based, the main difference being object representation.

Geometry-based object recognition systems model objects and input images using a discrete set of geometric features (as to say volume or surface elements embedded in 2D or 3D spaces). The recognition step consists in matching the model and scene features (Ponce et al., 1996). These systems typically use geometric constraints in order to avoid inconsistent matches. Appearance-based object recognition is an alternative approach to the geometry-based methods: the objects are modeled by a set of images, and recognition is performed by matching directly the input image to the model set. The matching process is guided by some measure of similarity between images that may be based on intensity, geometry, topology or a combination of these.

Although much attention in high-level vision has been devoted to the problem of individual object recognition, an equally important but less researched problem is that of recognizing entire scenes. In the rest of this Section we review the state of the art in geometry-based and appearance-based object recognition systems (Section 2.1.1-2.1.2) and in scene recognition (Section 2.1.3).

2.1.1 Geometry-based Object Recognition Systems

The first object recognition systems presented in literature were geometry-based. One of the first general-purpose vision systems performing object recognition was the SRI vision module (Agin, 1980). It used binary images and was based on connectivity analysis. This is a procedure that breaks a binary image into its connected components. While extracting connected components, the connectivity program extracts information about the component that will be used later on, such as the maximum limits of its extent, area, perimeter and coordinates of the points on the perimeter. The SRI vision system had two ways to recognize objects: a nearest neighbor technique and a binary decision tree procedure.

The first object recognition system which was designed to operate on noisy and incomplete image representations was ACRONYM (Brooks, 1981). ACRONYM was meant to be a general vision system. It has been used (Binford, 1982) as the basis for a simulator for robot systems and for automated grasping of objects, with a basic rule for determining which surfaces are accessible in the initial position, which surfaces are accessible in the final position, and ways to grasp with maximum stability.

Ayache and Faugeras (Ayache et al., 1986) developed the HYPER object recognition system. It was designed for the recognition of objects lying on a flat surface from 2D images, thus performing 2D from 2D recognition. The recognition process was structured as a search for consistent set of models and image features. The
shape of 2D objects was represented by polygonal approximations of their borders. Although this description is simple, it is compact, general and insensitive to variations in position and orientation, but it is difficult to apply it to 3D from 2D because of its extensive use of distance and angle.

Transforming clustering is another method used in object recognition (Grimson et al., 1990). In it, independent pieces of evidence are accumulated for each match. Each pair of model and image features (edges for instance) defines a range of possible transformations from a model to an image. Then, the pairs which are part of the same correct match of a model to an image, will result in approximately the same transformation. Thus, a cluster of similar transformations is assumed to correspond to a correct match, as random pairs of model and image features will result in randomly distributed transformations.

Object recognition systems can cluster transformations using the generalized Hough transform (Ballard, 1981). The generalized Hough transform permits to find arbitrary curves in a given image, without a need for the parametric equation of a curve. The method consists in constructing a parametric curve description based on simple situations detected in the learning stage. The generalized Hough transform can detect arbitrary shapes but requires complete specification of the exact shape of the target to achieve precise segmentation.

Grimson and Huttenlocher analyzed the generalized Hough transform as a method for recognizing objects from noisy data in complex cluttered environments (Grimson et al., 1990). They showed that the Hough transform should be adequate for the recognition of objects when limited occlusion and moderate sensor uncertainty is present, using isolated points such as vertices as matching features. The method scales poorly when applied to complex, cluttered scenes, or when using extended features, such edges, which are subject to partial occlusion. In these cases however, the generalized Hough transform may still be useful for identifying matches that will be verified further.

In object recognition a particular view of an object may differ from all the previously seen images of the same objects. In order to compensate for these variations, systems may allow the models (or the viewed object) to undergo certain compensating transformations during the matching stage. This is called alignment approach since an alignment transformation is applied to the model (or the viewed object) prior to, or during, the matching stage (Ulman et al., 1991). The task in alignment method consists in finding the minimum amount of information that, for a possible position and orientation, is needed to solve the problem. At the same time, it has to be minimized the amount of search required in matching local model and image features. This method was investigated by Huttenlocher and Ulman, among others, who developed the ORA object recognition system (Huttenlocher et al., 1990). They showed that the correspondence of three non-collinear points is sufficient to determine position, 3D orientation, and scale of a rigid solid object with respect to a 2D image.

The method developed by Basri (Basri, 1996) combined alignment with indexing and performed recognition by prototypes. In this method, objects are divided into
classes, where a class contains objects that share a fair number of similar features. Categorization is achieved by aligning the image to prototype objects; then the identity of the object is determined by aligning the image to individual models of its class.

Another approach to alignment was developed by Ulman and Basri (Ulman et al., 1991): recognition by linear combination of models. The modeling of objects is based on the fact that for many continuous transformations of interest in recognition, such as rotation, translation, and scaling, all the possible views of the transforming object can be expressed as the linear combination of other views of the same object. Ulman and Basri proved that in the case of an object with sharp edges, two views are sufficient to determine the object’s structure within an affine transformation and three are required to recover the full 3D structure of a rigidly moving object. For objects with smooth boundaries, three images are required to represent rotations around a fixed axis and five images are required for general rotations in 3D space.

A novel method developed by Belongie et al (Belongie et al., 2001) measures similarities between shapes and uses them for object recognition. The approach has three stages: (1) solves the correspondence problem between two shapes, (2) uses the correspondence to estimate an aligning transform, and (3) computes the distance between the two shapes as a sum of matching errors between corresponding points, together with a term measuring the magnitude of the aligning transform. Recognition is then treated in a nearest-neighbor classification framework. The advantage of this method is that it can be used for a variety of shapes, such as silhouettes, trademarks, handwritten digits, and 3D objects.

Lamdan and Wolfson (Lamdan et al., 1988) approached the problem as a geometric hashing. The objects and the scenes are described by sets of interest points, invariant under rotation, translation and scale. Hence the model-based recognition becomes a point-set matching task. During matching, for each ordered pair of points in the scene, the coordinates of the other points are computed taking this pair as a basis. For each of these coordinates, the entry in the hash-table is checked, and for every record (model, basis-pair) appearing there, a vote is counted for the model and the basis pair as corresponding to the ones in the scene. The (model, basis-pair) records that score a large number of votes are taken as matching candidates and verified against the scene.

Another approach to indexing was taken by Stein and Medioni (Stein et al., 1992). They described a method for 2D structural indexing that can be extended to 3D object recognition. Object boundaries are approximated by polygons. This way some of the curvature information is preserved in the form of the angle between consecutive segments. A fixed number of adjacent segments is used to form super segments. During the recognition process super segments are extracted from the scene and used as keys to retrieve the matching hypotheses between the super segments of both the model and the scene. The next step is to cluster the consistent hypotheses, i.e. the hypotheses that represent super segments coming from instances of the same model.
2.1. STATE OF THE ART

Califano and Mohan (Califano et al, 1993) analyzed how the parameters of indexing based recognition systems affect their performance. They concluded the following: to keep the average number of votes for the correct hypothesis constant while the index dimensionality is increased, the index has to be more coarsely quantized. The probability of false positives decreases exponentially with an increase in the dimensionality of the index. To increase the accuracy of an indexing system in term of discrimination, the dimensionality of the index must be increased and the quantization of the index must be made proportionally coarser. And last, but not least, indexing based recognition systems for large databases should employ high-dimensional indexes.

More recently, Beis and Lowe (Beis et al, 1999) introduced a method for indexing without invariant features for 3D recognition. Instead of relying on invariant features, their method uses stored samples of the distributions in feature space to form smooth probability estimates that a given shape corresponds to a particular database object. The index structure is a kd-tree and a nearest-neighbor search algorithm is applied.

Perceptual organization refers to a basic capability of the human visual system to derive relevant groupings and structures from an image without prior knowledge of its contents (Lowe, 1985). In object recognition, the structures obtained by perceptual grouping can lead to a substantial decrease in the search space.

Perceptual organization has its origins in the Gestalt theory that was developed during the 1920’s and 1930’s. Elements were grouped based on proximity, similarity, continuation, closure, symmetry and familiarity. Another important contribution of the Gestalt theory was the general principle of simplicity, also known as the ‘minimum principle’, which was stated by Hochberg in 1957 as the principle that other things being equal, the perceptual response to a stimulus will be obtained which requires the least amount of information to specify (Lowe, 1985).

One of the first object recognition systems based on perceptual organization is the SCERPO system developed by Lowe (Lowe, 1987), that recognizes known three-dimensional objects in single gray-scale images. In this system objects are modeled as polyhedral and grouping is made on the basis of proximity, parallelism and collinearity of the edges.

To deal with noise and occlusion, as well as be able to do generic recognition, Havaldar, Medioni and Stein (Havaldar et al, 1996) used a perceptual grouping hierarchy. Groups are based on proximity, parallelism, parallel and skewed symmetry and closure. Similar groups are grouped further into sets. Representation and matching of these sets is done using graphs. The system can handle generic recognition and occlusion.

2.1.2 Appearance-based Object Recognition Systems

An appearance-based model of an object is a description of the object features that are detectable in images of the object (Shapiro et al, 1995). A feature is detectable if there is a computer program that can extract the feature from an image of the
object, by means of some given procedure. Appearance-based models can be full-object models including all the features that appear in any view of the object, or they can be view-class models in which an object is represented by a small set of characteristic views, each having its own distinctive feature set.

Swain and Ballard (Swain et al, 1991) proposed to represent an object by its color histogram. Objects are identified by matching a color histogram from an image region with a color histogram from a sample of the object. The matching is performed using histogram intersection. The method is robust to changes in the orientation, scale, partial occlusion and changes of the viewing position. Its major drawbacks are its sensitivity to lighting conditions, and that many object classes cannot be described only by color.

Schiele and Crowley (Schiele et al, 2000) generalized this method by introducing multidimensional receptive field histograms to approximate the probability density function of local appearance. The recognition algorithm calculates probabilities for the presence of objects based on a small number of vectors of local neighborhood operators such as Gaussian derivatives at different scales. The method obtained good object hypotheses from a database of 100 objects using a small number of vectors.

Also based on local characteristics, Schmid and Mohr (Schmid et al, 1996) developed a system that can recognize objects in the case of partial visibility, image transformations and complex scenes. The approach is based on the combination of differential invariants computed at key points with a robust voting algorithm and semi local constraints. The recognition is based on the computation of the similarity (represented by the Mahalanobis distance) between two invariant vectors. Matching is performed on discriminant points of an image, and a standard voting algorithm is used to find the closest model to an image.

Nelson (Nelson et al, 1998) proposed to represent the appearance of an object as a loosely structured combination of a number of local context regions keyed by distinctive features. Recognition is based on a Hough like evidence combination scheme. One limitation of the approach is that curves cannot be robustly extracted from image data.

Principal component analysis has been widely applied for appearance-based object recognition (Turk et al, 1991; Murase et al, 1995; Ohba et al, 1996; Rao et al, 1995). The attractiveness of the approach is due to the representation of each image by a small number of coefficients, which can be stored and searched efficiently. However, methods from this category have to deal with the sensitivity of the eigenvector representation to changes of individual pixel values, due to translation, scale changes, image plane rotation or light changes. Several extensions has been investigated in order to handle complete parameterized models of objects (Murase et al, 1995), to cope with occlusion (Rao et al, 1995) and to be robust to outliers and noise (Leonardis et al, 2000).

Recently, Support Vector Machines (SVM) and kernel methods have gained in interest for appearance based object recognition (Osuna et al, 1997). Pontil (Pontil et al, 1998) examined the robustness of SVM to noise, bias in the
registration and moderate amount of partial occlusions, obtaining good results. Roobaert et al. (Roobaert et al, 2001) examined the generalization capability of SVM when just a few number of views per objects are available. Barla, Odone and Verri (Barla et al, 2002) proposed to use a new class of kernels, especially designed for vision and inspired by the Hausdorff distance, for 3D object acquisition and detection. A common limitation of SVM and kernel methods proposed so far, is the heuristic in the choice of the kernel function, and in the choice of the kernel parameters; the performance of the algorithm depends heavily on these choices. Ferrari et al. (Ferrari et al, 2004) recently proposed a method for simultaneous object recognition and segmentation. The approach is based on local, viewpoint invariant features: first, it generates a set of feature correspondences; then, it builds on them and gradually explores the surrounding area, trying to increase the number of matching features. The resulting process covers the object with matches, and simultaneously separates the correct matches from the wrong ones. As a result, recognition and segmentation are achieved at the same time. A current limitation of the method is that it works only for object detection.

2.1.3 Scene Recognition Systems

Scene recognition underlies many other abilities, most notably navigation through complex environments. Most of systems developed for localization of robotic systems based on visual information focus on the analysis of 3D scene information and/or the location of visual landmarks like edges or interest points (see (Borenstein et al, 1996) for a review). A different approach for localization is used by research in wearable computing (Clarkson et al, 2000) in which the system uses information about the statistics of simple sensors (acoustic and visual) for identifying coarse locations and events. Besides navigation, many other perceptual abilities such as object localization also rely on scene recognition. This, in general, is a complex task. One way to reduce the complexity of the problem is by relying on prominent landmarks or distinctive markings in the environment. However, such localized cues may not always be readily available in all circumstances. A general-purpose scene recognition scheme has to be able to function without critically relying on distinctive objects. Recently, Torralba et al (Torralba et al, 1999; Torralba et al, 2001) presented an holistic approach to scene recognition. This scheme does not require the presence of specific landmarks, neither it needs a prior assessment of individual objects. Scenes are represented in terms of the spatial layout of spectral components. Their representation is then embedded into a probabilistic framework that can be used for scene recognition as well as for modeling the relationship between context and object properties.
2.2 The General Framework for Appearance-based Methods

Appearance-based methods model the objects by a set of images, and recognition is performed by matching directly the input image to the model set. The model set can consist of the original images, considered as feature vectors (as in (Pontil et al, 1998), see Section 2.1.2 for more details), or in features extracted from the original views, such as color (Swain et al, 1991), textural (Schiele et al, 2000) or geometric information (Nelson et al, 1998), which are representative of the appearance of the objects to be recognized. The rest of this Section introduces appearance-based approaches (Section 2.2.1) mathematically, and its probabilistic formulation (Section 2.2.2).

2.2.1 Appearance-based Methods: the General Formulation

Let \( x = [x_{ij}] \), \( i = 1, \ldots L, j = 1, \ldots M \) be an \( M \times L \) image, with the range of \( x_{ij} \) determined by the quantization of intensity values, or an \( ML \) feature vector. We will consider each feature vector \( x \in G \equiv \mathbb{R}^m, m = ML \) as representative of the corresponding view (in the case in which raw pixels are taken, \( x \) corresponds to the original view). Assume we have \( K \) different object classes \( \Omega_1, \Omega_2, \ldots, \Omega_K \), and that for each object class is given a set of \( n_k \) data samples, \( \omega_k = \{x_{k1}^1, x_{k2}^1, \ldots, x_{nk_k}^k\} \), \( k = 1, \ldots K \). The object classification procedure will be a discrete mapping that assigns a test image, showing one of the objects, to the object class the presented test image corresponds to (see Figure 2.1). How the object class \( \omega_k \) (relative to that object class), varies for different appearance-based approaches. Through this thesis we will concentrate our attention on probabilistic appearance-based methods.

2.2.2 Appearance-based Methods: the Probabilistic Approach

The probabilistic approach to appearance-based object recognition considers the image views of a given object \( \Omega_k \) as random vectors. Thus, given the set of data samples \( \omega_k \) and assuming they are a sufficient statistic \(^1\) for the pattern class \( \Omega_k \), the goal will be to estimate the probability distribution \( P_{\Omega_k}(x) \) that has generated them. Then, given a test image \( x \), the decision step will be achieved using a Maximum A Posteriori (MAP) classifier:

\[
k^* = \arg \max_{k=1}^{K} P_{\Omega_k}(x) = \arg \max_{k=1}^{K} P(\Omega_k | x),
\]

\(^1\)The expression 'sufficient statistic', here and in the rest of the thesis, refers to a set of training data which is representative enough of the visual object under consideration, so that it makes it possible to estimate correctly its probability density function. This is not the same as Fisher’s sufficiency concept.
2.2. THE GENERAL FRAMEWORK

Figure 2.1: Appearance-based object recognition: each object is represented by a set of images. The classification step assigns the test image to an object class. How images are represented and how they are classified changes for different appearance-based methods.

and, using Bayes rule,

\[ k^* = \arg\max_{k=1}^{K} P(x|\Omega_k)P(\Omega_k). \tag{2.1} \]

where \( P(x|\Omega_k) \) are the Likelihood Functions (LFs) and \( P(\Omega_k) \) are the prior probabilities of the classes. In the rest of the paper we will assume that the priors \( P(\Omega_k) \) are constant and the same for all object classes; thus the Bayes classifier (2.1) simplifies to

\[ k^* = \arg\max_{k=1}^{K} P(x|\Omega_k). \]

Probabilistic methods are philosophically optimal in the sense that with a posterior probability distribution over classes, selecting a maximum probability class will minimize the probability of error (see (Bishop, 1995) and references therein). This statement assumes that meaningful probabilities can be computed, in this case from the \( \omega_k \), modeling assumptions and prior probabilities. In practice there are several examples reported in the literature where it has been possible to determine such models obtaining good performances and robustness to degradation of the data such as noise and occlusions (see for instance (Schiele et al., 2000; Leonardis et al., 2000)). A major problem in these approaches is that the functional form of the probability distribution of an object class \( \Omega_k \) is not known a priori. Assumptions have to be made regarding the parametric form of the probability distribution, and parameters have to be learned in order to tailor the chosen parametric form to the
pattern class represented by the data $\omega_k$. The performance thus will depend on the goodness of the assumption for the parametric form, and on whether the data set $\omega_k$ is a sufficient statistic for the pattern class $\Omega_k$ and thus permits to estimate properly the distribution’s parameters.

### 2.3 Markov Random Fields

A possible strategy for modeling the parametric form of the probability function is to use Gibbs distributions within a Markov Random Field framework (MRF, (Li, 1995; Winkler, 1995)). MRF provides a probabilistic foundation for modeling spatial interactions on lattice systems or, more specifically, on interacting features. It considers each element of the random vector $\mathbf{x}$ (that in MRF terminology is called a configuration) as the result of a labeling of all the sites representing $\mathbf{x}$, with respect to a given label set. The sites are related to one another via a neighborhood system. Consider for instance the image of a group of statues shown in Figure 2.2. We can chose to model the probability density function of the statues’ appearance on the gray level values of the image (Figure 2.2, top), or on features extracted from the image like lines and edges (Figure 2.2, bottom). In the first case, the lattice system will consist of the pixel matrix; the sites will be the pixels, and the labels the intensity gray levels. In the second case, sites, labels and distance between sites will have to be defined by the user (we refer the reader to (Li, 1995) and references therein for a complete description and discussion of this case). If we call $\mathcal{S} = \{1, \ldots, m\}$ the discrete set of sites, a neighborhood system for $\mathcal{S}$ is defined as

$$\mathcal{N} = \{ N_i | \forall i \in \mathcal{S} : i \notin N_i, i \in N_j \iff j \in N_i \} \quad (2.2)$$

where $N_i$ is the set of sites neighboring $i$. For a given set $\mathcal{S}$, the neighbor set of $i$ is given by the set of nearby sites within a radius $r$:

$$N_i = \{ j \in \mathcal{S} | [\text{dist}(i, j) \leq r], j \neq i \}. \quad (2.2)$$

For regular $\mathcal{S}$, dist$(i, j)$ denotes the Euclidean distance between $i$ and $j$ and $r$ takes an integer value (Figure 2.2, top). For an irregular $\mathcal{S}$, the distance needs to be defined appropriately for non-point features (Figure 2.2, bottom). In general, the neighbor sets $N_i$ for an irregular $\mathcal{S}$ have varying shapes and sizes.

The set of random vectors $\{\mathbf{x}\}$ is defined as a MRF on $\mathcal{S}$ with respect to a neighborhood system $\mathcal{N}$ if

$$P(x_i | x_{\mathcal{S} \setminus \{i\}}) = P(x_i | x_{N_i}),$$

where $\mathcal{S} \setminus \{i\}$ is the set difference, $x_{\mathcal{S} \setminus \{i\}}$ denotes the set of labels at the sites in $\mathcal{S} \setminus \{i\}$ and $x_{N_i} = \{x_i' | i' \in N_i\}$ stands for the set of labels at the sites neighboring $i$. Note that every random field is a MRF when all different sites are neighbors. There are two approaches for specifying a MRF, that is in term of the conditional probabilities $P(x_i | x_{N_i})$ and in term of the joint probability $P(\mathbf{x})$. A theoretical result about
2.3. MARKOV RANDOM FIELDS

the equivalence between MRF and Gibbs distribution provides a mathematically tractable mean of specifying the joint probability of a MRF.

A set of random vectors \( \{x\} \) is said to be a Gibbs Random Field (GRF) on \( S \) with respect to \( N \) if its configurations obeys a Gibbs distribution:

\[
P(x) = \frac{1}{Z} \exp(-E(x)), \quad Z = \sum_{\{x\}} \exp(-E(x)). \tag{2.3}
\]

The normalizing constant \( Z \) is called the partition function, and \( E(x) = \sum_i f_i(x_i|x_{N_i}) \) is the energy function. Here each \( f_i \) is an arbitrary real valued function of \( x_i \) and its neighboring variables \( x_{N_i} \), taken in some fixed order. \( P(x) \) measures the probability of the occurrence of a particular configurations \( x \); the more probable configurations are those with lower energies. A MRF is characterized by its local property (the Markovianity) whereas a GRF is characterized by its global property (the Gibbs distribution). The Hammersley-Clifford theorem establishes the equivalence between MRF and GRF ((Li, 1995)):

**Theorem:** For a given neighborhood system \( N \) defined on the set of sites \( S \), the probability distribution \( P(x) \) is a Markov Random Field distribution with respect to \( N \) if and only if \( P(x) \) is a Gibbs distribution with respect to \( N \).

Two major tasks when modeling MRFs are how to define the neighborhood system for irregular sites, and how to choose the energy function for a proper encoding of constraints. The neighbor relation between sites is related to their regularity; in the irregular case (Li, 1995), the neighborhood system is mostly defined by means of a heuristic distance that is feature-dependent. Consider for instance high level vision tasks, like object recognition. They usually are modeled on irregular neighborhood systems, resulting from some feature extraction procedure. When it is possible to define features invariant to pose, MRF modeling gives excellent results;

Figure 2.2: Examples of regular (top) and irregular (bottom) sites for MRF modeling. In the case of irregular sites, the neighborhood relationship depends on the definition of the distance measure between sites.
see for instance (Li et al, 1998) as an example of 2D object recognition using MRFs. The trouble is that often features are not invariant to pose (consider for instance 3D object recognition). In this case, pose parameters must be incorporated into the energy formulation and in the neighbor relations definition, with a dramatic increase in complexity. Furthermore, due to mutual occlusion, neighborhoods change with pose parameters. The energy function is a quantitative cost measure of the quality of a solution, where the best solution is the minimum. The form of the energy function determines the form of the Gibbs distribution. In the case of irregular sites, the energy function’s formulation can become something of an art, as it is generally done manually. These problems are so relevant that until now MRF modeling was restricted to low level vision tasks (Szelinski, 1990) and just few MRF approaches have been proposed for high level vision problems such as 3D object recognition (Wheeler et al, 1995; Modestino et al, 1992), which should generally be modeled with irregular sites. To the best of our knowledge, there are no previous works on appearance-based object recognition using MRFs. The problem of the neighborhood definition can be avoided in a fully connected MRF: full connectivity eliminates the need to define distances between sites, but it increases the algorithm complexity (see for instance (Zhu, 1999; Zhu et al, 1998)).

2.4 Spin Glasses and Associative Memories

This Section is dedicated to a short overview on the basics of SG theory, and to a more detailed, although qualitative, description of a particular SG energy function. The mathematical formulation of equilibrium statistical mechanics of SG is the same as for MRF models. We will show in the next Chapter that the integration of SG results in a MRF framework provides a rigorous, elegant and effective way to skip all the obstacles related to the modeling of MRF on irregular sites. The key factors will be the full connectivity of the SG energy function, and the detailed knowledge of its properties, which is the result of more than 20 years of intensive research.

2.4.1 The General Spin Glass Model

The expression Spin Glasses (SGs) was introduced to describe materials in which the interactions between the spins are random and conflicting (Mezard et al, 1987). The attempt to understand the cooperative behavior of such systems has led to the development of concepts and techniques which have been finding applications and extensions in many areas such as attractor neural networks (Amit, 1989), combinatorial optimization problems and so on (Mezard et al, 1987); thus the expression SG has now taken on a wider interpretation. Here we are mostly interested in the mathematical structures arising from the study of SGs.

Disorder and frustration are two basic properties of SG. Disorder refers to constrained disorder in the interactions between the spins and/or their locations. The
2.4. SPIN GLASSES AND ASSOCIATIVE MEMORIES

Spin orientations themselves are variables (i.e., not constrained), governed by the interactions, external fields and thermal fluctuations, free to order or not as their dynamics or thermodynamics tell them. The SG phase is an example of spontaneous cooperative freezing (or order) of the spin orientations in the presence of the constrained disorder of the interactions or spin locations. It is thus order in the presence of disorder. *Frustration* refers to conflicts between interactions or other spin-ordering forces, such that not all can be obeyed simultaneously. These features are readily visualized in the following energy function

$$E = - \sum_{(i,j)=1}^{N} J_{ij} s_i s_j$$  \hspace{1cm} (2.4)

where the $s_i$ are random variables, $\mathbf{s} = (s_1, \ldots, s_N)$ and $\mathbf{J} = [J_{ij}], (i,j) = 1, \ldots, N$ is the symmetric connection matrix: $J_{ij} = J_{ji}$ which determines how the sites’ labels influence each other.

The probability distribution of the set of random vectors $\{ \mathbf{s} \}$ at equilibrium is given by (Mezard *et al*, 1987)

$$P_J(\mathbf{s}) = \frac{1}{Z} \exp(-E(\mathbf{s})); \hspace{1cm} Z = \sum_{\{ \mathbf{s} \}} \exp(-E(\mathbf{s})), \hspace{1cm} (2.5)$$

which is formally identical to equation (2.3); the main difference is that for SG systems the configuration space has always a very high dimensionality (approaching infinity). Thus, SG systems can be viewed as MRFs defined on an infinite number of sites. In the classical formulation, the model requires also limited labeling values ($\pm 1$); many extensions have been made that permits to use multiple discrete label sets (Mezard *et al*, 1987) or continuous but limited label set ($\mathbf{s} \in [-1, +1]^N, N \to \infty$, (Hopfield, 1984)). In the rest of this thesis, we will use a different notation for the configuration vectors of MRF defined on image views (that we will call *image view configuration vector*, $\mathbf{x} \in \mathcal{G} \equiv \mathbb{R}^m$) and configuration vector of MRF defined on SG systems (that we will call *SG configuration vector*, $\mathbf{s} \in [-1, +1]^N, N \to \infty$); this wants to underline that, although in both cases we are in a MRF framework, the configuration space where they are defined is different ($\mathbb{R}^m$ for the image view configuration vectors, and $[-1, +1]^N, N \to \infty$ for the SG configuration vector).

Different choices of the connection matrix $\mathbf{J} = [J_{ij}], (i,j) = 1, \ldots, N$ will lead to systems with very different behaviors. In order to see this, let us consider four spins $s_i$ which can assume values $\pm 1$, and which are placed at the four corners of a square; the lines connecting the spins indicates which spins are interacting to each other (Amit, 1989). The energy of a given configuration of spins $(s_1, s_2, s_3, s_4)$ is given by (2.4), where the sum is over all pairs which are connected by lines in Figure 2.3. $J_{ij}$ will assume here values $\pm 1$: a positive (negative) sign indicates that two neighboring spins prefer to have the same (opposite) sign. In this example $J_{ij} = J = +1$ (Figure 2.3, left).
Every pair of spins on an edge would lower the energy by having the product of its values equal to the sign of the corresponding $J_{ij}$. It is easy to verify that there are two configurations for which it is possible to arrange all four spins so that every single pair is ‘happy’ (Amit, 1989) (as to say, $J_{ij}s_is_j = 1 > 0$); these configurations are

$$S_1 = S_2 = S_3 = S_4 = +1,$$
$$S_1 = S_2 = S_3 = S_4 = -1.$$ 

For these configurations, equation (2.4) will reach its lowest possible value ($E = -4$). Note that the two equilibrium configurations are equivalent under a symmetry transformation $^2$.

Consider now a different choice for the value of $J_{ij}$ (Figure 2.3, right): $J_{12} = J_{23} = J_{34} = +1, J_{14} = -1$. It is easy to see that now there is no configuration of the four spins which can satisfy all four bounds: the bound $1 - 4$ would like its two spins to be of opposite signs, but all other bounds would like to align $1$ with $2$, $2$ with $3$ and $3$ with $4$. These three bounds would tend finally to align $1$ and $4$, leading to a contradiction. There is, therefore, no configuration which can give an energy as low as $-4$. Thus, the system is frustrated. The lowest energy value for this system is $E = -2$, and there are 8 configurations of lowest energy:

$$S_1 = S_2 = S_3 = S_4 = +1;$$
$$S_1 = S_2 = S_3 = +1, S_4 = -1;$$
$$S_1 = -1, S_2 = S_3 = S_4 = +1;$$
$$S_1 = S_2 = +1, S_3 = S_4 = -1;$$

$$S_1 = S_2 = S_3 = S_4 = -1;$$
$$S_1 = S_2 = S_3 = -1, S_4 = +1;$$
$$S_1 = +1, S_2 = S_3 = S_4 = -1;$$
$$S_1 = S_2 = -1, S_3 = S_4 = +1.$$

Note that the two groups of four states are equivalent under a symmetry transformation, while the four states inside each group are not; this means that frustration prevents the energy from becoming as low as in an unfrustrated system, but it creates diversity, a variety of ground states $^3$.

The problem of finding, for a given $J$, the set $\{s\}$ that minimizes (2.4) is an NP complete problem (Mezard et al, 1987). The SG theory approach uses analytical tools from equilibrium statistical mechanics to determine absolute minim of energy, in a mean field approximation (Mezard et al, 1987; Amit, 1989); these approximation techniques make iterative techniques unnecessary in many applications. The result of these investigations is that frustrated interactions lead to a low temperature energy surface with many valleys (Mezard et al, 1987). In the next Section we will concentrate on a particular choice of the $J$ which has been of particular interest for attractor neural networks; as we will show in Chapter 3, it will turn out that the same $J$ can be taken as starting point for building a SG model of a MRF.

---

$^2$For $N \to \infty$ this is a 2D Ising model (Amit, 1989).

$^3$For $N \to \infty$, $J_{ij} = \pm 1$ sampled from a given random distribution, $(i,j) = 1, \ldots, N$, this is an infinite range SG (Mezard et al, 1987).
2.4. SPIN GLASSES AND ASSOCIATIVE MEMORIES

Figure 2.3: Two simple examples of Ising-like (left, (a)) and SG-like (right, (b)) systems. The system (a) does not present frustration, the system (b) does.

2.4.2 A Particular Spin Glass Model

This Section introduces a connection matrix that, due to its structure, will permit to integrate SG theory results in a MRF framework. Let us consider the following choice for the connection matrix:

\[ J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi^{(\mu)}_i \xi^{(\mu)}_j, \]  

with \( \xi^{(\mu)} \in [-1, +1]^N, N \to \infty \) vector in the configuration space \( \mathcal{H} \). The \( \{\xi^{(\mu)}\}_{\mu=1}^p \) are certain particular SG configuration vectors of the system (that we call prototypes) having the following properties:

(a) \( \xi^{(\mu)} \perp \xi^{(\nu)} \quad \forall \mu \neq \nu; \quad (aa) \ p \ll N, \quad N \to \infty. \)

Under these assumptions it has been proved that ((Amit, 1989), Chapter 4-6) the \( \{\xi^{(\mu)}\}_{\mu=1}^p \) are the absolute minima of \( E \); for \( p \sim N \) there is a phase transition (Amit, 1989), and the \( \{\xi^{(\mu)}\}_{\mu=1}^p \) are no longer the absolute minima. This result is proved using the mean field approximation which is asymptotically valid. It has been repeatedly confirmed in simulations that it is well approximated for values of \( N \) typical for physics and neuroscience studies. The value of \( N \) where the result ceases to hold (i.e., where degradation sets in) depends on application and is related both to the dimension of the feature vector \( N \) and the typical number

\(^4\)Here, and in the rest of the thesis, ‘\( N \to \infty \)’ in equations should be read as ‘\( N \) very large’. 
of prototypes to be stored \((p)\). We have found that, for typical object recognition applications, the prototype vectors (corresponding to object's views) are not orthogonal to each other, and the typical dimension of features is not very large compared to the number of prototypes. These two points, combined together, cause the approximation to deteriorate for the object recognition domain (this point will be discussed further in Chapter 3).

The energy function resulting from this choice of the connection matrix was proposed by Hopfield and Little (Amit, 1989; Hopfield, 1984) in an attempt to explain properties of the nervous system. The thermodynamic properties of the Hopfield-Little (H-L) model were extensively studied ((Amit, 1989) and many others) within the context of SG theory, and they can be considered fully understood. The interested reader will find a complete, rigorous discussion of the properties of the H-L model in (Amit, 1989), and a signal to noise analysis in its Appendix. Here we prefer to summarize these results in a more intuitive manner, using a landscape metaphor. As our goal is to use this energy for appearance-based object recognition, we will map the theoretical results obtained for the H-L energy \((2.4)-(2.6)\) to the appearance-based framework. We will do this in spite of the fact that the very large dimension of the configuration space, where this energy is defined, is unsuitable for modeling the appearance of objects. How this severe obstacle can be overcome constitutes one of the major contributions of this thesis and will be the topic of Chapter 3.

Consider the energy as pictured in Figure 2.4. Each prototype represents a valley; condition \((a)\) means that each valley is completely different from the others (as to say we store as prototypes uncorrelated SG configuration vectors); condition \((aa)\) means that there is a big distance between each valley; from the moment we leave the basin of attraction of one valley, to the moment we enter in the basin of attraction of another, we have to walk a lot. Now, let us keep in mind that the space where we plot this landscape is the surface of an \(N\)-dimensional hyper sphere. Each prototype lies on the surface. We take as prototypes just orthogonal vectors (condition \((a)\), pg 24), and between two stored vectors there are lots of directions.

Figure 2.4: Energy function as a landscape. The dot line between valleys represents the infinite space between them.
in between (condition (aa), pg 24; we remind again the reader that we are on the surface of a very large dimensional, ray unity hypersphere). If just one of these two conditions is not satisfied, the equilibrium properties are lost. If the prototypes are not orthogonal, and/or if there’s not enough distance between valleys (number of valleys comparable to the dimension of the space), the valleys will get mixed, the landscape will change dramatically, and the system will lose the recall properties.

From the point of view of appearance-based object recognition, the set \( \{ \xi^\mu \}_{\mu=1}^p \) of prototypes would represent a set of significant image view configuration vectors of the considered object. If it would be possible to have very large dimensional image views, mutually uncorrelated, then equations (2.4)-(2.6) would provide a way to model MRF on a fully connected neighborhood system. The energy function would be defined independently from the considered object class. The probability distribution’s parameters would consist of the set of uncorrelated image view configuration vectors \( \{ \xi^\mu \}_{\mu=1}^p \), to be learned from the training data. Unfortunately, object views are of finite dimension and correlated with each other; thus equations (2.4)-(2.6) cannot be directly applied to appearance-based object recognition.

2.5 Kernel Methods

Kernel methods have become more and more popular in recent years within the computer vision and machine learning communities (Schölkopf et al, 2002; Cristianini et al, 2000; Vapnik, 1998). The main common feature of all those methods is the usage of a nonlinear kernel function in algorithms that depends from the data just via a scalar product. The introduction of this nonlinearity leads to great improvements in the performance of these algorithms for several applications. Although the first and most famous example of kernelized algorithm is the Support Vector Machine, the increasing number of techniques employing kernels, and their excellent performances in several different tasks, has created a growing attention on the usage of kernel methods for computer vision applications.

Consider a vector \( \mathbf{x} = (x_1, \ldots, x_m) \in \mathcal{G}, \mathcal{G} \subset \mathbb{R}^m \), representing a visual entity that we wish to analyze. A common preprocessing strategy is to change the representation of the data:

\[
\mathbf{x} = (x_1, \ldots, x_m) \mapsto \Phi(\mathbf{x}) = (\Phi_1(\mathbf{x}), \ldots, \Phi_N(\mathbf{x})).
\]

This step is equivalent to mapping the input space \( \mathcal{G} \) into a new space, \( \mathcal{H} = \{ \Phi(\mathbf{x}) | \mathbf{x} \in \mathcal{G} \} \). The fact that simply mapping the data into another space can greatly simplify the task has been known for a long time in the research community, and has given rise to a number of techniques for selecting the best representation of data. The task of choosing the most suitable representation \( \Phi \) is known as feature selection (Cristianini et al, 2000). Feature selection should be viewed as a part of the learning process itself, and should be automated as much as possible. On the other hand, it is a somewhat arbitrary step, which reflects our prior expectations.
on the particular task under consideration. Kernel methods offer a completely new approach to this problem (Cristianini et al, 2000; Schölkopf et al, 2002).

**Definition:** A kernel is a function \( K \), such that for all \( x, z \in \mathcal{G} \)

\[
K(x, z) = \langle \Phi(x), \Phi(z) \rangle
\] (2.7)

The name kernel is derived from integral operator theory, which underpins much of the theory of the relation between kernels and their corresponding feature spaces (Cristianini et al, 2000; Schölkopf et al, 2002). An important consequence of the dual representation is that the dimension of the feature space does not need to affect the computation. As one does not represent the feature vectors explicitly, the number of operations required to compute the inner product for evaluating the kernel function is not necessarily proportional to the number of features (Cristianini et al, 2000).

What properties must be satisfied by a function in order to be considered a kernel? Clearly, the function must be symmetric,

\[
K(x, z) = \langle \Phi(x), \Phi(z) \rangle = \langle \Phi(z), \Phi(x) \rangle = K(z, x)
\]

and satisfy the inequality that follow from the Cauchy-Schwarz inequality:

\[
K(x, z)^2 = \langle \Phi(x), \Phi(z) \rangle^2 \leq ||\Phi(x)||^2 ||\Phi(z)||^2 = \langle \Phi(x), \Phi(x) \rangle \langle \Phi(z), \Phi(z) \rangle = K(x, x)K(z, z).
\]

These conditions are, however, not sufficient to guarantee the existence of a feature space. In order to be considered a kernel, a function must satisfy the **Mercer’s theorem** (Schölkopf et al, 2002; Cristianini et al, 2000; Vapnik, 1998):

**Theorem 1** Let \( \mathcal{G} \) be a compact subset of \( \mathbb{R}^m \). Suppose \( K \) is a continuous symmetric function such that the integral operator \( T_K : L_2(\mathcal{G}) \rightarrow L_2(\mathcal{G}) \),

\[
(T_K f)(\cdot) = \int_{\mathcal{G}} K(\cdot, x)f(x)dx
\] (2.8)

is positive, that is

\[
\int_{G \times G} K(x, z)f(x)f(z)dxdz \geq 0
\] (2.9)

for all \( f \in L_2(\mathcal{G}) \). Equation (2.9) is known as the **Mercer condition**. Then we can expand \( K(x, z) \) in a uniformly convergent series in terms of \( T_K \)’s eigenfunctions \( \Psi_j \in L_2(\mathcal{G}) \), normalized in such a way that \( ||\Psi_j||_{L_2} = 1 \), and positive associated eigenvalues \( \lambda_j \geq 0 \),

\[
K(x, z) = \sum_{j=1}^{\infty} \lambda_j \Psi_j(x)\Psi_j(z)
\]
With a degenerate kernel, the number of eigenfunctions can be finite, but in general the number is infinite. Call the number $N$. If vector $x$ is mapped to $N$-dimensional vector $\Phi(x)$ so that its $i$:th component is $\sqrt{\lambda_i} \Psi_i(x)$, then the dot product $\Phi(x) \cdot \Phi(y)$ is $K(x, y)$. The mapping $\Phi(x)$ is $x = (x_1, \ldots, x_m) \mapsto \Phi(x) = (\sqrt{\lambda_j} \Psi_j(x))_{j=1,\ldots,N}$.

The interested reader can find a complete discussion about kernels in (Schölkopf et al., 2002; Cristianini et al., 2000). Here we conclude the Chapter with the following proposition, that will be useful through the rest of the thesis:

**Proposition 2** Let $K_1$ and $K_2$ be kernels over $\mathcal{G} \times \mathcal{G}, \mathcal{G} \in \mathbb{R}^m, a \in \mathbb{R}^+, f(\cdot)$ a real-valued function on $\mathcal{G}$, $\Phi : \mathcal{G} \to \mathcal{H}$, with $K_3$ a kernel over $\mathbb{R}^m \times \mathbb{R}^m$ and $B$ a symmetric positive semi-definite $m \times m$ matrix. Then the following functions are kernels:

1. $K(x, z) = K_1(x, z) + K_2(x, z);$
2. $K(x, z) = aK_1(x, z);$
3. $K(x, z) = K_1(x, z)K_2(x, z);$
4. $K(x, z) = f(x)f(z);$
5. $K(x, z) = K_3(\Phi(x), \Phi(z));$
6. $K(x, z) = x^T B z.$

We will make use of kernel functions and kernel properties in the rest of the thesis. Indeed, we will show in next Chapters that results of Spin Glass theory can be integrated in Markov Random Fields modeling, and employed for object recognition, via kernel mapping.
Chapter 3

Spin Glass-Markov Random Fields

In this Chapter we show how MRF and SG theory can be combined together via Mercer kernel, giving rise to a new class of fully connected MRF that we call Spin Glass-Markov Random Fields (SG-MRFs). After a brief Introduction, Section 3.2 gives the mathematical formulation of the problem and discusses in details the obstacles for using the H-L energy in a MRF framework. In Section 3.3 we show that the chosen energy can be written as a function of the scalar product between configurations. This opens the possibility to use the kernel trick and overcomes the obstacles illustrated in the previous Chapter. The kind of kernel to be used must be chosen in order to satisfy theoretical constraints on the H-L energy, and those constraints lead to the choice of generalized Gaussian kernels (Section 3.4). Due to the particular properties of the mapping induced by these kernels, we can develop a proper ansatz for the choice of prototypes (Section 3.5). These three steps fully define the new model: in Section 3.6 we summarize the obtained results; we also show how it is possible to use SG-MRF in a Maximum A Posteriori (MAP) framework for appearance-based object recognition. Section 3.7 discusses the algorithm complexity and how to learn the kernel parameters, and Section 3.8 discusses some related methods presented in literature. Section 3.9 reports experiments, on different databases and using three different kinds of representations, that show the correctness and effectiveness of the proposed approach. The Chapter concludes with a summary discussion.
3.1 Introduction

As we pointed out in Section 2.4.1, there is an equivalence between MRF and SG-like energy function, where with the expression “SG-like energy” we intend energy of the form

\[ E = - \sum_{(i,j)=1}^{N} J_{ij} s_i s_j, \quad s \in [-1, +1]^N, \quad N \to \infty. \]

Between all the possible prescriptions for the matrix \( J \) we have concentrated our attention on those proposed in literature in the attempt to use the SG mathematical machinery for modeling brain function (Amit, 1989). A common feature of this big family of SG-like energies is the explicit modeling of the matrix \( J \) as a function of the set of prototypes. Our expectation, confirmed by theoretical and experimental results, is that the introduction of these energies in the MRF framework will provide an elegant and effective manner to skip MRF modeling problem related to irregular sites and to the energy definition, illustrated in Section 2.3. We concentrate our attention on the first (and simplest) SG-like energy proposed for attractor neural networks, as to say the H-L model reviewed in Section 2.4.2. We propose a strategy, based on the kernel trick (Schölkopf et al., 2002; Vapnik, 1998), that allows us to use this energy in a MRF framework. The key point consists in expressing the chosen energy as a function of scalar product between configuration vectors. This permits us to solve all the modeling problems related to:

1. The particular structure of the configuration space where the H-L energy is defined;
2. The theoretical conditions \((a), (aa)\) that leads to the H-L energy’s properties (see Section 2.4.2)

We want to emphasize that the same strategy is valid for every SG-like energy that can be written as a function of scalar product between configurations. We first concentrate the attention on the H-L energy (Chapter 3-4), and then we apply the very same strategy on a SG-like energy function that presents a hierarchical structure ((Amit, 1989), Chapter 5). These results show the generality of our strategy: hence, the name of SG-MRF for this model.

3.2 Problem Statement

As already described in Section 2.2.2, the probabilistic approach to appearance-based object recognition considers \( K \) different object classes \( \Omega_1, \Omega_2, \ldots, \Omega_K \), and for each class a set of data samples \( \omega_k = \{x_{1k}, x_{2k}, \ldots, x_{nk}\}, k = 1, \ldots, K \). Assuming they are a sufficient statistic for the object class \( \Omega_k \), we want to estimate the probability distribution that has generated them. Thus, given a test image view
configuration vector $\mathbf{x}$, it will be assigned to an object class according to

$$k^* = \arg\max_{k=1}^K P(\mathbf{x}|\Omega_k).$$

Using MRF modeling, $P(\mathbf{x}|\Omega_k)$ will be (equation (2.3))

$$P(\mathbf{x}|\Omega_k) = \frac{1}{Z} \exp(-E(\mathbf{x}|\Omega_k)), \quad Z = \sum_{\{\mathbf{x}\}} \exp(-E(\mathbf{x}|\Omega_k)),$$

where the dependence of $P(\mathbf{x}|\Omega_k)$ from the data set will be given by the functional form of $E$. The use of equation (2.4)-(2.6) would solve the modeling problem for MRF illustrated in Section 2.3 for irregular sites and the choice of the energy. Full connectivity would make the neighborhood definition irrelevant, and the energy function would be defined independently of the considered application, but modeled on the observed data via the connection matrix. The detailed analytical knowledge of the energy function should also make it possible to avoid the NP complete problem (as knowing the minims by construction would make unnecessary the use of a search algorithm for finding them, see Section 2.4.1). Unfortunately, the extension of the H-L model and, more generally, of a SG-like energy function to real-life applications seems to be not straightforward, due to the particular structure of the space where the energy is defined. As we have shown in Section 2.4.2, the energy function (2.4)-(2.6) is defined in a space $\mathcal{H} = [-1, +1]^N, N \to \infty$. This is not at all the standard situation in appearance-based object recognition, or computer vision applications. Particularly, the condition of very large dimension of the space is totally unacceptable, due to the curse of dimensionality (Bishop, 1995). As we always work with a limited amount of data, a high dimensionality of the data space leads to a sparse representation, and thus to a very poor capability to evaluate the underlying probability distribution, which is supposed to generate the data for a given visual entity. Hence the dimensionality reduction is usually performed in a pre-processing step (feature extraction, see Section 2.5). These observations lead to a dichotomy: if we want to work on real-life applications we need a finite dimensional space; if we want to use SG-like energy functions we need an very large dimensional space. The solution to this dilemma is to actually take two different spaces, one for the data and one for the energy, and to go from a lower dimensional space to a higher dimensional space with a non-linear mapping. How to define this mapping will be the topic of the next Section.

### 3.3 Kernel Associative Memories

As shown in Section 2.4.2, the H-L energy is given by equations (2.4)-(2.6), that we rewrite here for convenience:

$$E = - \sum_{(i,j)=1}^N J_{ij} s_i s_j, \quad J_{ij} = \frac{1}{N} \sum_{\mu=1}^p \xi_{ij}^{(\mu)} \xi_j^{(\mu)}.$$
with \( \{\xi^{(\mu)}\}_{\mu=1}^{p} \in [-1, +1]^N \) (note that here and in all the rest of the thesis we use the extension of the H-L model to a continuous and limited space \( \mathcal{H} \equiv [-1, +1]^N, N \to \infty \)) stored prototypes having the following properties:

\[(a) \quad \xi^{(\mu)} \perp \xi^{(\nu)} \quad \forall \mu \neq \nu; \quad (aa) \quad p \ll N, \quad N \to \infty.\]

As we already pointed out, the different notation for the SG configuration vector \((s \in \mathcal{H} \equiv [-1, +1]^N, N \to \infty)\) and the image view configuration vector \((x \in \mathcal{G} \equiv \mathbb{R}^m)\), wants to underline that, generally speaking, \( \mathcal{H} \) can never be assumed to be coincident with \( \mathcal{G} \). Furthermore, the condition of orthogonality between prototypes \((a)\) is in contrast with the well known fact that views representing the same object are correlated with each other.

Nevertheless, equations (2.4)-(2.6) can be used to model a MRF. Note first that the energy function (2.4), due to the choice of the connection matrix (2.6), can be rewritten as a function of the scalar product between two SG configuration vectors:

\[
E = -\frac{1}{N} \sum_{i,j=1}^{N} \sum_{\mu=1}^{p} \xi^{(\mu)}_i \xi^{(\mu)}_j s_i s_j = -\frac{1}{N} \sum_{\mu=1}^{p} \sum_{i=1}^{N} (\xi^{(\mu)}_i s_i) \sum_{j} (\xi^{(\mu)}_j s_j) = \frac{1}{N} \sum_{\mu=1}^{p} (\xi^{(\mu)} \cdot s)^2.
\]  

(3.1)

Now suppose we map the data from \( \mathcal{G} \) to \( \mathcal{H} \), using a mapping \( \Phi \) (see Figure 3.1, (Schölkopf et al, 2002; Vapnik, 1998)):

\[
\Phi : \mathcal{G} \equiv \mathbb{R}^m \mapsto \mathcal{H} \equiv [-1, +1]^N, N \to \infty.
\]

Equation (3.1) depends on the data through scalar products in the space \( \mathcal{H} \), that is, on functions of the form \( \Phi(x_1) \cdot \Phi(x_2) \). If we can find a kernel function \( K \) such that

\[
K(x_1, x_2) = \Phi(x_1) \cdot \Phi(x_2),
\]  

(3.2)

which satisfies the conditions

\[(j) \quad K(x, x) = 1, \quad \forall x \in \mathcal{G}; \quad (jj) \quad \dim(\mathcal{H}) = N, \quad N \to \infty,\]

we could substitute equation (3.2) in equation (3.1) without explicitly knowing \( \Phi \). The Mercer’s condition ((Schölkopf et al, 2002), Section 2.5) tells us for which kernels there exist a pair \( \{\mathcal{H}, \Phi\} \). The kernel trick thus allows to use the H-L energy function in a MRF framework: the energy function (3.1) becomes

\[
E_{KAM} = -\frac{1}{N} \sum_{\mu=1}^{p} \left[ K(x, \bar{x}^{(\mu)}) \right]^2,
\]  

(3.3)

where \( \{\bar{x}^{(\mu)}\}_{\mu=1}^{p} \) is the set of prototypes in the space \( \mathcal{G} \), to be determined such that \( \Phi(\bar{x}^{(\mu)}) = \xi^{(\mu)}, \mu = 1, \ldots, p \). Once the kernel \( K \) and the prototypes are determined,
3.4. CHOICE OF KERNELS

the energy (3.3) will be fully determined. It will depend on the image view configuration vectors, and it will be possible to use it in equation (2.3). How to choose the kernel and the prototypes will be the topic of the next two Sections.

\[ \mathcal{G} \equiv \mathbb{R}^m \]

\[ \mathcal{H} \equiv [-1, +1]^N, N \to \infty \]

Figure 3.1: The kernel trick maps the data from a lower dimension space \( \mathcal{G} \equiv \mathbb{R}^m \) to a higher dimension space \( \mathcal{H} \equiv [-1, +1]^N, N \to \infty \). This permits to use the H-L energy in a MRF framework.

3.4 Choice of Kernels

Many algorithms which make use of the kernel trick do not provide criteria in order to choose the kernel type, in spite of the fact that the choice of a certain kernel instead of another may lead to a poor performance of the algorithm; this is the case for example of SVM (Schölkopf et al., 2002). On the contrary, SG-MRF kernel’s choice must satisfy criteria (j)-(jj), pg 32. Two possible candidates are the polynomial kernel (Vapnik, 1998)

\[ K_n(x, y) = (x \cdot y)^n \]  \hspace{1cm} (3.4)

and the Gaussian kernel (Vapnik, 1998; Schölkopf et al, 2002)

\[ K_{Gauss}(x, y) = \exp\{-\rho \|x - y\|^2\}. \] \hspace{1cm} (3.5)

They both satisfy Mercer’s condition. It is proved that the polynomial kernel map the data in a space \( \mathcal{H} \) of dimension

\[ \dim(\mathcal{H})_{K_n} = N = \binom{m + n - 1}{n}, \]

where \( m \) is the dimension of the data space \( \mathcal{G} \) and \( n \) is the polynomial degree. Even if \( N \) is not infinite for polynomial kernels, it grows very quickly, thus condition (jj), pg 32 can be satisfied with a proper choice of \( n \), given \( m \). In general the polynomial
kernel (3.4) does not satisfy conditions \((j)\), pg 32 unless the original data are already satisfying it:
\[
\forall \Phi(x) : \Phi(x) \cdot \Phi(y) = K_n(x \cdot y) = (x, y)^n,
\]
\[
\Phi(x) \in [-1, +1]^N, N \to \infty \iff x \in [-1, +1]^m.
\]
Even if this is true for gray level image data, in real-life computer vision and pattern recognition problems this is not generally the case. This, united with the not general fulfillment of condition \((jj)\), pg 32, excludes polynomial kernels as suitable kernels for SG-MRF\(^1\).

Gaussian kernels (3.5) satisfy both conditions \((j)\), \((jj)\): it is demonstrated that (Vapnik, 1998; Schölkopf \textit{et al}, 2002)
\[
\dim(\mathcal{H})K_{Gauss} = N = \infty
\]
and it holds
\[
K_{Gauss}(x, x) = 1, \quad \forall x \in \mathbb{R}^m,
\]
thus the space \(\mathcal{H}\) is an infinite dimension hyper-sphere of radius unity on which all vectors \(x\) are mapped (Figure 3.2, (Schölkopf \textit{et al}, 2002)). Note that this particular geometrical structure of the space implies that, given two different image view configuration vectors \(x, y\), it is always possible to make them orthogonal in the space \(\mathcal{H}\), via a proper choice of the parameter \(\rho\):
\[
K_{Gauss}(x, y) = \exp\{-\rho||x - y||^2\} = \Phi_{Gauss}(x)^T \cdot \Phi_{Gauss}(y) = \frac{||\Phi_{Gauss}(x)|| \cdot ||\Phi_{Gauss}(y)|| \cdot \cos \gamma}{||\Phi_{Gauss}(x)|| \cdot ||\Phi_{Gauss}(y)||},
\]
where \(\gamma\) indicates the angle between the vectors \(\Phi_{Gauss}(x)\) and \(\Phi_{Gauss}(y)\). As equation (3.6) holds, we get
\[
K_{Gauss}(x, y) = \exp\{-\rho||x - y||^2\} = \cos \gamma \quad (3.7)
\]
that, if \(x \neq y\), goes to zero as \(\rho \gg \Delta = ||x - y||^{-2}\). This property of the mapping \(\Phi_{Gauss}(x)\) will become crucial for the choice of prototypes (see Section 3.5). This kernel can be seen as a particular case of a generalized Gaussian kernel (Vapnik, 1998):
\[
K_G(x, y) = \exp\{-\rho d_{a,b}(x, y)\}, \quad d_{a,b}(x, y) = \sum_i |x_i^a - y_i^a|^b. \quad (3.8)
\]
For \(a = 1\) and \(b = 1, 2\), equation (3.8), right, becomes an \(L_1\) (\(||x_i - y_i||\)) and \(L_2\) (\(||x_i - y_i||^2\)) distance measure, respectively. It is demonstrated that equation

\(^1\)While arguing with Gunnar Räsch (Canberra, November 2002), I did realize that condition \((j)\), pg 32 can be satisfied by any kernel, once the kernel is normalized in the kernel space. As to say, it is sufficient to compute \(K^*(x, y) = K(x, y)/\sqrt{K(x, x)K(y, y)}\). Then, the only issue to consider is the dimensionality of the mapped space \(\mathcal{H}\); this means that normalized polynomial kernels can be used in a SG-MRF framework. This possibility is not explored throughout this thesis, but it will in future work.
3.5. Choice of Prototypes

Kernel (3.8) also satisfies conditions (j)-(jj), pg 32 (Vapnik, 1998).

(3.8) satisfies Mercer’s condition if and only if $0 < b \leq 2$ (Vapnik, 1998); the exponentiation of $x_i$ by $a$ does not affect the validity of the Mercer’s condition, as it can be seen as a re-mapping of the input variables. Figure 3.2: Generalized Gaussian kernels map the data to an infinite dimensionhyper-sphere of radius unity. Thus, with a proper choice of $\rho$, it is possible to orthogonalize all the training data in that space.

### 3.5 Choice of Prototypes

Given a set of $n_k$ training examples $\{x^k_1, x^k_2, \ldots, x^k_{n_k}\}$ for the object class $\Omega_k$, the condition to be satisfied by the $p_k$ prototypes of pattern class $k$ is

$$\xi^{(\mu)} \perp \xi^{(\nu)} \quad \forall \mu \neq \nu, \quad \mu = 1, \ldots, p_k, \quad p_k << N$$

in the mapped space $\mathcal{H}$, that becomes

$$\Phi(x^{(\mu)}) \perp \Phi(x^{(\nu)}), \quad \forall \mu \neq \nu, \quad \mu = 1, \ldots, p_k, \quad p_k << \dim(\mathcal{H}) \quad (3.9)$$

in the data space $\mathcal{G}$, where $\{\tilde{x}^{(\mu)}\}_{\mu=1}^{p_k}$ denotes the set of prototypes to be determined.

#### 3.5.1 The Naive Ansatz

The number of training examples for object class $\Omega_k$ is always fixed; also, the dimension of the mapped space is infinite, $\dim(\mathcal{H}) = \infty$. This leads to the conclusion that the condition $p_k << \dim(\mathcal{H})$ is always satisfied. Moreover, the measure of orthogonality of the mapped patterns is the kernel function (3.2). Due to the particular properties of generalized Gaussian kernels (j)-(jj), pg 32, it has the effect to make orthogonal the patterns in the space $\mathcal{H}$, given a proper choice for the $\rho$ (see...
Section 3.4, equation (3.7)). Thus, equation (3.9) does not really give us any constraint. If we don’t want to introduce further criteria for the choice of prototypes, the natural conclusion is to take all the training samples as prototypes:

\[
\{x^k_1, x^k_2, \ldots, x^k_{n_k}\} = \{\tilde{x}^{(\mu)}\}_{\mu=1}^{p_k}, \quad n_k \equiv p_k, \quad \rho >> \Delta_{\text{min}}.
\]  

(3.10)

In this case the energy function will become

\[
E_{KAM} = -\frac{1}{N} \sum_{\mu=1}^{n_k=p_k} \left[ K_G(\tilde{x}^{(\mu)}, x) \right]^2.
\]  

(3.11)

We call this choice for the prototypes the naive ansatz. The naive ansatz can be inadequate when a large number of training examples is given. In this case it is possible that some of the training patterns are very similar to each other, and the only way to make them orthogonal is to choose a large value for \(\rho\). This problem can be solved if we take a threshold value for \(\Delta_{\text{min}}\), and then we perform a selection of the prototypes such that they respect condition (3.10).

### 3.5.2 The ICA Ansatz

In the limit \(N \to \infty\), the orthogonality condition on the prototypes relaxes to the condition of uncorrelation with respect to the distribution of \(\{\xi^{(\mu)}\}_{\mu=1}^{p_k}\) (see (Amit, 1989)), as to say

\[
\langle \langle \Phi(\tilde{x}^{(\mu)}_i), \Phi(\tilde{x}^{(\nu)}_i) \rangle \rangle = 0, \quad \forall \mu \neq \nu
\]  

(3.12)

where the average is intended with respect to the probability \(P_{\Omega_k}(x)\). Considering that, even with a fixed kernel function, the mapping \(\Phi\) is not uniquely determined, equation (3.12) tells us that the prototypes in the data space \(G\) are a set of configuration vectors statistically independent, to be determined by the training data. So, given

\[
F_k = [x^1_k \ldots x^{h_k}_k],
\]

matrix of the training examples for the object class \(\Omega_k\), where each observed data vector is a column of \(F_k\), the prototypes will be given by

\[
\hat{F}_k = W(F_k),
\]

where \(\hat{F}_k\) is the matrix of prototypes, as to say each column of \(\hat{F}_k\) is a prototype, and \(W()\) is the unknown transformation which operates on \(F_k\) in such a way that the columns of \(\hat{F}_k\) are statistically independent. In order to determine \(W()\) and \(\hat{F}_k\), we will make the hypothesis that \(W()\) is a linear transformation:

\[
\hat{F}_k = WF_k.
\]  

(3.13)

The implied assumption here is that the visual entities to be modeled by \(P_{\Omega_k}(x)\), of which \(\{x_i\}_{i=1}^{n_k}\) are a set of samples, manifests itself through a set of \(p_k\) statistically independent random vectors, and that these vectors are observed via a linear
process. The estimation of the matrices $\tilde{F}_k$ and $W_k$ is referred in literature as Independent Component Analysis (ICA, (Cardoso, 1998)); thus we refer to this choice for the prototypes as the ICA ansatz. The ICA ansatz avoids the problems related to the naive ansatz: the problem of storing pathological prototypes is automatically solved because they are statistically independent. Also the problem of storing a too high number of prototypes is solved; note indeed that the number of prototypes is equal to the number of columns of the matrix $\tilde{F}_k$. This is equal to the number of columns of the matrix $F_k$ (Cardoso, 1998); but, as we consider each column vector of $F_k$ as a linear combination of the columns of $\tilde{F}_k$, in the case of a high number of training data we can define a new matrix $F_{k1}$, with columns given by linear combinations of the original training samples, and performing ICA on $F_{k1}$ instead of $F_k$. In this way the number of prototypes is under control and the stability condition $(aa)$ is always satisfied. Nevertheless, the ICA ansatz presents also limitations. First, it is not known a priori how many independent components characterize a given visual object, and a wrong choice can affect the performance of the algorithm. Second, the hypothesis of linearity (3.13) can be incorrect for many classes of visual entities. We want to point out that these problems are common to every method which try to make use of linear ICA on visual problems, and are not especially related to SG-MRF modeling. Still, they are there.

3.6 A Spin Glass-Markov Random Fields

Now we are ready for setting out the SG-MRF model and its application to probabilistic appearance-based object recognition.

Given an object class $\Omega_k$ and a set of training samples $\omega_k = \{x^k_1, x^k_2, \ldots, x^k_{n_k}\}, k = 1, \ldots K$, $P_{\Omega_k}(x)$ will be, using a Spin Glass-Markov Random Field model,

$$
P(x|\Omega_k) = \frac{1}{Z} \exp \left( \frac{1}{N} \sum_{\mu=1}^{p_k} \left[ K_{d-G}(x, \bar{x}^{(\mu)}) \right]^2 \right);
$$

with $K_{d-G}$ generalized Gaussian kernel

$$
K_{d-G} = \exp\{-\rho d_{a,b}(x, y)\}, \quad d_{a,b} = \sum_{i=1}^{m} |x_i^a - y_i^a|^b
$$

and prototypes given by the naive ansatz:

$$
\{\bar{x}^{\mu}\}_{\mu=1}^{p_k = n_k} = \{x^k_1, \ldots, x^k_{n_k}\}, \quad \rho >> \Delta_{\text{min}}.
$$

$Z$ can be expressed using a saddle point approximation (see its Appendix for a sketch of the demonstration, and (Amit, 1989), for the complete one):

$$
Z = \sum_{\{x\}} \exp \left( \frac{1}{N} \sum_{\mu=1}^{p_k} \left[ K_{d-G}(x, \bar{x}^{(\mu)}) \right]^2 \right) \simeq
$$
\[
\sum_{\{\tilde{x}^{(\mu)}\}} \exp \left( \frac{1}{N} \sum_{\mu=1}^{p_k} \left[ K_{d-G}(x, \tilde{x}^{(\mu)}) \right]^2 \right) = \\
\exp \left( \frac{1}{N} \sum_{\mu=1}^{p_k} \left[ K_{d-G}(\tilde{x}^{(1)}, \tilde{x}^{(\mu)}) \right]^2 \right) + \ldots + \exp \left( \frac{1}{N} \sum_{\mu=1}^{p_k} \left[ K_{d-G}(\tilde{x}^{(p_k)}, \tilde{x}^{(\mu)}) \right]^2 \right) = \\
\exp \left( \frac{1}{N} \right) + \ldots + \exp \left( \frac{1}{N} \right) = p_k \exp \left( \frac{1}{N} \right),
\]
due to the property of generalized Gaussian kernels and to the orthogonality between prototypes. We want to stress that the explicit calculation of \( Z \) is problematic for many MRF models (Li, 1995); here again we can skip the obstacle using tools of statistical mechanics commonly employed in SG theory (Mezard et al, 1987).

SG-MRF can be used in the MAP classifier (2.1) for probabilistic appearance-based object recognition:

\[
k^* = \arg\max_{k=1}^{K} P(x|\Omega_k) = \\
\arg\max_{k=1}^{K} \frac{1}{Z} \exp \left( \frac{1}{N} \sum_{\mu=1}^{p_k} \left[ K_{d-G}(x, \tilde{x}^{(\mu)}) \right]^2 \right) = \\
\arg\max_{k=1}^{K} \frac{1}{\exp \left( \frac{1}{N} \right) p_k} \exp \left( \frac{1}{N} \sum_{\mu=1}^{p_k} \left[ K_{d-G}(x, \tilde{x}^{(\mu)}) \right]^2 \right).
\]

If the number of prototypes is the same for all classes, then \( p_k = p \) and the Bayes classifier becomes

\[
k^* = \arg\min_{k=1}^{K} \left( -\sum_{\mu=1}^{p_k} \left[ K_{d-G}(x, \tilde{x}^{(\mu)}) \right]^2 \right). \tag{3.17}
\]

If the number of prototypes differs from class to class, we get

\[
k^* = \arg\min_{k=1}^{K} \ln(p_k) \left( \sum_{\mu=1}^{p_k} \left[ K_{d-G}(x, \tilde{x}^{(\mu)}) \right]^2 \right). \tag{3.18}
\]

### 3.7 The Algorithm: Learning the Kernel Parameter

Generalized Gaussian kernels are uniquely determined once the kernel parameters \((\rho, a, b)\) are chosen. Their choice is crucial, as it determines the performance of the SG-MRF-MAP classifier (3.17). In the rest of the Section we will describe two strategies for learning these parameters from the training data. To the best of our knowledge, SG-MRF is the only kernel method that allows to learn the kernel parameters from the training data.
• **Learning** $\rho$. We have seen in Section 3.5 that $\rho$ must satisfy condition (3.10), that we rewrite here for convenience:

$$\rho >> \Delta_{\text{min}}, \quad \Delta(a, b) = \sum_{i=1}^{m} |x_i^a - y_i^a|^b.$$ 

If we assume that $x_i = y_i + \epsilon$, with $\epsilon$ very small (which is reasonable here as we are interested in $\Delta_{\text{min}}$, then it holds:

$$\Delta_{\text{min}}(a, b) \geq \Delta_{\text{min}}(a = 1, b = 2), \quad (x_i, y_i) \in [0, 1]. \quad (3.19)$$

This means that we can learn $\rho$ before we learn $a, b$: $\rho$ will be learned with respect to $\Delta_{\text{min}}(a = 1, b = 2)$, and in virtue of the inequality (3.19), it will satisfy the inequality (3.10) for every possible value of $(a, b)$. A pseudo algorithm for learning $\rho$ is shown in Table 3.1.

<table>
<thead>
<tr>
<th>Algorithm 1 Learning of the $\rho$ parameter for SG-MRF-MAP classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Require</strong></td>
</tr>
<tr>
<td><strong>Initialize</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Iterate</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Until</strong></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

| Table 3.1: Pseudo algorithm for learning $\rho$. |

The parameter $T$ controls the final number of prototypes and thus the sparseness of the algorithm; the parameter $K$ controls the orthogonality between prototypes. Note that $T$ is essential in order to avoid identical (or quasi identical) prototypes. Indeed, if we would just determine $\Delta_{\text{min}}$ and simply set $\rho = k\Delta_{\text{min}}$, we could have cases in which $\Delta_{\text{min}} = 0$, and, no matter which magnifying value $k$ we would use, there would be prototypes never orthogonal between each other.

• **Learning** $(a, b)$. SG theory doesn’t provide any indication on how to learn $(a, b)$. Thus we propose to learn them with respect to the classification performance of the SG-MRF-MAP classifier (3.17). Once the $\rho$ parameter has been learned, and once for each object class the prototypes has been selected, we built the SG-MRF-MAP classifier for each class, and we learned $(a, b)$ using a *leave-one-out* strategy. Note that the learning of $(\rho, a, b)$ is done off
line, thus its computational time is irrelevant from the point of view of the final classification.

We conclude this Section analyzing the algorithm complexity of the Bayes classifier given by equation (3.17). The SG-MRF-MAP classifier is a memory-based algorithm, as to say the recognition is based on the matching of the given image with respect to a set of views preliminary stored, in this case the prototypes. Thus, the computational complexity of the algorithm is given by (Duda et al, 2001)

\[ \text{computational cost} \sim \text{number of prototypes} \times \text{distance complexity} \]

The advantage with respect to other memory-based algorithms such as Parzen windows and nearest neighbor classifier is that the number of stored prototypes can be taken under control via the parameter \( T \).

3.8 Related Methods

3.8.1 Kernel Parzen Windows

SG-MRF in a MAP classifier are related to kernelized Parzen windows. First we notice that the square operation on the kernel \( K_{d-G}(x, \tilde{x}(\mu)) \) in (3.17) can be seen as a polynomial kernel of degree 2, operating on kernel \( K_{d-G}(x, \tilde{x}(\mu)) \); as the kernel of a kernel is still a kernel (Schölkopf et al, 2002; Vapnik, 1998), we can rewrite the energy as:

\[
E = \sum_{\mu=1}^{p_j} \tilde{K}(x, x(\mu)), \quad \tilde{K} = [K_{d-G}]^2. \tag{3.20}
\]

If the following hypotheses are true:

1. The number of prototypes is the same for all classes
2. The kernel \( \tilde{K}(x, x(\mu)) \), when one of the two arguments is fixed, is positive and has integral one (Schölkopf et al, 2002):
   \[
   \int \tilde{K}(x, x(\mu)) dx = 1, \tag{3.21}
   \]
3. The smoothing parameter \( \rho \) is chosen accordingly to condition (3.10):
   \[
   \rho \gg \Delta_{min}.
   \]

Then the quantity (3.20) can be interpreted as Parzen window estimator of the class densities in the space \( \mathcal{H} \).

This interpretation can help the reader to have a most intuitive feeling on how the SGMRF classifier works; but it should be kept in mind that in general the two algorithms differ, for the following reasons:
The underlying assumption here is that we use the Hopfield energy (2.4)-(2.6); there are many other SG-like energies that can be kernelized and used in the SG-MRF framework described in Section 3.6 (see for instance Chapter 5 and (Caputo et al, 2002e; Caputo et al, 2002g)); for these energies the argument drops.

The second hypothesis is true for generalized Gaussian kernels, but it is not true in general for all the Mercer kernels that satisfy criteria (j)-(jj), pg 32; for instance:

\[ K(x, y) = \frac{(x \cdot y + 1)^p}{\sqrt{(x \cdot x + 1)^p(y \cdot y + 1)^p}}, \quad p > 3 \]

is a Mercer kernel that satisfies the criteria, but not the second hypothesis.

The first and third hypotheses are strictly related: the \( \rho \) value determines, for each object class, the number of prototypes, selected from the training set. This number will vary in general from object class to object class, thus violating the first hypothesis.

The third hypothesis implies a specific method for selecting \( \rho \). The proper choice of the smoothing parameter \( \rho \) is a major problem in classification methods based on Parzen windows. For a too small value of \( \rho \) the density is smoothed and details are lost, for a too large value it becomes spiky and the generalization performance is poor; the choice is done with heuristic criteria (Cao et al, 1994), in contrast with SGMRF where it is determined by theoretical arguments.

In conclusion, we can say that SGMRFs and Parzen window estimators present some analogies, and for some specific choices of the free parameters they are equivalent; but as the discussion above has shown, they are two different methods.

### 3.8.2 Support Vector Machines

The first and most famous kernel method presented in literature has been the Support Vector Machine (SVM, (Schölkopf et al, 2002)) algorithm. In this Section we give a brief introduction to SVMs, and we underline the main differences between them and SG-MRF. We refer the reader to (Schölkopf et al, 2002; Vapnik, 1998; Cristianini et al, 2000) for a more comprehensive discussion on SVM.

Let \((x_i, y_i)_{1 \leq i \leq n}\) be a set of training examples, each example \(x_i \in \mathbb{R}^m\), \(m\) being the dimension of the input space, belongs to a class labeled by \(y_i \in \{-1, 1\}\). The aim is to define a hyperplane which divides the set of examples such that all the points with the same label are on the same side of the hyperplane. This amounts to finding \(w\) and \(b\) so that

\[ y_i(w \cdot x_i + b) > 0, \quad i = 1, \ldots, n. \quad (3.22) \]
If there exists a hyperplane satisfying eq. (3.22), the set is said to be linearly separable. In this case, it is always possible to rescale \( w \) and \( b \) so that
\[
\argmin_{i=1}^{n} y_i (w \cdot x_i + b) \geq 1, \quad i = 1, \ldots, n,
\]
as to say so that the closest point to the hyperplane has a distance of \( 1/|w| \). Then, eq. (3.22) becomes
\[
y_i (w \cdot x_i + b) \geq 1.
\] (3.23)

Among the separating hyperplanes, the one for which the distance to the closest point is maximal is called Optimal Separating Hyperplane (OSH). Since the distance to the closest point is \( 1/|w| \), finding the OSH amounts to minimizing \( |w|^2 \) under constraint (3.23).

The quantity \( 2/|w| \) is called the margin, and thus the OSH is the separating hyperplane which maximizes the margin. The margin can be seen as a measure of the generalization ability: the larger the margin, the better the generalization is expected to be (Schölkopf et al., 2002; Cristianini et al., 2000; Vapnik, 1998). Since \( |w|^2 \) is convex, minimizing it under linear constraints (3.23) can be achieved with Lagrange multipliers. If we denote by \( \alpha = (\alpha_1, \ldots, \alpha_n) \) the \( n \) non negative Lagrange multipliers associated with constraints (3.23), our optimization problem amounts to maximizing
\[
W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i \cdot x_j
\] (3.24)
with \( \alpha_i \geq 0 \) and under constraint \( \sum_{i=1}^{n} y_i \alpha_i = 0 \). This can be achieved by the use of standard quadratic programming methods (Schölkopf et al., 2002; Vapnik, 1998; Cristianini et al., 2000). Once the vector \( \alpha^0 = (\alpha_1^0, \alpha_2^0, \ldots, \alpha_n^0) \) solution of the maximization problem (3.24) has been found, the OSH \((w_0, b_0)\) has the following expansion:
\[
w_0 = \sum_{i=1}^{n} \alpha_i^0 y_i x_i.
\] (3.25)

The support vectors are the points for which \( \alpha_i^0 \) satisfy equation (3.23) with equality. Considering the expansion (3.25) of \( w_0 \), the hyperplane decision function can thus be written as
\[
f(x) = \text{sign} \left( \sum_{i=1}^{n} \alpha_i^0 y_i x_i \cdot x + b^0 \right).
\] (3.26)

The algorithm can be made non linear via the so called kernel trick: the input data is mapped into a high-dimensional feature space through some nonlinear mapping chosen a priori (Cristianini et al., 2000; Vapnik, 1998). In this feature space, the OSH is constructed. If we replace \( x \) by its mapping in the feature space \( \Phi(x) \), eq.
(3.24) becomes:

\[ W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \Phi(x_i) \cdot \Phi(x_j). \]

If we have \( K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j) \), then only \( K \) is needed in the training algorithm and the mapping \( \Phi \) is never explicitly used. Conversely, given a symmetric positive kernel \( K(x, y) \), Mercer's theorem (Schölkopf et al., 2002; Cristianini et al., 2000; Vapnik, 1998) indicates us that there exists a mapping \( \Phi \) such that \( K(x, y) = \Phi(x) \cdot \Phi(y) \). Once a kernel \( K \) satisfying Mercer's condition has been chosen, the training algorithm consists of minimizing

\[ W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j), \quad (3.27) \]

and the decision function becomes

\[ f(x) = \text{sign} \left( \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \right). \quad (3.28) \]

There are many differences between SVMs and SG-MRFs: first of all, SVM is a two-class classifier and SG-MRF is a framework for modeling probability density functions; it can be used for classification purposes in a MAP classifier, but in general the separating margin between classes will be different for SVMs and SG-MRF-MAP. Second, SVM can use any type of Mercer’s kernel, while SG-MRF is limited to the use of generalized Gaussian kernel. This limitation is compensated, for SG-MRF, by the possibility to learn from the training data the specific kernel to be used; for SVMs the choice of the kernel is heuristic (Schölkopf et al., 2002).

### 3.8.3 Synergetic Pattern Recognition

There are some analogies between SG-MRF and the synergetic approach to pattern recognition, which was first described by Haken (Haken, 1991). Haken’s synergetic approach classifies a pattern directly within the visual domain, by establishing a parallel between pattern recognition and pattern formation. The concept is to consider the visual input as the initial conditions for a pattern-formation process. This process is designed to terminate when the initial pattern has evolved into any of the user-defined prototype patterns. Each prototype represents a certain class, and the input is categorized as belonging to the class of the recalled prototype.

Haken developed a pattern-formation model defined by a system of differential equations. He also showed that, for a particular choice of parameters, the final solution of the dynamic system can be predicted without integrating the system of differential equations.
Being the concept of synergetic pattern recognition inspired to model of statistical physics, it is no surprise that there are analogies with SG-MRF. Here we want instead underline the differences: first, prototypes in synergetic pattern recognition are selected by the user, while in SG-MRF they are learned from the training data on the basis of theoretical constrains. Second, in synergetic pattern recognition, the classification process is determined by the integration of a set of differential equations, and just for some particular assumption the final solution can be predicted from the initial conditions, so that integrating the system is unnecessary for classification (Haken, 1991). In SG-MRF, once the prototypes are selected and the kernel parameters are learned from the training data, the pdf of the object class is estimated, and classification can be performed. Finally, the decision boundary between two classes is linear for synergetic pattern recognition; for SG-MRF, the decision boundary will be in general not linear, because of the presence of the kernel function into the algorithm. Moreover, as the kernel is learned from the training data, the non linearity is adapted to the considered application in the training stage.

3.8.4 The FRAME Model

The SG-MRF model presents many similarities with the FRAME model proposed by Zhu (Zhu et al, 1998) for texture modeling. It characterizes an ensemble of images $\mathbf{I}$ with the same texture appearance by a probability distribution $f(\mathbf{I})$ on a random field; the goal is to make inference about $f(\mathbf{I})$, given a set of observed texture examples. Texture modeling is achieved in two steps: first, a set of filters is selected from a general bank to capture features of the textures and then applied to the observed examples. The histograms of the filtered images are estimates of the marginal distributions of $f(\mathbf{I})$. Second, the maximum entropy principle is employed to obtain a distribution $P(\mathbf{I})$, restricted to have the same marginal distributions found in the first step; thus $P(\mathbf{I})$ is considered as an estimate of $f(\mathbf{I})$, and via Lagrange multipliers is given by

$$P(\mathbf{I}, \beta) = \frac{1}{Z(\beta)} \exp \left\{ - \sum_{\kappa=1}^{k} \beta_{\kappa} H_{\kappa}(\mathbf{I}) \right\} = \frac{1}{Z(\beta)} \exp\{-\beta \cdot H(\mathbf{I})\}, \quad (3.29)$$

where $P$ is specified once given the parameter $\beta = (\beta_{1}, \ldots, \beta_{k})$, which are determined by the constraints on the marginals. Equation (3.29) is a MRF model, and the energy function associated with the FRAME Gibbs distribution (3.29) is

$$E_{\text{FRAME}} = - \sum_{\kappa=1}^{k} \beta_{\kappa} H_{\kappa}(\mathbf{I}) = -\beta \cdot H(\mathbf{I}), \quad (3.30)$$

as to say an energy function which depends only by the scalar product between configurations, on a fully connected MRF. Equation (3.30) can be written as

$$E_{\text{FRAME}} = -K(\beta, H(\mathbf{I})), \quad K(x, y) = (x \cdot y)^{n}, \quad n = 1.$$
Thus, the FRAME energy has the same functional form of the SG-MRF one \(^2\), but for SG-MRF the scalar product in the energy function is defined via the kernel; thus several metrics are allowed in the probability model and not just the Euclidean one.

## 3.9 Experiments

In this Section we present experiments that show the effectiveness of SG-MRF for appearance-based object recognition. To this purpose, we ran several series of experiments on different databases, using different representations. We used three different kinds of representations, namely:

- **Raw Pixels Data (RPD).**
  
  With this representation, each view is represented by its pixels; usually, the spatial resolution of the original view is reduced to a fixed new resolution. In the experiments we ran using this representation, the final resolution was of 32×32 pixels, obtained using nearest neighbor interpolation. The choice of this representation is supported by many reasons: it does not require feature extraction, and it can be applied directly to images without pose estimation. Furthermore, many experiments have been presented in literature on publicly-available database, using this representation with the same reduced spatial resolution (see for instance (Pontil et al., 1998)). This offers the possibility to compare our results with those.

- **Multidimensional receptive Field Histograms (MFHs).**
  
  This representation was proposed by Schiele (Schiele et al., 2000) in order to extend the color histogram approach of Swain and Ballard (Swain et al., 1991). The main idea is to calculate multidimensional histograms of the response of a vector of receptive fields:

  \[
  MFH = \text{Prob}[X_1 = x_{ij}, X_2 = x_{ij}, \ldots], \quad i = 1, \ldots L, j = 1, \ldots M,
  \]

  with \(X_1, X_2, \ldots\) responses of the original image \(x\) to a set of receptive fields filters:

  \[
  X_1 = \hat{X}_1(x), X_2 = \hat{X}_2(x), \ldots \ldots .
  \]

  A MFH is determined once we chose the local property measurements (i.e., the receptive field functions), which determine the dimensions of the histogram, and the resolution of each axis. We chose for our experiments the following local characteristics, based on Gaussian derivatives:

  \[
  D_x = -\frac{x}{\sigma^2} G(x, y), \quad D_y = -\frac{y}{\sigma^2} G(x, y), \quad \text{Lap} = G_{xx}(x, y) + G_{yy}(x, y)
  \]

\(^2\)This is no surprise at all, since the Hopfield model can be derived from the maximum entropy principle as well.
where

\[ G(x, y) = \exp \left( -\frac{x^2 + y^2}{2\sigma^2} \right) \]

is the Gaussian distribution and

\[ G_{xx} = \left( \frac{x^2}{\sigma^4} - \frac{1}{\sigma^2} \right) G(x, y), \quad G_{yy} = \left( \frac{y^2}{\sigma^4} - \frac{1}{\sigma^2} \right) G(x, y) \]

are the second order derivatives with respect to \( x \) and \( y \). We used several different combinations of the local characteristics, chosen heuristically for different experiments and databases. We will specify them later for each specific experiment.

- **Color Histograms (CH).**
  A color histogram denotes the joint probabilities of the intensities of the three color channels (Swain et al., 1991). The color histogram is defined as

\[ h_{R,G,B}(r, g, b) = m \cdot \text{Prob}[R = r, G = g, B = b], \]

where \( R, G \) and \( B \) are the three color channels and \( m \) is the number of pixels in the image. The color histogram is computed by discretizing the colors within the image and counting the number of pixels of each color. Since the number of colors is finite, it is usually more convenient to transform the three channels histogram into a single variable histogram. Given an RGB image, one transform is given by

\[ t = r + m_r g + m_r m_g b, \]

where \( m_r, m_g \) and \( m_b \) are the number of bins for colors red, blue and green, respectively. This gives the single variable histogram

\[ h[t] = m \cdot \text{Prob}[T = t]. \]

As we used different representations for different experiments, we will specify them later, for each specific experiment.

We used SG-MRF in the MAP framework described in Section 3.6. For all the experiments we performed, we benchmarked SG-MRF recognition results with at least another method. All the experiments described in this Chapter, as all the experiments described in this thesis, were ran on a pc Pentium III 600 MHz, 512 MB. If not specified differently, we used a C++ implementation under Linux.

### 3.9.1 Columbia Database Experiments.

We ran a first set of experiments on the Columbia database (Nene et al., 1996), which can be seen as a benchmark for object recognition algorithms. It consists of 7200 color images of 100 objects (72 views for each of the 100 objects); each image is of 128 x 128 pixels. The images were obtained by placing the objects on a turntable and taking a view every 5°. Figure 3.3 shows some of the objects contained in the Columbia database, and Figure 3.4 shows some of the views per object.
Figure 3.3: An example of 20 objects of the 100 contained into the Columbia database.

Figure 3.4: Columbia database, Examples of different views.
3.9.1.1 ICA Ansatz versus Naive Ansatz Experiments

We ran a preliminary experiment on a subset of the Columbia database consisting of 20 objects (COIL database, (Nene et al, 1996)), for testing the naive ansatz versus the ICA ansatz. In the series of experiments we performed, we divided the database in training and test set differently, thus we will describe it below. The use of the ICA ansatz implies the estimation of the independent components of the considered visual classes, which is usually performed by formulating an objective function and then maximizing or minimizing it (Cardoso, 1998). Here we decided to use the infomax algorithm derived by Bell and Sejnowsky (Bell et al, 1995); this algorithm has been successfully applied to many signal separation problems (Bell et al, 1995); it has also been applied for finding image filters that give independent outputs from natural scenes (Bell et al, 1997), employing the same architecture we use here for finding the SG-MRF’s prototypes.

We ran three set of experiments on a RPD representation. We performed Principal Component Analysis (PCA, (Duda et al, 2001)) on the training dataset, and the data were represented on the principal component axes; the number of axes changed for each experiment and will be specified later. For all the experiments the infomax algorithm \(^3\) was ran for 1600 iterations, with a learning rate initialized at 0.0002 and annealed down to 0.00002. The SG-MRF-MAP classifier was ran using a MATLAB implementation.

We did a first set of experiments for evaluating SGMRF’s performance using different kernels (thus in this experiments we skipped the learning of kernel parameters \((a, b)\)). We ran experiments on a training database of 720 views (36 views per object) and a PCA representation of 30 components, for the following kernel parameters: \(a = 1, b = 2, 1, 0.5\). We tested the performance on the remaining 720 views; results are reported in Table 3.2. We see that the naive ansatz leads to the best performance for every kernel value. We see also that, when recognition rate decreases for \(b \rightarrow 0\), the ICA ansatz suffers much more than the naive ansatz.

A second set of experiments was ran for comparing the ansatzs’ performance as the training dataset is reduced. We performed experiments with 36, 24, 18 and

\(^3\)We gratefully acknowledge Scott Makeing who has made available the infomax algorithm at http://www.cnl.salk.edu/.
3.9. EXPERIMENTS

<table>
<thead>
<tr>
<th>Kernel parameters</th>
<th>ICA ansatz</th>
<th>Naive ansatz</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a = 1, b = 2$</td>
<td>94.17</td>
<td>99.58</td>
</tr>
<tr>
<td>$a = 1, b = 1$</td>
<td>70.28</td>
<td>91.38</td>
</tr>
<tr>
<td>$a = 1, b = 0.5$</td>
<td>65.56</td>
<td>87.36</td>
</tr>
</tbody>
</table>

Table 3.2: Performance evaluation of the naive ansatz versus the ICA ansatz, with 30 PC components and 36 prototypes per object class. We report recognition results when kernel parameters vary.

12 views per object, PCA representation of 30 components and kernel parameters $a = 1, b = 2$. Performance was tested respectively on 36, 48, 54 and 60 views per object; we report results in Table 3.3. We see that, for both ansatz, performance decreases as the number of prototypes does. Again, the naive ansatz obtains the highest recognition rate.

Finally, we ran a set of experiments for exploring the performance’s dependence from the components of the PCA representation. We ran experiments for PCA components $= 10, 20, 30, 40$ for 12 views per object, kernel parameters $a = 1, b = 2$. Classification results, on a test set of 60 views per object, are reported in Table 3.4. Although once again we have the best performance with the naive ansatz, we observe a different behavior for the two ansatz: for the ICA ansatz, the performance increases as the number of PC components does. For the naive ansatz, performance reaches its maximum for 30 PCs, and then starts to decrease.

In conclusion, we can say from these experiments that the naive ansatz gives the best performance, and thus will be used through this thesis. Regarding the possible reasons for this behavior, we give the following interpretation: first, as PCA decorrelates vectors on the whole training set, this probably explains the good performance of the naive ansatz. Second, the ICA ansatz presents many heuristic parameters (the linearity assumption, the choice of the algorithm for performing ICA, the number of prototypes, the number of PCA components) that are not related with the statistical independence of the prototypes, but with the way we decide to extrapolate the independent components, for a given visual class.

3.9.1.2 RPD Experiments

Once settled on the naive ansatz, we performed a first set of experiments using a RPD representation. We chose generalized kernels with $a = 1, 0.5$ and $b = 2, 1, 0.5$, in order to see how the performance varies when kernel parameters do (thus we skipped the learning of $(a, b)$ in the training stage). The training set consisted of 36 views per object, test set of the remaining 36 views. The obtained recognition rates

---

4This suggests also that PCA can be fruitfully used as representation for SG-MRF classifiers. We will explore this possibility in the future.
Table 3.3: Performance evaluation of the naive ansatz versus the ICA ansatz, with 30 PC components and kernel parameters \(a = 1, b = 2\). We report recognition results when the number of prototypes varies.

<table>
<thead>
<tr>
<th>Number of Prototypes</th>
<th>ICA ansatz</th>
<th>PCA ansatz</th>
</tr>
</thead>
<tbody>
<tr>
<td>p=36</td>
<td>94.17</td>
<td>99.58</td>
</tr>
<tr>
<td>p=24</td>
<td>92.43</td>
<td>99.24</td>
</tr>
<tr>
<td>p=18</td>
<td>91.74</td>
<td>98.82</td>
</tr>
<tr>
<td>p=12</td>
<td>87.78</td>
<td>97.29</td>
</tr>
</tbody>
</table>

Table 3.4: Performance evaluation of the naive ansatz versus the ICA ansatz, with 12 prototypes per object and kernel parameters \(a = 1, b = 2\). We report recognition performance when the number of principal components varies.

<table>
<thead>
<tr>
<th>Number of PC components</th>
<th>ICA ansatz</th>
<th>Naive ansatz</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC= 10</td>
<td>76.46</td>
<td>93.33</td>
</tr>
<tr>
<td>PC=20</td>
<td>84.38</td>
<td>96.11</td>
</tr>
<tr>
<td>PC=30</td>
<td>87.78</td>
<td>97.29</td>
</tr>
<tr>
<td>PC=40</td>
<td>93.75</td>
<td>97.15</td>
</tr>
</tbody>
</table>

for NNC, \(\chi^2\), SVM with several kernels \(^5\) and SG-MRF for the six chosen kernels are reported in Tables 3.5-3.6. Recognition time per view goes from a minimum of 0.15 sec for kernel parameters \(a = b = 1\) to a maximum of 0.37 sec for kernel parameters \(a = b = 0.5\).

<table>
<thead>
<tr>
<th>NNC</th>
<th>SG-MRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.29</td>
<td>(b = 2) (99.29) (b = 1) (99.76) (b = 0.5) (99.86)</td>
</tr>
<tr>
<td>(a = 1)</td>
<td>(a = 0.5)</td>
</tr>
</tbody>
</table>

\(b = 1\) \(99.61\) \(b = 0.5\) \(99.72\)

Table 3.5: Classification results using RPD representation for NNC and SG-MRF and SVM.

We see that SVM with Gaussian kernel and a specific value of \(s\) gives the best performance, followed by SG-MRF (\(-0.11\%\)). It is worth to note anyway that all methods give a very good performance (99\% and above) using this representation. Moreover, although SG-MRF doesn’t give the best absolute performance, its results are much more stable compared with those of SVM as the kernel function changes.

\(^5\)We gratefully acknowledge R. Collobert that has made available the SVMTorch software on the web (http://kernel/methods.org.)
### 3.9. EXPERIMENTS

#### 3.9.1.3 MFH Experiments

A second set of experiments was run using a MFH representation with $DxDy$ Gaussian derivatives, $\sigma = 1.0$ and resolution of the histogram bin per axis 16; we call this choice of the MFH parameters 2D-MFH, and we will use it often through the thesis, as we have heuristically found it quite effective. We used again half of the views as training and half as test set, as for the previous experiments. Experiments were performed as described above for the RPD representation. Recognition time per view are analogous to those reported for the RPD representation. We report results in Tables 3.7-3.8: in this case SG-MRF gives the best performance. We note once again the great stability in SG-MRF’s performance as the kernel parameters vary, with respect to the SVM’s performance when the kernel function changes.

#### Table 3.7: Classification results for NNC and SG-MRF on the COIL100 database using a MFH representation.

<table>
<thead>
<tr>
<th>NNC</th>
<th>SG-MRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>94.43</td>
<td>$a = 1$</td>
</tr>
<tr>
<td></td>
<td>$b = 1$</td>
</tr>
<tr>
<td></td>
<td>$b = 0.5$</td>
</tr>
</tbody>
</table>

#### Table 3.8: Classification results for $\chi^2$ and SVM on the COIL100 database using a MFH representation.

<table>
<thead>
<tr>
<th>$\chi^2$</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.52</td>
<td>$K_1 = 93.01$</td>
</tr>
<tr>
<td></td>
<td>$K_3 = 1.00$</td>
</tr>
<tr>
<td></td>
<td>$K_5, s = 4000$</td>
</tr>
</tbody>
</table>
3.9.1.4 CH Experiments

A third set of experiments was ran using CH. We chose a Hue CH representation (Swain et al, 1991), resolution per bin axis 8, training and test set as in previous experiments. We used a generalized Gaussian kernel with $a = 0.5$ and $b = 2$. Performance of SG-MRF was compared with those of a NNC and of a $\chi^2$ distance. Recognition time per view is 0.08 sec per view. The obtained results are reported in Table 3.9; in this case both SG-MRF and $\chi^2$ give the best performance.

3.9.2 NELSON Database Experiments

The NELSON database (Nelson et al, 1998) is composed by 59 objects: 11 cups, 5 dolls, 6 planes, 6 fighter jets, 9 lizards, 5 spoons, 8 snakes and 9 sport cars. Some examples are shown in Figure 3.6. Each object is represented in the training set by a collection of views taken approximatively every 20 degrees on a sphere; this amounts to 106 views for a full sphere, and 53 for a hemisphere. The test set consists of 48 (24) views, positioned in between the training views, and taken under the same conditions. Cups, dolls, fighters, planes, spoons are represented by 106 views in the training set and 48 views in the test set; lizards, snakes, sport cars are represented by 53 views in the training set and 24 views in the test set. We performed a set of experiments using a MFH representation, with Lap Gaussian derivatives, $\sigma_1 = 1.0$ and $\sigma_2 = 1.6$, resolution of the histogram bin per axis 16, parameters chosen heuristically; we call this choice of the MFH parameters Lap$2\sigma$-MFH. We chose generalized Gaussian kernels with $a = 0.5$ and $b = 2, 1.5, 1, 0.5$. The performances of SG-MRF were compared with those of a NNC, of a $\chi^2$ distance and of an intersection ($\cap$, (Schiele et al, 2000)) similarity measure. Recognition results are reported in Table 3.10, and we can see that once again the best recognition rate is obtained by SG-MRF. It would be very interesting to compare these results with those obtained using the Nelson algorithm (Nelson et al, 1998), for which the database was especially built. Unfortunately, experimental results are published just on subsets of the database, from 6 up to up to 24 objects (Selinger, 2001). Recognition rate goes from 98.5 As on the same database, but with 59 objects, we obtain a recognition rate of 97.74%, we can conclude that SG-MRF performs better for this task. Recognition time per view is analogous to that reported for the Columbia experiment.
3.10. SUMMARY

53

Figure 3.6: Examples of objects in the NELSON database.

<table>
<thead>
<tr>
<th>NNC</th>
<th>$\chi^2$</th>
<th>$\cap$</th>
<th>SG-MRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>80.93</td>
<td>97.19</td>
<td>92.89</td>
<td>$a = 0.5, b = 2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$a = 0.5, b = 1.5$</td>
</tr>
</tbody>
</table>

Table 3.10: Classification results for NNC, $\chi^2$, $\cap$ and SG-MRF on the NELSON database using a MFH representation.

3.9.3 Discussion

The reported results show that, for every database, for every representation, there is at least one kernel that gives the best performance in the SG-MRF framework. These results confirm the correctness and the effectiveness of SG-MRF for appearance-based object recognition. As a final remark, we observe that all the experiments performed indicates that the performance of SG-MRF improves when the product $ab \to 0$ (Chapelle et al, 1999). This suggest that a good choice for kernel parameters is $a = 0.5, b = 1.0$, for those cases in which one prefers to avoid the learning stage.

3.10 Summary

In this Chapter we derived Spin Glass-Markov Random Fields, a new class of graphical model connecting together Markov Random Field modeling and results of stat-
istical physics of Spin Glasses. This result is achieved via kernel functions; thus Spin Glass-Markov Random Fields can be considered a new kernel method. We presented several experiments that proves the correctness of the derived model and its effectiveness. In the following Chapters we will explore in more details the properties of Spin Glass-Markov Random Fields. We will particularly focus the attention on its robustness properties for appearance-based object recognition; this will be the topic of the next Chapter.
Chapter 4

Robustness of Spin Glass-Markov Random Fields

In this Chapter we explore robustness of Spin Glass-Markov Random Fields for object recognition. We present a theoretical analysis of the algorithm behavior in presence of degraded visual information. We consider the case of degradation in the training and test set; we find that adding degraded views into the training set leads to a higher robustness (Section 4.2-4.3). We verify this theoretical result by an extensive set of experiments: we tested robustness of Spin Glass-Markov Random Fields to noise (Section 4.4.1), occlusion (Section 4.4.2), decreasing number of training views (Section 4.4.3) and heterogeneous background (Section 4.4.4). Experiments agree with theoretical predictions.
4.1 Introduction

Robust appearance-based object recognition is a challenging problem. This is due partially to the intrinsic difficulty of the task, and partially to the difficulties induced by the appearance-based approach. An ideal robust object recognition system should be able to recognize a wide variety of objects, under different lighting conditions, in different environments. It should be able to recognize objects even when they are just partially visible, or when their images are blurred by dust or some other agents. Although object recognition is a much researched area in computer vision, these results are still far from being achieved.

The ultimate goal of computer vision is to make systems that can autonomously interpret the visual environment under almost any operating condition (Meer, 2000). This ambitious goal makes difficult to define the term *robustness* in a unique and precise way. Within the context of computer based image understanding, for example, different definitions emerge at different levels. At the top level a robust vision system should be able to recognize new objects based exclusively on previous examples from the same class of functionality (Meer, 2000). A primary concern at the low level is to have robust feature extraction techniques that tolerate the presence of outliers (Pinzon et al, 1998). Changes in the visual environment give rise to complex visual events; their robust interpretation requires separating the external causes from the intrinsic properties in the appearance of each object (Meer, 2000). This abstract definition of robustness at the intermediate level is mostly used in a more concrete way, when a certain system or model is examined for its robustness. If we consider the field of 3D object recognition, the term robustness refers to the fact that results remain stable, in the presence of various types of noise and certain portions of outliers, such as occlusion and changes of scale, viewpoint or illumination (Leonardis et al, 2000). The use of probabilistic models in this field was motivated, among others, by the fact that they tolerate in various degrees the presence of data points, or features, that do not obey to the assumed model. Different probabilistic models such as histograms (Schiele et al, 2000), Support Vector Machines (SVM) (Pontil et al, 1998) or eigenspace approach (Leonardis et al, 2000) have been examined for their robustness against noise, partial occlusion and other types of changes such as scale changes or viewpoint alteration (Schiele et al, 2000); only a few characterize robustness theoretically (based on methods of robust statistics) (Leonardis et al, 2000). All of them investigate the stability of the results when different forms of degradation are present in the test set. Such an approach corresponds to the ideal case in which the training stage is done on a set of data not affected by any degradation. This implies the possibility to have a sort of control on the quality of the training set. But this is not always the case: the training set itself can contain samples affected by noise or occlusion. Consider for example the case in which the training data set is acquired from real environments like streets, interior of buildings and so on. Dust, changes in lighting conditions, sudden and unpredictable changes in the environment can have, as final effect, the acquisition of degraded training views. A possible way to overcome this problem is of course, to
4.2. ROBUSTNESS OF SPIN GLASS-MARKOV RANDOM FIELDS: A STATISTICAL MECHANICS VIEW

reject degraded samples from the training set. Here we propose a different strategy, as to say we assume that the training set can contain degraded samples, and we study the robustness of SG-MRF to this kind of perturbation of the data. This implies an extension of the concept of robustness to the capability of a model to generalize correctly from degraded data.

The rest of the Chapter is organized as follows: in Section 4.2 we will analyze the robustness properties of Spin Glass-Markov Random Fields from a statistical mechanics viewpoint. We will show that theoretical results from SG theory make us expect SG-MRF to be highly robust. In Section 4.3 we will give a more intuitive explanation on the robustness of SG-MRF, based on kernel properties. Section 4.4 reports several series of experiments designed to check the robustness of SG-MRF to Gaussian noise, to two kinds of occlusion, to decreasing number of prototypes and to heterogeneous background. The Chapter concludes with a summary discussion.

4.2 Robustness of Spin Glass-Markov Random Fields: A Statistical Mechanics View

Robustness is a remarkable property of the H-L energy (2.4)-(2.6), that we rewrite here for sake of clarity:

\[ E = - \sum_{(i,j)=1}^{N} J_{ij} s_i s_j, \quad s \in [-1, +1]^N, \quad N \to \infty. \]

\[ J_{ij} = J_{ji}, \quad J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi^{(\mu)}_i \xi^{(\mu)}_j. \]

It can be demonstrated that:

1. it correctly classifies degraded versions of the prototypes \( \{\xi^{(\mu)}\}_{\mu=1}^{p} \);

2. it correctly generalizes from degraded versions of the prototypes \( \{\xi^{(\mu)}\}_{\mu=1}^{p} \).

For the first point, consider Figure 4.1, and let us call \(2w\) the vector representing the width of the basin of attraction of the \(\mu\)-th prototype \(\xi^{(\mu)}\) (all prototypes have basins of attraction of the same width, (Amit, 1989)). Every SG configuration vector \(s = \xi^{(\mu)} \pm aw, a < 1\) will be inside the basin of attraction of the prototype \(\xi^{(\mu)}\), that thus will be the dominant attractor for \(s\). The width of the basin of attraction for the H-L energy is quite large ((Amit, 1989), Chapter 4), thus we can say that it correctly classifies degraded versions of the prototypes \( \{\xi^{(\mu)}\}_{\mu=1}^{p} \). From the point of view of appearance-based object recognition, this would mean that the H-L energy classifies correctly views, representing the considered object, that differs until a certain level from the stored one; the difference can be due to a change of pose parameters, to noise or occlusions.
Regarding the second point, it means storing degraded versions of the original prototypes, namely a set of prototypes \( \{ \xi^\mu \}_{\mu=1}^p \), with

\[
\tilde{\xi}^\mu = \xi^\mu + \delta^\mu,
\]

with \( \delta^\mu \) random vectors. This is equivalent to choose as connection matrix

\[
T_{ij} = J_{ij} + \eta_{ij}. \tag{4.1}
\]

The properties of the energy function (2.4), obtained using the connection matrix (4.1) have been studied in details by Sompolinski (Sompolinski, 1986), using tools of statistical mechanics. We summarize again the results with a landscape metaphor. Figure 4.1 shows the effect of the degradation on the stored prototypes (dashed line) with respect to the energy landscape when the original prototypes (continuous line) are stored. We see that the \( \{ \xi^\mu \}_{\mu=1}^p \) are the absolute minima of both energies, and the valleys have the same depth. But the width of the basin of attraction increases when there is a perturbation in the set of prototypes. This means that 1) the system correctly generalizes from degraded versions of the original prototypes, and 2) its robustness augments when it stores degraded prototypes. The problem is that, as the width of the basin of attraction augments, the number of prototypes that can be stored in the energy becomes a function of the amount of perturbation \( \delta^\mu \), and it decreases as it increases. This can lead to a phase transition, and to a loss of the recall properties of the energy. From the point of view of appearance-based object recognition, this would mean that storing degraded versions of the set of original prototypes would lead to a higher robustness with respect to degradations such as noise and occlusions.

### 4.3 Robustness of Spin Glass-Markov Random Fields: a Kernel View

We shown in the previous Section that, due to the properties of the H-L energy, SG-MRF are expected to be highly robust with respect to noise and occlusion. The reader may wonder how a fully connected model, thus a global representation of the object view, is expected to show robustness to occlusion, a property that relies on the capability to model local features of the object views. Things get clearer if we rewrite the kernelized SG-MRF energy as follows:

\[
E = -\sum_{\mu=1}^{p_j} [K_{d-C}(x, \tilde{x}^\mu)]^2 = -\sum_{\mu=1}^{p_j} [\Phi(x)^T \cdot \Phi(\tilde{x}^\mu)]^2 = -\sum_{\mu=1}^{p_j} \Phi(x)^T \cdot \Phi(\tilde{x}^\mu) \Phi(\tilde{x}^\mu)^T \cdot \Phi(x);
\]
4.3. ROBUSTNESS OF SPIN GLASS-MARKOV RANDOM FIELDS: A KERNEL VIEW

Figure 4.1: Energy function as a landscape, when the stored prototypes are with (bottom) or without (top) added noise. The dot line between valleys represents the space between them. When noisy prototypes are stored, it results in an enlargement of the width of the basin of attraction (bottom).

\[
E = -\sum_{i,j=1}^{m} \Phi(x_i)^T C_{ij} \Phi(x_j), \tag{4.2}
\]

where

\[
C_{ij} = \sum_{\mu=1}^{p_j} \Phi(\tilde{x}_i^\mu)^T \Phi(\tilde{x}_j^\mu) \tag{4.3}
\]

is the kernelized connection matrix. Equation (4.2) shows that the SG-MRF energy function is given by the sum, on all the components of the image view configuration vector, of all the possible coupled interactions between components, weighted by a factor given by the corresponding element of the connection matrix (4.3). In order to fix ideas, let us consider the case when the \( \mathbf{x} \) is the image view (and thus \( x_i \) is the pixel); the energy (4.2) models the pdf generating \( \mathbf{x} \) considering all the possible interactions between couples of pixels, weighted by a factor \( C_{ij} \). Thus, the energy is fully connected, which makes the representation global. But this globality is achieved via the sum of all the possible local interactions, weighted by a coefficient (the matrix element \( C_{ij} \)) which is learned by the training data (note that the connection matrix is determined by the set of prototypes, equation (4.3)). This means that many of those contributions can be zeros; thus we achieve a globality that is the result of the contribution of all the significant localities for the considered object. As a consequence of this property of SG-MRF we expect them to be robust to noise (due to the globality) as well as to occlusion (due to the locality).
4.4 Robustness of SG-MRFs: Experiments

In this Section we present experiments that check the robustness of SG-MRF. In order to investigate whether the experimental performance is in agreement with the theoretical analysis presented in Section 4.2-4.3 we ran four series of experiments designed to measure robustness to independent additive Gaussian noise, to two kinds of partial occlusion, to decreasing number of prototypes and to heterogeneous background.

We performed experiments on the COIL and Columbia (Nene et al., 1996) databases, described in Section 3.9. For all these experiments, we used a MFH representation.

4.4.1 Robustness to Noise

Most of the experiments presented in literature on the robustness of object recognition systems against noise use ideal views for training and introduce noise in the test set (Pontil et al., 1998; Schiele et al., 2000). This only demonstrates the capability of a system to recognize corrupted images but not its ability to generalize from degraded training images. In this Section we report experiments we ran to investigate this second property. Motivated by theoretical results (see Section 4.2), we expect that, with an appropriate set of prototypes and a reasonable amount of noise, the system would retain its storage capacity and hence be robust to this kind of degradation. Noise can be considered as global degradation of an image, since it partially affects the gray values of all picture elements.

We performed several series of experiments on the COIL and Columbia databases using 36 prototypes per object in the training set. Test set consisted of the whole set of views. In order to explore robustness of SG-MRF to noise, we added independent Gaussian noise with \( \sigma_{\text{noise}} \in \{10, 50, 80, 120\} \). Note that since the image gray levels are bound to be between 0 and 255, adding Gaussian noise means that the noisy images were actually rescaled within the range \([0, 255]\) (Pontil et al., 1998). Some examples of noisy views are shown in Figure 4.2. For each level of noise, we generated four training sets containing respectively 25%, 50%, 75% and 100% of degraded views with respect to the original training set. For example, a 25% training set - for a given noise level, for a given object - contains 9 noisy and

Figure 4.2: COIL database, examples of noisy images: from left to right, \( \sigma_{\text{noise}} \in \{0, 10, 50, 80, 120\} \)
27 original views. We proceeded in the same way to generate four test sets with
the same percentages of degraded images. The views to be corrupted were picked
out randomly \(^1\). For each noise level, for each test set the training was performed
respectively on the generated training sets including the ideal one. The views were
represented using 2D-MFH, with \(D_xD_y\) Gaussian derivatives, \(\sigma = 1.0\), resolution
of histogram axis of 16 bins. The kernel parameters were \(a = 0.5\) and \(b = 1\). Recogni-
tion time per view was 0.02 sec for the COIL database and 0.23 sec for Columbia
database.

Results for \(\sigma_{\text{noise}} = 10\) Figure 4.3 shows the recognition rates obtained when
the training set contains 0% of noisy views, as a function of increasing number of
noisy views in the training set. We see that for both databases and all classifiers,
performance decreases as the percentage of noisy views in the test set increases.
Still, SG-MRF performs significantly better than \(\chi^2\) and NNC. Figure 4.4 shows

![Figure 4.3: Recognition rates for the COIL (left) and Columbia (right) database,
for the noise level \(\sigma_{\text{noise}} = 10\) and 0% of degraded views into the training set.
Results are reported as a function of the percentage of degraded views in the test
set, for SG-MRF, \(\chi^2\) and NNC.](image)

the recognition rates obtained when the training set contains 50% of noisy views,
as the number of noisy views in the test set increases. Here we see a spectacular
change of behavior for SG-MRF and \(\chi^2\): for both classifiers, adding noisy training
views leads to a remarkable robustness. Although the behavior is similar for the
two classifiers, SG-MRF still performs better than \(\chi^2\), particularly in the case of
a high number of objects (Columbia database, Figure 4.4, right). It is interesting
to note that NNC doesn’t benefit from the degradation into the training set. With
respect to the case of 0% noisy views in the training set, we see that recognition
rate drops for 0% noisy views in the test set (see Figures 4.3 and 4.4). Then, as the

\(^1\)A pilot experiment with a uniform distribution of noisy views led to comparable outcomes.
We conclude that the chosen mixture does not affect the general behavior of the results.
percentage of noisy views in the test set increases, the performance increases as well, in contrast with what happened in the experiment described above. This behavior is more evident for the Columbia database (see Figures 4.3 and 4.4, right). For any percentage of noisy views in the test set, NNC performs worse than SG-MRF and $\chi^2$. Figure 4.5 shows recognition performances when the training set consists of all noisy views. With respect to NNC, the behavior is similar to that observed in the previous experiments. We see that $\chi^2$ recognition performance decreases as the test set doesn’t contain noisy views. This behavior is very evident for the Columbia database. (Figure 4.5, right). On the contrary, SG-MRF maintains its robustness properties for both databases, obtaining the best overall performance.

Results for $\sigma_{\text{noise}} = 50$ Figure 4.6 shows the recognition performance for 0% of noisy views in the training set, as the percentage of noisy views in the test set increases. We see that all classifiers behave similarly, and that performance decreases as the percentage of noisy views in the test set increases. For COIL database experiments, we observe that $\chi^2$ achieves the best performance for a test set of all noisy views. But as this performance is very poor anyway, we claim that this result doesn’t show a superiority of $\chi^2$ on the two other classifiers. Figure 4.7 shows the recognition performance for 50% of noisy views in the training set. We see that, for all classifiers, performance increases considerably and becomes more robust. SG-MRF and $\chi^2$ obtain basically the same performance; NNC had the same behavior but worse performance, particularly for the Columbia database.

Figure 4.8 presents the recognition performance when the training set contains 100% of noisy views. We see that, for all classifiers, robustness properties are lost, and performances are basically the same. This behavior should be compared with
4.4. ROBUSTNESS OF SG-MRFS: EXPERIMENTS

Figure 4.5: Recognition rates for the COIL (left) and Columbia (right) database, for the noise level $\sigma_{\text{noise}} = 10$ and 100% of degraded views into the training set. Results are reported as a function of the percentage of degraded views in the test set, for SG-MRF, $\chi^2$ and NNC.

Figure 4.6: Recognition rates for the COIL (left) and Columbia (right) database, for the noise level $\sigma_{\text{noise}} = 50$ and 0% of degraded views into the training set. Results are reported as a function of the percentage of degraded views in the test set, for SG-MRF, $\chi^2$ and NNC.

the one obtained with a lower level of noise (Figure 4.7); in that case, SG-MRF still presented robustness properties, while the other two classifiers didn’t.

Performances for $\sigma_{\text{noise}} = 80$ and $\sigma_{\text{noise}} = 120$ are analogous to those obtained with $\sigma_{\text{noise}} = 50$. We refer the interested reader to Appendix C.1 for more details on these experiments.
CHAPTER 4. ROBUSTNESS OF SG-MRF

Figure 4.7: Recognition rates for the COIL (left) and Columbia (right) database, for the noise level $\sigma_{\text{noise}} = 50$ and 50% of degraded views into the training set. Results are reported as a function of the percentage of degraded views in the test set, for SG-MRF, $\chi^2$ and NNC.

Figure 4.8: Recognition rates for the COIL (left) and Columbia (right) database, for the noise level $\sigma_{\text{noise}} = 50$ and 50% of degraded views into the training set. Results are reported as a function of the percentage of degraded views in the test set, for SG-MRF, $\chi^2$ and NNC.

Discussion

Diagrams in Figures 4.3-4.8 show the impact of increasing the amount of noise in the test views on the performance of SG-MRF, $\chi^2$ and NNC when the training set consists exclusively of ideal images; for small level of noise ($\sigma_{\text{noise}} = 10$), SG-MRF gives considerably better results than $\chi^2$, while NNC performance degrades already at this low level of degradation. As the corruption of the test views increases, with respect to the quantity of the degraded views in the test set and with respect to the
amount of degradation, all classification performances degrade toward a common (low) recognition rate (Caputo et al, 2002a). The introduction of noisy prototypes in the training set leads to better classification performances; in particular for 50% of noisy views in the training sets we observe an amazing stability with respect to the percentage of degraded views in the test set and with respect to small levels of noise. This stability persists only for SG-MRF, when the training set contains exclusively degraded views, provided the amount of noise is small. For higher level of noise, the recognition rate increases as the number of corrupted images in the test set grows. These experiments confirm theoretical expectations and show robustness of SG-MRF to this kind of global degradation. Almost the same tendency is observed for the three approaches with the Columbia database: for small level of noise, robustness is achieved when noisy patterns are added to the training set. It is remarkable to observe that results obtained with this database are not always as stable as with the smaller data set. This could have the following explanation: as adding noise to the prototypes results in an enlargement of their basin of attraction and by increasing the number of prototypes, the risk that the basins of attraction get mixed would become higher and the system would lose its recall capability; this could be a sign of a phase transition, which would also explain the decrease in the recognition rate for the higher levels of noise also with the COIL database.

4.4.2 Robustness to Occlusion

Occlusion is another kind of degradation that can affect the performance of a system; it manifests itself differently from noise (studied in the previous Section), in the sense that occlusion is a total degradation of a part of the image. Thus, it can lead to significant degradation in the performance if salient parts of the object are occluded. In order to investigate robustness of SG-MRF against occlusions, we carried out two series of experiments. In the first we occluded a certain number of rows and columns assigning to them a random gray value. In the second we reduced gradually the amount of visible portion of the object; in order to be sure that the windows include always portions of the objects, we took the windows centered to the whole image. The same procedure was used by Schiele (Schiele et al, 2000) for a similar experiment. All these experiments were run on the COIL and Columbia databases using 36 prototypes per object for training. Test set consisted of the whole set of views per object. In both series of experiments we followed a similar procedure as for the noise experiments to generate four test and training sets that contain an increasingly percentage of occluded images.

First Kind of Occlusion

For the first kind of occlusion, we chose \( c \) columns and \( r \) rows in the images such as \((c, r) \in \{(0, 0), (8, 8), (16, 16), (32, 32), (64, 0)\}\) and assigned them a random gray value between 0 and 255. Raws and columns were merged together and positioned uniformly at random within the images. Figure 4.9 shows some examples of
CHAPTER 4. ROBUSTNESS OF SG-MRF

occluded images. Views were represented using 2D-histograms, with Gaussian de-

Figure 4.9: Examples of occluded images; from left to right: \((c, r) \in \{(0, 0), (8, 8), (16, 16), (32, 32), (64, 0)\}\)

rivatives \(D_x D_y, \sigma = 1.0\), resolution per histogram axis 16 bins. Kernel parameters were \(a = 0.5\) and \(b = 1\). The task is to classify all test sets for each of the training sets and for every level of occlusion. Results are compared with NNC and the \(\chi^2\) distance measure.

Results for \(r/c = (8, 8)\) Figure 4.10 shows recognition rates when the test set contains an increasing number of occluded views. We see that SG-MRF and \(\chi^2\) performs similarly for COIL database, and they are both robust to the degradation of the test set (Figure 4.10, left). On the contrary, NNC’s performance decreases as the occluded views in the test set increases. Things are different for the Columbia database (Figure 4.10, right): in this case SG-MRF is clearly more robust that \(\chi^2\) and NNC.

Figure 4.10: Recognition rates for the COIL (left) and Columbia (right) database, for the occlusion level \(r/c = 8\) and 0% of degraded views into the training set. Results are reported as a function of the percentage of degraded views in the test set, for SG-MRF, \(\chi^2\) and NNC.

Figure 4.11 shows recognition results for 50% of occluded views in the training set. For both databases, NNC performs as in the previous experiment: it doesn’t
4.4. ROBUSTNESS OF SG-MRFS: EXPERIMENTS

Figure 4.11: Recognition rates for the COIL (left) and Columbia (right) database, for the occlusion level $r/c = 8$ and 50% of degraded views into the training set. Results are reported as a function of the percentage of degraded views in the test set, for SG-MRF, $\chi^2$ and NNC.

Figure 4.12: Recognition rates for the COIL (left) and Columbia (right) database, for the occlusion level $r/c = 8$ and 100% of degraded views into the training set. Results are reported as a function of the percentage of degraded views in the test set, for SG-MRF, $\chi^2$ and NNC.

show any robustness as the amount of degradation in the test set increases. Furthermore, the recognition performance for 0% occluded view in the test set decreases considerably with respect to the case when there are no occluded views in the training set (Figures 4.10-4.11, dotted lines). SG-MRF and $\chi^2$ both show robustness properties; for the Columbia database, again SG-MRF performs better than $\chi^2$, but loses robustness with respect to the case of 20 objects (COIL database, Figure 4.11, left). Figure 4.12 shows recognition results for 100% of occluded views in the
training set. For the COIL database, SG-MRF and $\chi^2$ have a similar performance and both show robustness; NNC yields the worst performance. For the Columbia database, all three classifiers perform better as the amount of occluded views in the test set increases. This indicates that none of them is generalizing, as the performance increases as the test set becomes more and more similar to the training set. In spite of this general trend, SG-MRF is the most robust and best performing classifier. Particularly, for 0% of occluded views in the test set, it achieves a recognition rate of 81.02%, to be compared with the 68.22% of $\chi^2$ and 32.73% of NNC. Results for level of occlusion $c/r = 16$ are analogous to those discussed above, thus we skip their discussion here. The interested reader can find them in Appendix C.1.

Results for $c/r = (32, 32)$ Figure 4.13 shows recognition results for 0% of occluded views in the training set. We can identify two behaviors: (1) SG-MRF performs better than $\chi^2$ and NNC as the amount of occluded views in the test set increases; (2) robustness decreases dramatically for all classifiers as the number of objects grow. Figure 4.14 shows recognition results for 50% of occluded views in the training set. For COIL database, we see that robustness improves both for SG-MRF and $\chi^2$ with respect to the case of 0% occluded training views. Still, SG-MRF performs significantly better. NNC doesn’t show robustness properties and performs slightly worse than the other two classifiers. Still we see that dealing with a large number of objects (20 for COIL, 100 for Columbia database) leads to a loss in performance. Figure 4.15 shows results for 100% of occluded views in the training set. NNC and $\chi^2$ don’t show robustness for both databases, whereas SG-MRF presents a high robustness for COIL, and loses it for Columbia database, showing once again the effect of increasing the amount of objects to be classified. Results with level of occlusion $r/c = (64, 64)$ are analogous to these, thus they are reported just in Appendix C.1.

Second Kind of Occlusion

In the second series of the occlusion experiments we reduced gradually the amount of visible portion of the object, while keeping it centered within the image. This corresponds to the ideal case in which the location of the object is known. Let $o$ be the visible object portion; we computed the recognition rates for

$$o \in \{20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%\};$$

the resulting object portions. Views were represented using 3D-histograms, with Gaussian derivatives $D_x D_y \text{Lap}, \sigma = 1.0$, resolution per histogram axis 16 bins. Kernel parameters were $a = 0.5$ and $b = 2$. We restrict here the representation of the results to those achieved with SG-MRF. Figure 4.17 show classification results of occluded images with regard to the visible portion of the object on the COIL and Columbia databases. Once again, we see that adding degraded views into the
4.4. ROBUSTNESS OF SG-MRFS: EXPERIMENTS

Figure 4.13: Recognition rates for the COIL (left) and Columbia (right) database, for the occlusion level \( r/c = 32 \) and 0% of degraded views into the training set. Results are reported as a function of the percentage of degraded views in the test set, for SG-MRF, \( \chi^2 \) and NNC.

Figure 4.14: Recognition rates for the COIL (left) and Columbia (right) database, for the occlusion level \( r/c = 32 \) and 50% of degraded views into the training set. Results are reported as a function of the percentage of degraded views in the test set, for SG-MRF, \( \chi^2 \) and NNC.

training set leads to a considerable increase in the recognition stage. We observe also that the higher the number of objects to be recognized, the sooner the performance decreases (with respect to the amount of degradation).

Discussion

For the first experiment on occlusion, three general remarks can be done: first, with respect to NNC and \( \chi^2 \), SG-MRF has definitively the best performance and
Figure 4.15: Recognition rates for the COIL (left) and Columbia (right) database, for the occlusion level $r/c = 32$ and 100% of degraded views into the training set. Results are reported as a function of the percentage of degraded views in the test set, for SG-MRF, $\chi^2$ and NNC.

Figure 4.16: Examples of the second kind of occlusion; from left to right: visible object portions $o \in \{80\%, 70\%, 60\%, 50\%, 40\%, 30\%, 20\%\}$

obtains more stable results than both approaches. This is particularly evident for the two first levels of occlusion and with respect to all perturbations introduced in the training set. Second, we observe for the occlusion experiments almost the same tendency as for the noise experiments: for small amount of occlusion, adding images into the training set has the effect to make more stable the recognition results. For higher amounts of degradation all approaches lose their stability and we observe an increase in the recognition rates just when the similarity between test and training views is higher. Third, robustness achieved against occlusion is higher than robustness against noise (for almost all percentages of occlusion in the training set, the recognition rates for the two first levels remain stable for increased number of corrupted views in the test set). This confirms the theoretical expectations that SG-MRF is more robust to this kind of local degradation than to global degradation such as Gaussian noise (Section 4.2, (Caputo et al, 2002a)).

Results achieved on the Columbia database show almost the same behavior, but not always with the same stability and not at a very high recognition rate, as those reached on the COIL database. For the same arguments reported in Section 4.4.1, this could indicate the observation of a phase transition, which is due to the high
4.4. ROBUSTNESS OF SG-MRFS: EXPERIMENTS

Figure 4.17: Recognition rates for the COIL (left) and Columbia (right) database as a function of the portion of visible object in the test set with respect to the different percentages of occlusion in the training data. The test set consists of exclusively occluded images.

number of stored prototypes with respect to the amount of introduced degradation. Concerning the second kind of occlusion, we notice that, when the model is trained on ideal images, the recognition rate gets worse, the smaller is the visible object portions. Mixing the prototypes with portions of the objects leads to a spectacular stability even with just 40% of visible object portion; we obtain a recognition rate above 98%. This stability is especially observed when the training set includes 50% of occluded images. For this quantity of degradation in the prototypes, there is a remarkably improvement in the recognition rates with respect to the portion of visible object, when the test set contains just occluded images (Figure 4.17). The diagrams in Figure 4.17 confirm the theoretical expectation for the behavior of SG-MRF against occlusion not only when the pixels are occluded, but also when they are cut out of the image.

4.4.3 Robustness to Decreasing Training Set

In this Section we investigate robustness of SG-MRF with respect to a decreasing number of prototypes. We were interested in determining the smallest number of prototypes, for which the performance of SG-MRF degrades. The number of training views was uniformly step-wise reduced from 36 views per object for COIL database to 4 views per object; the remaining views in the database (which are 36 to 68 views per object) were used as test set. Views were represented using 2D-histograms, Gaussian derivatives $D_x D_y$, $\sigma = 1.0$, resolution per histogram axis of 16 bins. As kernel parameters we chose $a = 0.5$, $b = 1$ and the $\rho$ was recomputed for each training set. The obtained results for SG-MRF, $\chi^2$ and NNC are shown in Figure 4.18. We see that SG-MRF obtains a recognition rate of $\sim 82\%$, with
a small number of prototypes. Actually, up to 8 prototypes per object, we have a recognition rate above 98%. SG-MRF and $\chi^2$ perform similarly, showing a good generalization capability. Once again, NNC obtains the worst performance.

![Figure 4.18: Robustness to decreasing training set; experimental results for COIL database, for SG-MRF, $\chi^2$ and NNC.](image)

These results can be improved by changing the combination of local characteristics for the histogram representation and also by learning the kernel parameters. The best result reached with the COIL database is obtained with the generalized kernel $a = 1, b = 1$ and 4D-histograms (Gaussian derivatives $D_y^L, \sigma_1 = 3.5, \sigma_2 = 7.0$, resolution per histogram axis of 16 bins) as feature representation. With this representation, it is possible to obtain, with just 4 views per object, a recognition rate of 95.52%, to be compared with the performance obtained using the 2D representation (83.75% see Figure 4.19).

### 4.4.4 Robustness to Heterogeneous Background

Many experiments presented in literature on appearance-based object recognition considers objects as located in a homogeneous, contrasting background ((Belongie et al, 2001; Hornegger et al, 1995) and many others). The main motivation beyond this assumption is that, if we want to model the appearance of an object, we need information regarding that object, and that object only. On the other side, in our every day experience we meet objects in several different backgrounds; it is thus essential for every object recognition system to be robust with respect to every kind of background. Although heterogeneous background can be seen as a degradation of a given view just in a loose sense, the theoretical and experimental results shown
in previous Sections induce us to investigate robustness of SG-MRF with respect to heterogeneous background, using the same procedure employed for noise and occlusions.

We ran experiments on the COIL and Columbia database, and tested how the robustness of SG-MRF increases when we add in the training set views containing heterogeneous background. In both experiments, we used a MFH representation.

### 4.4.4.1 COIL/Columbia Experiments

In this set of experiments, objects from the COIL and Columbia databases were cut from the original images and put on images with heterogeneous background, having a resolution of $128 \times 128$. We chose 4 different backgrounds shown in Figure 4.20: one is homogeneous and the others are heterogeneous representing three different scenes. As a referring point, we performed a set of experiments training the system with views on homogeneous background. In a second step, we mixed the training data with views having heterogeneous background: 25% of each background was present in the training set; the distribution of the mixture of backgrounds was uniform. In both experiments, we performed the testing on each background separately. Features were extracted in all these experiments using 4D-MFH, with $D_x D_y$ Gaussian derivatives, $\sigma_1 = 3.5$ and $\sigma_2 = 7.0$; kernel parameters were set to $a = 1, b = 1$. The training set contained 36 prototypes per object and the test set consisted of the whole database. Classification results achieved with the homogeneous training set are reported in Table 4.1. The results obtained with the mixed training set are reported in Table 4.2 for both databases.

The recognition results reported in Table 4.1 show that when the system is trained on ideal images, there is a dramatic decrease in the recognition rate when the objects are put on heterogenous background. By mixing the training set with
Figure 4.20: Four different backgrounds involved in the localization experiments and their labeling: hbg0 is homogeneous, hbg1, hbg2 and hbg3 represent three different scenes.

<table>
<thead>
<tr>
<th></th>
<th>hbg0</th>
<th>hbg1</th>
<th>hbg2</th>
<th>hbg3</th>
</tr>
</thead>
<tbody>
<tr>
<td>COIL20</td>
<td>100</td>
<td>85.00</td>
<td>87.36</td>
<td>87.29</td>
</tr>
<tr>
<td>COIL100</td>
<td>100</td>
<td>59.00</td>
<td>60.75</td>
<td>66.65</td>
</tr>
</tbody>
</table>

Table 4.1: Recognition results on the COIL20 and the COIL100 databases, when the train set contains objects with homogeneous background

<table>
<thead>
<tr>
<th></th>
<th>hbg0</th>
<th>hbg1</th>
<th>hbg2</th>
<th>hbg3</th>
</tr>
</thead>
<tbody>
<tr>
<td>COIL20</td>
<td>99.93</td>
<td>99.93</td>
<td>99.93</td>
<td>99.93</td>
</tr>
<tr>
<td>COIL100</td>
<td>99.53</td>
<td>99.37</td>
<td>99.16</td>
<td>99.15</td>
</tr>
</tbody>
</table>

Table 4.2: Recognition results on the COIL20 and the COIL100 databases, when the train set contains objects with a mixture of homogenous and heterogenous backgrounds

these backgrounds we achieved stability and robustness in the recognition rates (see Table 4.2) for both databases. This confirms once again the theoretical expectations on the robustness of SG-MRF.

4.5 Summary

In this Chapter we extended the concept of robustness, often used in literature as the capability of a system to correctly recognize corrupted images from ideal views, to include also the capability to generalize from degraded patterns. This means that we assumed that the training set can also contain degraded samples and we studied the robustness of SG-MRF to this kind of perturbations in the data. We presented a theoretical analysis and a broad set of experiments which show that SG-MRF are robust with respect to noise, occlusion and heterogeneous background, when these degradations are contained in the training and in the test set. Benchmark evaluation with other classification methods shows the higher robustness of SG-MRF.
Chapter 5

Ultrametric Spin Glass -Markov Random Fields

This Chapter introduces a new Spin Glass-like energy function, and develops the corresponding Spin Glass-Markov Random Field model. This energy consists of a hierarchical organization of Hopfield energies, and makes it possible to store non mutually orthogonal prototypes. It was originally designed on the basis of the ultrametric structure of the Spin Glass configuration space (Amit, 1989). Thus, we will call Ultrametric Spin Glass-Markov Random Field the corresponding model. After the theoretical derivation of the model (Section 5.2), we apply it to hierarchical object identification (Section 5.3), multi-cue integration (Section 5.4) and object identification via probabilistic scene modeling (Section 5.5). Results obtained in all these cases motivates the introduction of this new model.
5.1 Introduction

In previous Chapters we defined Spin Glass-Markov Random Fields, via the Hopfield energy function, and we applied it to appearance-based object recognition. We presented theoretical and experimental results that show how SG-MRF are effective for object identification (Chapter 3), and we explored the robustness of the model, obtaining remarkable results (Chapter 4). As pointed out in Chapter 3, the same strategy used for integrating the Hopfield energy with MRF modeling can be used for any other SG-like energy function, provided that it can be kernelized. This Chapter illustrates this concept introducing a new SG-like energy function, and developing the corresponding SG-MRF model. One of the main characteristics of this model is that it allows us to store non mutually orthogonal prototypes; it basically consists in a hierarchical organization of Hopfield energies. As this energy was originally derived taking into account the ultrametric properties of the SG configuration space, we will refer to it as the ultrametric energy, and we will call the deriving model Ultrametric Spin Glass-Markov Random Field (USG-MRF). From a vision point of view, the two key properties of the model are: (1) it allows to store non-orthogonal prototypes, (2) it induces a hierarchical organization of the data, (3) it allows to use different kernels at different levels of the hierarchy. These three properties will make it possible to use USG-MRFs for several visual applications, showing remarkable performances.

The rest of the Chapter is organized as follows: Section 5.2 will present USG-MRF: we will introduce the ultrametric energy and we will show how it can be kernelized (Section 5.2.1); we will derive the ultrametric classifier to be used for visual applications (Section 5.2.2) and we will describe the learning procedure for the algorithm (Section 5.2.3). The rest of the Chapter describes three different visual applications in which USG-MRF shows its effectiveness. Section 5.3 will present how USG-MRF can be successfully applied to hierarchical appearance-based object recognition (Caputo et al, 2002f); Section 5.4 will discuss how to combine effectively color and shape information for appearance-based object recognition (Caputo et al, 2002g), and how USG-MRF can be employed successfully for such a task. In Section 5.5 we will present a probabilistic model of a scene that performs appearance-based object recognition using the contextual information of the scene; such a model is based on USG-MRF. The Chapter concludes with a summary discussion.

5.2 Ultrametric Spin Glass-Markov Random Field

5.2.1 The Ultrametric Energy

Consider the energy function

\[ E = - \sum_{(i,j)=1}^{N} J_{ij} s_i s_j \]  

(5.1)
5.2. ULTRAMETRIC SPIN GLASS-MARKOV RANDOM FIELD

with the following connection matrix:

\[ J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_{i}^{(\mu)} \xi_{j}^{(\mu)} + \frac{1}{N(1-a_{\mu}^2)} \sum_{\mu=1}^{p} \xi_{i}^{(\mu)} \xi_{j}^{(\mu)} \sum_{\nu=1}^{q_{\mu}} (\eta_{i}^{(\mu\nu)} - a_{\mu})(\eta_{j}^{(\mu\nu)} - a_{\mu}) \]  

(5.2)

with

\[ \xi_{i}^{(\mu\nu)} = \xi_{i}^{(\mu)} \eta_{i}^{(\mu\nu)}, \quad a_{\mu}^2 = \frac{1}{N} \sum_{i=1}^{N} \eta_{i}^{(\mu\nu)} \eta_{i}^{(\mu\lambda)} . \]

This energy induces a hierarchical organization of stored prototypes (\textit{Amit}, 1989). The set of prototypes \( \{ \xi_{i}^{(\mu)} \}_{\mu=1}^{P} \) are stored at the first level of the hierarchy and are usually called the ancestors. Each of them will have \( q \) descendants \( \{ \xi_{i}^{(\mu\nu)} \}_{\nu=1}^{q_{\mu}} \). The parameter \( \eta_{i}^{(\mu\nu)} \) measures the similarity between ancestors and descendants; the parameter \( a_{\mu} \) measures the similarity between descendants. \( \Delta(a_{\mu}) \) is a normalizing parameter, that guarantees that the energy per site is finite. In the rest of the Chapter we will limit the discussion to the case

\[ a_{\mu}^2 = a^2 . \]

The connection matrix thus becomes:

\[ J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_{i}^{(\mu)} \xi_{j}^{(\mu)} + \frac{1}{N(1-a^2)} \sum_{\mu=1}^{p} \xi_{i}^{(\mu)} \xi_{j}^{(\mu)} \sum_{\nu=1}^{q_{\mu}} (\eta_{i}^{(\mu\nu)} - a)(\eta_{j}^{(\mu\nu)} - a) \]

\[ = \frac{1}{N} \sum_{\mu=1}^{p} \xi_{i}^{(\mu)} \xi_{j}^{(\mu)} + \frac{1}{N(1-a^2)} \sum_{\mu=1}^{p} \xi_{i}^{(\mu)} \xi_{j}^{(\mu)} \sum_{\nu=1}^{q_{\mu}} (\eta_{i}^{(\mu\nu)} - a)(\eta_{j}^{(\mu\nu)} - a) \]

\[ T_1 \quad T_2 \]

\( T_1 \) is the Hopfield connection matrix (2.6); \( T_2 \) is a new term that allows us to store as prototypes patterns correlated with the \( \{ \xi_{i}^{(\mu)} \}_{\mu=1}^{P} \), and correlated between each other. The energy resulting by using the connection matrix (5.2) will have \( p + \sum_{\mu=1}^{p} q_{\mu} \) minims, of which \( p \) absolute (ancestor level) and \( (\sum_{\mu=1}^{p} q_{\mu}) \) local (descendant level). When \( a \rightarrow 0 \), the ultrametric energy reduces to a hierarchical organization of Hopfield energies; in this case the prototypes at each level of the hierarchy must be mutually orthogonal, but they can be correlated between different levels. Note also that we limited ourselves to two levels, but the energy can be easily extended to three or more. For a complete discussion on the properties of this energy, we refer the reader to (\textit{Amit}, 1989). Here we are interested in using this energy in the SG-MRF framework. To this purpose, we show that the energy (2.4),

\(^{1}\text{Considering the general case would not add anything from the conceptual point of view and would make the notation even heavier.}\)
with the connection matrix (5.2), can be written as a function of scalar product between configurations:

\[ E = -\sum_{(i,j)=1}^{N} \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^{(\mu)} \xi_j^{(\mu)} s_i s_j - \]

\[ -\frac{1}{N(1-a^2)} \cdot \sum_{(i,j)=1}^{N} \sum_{\mu=1}^{p} \sum_{\nu=1}^{q_\mu} (\eta_i^{(\mu\nu)} - a)(\eta_j^{(\mu\nu)} - a) s_i s_j = \]

\[ = -\frac{1}{N} \sum_{\mu=1}^{p} (\xi^{(\mu)} \cdot s)^2 - \frac{1}{N(1-a^2)} \sum_{\mu=1}^{p} \sum_{\nu=1}^{q_\mu} [a^2 (\xi^{(\mu)} \cdot s)^2 + \]

\[ \sum_{(i,j)=1}^{N} s_i s_j \xi_i^{(\mu)} \xi_j^{(\mu)} \eta_i^{(\mu\nu)} \eta_j^{(\mu\nu)} - a \sum_{(i,j)=1}^{N} s_i s_j \xi_i^{(\mu)} \xi_j^{(\mu\nu)} \eta_i^{(\mu\nu)} + \eta_j^{(\mu\nu)}] = \]

\[ = -\frac{1}{N} \sum_{\mu=1}^{p} (\xi^{(\mu)} \cdot s)^2 - \frac{1}{N(1-a^2)} \sum_{\mu=1}^{p} \sum_{\nu=1}^{q_\mu} [a^2 (\xi^{(\mu)} \cdot s)^2 + \]

\[ + \sum_{(i,j)=1}^{N} s_i s_j \xi_i^{(\mu\nu)} \xi_j^{(\mu\nu)} - a \sum_{(i,j)=1}^{N} [s_i s_j \xi_i^{(\mu\nu)} \xi_j^{(\mu\nu)} + s_i s_j \xi_i^{(\mu\nu)} \xi_j^{(\mu\nu)}] = \]

\[ E = -\frac{1}{N} \sum_{\mu=1}^{p} (\xi^{(\mu)} \cdot s)^2 - \frac{1}{N(1-a^2)} \sum_{\mu=1}^{p} \sum_{\nu=1}^{q_\mu} (\xi^{(\mu\nu)} \cdot s)^2 + \]

\[ \frac{2a}{N(1-a^2)} \sum_{\mu=1}^{p} \sum_{\nu=1}^{q_\mu} (\xi^{(\mu)} \cdot s)(\xi^{(\mu\nu)} \cdot s) - \frac{a^2}{N(1-a^2)} \sum_{\mu=1}^{p} \sum_{\nu=1}^{q_\mu} (\xi^{(\mu)} \cdot s)^2. \quad (5.3) \]

If we assume that \( a \to 0 \), as to say we impose orthogonality between prototypes at each level of the hierarchy, the energy reduces to

\[ E = -\frac{1}{N^2} \sum_{\mu=1}^{p} (\xi^{(\mu)} \cdot s)^2 + \sum_{\mu=1}^{p} \sum_{\nu=1}^{q_\mu} (\xi^{(\mu\nu)} \cdot s)^2. \quad (5.4) \]

The ultrametric energy, in the general form (5.3) or in the simplified form (5.4) can be kernelized as done for the Hopfield energy and thus can be used in a MRF framework. We call the resulting new MRF model Ultrametric Spin Glass-Markov
5.2. ULTRAMETRIC SPIN GLASS-MARKOV RANDOM FIELD

Random Fields (USG-MRF). In the SG-MRF framework the ultrametric energy becomes:

\[ E_{USG} = - \sum_{\mu=1}^{p_k} [K_a(\bar{x}^{(\mu)}, x)]^2 - \sum_{\mu=1}^{p_k} \sum_{\nu=1}^{q_\mu} [K_d(\bar{x}^{(\mu\nu)}, x)]^2, \]

(5.5)

where \( \{x^{(\mu)}\}_{\mu=1}^{p_k} \) will be the set of prototypes relative to the ancestor level, and \( \{x^{(\mu\nu)}\}_{\nu=1}^{q_\mu}, \mu = 1, \ldots, p_k \) the set of prototypes at the descendant level. \( K_a \) is the generalized Gaussian kernel at the ancestor level, and \( K_d \) is the generalized Gaussian kernel at the descendant level.

There are two main points to be underlined:

1. **Within each level, prototypes must satisfy the orthogonality constraint; between each level, prototypes have no constraints to be satisfied.**

   This constitutes an important improvement of the ultrametric energy with respect to the Hopfeld one. As discussed in Chapter 3, the condition of orthogonality on the set of prototypes for the Hopfield energy can be an issue. This condition is a legacy from the theoretical derivation of the Hopfield energy, and for some particular applications can become a problem. Consider for instance the task of classifying an object using multiple cues, such as shape and color, or using the contextual information relative to the particular scene where the object is located. In these examples, we can expect to have set of training examples highly correlated between each other; imposing the orthogonality condition on the prototypes, with the strategy illustrated in Section 3.7, can lead to poor results. The hierarchical structure of USG-MRF leads to more relaxed constrains on the set of prototypes and constitutes the natural framework for these and other vision applications. This point will be illustrated more effectively in next Sections, on three different visual applications.

2. **The kernel must be the same at each level of the hierarchy, but can be different between levels (as to say between ancestor and descendants).**

   We have shown in Section 3.9 that the classification performance of SG-MRF varies as the kernel parameters \((a, b)\) do. This is true for every kernel method: the algorithm performance will strongly depend from the choice of the kernel function. We already pointed out more than once in this thesis that, although SG-MRFs are restricted to some specific kernel families (such as Gaussian kernels), the model is still rich enough to capture the significant information for recognition purposes. Thus, what is in principle a limitation turns out to be an advantage (Schölkopf et al., 2002; Barla et al., 2002). But the ultrametric energy gives us the extra power to use a different kernel at each level of the hierarchy. We will show in the rest of this Chapter that this feature of the model will lead to remarkable performances for several visual recognition applications.
5.2.2 Ultrametric Bayes Classifier

The ultrametric energy (5.5) derived in the previous Section can be used in the SG-MRF framework for modeling the probability distribution of the considered visual classes. Thus, we can derive for USG-MRF a Bayes classifier as made in Section 3.6 for SG-MRF. As the main interest of this thesis is to develop techniques for object recognition, we will illustrate the USG-MRF-MAP classifier with respect to this application; of course the model can be applied to any pattern recognition problem.

Consider the probabilistic appearance-based framework described in Section 2.2.2. Given a view \( x_n^k \), we will suppose to extract two or more feature vectors from it: for instance \( z(1)_n^k = T_1(x_n^k) \) will be the feature vector extracted from the view \( x_n^k \) using the feature extractor \( T_1 \); \( z(2)_n^k = T_2(x_n^k) \) will be the feature vector extracted from the view \( x_n^k \) using the feature extractor \( T_2 \); and so on. The feature extractors \( T_1, T_2, T_3, \ldots \) will be different between each other and thus will produce different feature vectors: for example, \( T_1 \) can extract shape information, \( T_2 \) color and \( T_3 \) scale information; or \( T_1 \) can extract contextual information and \( T_2 \) object information; and so on. If we use \( g \) feature extractors, we will end up with \( g \) training sets, derived from the original one:

\[
\begin{align*}
T_1\{x_1^k \ldots x_n^k \} &\rightarrow \{z(1)_1^k \ldots z(1)_n^k \} \\
T_2\{x_1^k \ldots x_n^k \} &\rightarrow \{z(2)_1^k \ldots z(2)_n^k \} \\
&\vdots \\
T_g\{x_1^k \ldots x_n^k \} &\rightarrow \{z(g)_1^k \ldots z(g)_n^k \}.
\end{align*}
\]

We can expect to have a strong correlation between feature vectors obtained from the same view using different feature extractors (for example \( z(1)_1^k \) and \( z(2)_1^k \)); but with respect to the feature vectors obtained from the original views using a given feature extractor (for example \( z(1)_1^k \) and \( z(1)_2^k \)), the situation is the same as for SG-MRF.

We will have as many hierarchical levels as the number of feature extractors used. The Bayes classifier using the USG energy on the structure of data described above will be:

\[
k^* = \arg\min_{k=1}^{K} \{ - \sum_{\mu_1=1}^{p_k} [K_1(\tilde{z}(1)_{k}^{\mu_1}, z(1))]^2 - \sum_{\mu_1=1}^{p_k} \sum_{\mu_2=1}^{q_{\mu_1}} [K_2(\tilde{z}(2)_{j}^{\mu_1\mu_2}, z(2))]^2 - \ldots \\
- \sum_{\mu_1=1}^{p_k} \sum_{\mu_2=1}^{q_{\mu_1}} \ldots \sum_{\mu_g=1}^{q_{\mu_{g-1}}} [K_g(\tilde{z}(g)_{j}^{\mu_1\mu_2\ldots \mu_g}, z(g))]^2 \},
\]

where:
1. $K_1(), K_2(), \ldots, K_g()$ are the generalized Gaussian kernels used at each level of the hierarchy. As the index indicates, they are relative to each feature extractor, and can by principle be different for each hierarchical level. We will discuss this point in more details in the next Section.

2. $\{\tilde{z}(1)_{\mu_1}, \ldots, \tilde{z}(1)_{\mu_k}\}$ are the set of prototypes at the ancestor level; $\{\tilde{z}(2)^{\mu_1\mu_2}, \ldots, \tilde{z}(2)_{\mu_k\mu_l}\}$ are the set of prototypes at the first descendant level, and so on. At each level of the hierarchy, the prototypes are selected from the training set via the orthogonality condition, as for SG-MRF. The indexes $\mu_1\mu_2, \mu_1\mu_2\mu_3, \ldots, \mu_1\mu_2\mu_3 \cdots \mu_g$ indicates that between levels, the prototypes are correlated. The parameters controlling the degree of correlation are the $\rho_1, \rho_2, \ldots, \rho_g$, inside the generalized Gaussian kernel. We will discuss how to learn it from the training data in the next Section.

The USG-MRF-MAP classifier (5.6) is especially designed for applications and data that have a hierarchical structure, but can also be used very effectively for vision applications that are not usually formulated in a hierarchical way, such as object recognition with multiple cues; in that case the hierarchical structure will be imposed on the data. We will illustrate this point in Section 5.4.

5.2.3 Learning The Kernel Parameters

The USG-MRF-MAP classifier can be used once the prototypes are selected from the training vectors and the kernel parameters are learned from the training data, at each level of the hierarchy. This is achieved in two steps:

- **Step1: learning within hierarchical levels**
  If we look at each hierarchical level separately, it is equivalent to a SG-MRF-MAP classifier with Hopfield energy. Thus, for each hierarchical level the kernel parameters $a_q, b_q, \rho_q$ are learned and the set of prototypes are selected for each object class as described in Section 3.7. This implies that, for each level of the hierarchy, the kernel can be different.

- **Step2: learning between hierarchical levels**
  As described in Section 3.7, the $\rho_q$ parameter for the kernel $K_q$ determines the orthogonality between prototypes, and its value controls the sparseness of the algorithm with respect to the number of stored prototypes for each object class. Its value thus is not uniquely determined: indeed, the condition that $\rho_q$ must satisfy is (equation 3.10)

  $$\rho_q \gg \Delta_{\text{min}}$$

  This will be true for all the hierarchical levels; we can use this soft constraint on $\rho_q$ for optimizing the performance of the USG-MRF-MAP classifier with respect to the recognition rate. This is done with the same leave-one-out strategy used for learning the kernel parameters $a_q, b_q$ at each hierarchical
level (see Section 3.7). In analogy with \( a_q, b_q \), that must vary within a range of permitted values, the set of kernel variances \( (\rho_1, \rho_2, \ldots, \rho_q) \) will be subjected to constraints regarding their possible values; these constraints are determined, for each hierarchical level, in step 1, via condition (3.10).

Finally, the complexity of the classification algorithm is given by:

\[
\text{comp. complexity} \sim \# \text{ of prototypes} \times \text{comp. of metric} \times \# \text{ of hierarchical levels}
\]

When there are many hierarchical levels, and many classes at each level, the learning stage can become computationally heavy. This learning stage is anyway done off line, thus doesn’t affect the classification time. Moreover it is possible to consider, for a given level of the hierarchy, just the \( n \) best matches, or the matches with energy value beyond a fixed threshold value. This is quite effective for some applications, and allows to keep under control the classification time.

5.3 Hierarchical Appearance-based Object Recognition.

5.3.1 Problem Statement

Most of work on object recognition tries to answer the following question: given a collection of objects, can we recognize correctly one of them among the others? This problem is faced in a wide variety of situations: in cluttered or heterogeneous background (Leonardis et al., 2000; Nelson et al., 1998), under different lighting conditions (Schiele et al., 2000), in presence of noise or occlusion (Schiele et al., 2000; Leonardis et al., 2000; Caputo et al., 2002a), and so on. The success of an algorithm is then measured in terms of recognition rates, as to say how many times the object was recognized successfully. Although the recognition rate is undoubtedly an important indicator of the performance of an algorithm, it cannot be the only one. In other words, not all errors can be considered in the same way. This is something we experience everyday: we know that not all errors have the same consequences, and that there are mistakes that we cannot do more than once in life. If we ask someone “Please give me my pen”, and (s)he makes a mistake, it is not the same if we get a different pen, or a cup. Many times, a pen (even if not ours) will do, but a cup won’t. If we walk in the forest and hear a noise that makes us realize that there is an animal close to us, the point is not to recognize whether it is a lion, a tiger, an antelope or a rabbit, but to decide whether it is dangerous or not, and react consequently.

Many experiments (Brooks, 1981) show that biological systems tackle this problem using a hierarchical organization of visual information, based on visual categories.\(^2\) Coding by category is fundamental to mental life because it greatly reduces

\(^2\)We are aware that some researchers in the computer vision community argue that such parallel are unneeded. Although we understand their motivations, it should be firmly kept in mind that “biological vision is currently the only indication we have that the general vision problem is even open to solution” (Brooks, 1981).
the demands on perceptual processes, storage space, and reasoning processes. This also induces a hierarchical classification system: a visual pattern is recognized first as a member of a visual category, then as a member of a relative sub-category, and so on until it is recognized as individual (if known).

In order to use this strategy in computer vision algorithms for object recognition, we must satisfy two major requirements:

1. We need an effective and robust way to recognize visual categories. Note that when we say robust we intend it in the common way it is intended for object recognition (robustness to noise, occlusion, cluttered background, varying lighting conditions, scale changes, and so on), and also robust with respect to the variations within different members of a same visual category.

2. We need an effective hierarchical organization of the object information.

In this Section we propose to use USG-MRF for hierarchical appearance-based object recognition. From the description of the model made in Section 5.2, it is clear that USG-MRF satisfies the second requirement. With respect to the first requirement, we will not look for a new representation, designed so to capture the visual category information. On the contrary, we will use the same representation for visual categories and for members of the visual categories as well; this information will be treated differently at each level of the hierarchy through the use of different kernels. The experiments we performed gave promising results.

5.3.2 The Ultrametric Approach.

Let us suppose to have a set of objects \( \Omega^h_k, h = 1, \ldots H, k = 1, \ldots K \), each represented by a training set \( \omega^h_k \). Suppose also that the set of \( K \) objects can be divided in \( K \) subset of visual object categories \( C_h \):

\[
C_1 = \{ \Omega^1_k \}_{k=1}^K \\
C_2 = \{ \Omega^2_k \}_{k=1}^K \\
\vdots \\
C_H = \{ \Omega^H_k \}_{k=1}^K \\
C_1 \cup C_2 \cup \ldots \cup C_H = \{ \Omega^h_k \}_{k=1}^H, h = 1, \ldots H 
\]

Our goal is to classify a given data sample \( \mathbf{x} \) as belonging to one of the \( \Omega^h_k \) objects and to one of the \( C_h \) visual categories. To this purpose, we will apply the USG-MRF-MAP classifier (5.5) as follows:

- **Step 1: representation**
  We will represent the training views of each object via a robust feature representation; in this thesis we decided not to concentrate the attention on
developing new image descriptors, thus we will assume to have already available a “good” representation, where with this name we intend any of the many representations proposed in literature for robust object recognition ((Schiele et al, 2000) and many others).

• **Step 2: partition of the training set**
  From the original training set, consisting of \( M \) sets of views (now represented by the extracted features), we will build two training sets, one for the ancestor level and one for the descendant level. The training set for the ancestor level will consist of \( \{ C_h \}_{h=1}^{t} \) classes (the visual categories), each represented by a set of training views containing all the training views relative to the objects \( \Omega^h \) belonging to that specific visual category. The training set for the descendant level will consist of the original set of objects \( \Omega^h \), with their relative training set \( \omega^h_k \). Note that the total set of views on which the training is done at each level are the same. What changes at each level of the hierarchy is the partition of the data on which the training will be done.

• **Step 3: learning of kernels parameters**
  We will apply the USG-MRF-MAP classifier on the data prepared as described above; we will learn the kernel parameters and we will select the prototypes at each level, as described in Section 5.2.3. Note that, although the total number of views in the training set is the same at each level, the kernel parameters and the selected prototypes will be, in general, different at each level.

A main feature of this approach is that it does not require to develop a new representation for visual categories, but allows to benefit from the properties of representations developed for object recognition. It will be the specific kernel at each level to give more weight to the information of interest contained in the representation vector, for the task relative to that hierarchical level \(^3\). This means that we try to tackle the problem of how to recognize visual categories not developing new representations, but learning new ways to combine the existing information.

The USG-MRF-MAP classifier will have two levels, and will be given by:

\[
k^* = \arg\min_{k=1}^{K} \left\{ -\sum_{\mu=1}^{p_k} \left[ K_C(\hat{h}^{(\mu)}_C, h) \right]^2 - \sum_{\nu=1}^{q_k} \sum_{\mu=1}^{p_k} \left[ K_{\Omega}(\hat{h}^{(\mu\nu)}_{\Omega}, h) \right]^2 \right\},
\]

where \( K_C, K_{\Omega} \) are the kernel learned for the visual category and object level, \( \{ \hat{h}^{(\mu)}_C \} \) are the set of prototypes for the visual categories level, and \( \{ \hat{h}^{(\mu\nu)}_{\Omega} \} \) are the set of prototypes for the object classes. Kernels and set of prototypes are all learned from the training data.

\(^3\)The implicit assumption here is that each representation vector contains information relative to the specific object and to the relative visual category. This is a very mild assumption to do: for example, a picture of my dog contains information of it as a specific dog and as a member of the visual category “dog”.

\[
\text{(5.7)}
\]
5.3. HIERARCHICAL APPEARANCE-BASED OBJECT RECOGNITION.

5.3.3 Experiments

We performed a series of experiments in order to test our model. We ran all the experiments on the NELSON database (Nelson et al., 1998), see Section 3.9.2, that presents a hierarchical structure. As described in Section 3.9.2, the database is composed by 11 cups, 5 dolls, 6 planes, 6 fighter jets, 9 lizards, 5 spoons, 8 snakes and 9 sport cars. Some examples are shown in Figure 5.1. We chose MFH as representation, both for its robustness properties (emphasized in combination with SG-MRF, see Chapter 4), both because views into the Nelson database have different sizes. We used three different kinds of MFH representation, selected heuristically so to capture an increasing amount of information: the first consisted in Gaussian derivatives along $x$ and $y$ directions and with $\sigma = 1.0 (D_xD_y)$. The second consisted in Laplacian Gaussian operator with $\sigma_1 = 1.6$ and $\sigma_2 = 3.2 (Lp2\sigma)$. The third consisted in Laplacian Gaussian operator with $\sigma_1 = 1.6$, $\sigma_2 = 3.2$ and $\sigma_3 = 6.4 (Lp3\sigma)$. For all these representations, the resolution for histogram axes was of 16 bins.

The structure of this database is “naturally” hierarchical: although it is composed by 59 objects, they can be naturally separated in subgroups with respect to the visual category they belong to (cup, doll, fighter, lizard, spoon, sport car). The goal of these experiments is to check whether the use of the USG-MRF hierarchical structure can lead to a) a higher recognition rate, b) a lower recognition time, c) a lower category error rate.

Thus, we ran a first set of experiments on the complete database of 59 objects, using SG-MRF in a MAP-Bayes classifier, for all the three representations described above. These results constitute a benchmark for those obtained using USG-MRF, and are reported in Table 5.1.

We see that the best recognition rate, obtained with the $Lp3\sigma$ representation, does not correspond to the lower category error. Moreover, the higher recognition rate is obtained with a 3D histogram representation, which leads to a remarkable increase of the computational time.

Then we ran a second set of experiments, consisting in the recognition of the
8 visual categories mentioned above. For each visual category, the training (test) set consisted of all the training (test) views of the object belonging to the category itself. We used a $D_xD_y$ and $Lp2\sigma$ representation, and SG-MRF as described in the previous experiment. Results are reported in Table 5.2.

<table>
<thead>
<tr>
<th></th>
<th>$D_xD_y$</th>
<th>$Lp2\sigma$</th>
<th>$Lp3\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rec. Rate (%)</td>
<td>90.82</td>
<td>97.74</td>
<td><strong>98.23</strong></td>
</tr>
<tr>
<td>Rec. time (sec)</td>
<td><strong>0.67</strong></td>
<td>1.35</td>
<td>6.77</td>
</tr>
<tr>
<td>n. misclass.</td>
<td>217</td>
<td>50</td>
<td><strong>39</strong></td>
</tr>
<tr>
<td>category errors</td>
<td>42</td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 5.1: Classification results using SG-MRF, for 59 objects

We see that, using the $Lp2\sigma$ representation, we achieve a recognition rate of 100%, which of course corresponds in this case to 0 category errors. This inspired us to perform two hierarchical experiments. In the first experiment, we used the USG-MRF classifier as described in Section 5.2.2, and the $D_xD_y$ representation at the ancestor level (visual category level) and the $Lp3\sigma$ at the descendant level (object class level). Results are reported in Table 5.3. In the second experiment, we used SG-MRF and $Lp2\sigma$ for the visual category classification; we considered this result prior knowledge and then we used again SG-MRF for the recognition of each object class. This procedure allows to use, for each group of objects belonging to the same visual category, a different kernel. Results are reported in Table 5.3.

<table>
<thead>
<tr>
<th></th>
<th>$D_xD_y$</th>
<th>$Lp2\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rec. Rate (%)</td>
<td>98.23</td>
<td><strong>100</strong></td>
</tr>
<tr>
<td>category errors</td>
<td>39</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5.2: Classification results using SG-MRF, for 8 object categories

<table>
<thead>
<tr>
<th></th>
<th>$D_xD_y$-Lp3$\sigma$</th>
<th>$Lp2\sigma$- Lp3$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rec. Rate (%)</td>
<td>98.28</td>
<td><strong>98.59</strong></td>
</tr>
<tr>
<td>Rec. time (sec)</td>
<td><strong>3.18</strong></td>
<td>4.94</td>
</tr>
<tr>
<td>n. misclass.</td>
<td>38</td>
<td><strong>30</strong></td>
</tr>
<tr>
<td>category errors</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5.3: Classification results using USG-MRF, for 59 objects

The best recognition rate is obtained with the second experiment, which gives also (not surprisingly) the lowest category error. Nevertheless, results obtained
5.4. COMBINING SHAPE AND COLOR INFORMATION FOR OBJECT RECOGNITION.

Using USG-MRF are impressive: it is faster with respect to the second experiment, with just 8 more misclassifications, 2 of which category errors. This last result is particularly impressive, because we used the $D_x D_y$ representation at the ancestor level, which would give alone 39 category errors (see Table 5.2). This, with the awareness that the second experiment relies heavily on the performance at the first level (that gives a 100% recognition rate in this case, but that by principle can be different for different databases), makes us conclude that USG-MRF is the most promising strategy for hierarchical appearance-based object recognition.

Finally, we compare USG-MRF results with those obtained with the first set of experiments. We see from Table 5.1 and Table 5.3 that USG-MRF gives the highest recognition rate and the lowest category error; it pays a price in terms of computational time with respect to the $L_p 2\sigma$ representation (second column, Table 5.1), but this is well compensated by the higher recognition rate and the lower category error. We can conclude from these experiments that USG-MRF is an effective probabilistic method for hierarchical appearance-based object recognition.

5.4 Combining Shape and Color Information for Object Recognition.

5.4.1 Problem Statement

Although it is generally acknowledged that both color and texture/shape information are important for object recognition (Mel, 1997; Slater et al, 1995), few systems employ both. This is because no single representation is suitable for both types of information. Traditionally, the solution proposed in literature consists of building up a new representation, containing both color and texture/shape information (Mel, 1997; Slater et al, 1995; Matas et al, 1995). Systems using this kind of approach show very good performances (Mel, 1997; Slater et al, 1995; Matas et al, 1995). This strategy solves the problems related to the common representation; a major drawback is that the introduction of a new representation does not permit the use of just color or just textural/geometrical information alone, depending on the task considered. A huge literature shows that color only, or texture/shape only representations work very well for many applications (see for instance (Jain et al, 1993; Leonardis et al, 2000; Schiele et al, 2000; Ballard, 1981)). Thus, the goal should be a system that uses both forms of information while keeping them distinct, allowing the flexibility to use the information sometimes combined, sometimes separate, depending on the application considered.

Another important point is the dimension of the feature vector relative to the new representation. If it carries as much information about color and texture/shape as separate representations do, then we must expect the novel representation to have more parameters than each separate representation alone, with all the risks of a curse of dimensionality effect. If the dimension of the new representation vector
is kept under control, this means that the representation contains less color and texture/shape information than single representations.

In this Section we propose a new strategy to this problem. Given a texture/shape only and color only representation, we focus the attention on how they can be combined together as they are, rather than define a new representation. We will achieve this result using USG-MRF, and treating the information as if it would have a hierarchical structure. We will show that the USG-MRF model provides a natural framework for combining texture/shape and color representation together, without any need to define a new representation. The advantages of this approach are:

- It permits us to use existing and well tested representations both for texture/shape and color information;
- It permits us to use this knowledge in a flexible manner, depending on the task considered.

### 5.4.2 The Ultrametric Approach

Assume we have $K$ different classes $\Omega_1, \Omega_2, \ldots, \Omega_K$ of objects, and that for each object is given a set of $n_k$ data samples, $\omega_k = \{x_1^k, x_2^k, \ldots, x_{n_k}^k\}, k = 1, \ldots, K$. How the object class $\Omega_k$ is represented, given a set of data samples $\omega_k$ (relative to that object class), varies for different appearance-based approaches: it can consider texture/shape information only, or color information only, or both. This is equivalent to consider a set of features $\{h_1^k, h_2^k, \ldots, h_{n_k}^k\}, k = 1, \ldots, K$, where each feature vector $h_{n_k}^k$ is computed from the image $x_{n_k}^k, h_{n_k}^k = T(x_{n_k}^k)$. Assuming that the data sample $\omega_k$ are a sufficient statistic for the pattern class $\Omega_k$, the goal will be, given a test image $x$ and its associate feature vector $h$, to classify it as belonging to one of the object classes $\Omega_k$. To this purpose, we will use the USG-MRF-MAP classifier as follows: from each data sample in $\omega_k$ we will extract two feature vectors, $h(s)_{n_k}$ containing texture/shape information and $h(c)_{n_k}$ containing color information. We will consider $\{h(c)_1, h(c)_2, \ldots, h(c)_{n_k}\}$ as the training set at the ancestor level and $\{h(s)_1^k, h(s)_2^k, \ldots, h(s)_{n_k}^k\}$ as the training set at the descendant level. The USG-MRF energy function will be in this case:

$$E_{USG-MRF} = - \left\{ \left[ \sum_{\mu=1}^{p_k} K_c(\tilde{h}(c)^{(\mu)}, h(c)) \right]^2 + \left[ \sum_{\mu=1}^{p_k} \sum_{\nu=1}^{q_k} K_s(\tilde{h}(s)^{(\mu\nu)}, h(s)) \right]^2 \right\},$$

(5.8)

that leads to the Bayes classifier

$$k^* = \arg\min_{k=1}^{K} \left\{ - \left[ \sum_{\mu=1}^{p_k} K_c(\tilde{h}(c)^{(\mu)}, h(c)) \right]^2 + \left[ \sum_{\mu=1}^{p_k} \sum_{\nu=1}^{q_k} K_s(\tilde{h}(s)^{k}_{n_k}, h(s)) \right]^2 \right\}. $$

(5.9)
5.4. COMBINING SHAPE AND COLOR INFORMATION FOR OBJECT RECOGNITION.

Figure 5.2: Examples of different views.

\( K_c, K_s \) are the kernel learned for the texture/shape and for the color level, \( \{ \tilde{h}(c)^{(\mu)} \} \), \( \mu = 1, \ldots, p_k \) are the set of prototypes for the color level, and \( \{ \tilde{h}(s)^{(\mu, \nu)} \} , \mu = 1, \ldots, p_k, \nu = 1, \ldots, q_k \) are the set of prototypes for the texture/shape level. Kernels and set of prototypes are all learned from the training data.

5.4.3 Experiments

In order to show the effectiveness of USG-MRF for appearance-based object recognition, we performed several sets of experiments. All of them were ran on the Columbia database (Nene et al., 1996). For all the experiments, the training set consisted of 12 views per object (one every 30°). The remaining views constituted the test set.

Among the many representations proposed in literature, we chose a texture/shape only and color only representation, and we ran experiments using these representations separated, combined together in a common feature vector and combined together in the USG-MRF. The purpose of these experiments is to prove the effectiveness of the USG-MRF model rather than select the optimal combination for the shape and color representations. Thus, we limited the experiments to one texture/shape only and one color only representations.

As color only representation, we chose two dimensional \( rg \) Color Histogram (CH), with resolution of bin axis equal to 8 (Ballard, 1981). The CH was normalized to 1. As texture/shape only representation, we chose two dimensional MFH, \( D_x D_y \), with \( \sigma = 1.0 \) and resolution of bin axis equal to 8. The histograms were normalized to 1.

These two representations were used for performing the following sets of experiments:

1. **Texture/Shape experiments**: we ran experiments using the texture/shape features only. Classification was performed using SG-MRF with the kernelized Hopfield energy (2.4)-(2.6). Kernel parameters \( (a, b, \rho) \) were learned using a leave-one-out strategy. Results were benchmarked with those obtained with
a $\chi^2$ and $\cap$ similarity measures, which proved to be very effective for this representation, and with SVM with Gaussian kernel, $s \in [0.001, 10]$.

2. **Color experiments**: we ran experiments using the color features only. Classification and benchmarking were performed as in the texture/shape experiment.

3. **Color-Texture/Shape experiments**: we ran experiments using the color and texture/shape features combined together to form a unique feature vector. Again, classification and benchmarking were performed as in the texture/shape experiment.

4. **Ultrametric experiment**: we ran a single experiment using the texture/shape and color representation disjoint in the USG-MRF framework. The kernel parameters relative to each level ($a_s, b_s, \rho_s$ and $a_c, b_c, \rho_c$) are learned with the leave-one-out technique. Results obtained with this approach cannot be directly benchmarked with other similarity measures. Anyway, it is possible to compare the obtained results with those of the previous experiments.

Table 5.4 reports the error rates obtained for the 4 sets of experiments. With respect to SVM, we report just the best and the worst results. As the $s$ varies for each result, we don’t report the values here: the interested reader will find a more detailed report on this experiment in Appendix C.5. We see that for almost all series of experiments, for all representations, SG-MRF always gives the best recognition result. Moreover, the overall best recognition result is obtained with USG-MRF. USG-MRF has an increase of performance of $+2.73\%$ with respect to SG-MRF, best result, and of $+5.92\%$ with respect to $\chi^2$ (best result obtained with a non SG-MRF technique). Table 5.5 shows some examples of objects misclassified by SG-MRF and correctly classified by USG-MRF. We see that USG-MRF classifies correctly in cases where shape only or color only gives the right answer (but not both, and not in the concatenated representation; Table 5.5, left and middle column), and also in cases where color only and shape only don’t classify correctly (Table 5.5, right column). These examples show clearly that the better performance of USG-MRF
is due to its hierarchical structure that permits to use different kernels on different features, thus to weight their relevance in a flexible manner with respect to the considered application. Recognition time per view was 0.08 sec for the color only and texture/shape only representation, 0.17 sec for the concatenated representation, and 0.16 for the USG-MRF classifier.

As a last remark on Table 5.4, the fact that the error rates for the color experiments are all above 20% is an indicator that the color representation we chose is far from being optimal. These results confirm our theoretical expectation and show the effectiveness of USG-MRF for color and texture/shape appearance-based object recognition.

<table>
<thead>
<tr>
<th></th>
<th>1st match</th>
<th>1st match</th>
<th>1st match</th>
</tr>
</thead>
<tbody>
<tr>
<td>USG-MRF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SG – MRF_s</td>
<td>2nd match</td>
<td>1st match</td>
<td>3rd match</td>
</tr>
<tr>
<td>SG – MRF_c</td>
<td>1st match</td>
<td>2nd match</td>
<td>7th match</td>
</tr>
<tr>
<td>SG – MRF_sc</td>
<td>3rd match</td>
<td>2nd match</td>
<td>5th match</td>
</tr>
</tbody>
</table>

Table 5.5: Examples of objects misclassified with shape (color) only representation, and correctly classified using USG-MRF.

5.5 A Probabilistic Model of a Scene

Scene recognition and modeling has always been a much researched field in computer vision, and in last years it has gained even more attention (Borenstein et al, 1996; Torralba et al, 1999). Most of the visual information our eyes processes every day is constituted by scenes: interior scenes such as offices, restrooms, shops, our apartments, hotel rooms and so on; exterior scenes such as city streets, parks, mountain landscapes, sea landscapes, country fields and more. A very popular strategy for scene recognition in computer vision is to consider a scene as the sum of the objects that compose it. In this framework, scene recognition will consist of two steps: (1) recognize each single object, and (2) model the spatial relationship between them. It is clear that object recognition has a primary role in this strategy, which was explored by many researchers (Borenstein et al, 1996; Clarkson et al, 2000). This approach suffers from many problems: first, the computational complexity will be equal to the sum of the computational complexities for the recognition of each single object (plus the computational complexity of the relationships between objects). This leads to a model computationally very expensive, thus problematic for real life applications. Second, it relies heavily on the capability
of the algorithm of recognizing in a reliable and robust manner individual objects. As we have seen through all this thesis, this is a challenging task that cannot be considered solved for all the possible objects, environments and situations. Third, a scene recognition model based on the recognition of individual objects will suffer from robustness problems. Consider as an example the task of recognizing a specific scene - an office, for example. An office is composed by many objects: a desk, one or more chairs, one or more computers, bookshelves, pens, .... During the day, some of these objects change their position within the room - some leave the room, but the scene is still an office scene. A scene recognition system as described above should be robust to these variations, a non trivial task. Finally, there is an implicit fragility in this approach: as it assumes that the scene is the sum of several objects, the scene is recognized after that all (or most of) the objects have been recognized. This approach fords the use of contextual information for object recognition, in spite of the fact that context is an important source of information for object identification and classification.

In this thesis we take a different approach. We consider a scene as a new individual visual entity, and not a sum of visual objects. We try to recognize a scene as a whole, and to use this information for recognizing objects inside the scene. The idea of considering a scene as a whole is not new, as it was recently proposed with the name of “holistic representation” (Torralba et al, 1999; Torralba et al, 2001). Compared to the approach described above, the holistic representation presents several advantages: first, the complexity of the algorithm will be equivalent to that of an object recognition system, thus much faster. Second, as this approach considers a scene as a visual object to be recognized as a whole, we can expect that a lot of knowledge developed for object recognition and recognition of visual categories will be of interest for scene recognition. Finally, this approach consents to use the scene recognition information as prior knowledge for the recognition of specific classes of objects, in certain positions of the scene. In this Section we will concentrate the attention on how to use scene recognition as prior knowledge, as to say as contextual information, for object recognition. Thus, we will consider scenes and we will consider objects within a scene, but in this last case we will consider as known the position of the object into the scene (an example of how the contextual information can be used for determining the most probable locations of an object in a given scene can be found in (Torralba et al, 2001)). We will apply the USG-MRF framework to this problem (Section 5.5.2) and we will present experiments that show how considering the contextual information highly increases the recognition performance (Section 5.5.3).

5.5.1 Contextual Information versus Heterogeneous Background

The appearance-based approach to object recognition assumes that the identity of an object can be determined from its visual appearance; thus, when we say that a set of training views \( \omega_k = \{ x_{1k}, x_{2k}, \ldots, x_{nk} \} \) relative to the object \( \Omega_k \) are a sufficient statistic for \( \Omega_k \), we mean that the views in \( \omega_k \) give us all the information we need
in order to estimate the relative probability density function. Now, an object view will contain information regarding the object of interest, and information regarding where the picture was taken. Shall we use this information in the object recognition process? In other words, when what is inside the view and is not object can be considered as information relevant to the recognition process (in this case we will call it contextual information), and when can be considered as noise (in this case we will call it heterogeneous background)? We will take into account two factors:

1. **Amount of the non-object information**
   A view representing an object will have a given size, of which the object will occupy a percentage. In an appearance-based approach to object recognition, we can expect that the object will be contained wholly inside the view; but the amount of the view representing the scene where the object is contained can vary a lot from view to view. Consider for instance the objects shown in Figure 5.3: the view on right contains almost no information about where the image was taken, while the view on left allows quite easily to recognize where the picture was taken (an office desk). If the amount of the non-object information is not enough to identify it, we will treat it as noise and we will call it heterogeneous background; if the amount of the non-object information is enough to identify it, we will treat it as prior knowledge and we will call it contextual information.

2. **Statistical significance of the non-object information**
   If we want to use the contextual information as prior knowledge in the object recognition procedure, we need a sufficient statistic for the context, exactly as we need a sufficient statistic for the object. This means in practice that each view in the training set must contain enough information about where the picture was taken, and there must be enough views taken in a given scene. This can mean that all the views were taken in a specific scene, or in two or three; but if all the views would be taken in different scenes, we wouldn’t have a sufficient statistic for modeling the contextual information, and we couldn’t use it.

In what follows, we will make the two following assumptions: (1) we will have training sets containing in each view enough information about the object to be recognized and about the scene(s) where the object picture was taken; (2) we will have training sets constituting a sufficient statistic for the considered object and for the scene(s) where the pictures were taken. We consider all that is not included in this framework as heterogeneous background and, as such, as a form of degradation of the view, to be tackled via robustness properties. We refer the reader to Chapter 4 with respect to this last point.

In the next Section instead we will present an object recognition system that takes advantage of the prior knowledge regarding the context using USG-MRF. As experiments will show, it constitutes an effective manner for modeling the scene information in a probabilistic framework for object recognition.
5.5.2 The Ultrametric Organization of Contextual Information

Assume to have a set of objects $\Omega_k, k = 1, \ldots, K$, each represented by a training set $\omega_k$. Suppose also that the set of $\{\omega_k\}_{k=1}^K$ training views can be divided in $B$ subsets, with respect to the scene where the picture views were taken:

$$S_1 = \{\{\omega_k\}_{k=1}^K \mid \text{picture view taken in } S_1\}$$

$$S_2 = \{\{\omega_k\}_{k=1}^K \mid \text{picture view taken in } S_2\}$$

$$\cdots$$

$$S_1 \cup S_2 \cup \cdots = \{\omega_k\}_{k=1}^K.$$

Our goal is to classify a given data sample $x$, representing one of the objects $\Omega_k$ in one of the scenes $S_B$, as belonging to one of the scenes $S_B$, and then to use this information for the object recognition step. To this purpose, we will make use of the USG-MRF framework and of the robustness properties of SG-MRF as follows:

1. **Extraction of Region Of Interests (ROIs)**

   This step is relevant just for the scenes datasets $(S_1, S_2, \ldots, S_B)$. The views in $\{\omega_k\}_{k=1}^K$ represent all an object, surrounded by a scene (this is the consequence of: (1) assuming that the object views are a sufficient statistic for the objects, and (2) assuming that the object location is known). Thus, we can reasonably assume that, for each view, the object will be roughly located in the center. We will extract, from each view, a portion of image not centered; an example can be a corner window of fixed size. As we see from Figure 5.4, this procedure leads to a set of views (Region Of Interests, ROIs) in which the dominant information is about the scene to be modeled and then classified. Note that the object is still partially visible in some views, but it is not anymore the dominant information.
2. **Partition of the training set**

Our goal is to organize the database in a hierarchical fashion, so to apply the USG-MRF framework. We will have the scenes at the ancestor level, and objects at the descendant level; at each level, the training set will be:

- **ancestor level training set:** there will be $S_1, S_2, \ldots S_B$ training sets, each relative to one of the scenes to be recognized. For each scene, each training set is composed by the ROIs extracted from the training object views which were taken in that scene.

- **descendant level training set:** for each object, for each scene, we will have a set of training views $\omega_{Bk}$: for example, $\omega_{1k}$ will be the training set relative to the object $\Omega_k$ in the scene $S_1$; $\omega_{2k}$ will be the training set relative to the object $\Omega_k$ in the scene $S_2$; and so on. We will assume that, for at least one scene, there is a training set which is a sufficient statistic for the object $\Omega_k$. Then, we will build $B$ training sets for each object: for example, if the scene for which we have a sufficient statistic for $\Omega_k$ is $S_1$, $\omega_{1k}$ will be a set of views all taken in the scene $S_1$. With respect to $S_2$, the training set $\omega_{2k}$ will have all the views taken in the scene $S_1$ plus some views taken in the scene $S_2$; their number must be between 25% and 75% of the total views of the object in the scene $S_1$; and so on.

We organize the database in this apparently complicated way because we want to use the USG-MRF framework combined with the robustness properties of SG-MRF. The idea is to build a USG-MRF-MAP classifier that recognizes the scene at the ancestor level, and that classify the object at the descendant level. We have seen in Section 4.4.4 that the capability of SG-MRF of recognizing objects in heterogeneous background increases when the training is made on a mixture of homogeneous and heterogeneous background. Here we want to push further this property and to make use of the contextual information. Hence, the described structure of the database at the descendant level.
3. **Representation**
   As we have done through all the thesis, we will represent each training view
   with a robust feature representation.

4. **Learning of the kernel parameters**
   We will apply the USG-MRF-MAP classifier on the data prepared as described
   above, and we will learn the kernel parameters and select the prototypes at
   each level, as described in Section 5.2.3. Once learning is completed, we will
   force the ancestor level to consider just the first match; the prototypes to be
   used at the descendant level will be relative to the scene recognized at the
   first match at the the ancestor level.

5.5.3 **Experiments**

We ran an experiment on a database of 6 objects (Nelson6, (Nelson et al, 1998)):
a cup, a fighter, a plane, a car, a toy rabbit and a toy bear (Figure 5.5). The
database originally consists of a training set of 106 views (53 for the car) taken
on a sphere (hemisphere for the car), approximately every 20 degrees, in black
homogeneous background. There are three different test sets: one taken in black
homogeneous background, one on a white marble table and one on a poster with
Christmas decorations. Some examples are shown in Figure 5.6

![Figure 5.5: Objects from the the Nelson6 database; the objects views are shown in the poster background.](image)

Each of these test sets has 53 (24 for the car) views, positioned in between
the training views. The test views are taken at the same scale of those in the
training set, but the illumination conditions changes for the different backgrounds.
Although we cannot say that the objects in this database are in a scene like an
office, this database gives a sufficient statistic with respect to the three scenes, the
images are taken in realistic environments and are not the result of a simulation (as
Figure 5.6: An example of an object from the Nelson6 database in three different scenes

done for the experiments presented in Section 4.4.4). Thus, we decided to apply the method described in the previous Section on this database.

We proceeded as follows: from all the test views of all objects, for the three different scenes, we extracted a window of dimension $128 \times 128$ from the upper left corner; some examples are shown in Figure 5.4. Each scene view contains mostly scene information, plus sometimes a small portion of the object. These scene views constituted the training set for the ancestor level, which thus has three scene classes, and a training set for class of 583 views, all of dimension $128 \times 128$.

For each scene, we built the corresponding training set: for example, for the scene “poster”, the training set at the descendant level consisted of the original training views (in homogeneous background), plus all minus one test views in the poster background. The single test view not included in the training set is used for testing; the procedure is repeated for all objects, for the three scenes and for all the original test view with a leave-one-out procedure. We used as representation (heuristic choice) MFH, $D_xD_y, \sigma = 1.0$, resolution for histogram axis 16 bins.

For each training set, for each test view kept apart, we learned kernel parameters and selected prototypes as described in Section 5.2.3. We performed 4 different experiments:

1. **SG-MRF train homo**
   We used a SG-MRF-MAP classifier, with training set consisting, for each object, of the training set relative to the “homogeneous background” scene. The model was tested, with the leave-one-out strategy, on the test set in homogeneous scene, in “white” scene and in “poster” scene.

2. **SG-MRF train white**
   We performed the same experiment as described above, but with the training set relative to the “white” scene.

3. **SG-MRF train poster**
   We ran the same experiment as described above, but with the training set relative to the “poster” scene.

4. **USG-MRF**
   We first recognized the test view, representing an object in a scene, as belonging to one of the three scene classes (ancestor level); then, we recognized the
object in the test view as belonging to one of the object classes (descendant level). The prototypes for the descendant level contains, for each object, a mixture of views representing the “homogeneous” scene and views containing the scene recognized at the ancestor level. In this sense, the object recognition system we propose uses the context as prior knowledge in the recognition stage.

Recognition results are reported in Table 5.6. We see that the use of the contextual information increases remarkably the recognition performance; recognition time per view was 0.06 sec for SG-MRF and 0.12 for USG-MRF.

<table>
<thead>
<tr>
<th></th>
<th>test homo (%)</th>
<th>test white(%)</th>
<th>test poster(%)</th>
<th>total(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SG-MRF train homo</td>
<td>99.24</td>
<td>28.03</td>
<td>34.47</td>
<td>53.91</td>
</tr>
<tr>
<td>SG-MRF train white</td>
<td>98.86</td>
<td>93.56</td>
<td>42.80</td>
<td>78.84</td>
</tr>
<tr>
<td>SG-MRF train poster</td>
<td>95.07</td>
<td>28.79</td>
<td>90.91</td>
<td>71.59</td>
</tr>
<tr>
<td>USG-MRF</td>
<td><strong>99.24</strong></td>
<td><strong>93.56</strong></td>
<td><strong>90.91</strong></td>
<td><strong>94.57</strong></td>
</tr>
</tbody>
</table>

Table 5.6: Classification results for the recognition of objects in heterogeneous background, using (first three rows) and non-using (last row) the background as contextual information.

This database was built and used by Nelson and Selinger (Nelson et al., 1998); they developed an object recognition system based on loosely structured combinations of a number of local context regions, described by distinctive features. Thus, it is possible to compare their results with those obtained with USG–MRF (Table 5.7). We see that USG-MRF gives always the best performance, particularly obtaining in the case of the “poster” scene an increase of performance of +9.91%.

We can conclude that contextual information is an important source of knowledge for object recognition, and that USG-MRF, combined with the robustness properties of SG-MRF, is an effective approach for this problem.

<table>
<thead>
<tr>
<th></th>
<th>test homo (%)</th>
<th>test white(%)</th>
<th>test poster(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nelson train homo</td>
<td>98.50</td>
<td>90.90</td>
<td>81.00</td>
</tr>
<tr>
<td>USG-MRF</td>
<td><strong>99.24</strong></td>
<td><strong>93.56</strong></td>
<td><strong>90.91</strong></td>
</tr>
</tbody>
</table>

Table 5.7: Classification results for the recognition of objects in heterogeneous background, using USG-MRF and the Nelson approach.

5.6 Summary

In this Chapter we presented a new energy function that can be used in the SG-MRF framework. This energy consists mainly of a hierarchical organization of Hopfield
energies; as it was originally derived taking into account the ultrametric properties of the SG configuration space, we called the new model Ultrametric Spin Glass-Markov Random Fields. We described the new energy, how it can be kernelized and the corresponding MAP classifier. We presented three vision applications in which the new model obtains remarkable results: we showed how it can be used for hierarchical object recognition, for combining together shape and color information for recognizing objects, and for a probabilistic model of a scene in order to use contextual information in the object recognition process. In all those examples, the results obtained confirm the effectiveness of the new model.
Chapter 6

Moving Beyond

Previous Chapters have shown the potential of SG-MRF for object recognition. In this Chapter we explore future directions of research in three different fields: computer vision, machine learning and statistical mechanics. SG-MRF has been developed on the basis of results obtained in all these fields. Thus, we can expect that new research, developed on the basis of SG-MRFs, will be of interest for these three fields.

Section 6.2 presents a novel kernel classifier that permits to use different features for different classes. This constitutes an important contribution on how to select the proper set of features for a given application. It consists in learning the features (as to say the kernel) specifically for each class under consideration.

In Section 6.3 we report a possible strategy for the recognition of visual categories, using SG-MRF. Recognition of visual categories is an important topic of research in computer vision, which has been gaining increasing attention in the last few years. We propose an approach based on kernel properties, which shows promising results.

Section 6.4 presents a study on the properties of kernel associative memories from the point of view of Ising-like systems. We report a first analysis on the storage capacity of the model, using a replica method approach.
6.1 Introduction

Spin Glass-Markov Random Field is a probabilistic model that uses results and concepts developed in three different research areas: computer vision, machine learning and statistical physics of spin glasses. As such, several research lines can be developed from it, of interest for all the three above mentioned communities. This Chapter describes three possible directions for future research, and presents some preliminary results.

- **Machine Learning: Kernel Class Specific Classifiers**
  Feature selection is a crucial issue in pattern recognition. Within the context of multi-class object recognition, it faces two big issues: (1) with respect to the single object (or category) to be recognized, how to select the optimal set of features; (2) how to find a common optimal feature set, for all the objects (categories) to be recognized. Alternatively, how to use different optimal features for different objects, in a common framework. Section 6.2 presents a novel kernel classifier that permits to use different features for different classes. This result is achieved combining SG-MRF with class specific classifier (Baggenstoss et al., 2000). It addresses the problem of how to select the proper set of features for a given application, learning the features (as to say the kernel) specifically for each class under consideration.

- **Computer Vision: Recognition of Visual Categories**
  Through most of this thesis, we focused on object identification. Recognition of visual categories is an equally crucial topic, which has attracted more and more attention during the last years (Fergus et al., 2003; Leibe et al., 2004; Weber et al., 2000b). Categorization plays a key role for coding of experience in our daily activities. Many categories tend to maximize the physical (and thus visual) similarity between its members, and at the same time maximize dissimilarities between its members and members of other categories. This process implies the choice of an optimal representation and similarity measure. Here we propose to use SG-MRF for recognition of visual categories: being a probabilistic method, it allows to take into account variabilities within a specific category. Being a kernel method and allowing to learn the kernel from training data, it permits to find the optimal way of measuring similarities via the kernel function. Section 6.3 reports how to formulate the problem in terms of SG-MRF; experiments show promising results.

- **Statistical Mechanics: Kernel Associative Memories**
  Section 6.4 presents a study on the properties of kernel associative memories from the point of view of Ising-like systems. Kernel associative memories are a nonlinear generalization of the Hopfield energy never studied before. Thus, it is of interest for the attractor neural network and spin glass research community to investigate the properties of these energies. Here we present
the first analysis on the storage capacity of the model, using a replica method (Mezard et al, 1987) approach.

6.2 Kernel Class Specific Classifier

Consider any multi-class pattern recognition problem. The possibility to choose, for a given set of classes, a different representation for each class, is in principle very powerful. Consider for example the problem of classifying a vegetable as a member of 5 possible classes: tomatoes, carrots, zucchini, onions and pumpkins. If we represent these classes using the color representation, tomatoes will be identified with no ambiguities, but carrots and pumpkins will be mixed. If we choose instead a shape representation for all classes, tomatoes and onions could be confused, while pumpkins would be unambiguously identified; and so on. More generally, we can say that for each class there is at least one representation that captures at best the “essence” of that class, and thus makes that class easily distinguishable from others. This representation can be different for different classes. Assuming it is possible to determine it, a possible solution could be to take a unique, common representation which contains all the representations relative to the considered classes. In the example above, this could mean taking as representation color+shape+size. This solution is not feasible in general, due to the curse of dimensionality effect (Bishop, 1995).

Baggenstoss et al. (Baggenstoss et al, 2000) proposed recently a feature-based Class Specific Classifier (CSC) that allows the use of different features (thus representations) for different classes in a Bayes classifier. This result is obtained introducing a common reference hypothesis class and using results of statistical theory (Baggenstoss et al, 2000). An open point for CSCs is how to choose features for each class. The strategy to choose a different set of features for each class will be winning as long as the chosen features are the right ones, according to the need of that class. For most applications, the kind of features that can be used is, a priori, huge. Choices are usually made heuristically, and the truth is that, even when the performance of the final classifier is good, we cannot be sure that it wouldn’t improve with another set of features. How to choose features for a given set of classes is an open problem for all computer vision and pattern recognition applications (Bishop, 1995; Duda et al, 2001). For CSCs, which base their power on the possibility to choose several sets of features for different classes, the problem is more relevant.

In the last years, it has been proposed a possibly alternative strategy to the choice of a feature set. It consists of the use of the so-called kernel methods, the most popular of which is the Support Vector Machine algorithm (Schölkopf et al, 2002). Kernel methods apply to every algorithm that depends on the scalar product between data, and replace the scalar product with a kernel function, which can be interpreted as the scalar product between the original data in a higher dimensional space. This space is reached via a non linear mapping, that replaces (and is by
principle equivalent to) the feature extraction step. The power of the idea lies in the fact that the kernel function -and not the mapping -is explicitly known. Due to theoretical constraints, the functional form of these functions is known and limited (Schölkopf et al., 2002). Thus, the number of choices is limited compared to the choice of a set of features, although the criteria is still mostly heuristic.

Here we propose combining CSC with kernel methods via SG-MRFs. The use of SG-MRF in a CSC carries many advantages. First, it does allow to use the power of kernel functions for classification purposes, still leaving open the possibility to use different sets of features. Second, the class of possible kernel functions is determined a priori by theoretical constraints. As the choice of kernels is limited, it does not become ungovernable. At the same time is wide enough to reasonably guarantee the possibility to tailor each kernel according to each class needs. For each class, the kernel is learned by the training data with a leave-one-out strategy. We call this new method Kernel-Class Specific Classifier (K-CSC).

The rest of the Section is organized as follows: after the formulation of the general problem, we review CSC method (Section 6.2.2); Section 6.2.3 derives the K-CSC, and Section 6.2.4 presents experiments on appearance-based object recognitions. We conclude with a summary discussion.

6.2.1 Problem Statement

We are here concerned with any statistical method where it is needed to estimate the probability density function of the input data. We will focus the attention on an important and basic example of such methods: the so-called $K$-ary classifier. Let $H_k, k = 1, \ldots, K$ be $K$ different classes or statistical hypotheses: given a data sample $x$, produced by one of $K$ possible classes, our goal is to classify $x$ as a sample from $H_{k^*}$, one of the $H_k$ classes. It is well known that the optimal classifier, which results in the lowest probability of error is the Maximum A Posteriori (MAP) classifier:

$$k^* = \arg\max_{k=1}^{K} P(H_k|x) = \arg\max_{k=1}^{K} \left\{ P(x|H_k)P(H_k) \right\}; \quad (6.1)$$

using Bayes rule, where $P(x|H_k)$ are the Likelihood Functions (LFs) and $P(H_k)$ are the prior probabilities of the classes. Assuming that $P(H_k)$ are constant and equal, the Bayes classifier simplifies to

$$k^* = \arg\max_{k=1}^{K} P(x|H_k).$$

Although in well defined problems (such as simple detection or pattern matching problems involving known signal waveforms) the LFs can be written down (Duda et al., 2001), most problems of practical interest involve the detection or classification of natural disturbances or patterns which either are unknown or do not obey statistical models that can be easily described. Examples are human speech and computer vision. As a result, the LFs have traditionally had to be learned from data samples.
6.2. KERNEL CLASS SPECIFIC CLASSIFIER

6.2.2 The Class Specific Theorem and the Class Specific Classifier

The high dimension of the raw data usually precludes doing any kind of non-parametric PDF estimation. Therefore, the usual approach extracts a small number of information-bearing statistic, called features (Figure 6.3). Let \( z = T(x) \) be such a set of features. The Bayesian classifier based on \( z \) is

\[
k^* = \arg\max_{k=1}^{K} P(z|H_k).
\] (6.2)

Thus, the features replace the raw data. The hidden implication here is that \( z \) is a sufficient statistic for the classification problem:

\[
P(x|H_k) = g(T(x)|H_k)h(x), k = 1, \ldots, K.
\] (6.3)

This is the Neyman-Fisher factorization theorem (Duda et al., 2001). Conceptually, learning of LFs is simple. The difficulty in PDF estimation is that as the dimension of \( z \) increases, the complexity of the problem increases exponentially. Indeed, it has been shown that given that the PDF meets certain smoothness assumptions, the amount of training data required for nonparametric estimators\(^1\) rises exponentially with the dimension (Bishop, 1995); the rapid increase in complexity of systems has been termed the curse of dimensionality by Richard Bellman. As features are added, it is possible that algorithm performance may actually get worse in spite of information content of the added feature. Likewise, eliminating an information-bearing feature may actually improve performance.

Dimensionality reduction is often a necessary first step to PDF estimation and is the subject of much research currently and over the past decades (Cristianini et al., 2000; Bishop, 1995; Duda et al., 2001). The trouble is that projecting the data to a low-dimensional feature space requires knowing the nature or the structure of the mechanism; but knowing the structure requires knowing something about the PDF; and yet learning the PDF requires first knowing the structure. This frustrating dilemma has given rise to a seemingly infinite number of ingenious methods, each based on its own implicit assumptions. Various approaches include feature selection (Duda et al, 2001), projection pursuit (Bishop, 1995) and independence grouping (Duda et al, 2001; Bishop, 1995). Several other methods are based on projection of the feature vectors onto lower dimensional subspaces. A very large percentage of these methods can be characterized as a three-step process: (1) collect a number of information-bearing features, (2) reduce the dimension of this feature set, (3) estimate the PDF of the reduced-dimension feature-set conditioned on each data class.

\(^1\)By non-parametric, here we mean that we do not know the parametric form of the PDF, so we have to assume a standard form such as Gaussian mixture, then estimate the parameters of this assumed model with the understanding that the model we assume is just an approximation.
In all of these methods, there is an implicit approximation which limits the theoretical performance. For example, subspace methods project the data to low-dimensional subspaces, which may be inappropriate in some problems. A visual example is a 2-dimensional plane in which one data class is distributed in a ring enclosing the samples of a second data class. Any attempt to project the samples to a lower-dimensional space (a line) will fail to preserve the class separation. Only a non-linear transformation (i.e. to polar coordinates) resolves the classes. Unfortunately, the correct non-linear transformation cannot be easily discovered at high dimension, such as in the simple example just described. Another common characteristic of these methods is that the relationship between the features and the raw data is forgotten. The class-specific method to be described differs in both respect. Dimension reduction is accomplished at the theoretical level, before any practical approximations are made, and the relationship to the input data is preserved.

It was recognized for some time (Duda et al, 2001) that the \(K\)-ary classifier (6.1) can be constructed by knowing only the likelihood ratios

\[
\frac{P(x|H_2)}{P(x|H_1)} \cdot \frac{P(x|H_3)}{P(x|H_1)} \cdots \frac{P(x|H_M)}{P(x|H_1)}.
\]

Thus, these likelihood ratios are sufficient. Van Trees recognized (Duda et al, 2001; Bishop, 1995) that an additional class, \(H_0\), the “dummy” class, can be used in the denominator:

\[
\argmax_{k=1}^K \frac{P(x|H_k)}{P(x|H_0)}.
\]

The well known corollary of the Neyman-Fisher factorization theorem (6.3) is that any likelihood ratio is invariant when written in terms of a sufficient statistic. Thus, if \(z\) is a sufficient statistic,

\[
\frac{P(x|H_j)}{P(x|H_k)} = \frac{P(z|H_j)}{P(z|H_k)}.
\]

The \(K\)-ary classifier (6.1) can be constructed by knowing only the likelihood ratios; moreover, it is possible to use in the denominator an additional class, \(H_0\):

\[
k^* = \argmax_{k=1}^K \frac{P(x|H_k)}{P(x|H_0)}.
\]

Now, note that the likelihood ratio doesn’t change when we take two different mapping \(z_1 = T_1(x), z_2 = T_2(x)\):

\[
\frac{P(x|H_k)}{P(x|H_0)} = \frac{P(z_1|H_k)}{P(z_1|H_0)} = \frac{P(z_2|H_k)}{P(z_2|H_0)}.
\]

And thus we may write the Bayes classifier as (Baggenstoss et al, 2000)

\[
k^* = \argmax_{k=1}^K \frac{P(z_k|H_k)}{P(z_k|H_0)}.
\]
6.2. KERNEL CLASS SPECIFIC CLASSIFIER

where \( z_k = T_k(x), 1 \leq k \leq K \) are feature transformations that depend on the class being tested, thus they are class specific features. This is the feature-based Class Specific Classifier (CSC (Baggenstoss et al, 2000), Figure 6.4). CSCs major advantage is that they allow us to use different features for each class; a serious problem is that they do not provide any criteria on how to choose the best possible set of features for each class.

6.2.3 The Kernel Class Specific Classifier

In this Section we show how the combination of CSC and SG-MRF leads to a new kernel classifier which fully uses the power of both ideas. First of all, recall that a \textit{kernel} is a function \( K \) such that, for all \( x, y \in X \),

\[
K(x, y) = \Phi(x) \cdot \Phi(y),
\]
where $\Phi$ is a mapping from $X$ to an (inner product) feature space $F$ (Schölkopf et al., 2002). The SG-MRF energy function can be rewritten as:

$$E_{SG-MRF} = -\sum_{\mu=1}^{p_k} [K(x, \bar{x}^\mu)]^2 = -\sum_{\mu=1}^{p_k} \tilde{K}(x, \bar{x}^\mu),$$

where $\tilde{K}$ represents a new kernel function (Schölkopf et al., 2002). So we can write the SG-MRF energy as

$$E_{SG-MRF} = -\sum_{\mu=1}^{p_k} \tilde{K}(x, \bar{x}^\mu) = -\sum_{\mu=1}^{p_k} \Phi(x) \cdot \Phi(\bar{x}^\mu),$$

and the SG-MRF probability distribution becomes

$$P_{SG-MRF}(x|H_k) = \frac{1}{Z} \exp \left[ \sum_{\mu=1}^{p_k} \Phi(x) \cdot \Phi(\bar{x}^\mu) \right] . \tag{6.8}$$

Equation (6.8) tells that $P_{SG-MRF}$ depends on $x$ via a mapping $\Phi(x) = z$. Thus, we can use this probability in the CSC classifier (6.7), identifying the feature extraction operator $T_k(x)$ with the mapping $\Phi_k(x)$, as to say using a different mapping, and thus a different kernel for each class. We get:

$$k^* = \arg\max_{k=1}^{K} \frac{P_{SG-MRF}(\Phi_k(x)|H_k)}{P_{SG-MRF}(\Phi_k(x)|H_0)}$$

$$= \arg\max_{k=1}^{K} \left\{ \frac{1}{Z} \exp \left[ \sum_{\mu_k=1}^{p_k} \Phi_k(x) \cdot \Phi_k(\bar{x}^{\mu_k}) \right] \right\}$$

$$= \arg\max_{k=1}^{K} \left\{ \frac{1}{\exp\left(\frac{1}{N_k}\right)p_k} \exp \left[ \sum_{\mu_k=1}^{p_k} \Phi_k(x) \cdot \Phi_k(\bar{x}^{\mu_k}) \right] \right\}$$

where $\{\bar{x}^{\mu_k}\}, \mu_k = 1, \ldots p_k$ are the set of prototypes of class $H_k$; $\{\bar{x}^{\mu_0}\}, \mu_0 = 1, \ldots p_0$ are the set of prototypes of class $H_0$. $N_k$ is the dimension of the space where the pdf representing class $H_k$ is mapped by the mapping $\Phi_k$; $N_0$ is the dimension of the space where the pdf representing the reference hypothesis $H_0$ is mapped by the mapping $\Phi_k$. Note that, as the mapping $\Phi_k$ is the same for the numerator and denominator, $N_k = N_0, \forall k = 1, \ldots K$. Thus, it follows:

$$= \arg\max_{k=1}^{K} \left\{ \frac{p_k}{p_0} \exp \left[ \sum_{\mu_k=1}^{p_k} \Phi_k(x) \cdot \Phi_k(\bar{x}^{\mu_k}) - \sum_{\mu_0=1}^{p_0} \Phi_k(x) \cdot \Phi_k(\bar{x}^{\mu_0}) \right] \right\}$$
6.2. KERNEL CLASS SPECIFIC CLASSIFIER

\[
= \arg\min_{k=1}^{K} \left\{ \ln \frac{p_k}{p_0} \left[ - \sum_{\mu_k=1}^{p_k} \phi_k(x) \cdot \phi_k(\bar{x}^{\mu_k}) + \sum_{\mu_0=1}^{p_0} \phi_k(x) \cdot \phi_k(\bar{x}^{\mu_0}) \right] \right\}.
\]

Thus, the CSC united to SG-MRF gives a Kernel-Class Specific Classifier:

\[
k^* = \arg\min_{k=1}^{K} \left\{ \ln \frac{p_k}{p_0} \left[ - \sum_{\mu_k=1}^{p_k} [K_k(x, \bar{x}^{\mu_k})]^2 + \sum_{\mu_0=1}^{p_0} [K_k(x, \bar{x}^{\mu_0})]^2 \right] \right\}. \tag{6.9}
\]

If the number of prototypes is the same for all the classes \( \{ H_k \} \) and for the reference hypothesis \( H_0 \), then equation (6.9) becomes

\[
k^* = \arg\min_{k=1}^{K} \left[ - \sum_{\mu_k=1}^{p_k} [K_k(x, \bar{x}^{\mu_k})]^2 + \sum_{\mu_0=1}^{p_0} [K_k(x, \bar{x}^{\mu_0})]^2 \right]. \tag{6.10}
\]

Given a training set, the kernel \( K_k \) can be learned, for each class \( H_k \), as for SG-MRF (Section 3.7). Thus, K-CSC permits using for each class a different representation according to its needs, as CSC does. But as the K-CSC representation is bound to be a specific class of kernels, it solves the ambiguity of CSC regarding the choice of the representations and permits to learn them. Moreover, this permits to use a different kernel and a different set of features for each class.

6.2.4 Experiments

The application of the CSC method to vision problems is possible in principle, but has been very challenging until now (Baggenstoss et al., 2000); on the other side, we have shown in this thesis that SG-MRFs are very effective for vision applications, particularly for appearance-based object recognition (Caputo et al., 2002a; Caputo et al., 2002g). For these reasons, we decided to test the K-CSC performance on an appearance-based object recognition application. We ran two sets of experiments: the first on the Nelson6 database (see Section 5.5), the second on the Nelson database (see Section 3.9). For both experiments, we used as representation 2D MFH, with filters given by Gaussian derivatives along \( x \) and \( y \) directions and with \( \sigma = 1.0 \); resolution for histogram axis was 16 bins.

The K-CSC algorithm has two parameters which have to be decided by the user: the reference hypothesis \( H_0 \), and the range of the kernel parameters \((a, b, \rho)\) for the automatic learning of the class-specific kernel. With respect to the range of the kernel parameters, for both experiments we imposed the algorithm to learn the optimal kernel from the following ranges of kernel parameters: \( a \in [0.0, 0.1, \ldots, 1.0] \); \( b \in [0.1, 0.2, \ldots, 2.0] \); \( \rho \in [0.01, 100] \). The learning strategy is analogous to that described in Section 3.7.

As reference hypothesis, we decided to use one of the object classes, selected during the training step. This has the double advantage of limiting the field of possible choices for \( H_0 \), and not having to estimate an additional probability density function. Although we cannot claim that this strategy for the choice of \( H_0 \) is optimal,
we argue that it is reasonable. Indeed, the underlying principle in doing classification using probability ratios is to perform multi-class discrimination against a 'non-object' class. If this class is too broad (a common strategy is to define a 'background' class, see for instance (Fergus et al, 2003)) there are two possible risks:

1. If the estimate of the probability density function is done on a training set which is not representative enough of the reference class, the resulting probability ratio classifier will be biased toward the object class we are interested in, leading to a very high number of false positive.

2. If the class is too broad and it is estimated effectively, it might become a strong attractor for almost any class, and the resulting classifier might be biased toward the reference hypothesis, resulting in a very high number of false negative (this might happen for instance in the case of distinguishing objects like cups, books and so on, if we would choose as reference hypothesis the class ‘desk background’).

By choosing as reference hypothesis one of the objects we want to classify, we make sure that we will discriminate with respect to well-defined visual classes, and we have a reasonable hope that the training data from which we estimate the probability density functions are a sufficient statistic for the reference class $H_0$ (if this would not be the case, we would be in trouble anyway for the estimate of the pdf of that object class). Last but not least, this means that, for a $N$-class recognition problem, we must estimate $N$ probability density functions, and not $N + 1$.

Table 6.1 shows recognition results obtained on the Nelson6 database using SG-MRF, K-CSC and Nelson’s method (as in Section 5.5, we report results published in (Selinger, 2001)). The learned reference hypotheses for K-CSC was the object class ‘cup’; kernel parameters learned for SG-MRF were $a = 0.5, b = 1.4, \rho = 0.1$. Learning for K-CSC was computationally expensive (~2 hours), but recognition time per view was 0.1 sec, to be compared with recognition time per view of SG-MRF (0.06 sec). We see that K-CSC obtains the better result, followed by SG-MRF and finally by Nelson’s method.

A second experiment was run on the Nelson database described in Section 3.9.2; the procedure was the same as described above for the 6 class experiment. Learning was particularly expensive for K-CSC (~2 days), due to the fact that we are trying here to solve a 59 class experiment, and we are selecting the reference hypothesis $H_0$ out of 59 possible candidate. Recognition time per view was 0.08 sec for SG-MRF, and 0.34 sec for K-CSC. SG-MRF gave a recognition rate of 90.17% using the kernel parameters (learned during the training stage) $a = 0.5, b = 0.9, \rho = 0.1$; K-CSC gave a recognition rate of 98.06%, with reference hypothesis ‘cup1’. Thus, the computationally expensive learning stage for K-CSC is compensated by an increase in recognition rate of ~ 8%.
6.2. KERNEL CLASS SPECIFIC CLASSIFIER

<table>
<thead>
<tr>
<th></th>
<th>SG-MRF</th>
<th>K-CSC</th>
<th>Nelson</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cup</td>
<td>100</td>
<td>100</td>
<td>93.75</td>
</tr>
<tr>
<td>Fighter</td>
<td>100</td>
<td>100</td>
<td>95.83</td>
</tr>
<tr>
<td>Plane</td>
<td>100</td>
<td>100</td>
<td>97.92</td>
</tr>
<tr>
<td>Car</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Toy Rabbit</td>
<td>93.75</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Toy Bear</td>
<td>100</td>
<td>100</td>
<td>91.67</td>
</tr>
<tr>
<td>Tot</td>
<td>98.86</td>
<td>100</td>
<td>98.50</td>
</tr>
</tbody>
</table>

Table 6.1: Classification results for SG-MRF, K-CSC and Nelson algorithm. We report the recognition rates.

6.2.5 Discussion

We presented in this Section a new kernel classifier that permits us to use different kernels for different classes. We achieve this result combining SG-MRF with the feature-based class specific classifier. The novel classifier presents several advantages: first, it permits to learn the optimal kernel to be used for each class. This results in a remarkable increase in the recognition rate with respect to SG-MRF. Second, it offers a novel approach and a novel solution to the problem of how to select optimal features for a specific classification problem. Finally, it makes it straightforward to use the class specific classifier approach to visual pattern recognition, a challenging task until now.

We presented here just preliminary experiments: the method should be tested extensively on different databases and on visual and not visual pattern recognition problems. It should be benchmarked with other kernel methods like Support Vector Machines, and with other feature selection methods. It should also be noted that the increase of performance is obtained via a very expensive learning procedure, which might limit the use of the method for real-world applications with a large number of classes. A possible solution might be to use a faster classifier (as SG-MRF) in a first step to generate hypotheses, and then use K-CSC in a second step, on the subset of the best $n$ hypotheses; this would permit to take under control the number of classes for the K-CSC classifier. The number of hypotheses might be determined in a principled way considering the recognition time and the number of false positive that the task at hand can afford, as suggested by Viola and Jones in (Viola and Jones, 2001). Future work will explore this possibility.
CHAPTER 6. MOVING BEYOND

6.3 Recognition of Visual Categories using Spin Glass-Markov Random Field

We code most of our every day experiences using categories. For example, we can perceive some complex object as a kind of “phone”, remember it as a “phone”, describe it to others as a *phone*, and reason about it in the same way. A characteristic of many categories is that their members tend to be physically similar to one another, while being physically dissimilar from members of different categories (Brooks, 1981). Thus it is possible, by principle, to build an artificial visual system able to recognize categories.

Object categorization has received an increasing amount of attention in the last years. Weber et al. (Weber *et al.*, 2000a; Weber *et al.*, 2000b) proposed a method to learn models of object categories, imaged in cluttered scenes. These models represent objects as probabilistic constellations of features; variability within a class is represented by a joint probability density function on the shape of the constellation and the appearance of the parts. During training, first distinctive features are identified; then, the set of model parameters is learned using expectation maximization. Fergus et al. (Fergus *et al.*, 2003) extended the method so to be robust to scale variations, and introduced an entropy-based feature detector to select regions and their scale within the image. A further extension was proposed by Li et al. (Li *et al.*, 2003): they investigate the possibility to learn visual models of object categories from just few images. They incorporate generic knowledge which may be obtained from previously learned models of unrelated categories, operating in a variational Bayesian framework. Object categories are represented by probability models, and prior knowledge is represented as a probability density function on the parameters of these models. The posterior model for an object category is obtained by updating the prior in the light of one or more observations.

A different approach is proposed by Leibe et al. (Leibe *et al.*, 2004). The method uses an implicit shape model which integrates category recognition and segmentation in a common probabilistic framework. The method consists of two different steps: first, it learns a codebook of local appearance that contains information on which local structures may appear on objects of the considered category. Then, it learns the implicit shape model that specifies where on the object the codebook entries may occur.

A common limitation of all these methods is that they can perform category detection only.

In this Section we present a probabilistic method for the recognition of visual categories. The approach we propose is general and does not depend on the particular visual category under consideration. The probability distribution of each category is modeled by SG-MRFs, and then each category view is classified as belonging to one of the modeled categories. The key feature of this approach is that it allows to learn the kernel from the training data, due to the properties of SG-MRFs. This is equivalent to learn, from a given training set, the optimal representation.
We present experiments that show the effectiveness of our approach.

6.3.1 Problem Statement

In what follows we will define a category as a class of objects that belong together. We will consider categories in which their members tend to be physically similar to one another, and we will refer to them as *visual categories* (see for instance Figure 6.3). This choice has two motivations: first, many studies support the idea that physical similarity represents a guide to categorization (Brooks, 1981). Second, this assumption enables us to recognize categories on the basis of their visual appearance.

As in Section (5.3.2), we will assume to have a set of objects $\Omega_k^h, h = 1, \ldots, \mathcal{H}$, each represented by a training set $\omega_k^h$. We will also suppose that the set of $\mathcal{H}$ objects can be divided into $\mathcal{K}_\mathcal{H}$ subset of visual categories $C_h : C_1 = \{\Omega_k\}_{k=1}^{K_1}, \ldots, C_{\mathcal{H}} = \{\Omega_k\}_{k=1}^{K_{\mathcal{H}}}$, with $C_1 \cup C_2 \cup \ldots \cup C_{\mathcal{H}} = \{\Omega_k\}_{k=1}^{K}$, $h = 1, \ldots, \mathcal{H}$. Our goal is to develop a strategy that enables us, given a sample object not belonging to the training set, to classify it as a member of one of the $\{C_h\}_{h=1}^{\mathcal{H}}$ visual categories.

The reader will notice that there is a similarity between the problem we face here and the one we approached in Section (5.3.2). In both cases the goal consists in recognizing objects as belonging to some visual categories. The main difference is that in Section 5.3.2 we wanted to recognize instances of objects already seen before, as belonging to a given visual category as well as a specific object. Thus, the final goal was object identification, and the challenge was to use category information to improve it. On the contrary, here we aim to build a probabilistic model for visual categories with the final goal to recognize objects never seen before. Thus, we explore the generalization capabilities of SG-MRF with respect to categorization.

A possible approach is the following (Schiele et al, 1997; Zhu et al, 1996). Suppose we have a set of objects $\Omega_k^h, k = 1, \ldots, \mathcal{K}$, members of the visual category $C_h$. Assume that we have a *sufficient statistic* for the category $C_h$. Then we can extract,
for each member, the visual information shared by all the members of the category. That information characterizes the visual category itself. Thus, a sample object $\Omega$ will be classified as a member of $C_h$ if it presents the visual information associated with $C_h$. This approach presents several limitations: first, it is completely heuristic how to define the visual information that characterizes a given visual category. Second, it is heuristic how to represent this visual information. Finally, this approach will lead in the general case to each visual category characterized by different visual information, represented differently. This opens the problem of how to make comparable this information when the sample object $\Omega$ is given.

A possible alternative approach was proposed by Perona and others (Weber et al., 2000a; Weber et al., 2000b; Fergus et al., 2003). They proposed a method to learn heterogeneous models of object classes for visual recognition, where the training images contain a preponderance of clutter and learning is unsupervised. Their model represents objects as probabilistic constellations of rigid parts (features); the variability within a class is represented by a joint probability density function on the shape of the constellation and the appearance of the parts. In a first stage, the method automatically identifies distinctive parts in the training set by applying a clustering algorithm to patterns selected by an interest operator. It then learns the statistical shape model using expectation maximization. The method obtains good classification results on categories like human faces, motorbikes, cars and planes. The method has some problems: first, it is computational expensive; second, it is heuristic how many distinctive parts to choose for each visual category, and this number is critical for the good performance of the algorithm. Finally, the method was tested for category detection only, and without cross validation; in order to evaluate better its relevance, it would be important to test it on recognition of one visual category between others, and to repeat the experiments for different partitions of the training and test set (cross validation).

Here we take a different approach. We will concentrate our attention on how similarities are measured between members of a category and between members of different categories. We argue that this approach overcomes the limitations of the approach illustrated above, as it does not present any heuristic procedure at any stage. We propose to learn from the data samples the similarity between members of a visual category using a probabilistic approach. This makes the approach general and independent from the particular visual category under consideration. Moreover, with this approach the characterization of a category can be directly compared with the characterization obtained for others; thus the method can be easily used for visual categories classification and detection.

6.3.2 The Probabilistic Approach.

Consider a set of visual objects $\Omega^h_k, k = 1, \ldots, K$, members of the visual category $C_h$. Consider also a set of data samples, $\{x^h_{k1}, \ldots, x^h_{kn_k}\}, x \in \mathbb{R}^m$ for the visual object $\Omega^h_k$. We will assume that $\{x^h_{k1}, \ldots, x^h_{kn_k}\}$ are a sufficient statistic for $\Omega^h_k$ and $\{\Omega^h_k\}_{k=1}^K$ are a sufficient statistic for $C_h$. Given these assumptions, it follows
that \( \{ x_{k1}^h, \ldots, x_{km}^h \}_{k=1}^K \) is a sufficient statistic for \( C_h \). Thus, we can consider \( \{ x_{k1}^h, \ldots, x_{km}^h \}_{k=1}^K \) as a random sample from the probability distribution \( P_{C_h}(x) \) defined on \( \mathbb{R}^m \), relative to the visual category \( C_h \). If we consider \( \mathcal{H} \) different visual categories and we are given a data sample \( x \), we will classify it as belonging to one of the \( \{ C_h \}_{h=1}^\mathcal{H} \) visual categories according to:

\[
h^* = \arg\max_{h=1}^{\mathcal{H}} P(C_h|x) = \arg\max_{h=1}^{\mathcal{H}} \{ P(x|C_h)P(C_h) \} = \arg\max_{h=1}^{\mathcal{H}} P(x|C_h),
\]

where we used a Maximum A Posteriori (MAP) criterion, Bayes rule and we assumed all the visual categories to be equiprobable (\( P(C_h) \) constant and equal).

We propose to model the parametric form of the probability function using Spin Glass-Markov Random Fields. The main reason why we use it here is that it enables to use the power of kernels in a probabilistic framework. The SG-MRF probability distribution is given by

\[
P(x|C_h) = \frac{1}{Z} \exp \left[ -E_{SGMRF}(x|C_h) \right],
\]

with

\[
E_{SGMRF} = - \sum_{h=1}^{p_h} \left[ K(x, x(\mu)) \right]^2, K(x, y) = \exp \{ -\rho d_{a,b}(x, y) \},
\]

and \( d_{a,b}(x, y) = \sum_i |x_i^a - y_i^b|^2 \) (Vapnik, 1998); \( \{ x(\mu) \}_{\mu=1}^{p_h}, h \in [1, \mathcal{H}] \) set of prototypes. Thus, the MAP classifier will become

\[
h^* = \arg\min_{h=1}^{\mathcal{H}} \left\{ - \sum_{\mu=1}^{p_h} [K(x, \bar{x}(\mu))]^2 \right\}
\]

We expect this probabilistic classifier to be particularly suitable for the recognition of visual categories, due to the functional form of the energy (2.4). This can be explained similarly as we explained the robustness properties of the algorithm, from a kernel point of view (see Section 4.3); indeed, recognition of visual categories needs robust algorithms, able to cope with visual variations of the instances to be recognized. Consider equation (4.2), that we rewrite here for sake of clarity:

\[
E_{SGMRF} = - \sum_{i,j=1}^{m} \Phi(x_i) J_{ij} \Phi(x_j), \quad J_{ij} = \sum_{\mu=1}^{p_h} \Phi(\bar{x}_i^{(\mu)}) \Phi(\bar{x}_j^{(\mu)}).
\]

This equation shows that the SG-MRF energy function is given by the sum, on all the components of the vector, of all the possible coupled interactions between components, weighted by a factor given by the corresponding element of the connection matrix. As the parametric form of the kernel is known, it is possible, for a
given set of visual categories, to select the kernel parameters such that they classify optimally between categories. This will correspond to use a connection matrix that will properly weight the information contained into the vector $x$. Note that this procedure does not depend upon any prior assumption on the nature of the considered categories; the only request is to have a sufficient statistic for the members of each considered category.

### 6.3.3 Experiments

We ran all the experiments on the Nelson database ((Nelson et al, 1998), see Section 3.9). As representation (chosen heuristically), we used 2D MFH, with filters given by Gaussian derivatives along $x$ and $y$ combined into a Laplace operator with $\sigma_1 = 1.4$ and $\sigma_2 = 2.8$. Resolution for histogram axis was of 16 bins. For the classification step, we used SG-MRF in the MAP framework described above, with a naive ansatz, and the $\rho, a, b$ in the Gaussian kernel learned as described in Section 3.7.

We ran a first experiment on the complete database (as to say object identification on 59 objects, training and test set as described in Section 3.9) in order to check whether the MFH representation we chose contains enough information to distinguish between objects belonging to the same visual category. If this is the case, then the capability of our model of classifying correctly visual categories will depend just on the selection of the proper kernel parameters $a, b$. The result of this experiment is a recognition rate of 97.83% with kernel parameters $a = 0.5; b = 1.4$.

A second set of experiments was done for the classification of visual categories. From the original database, we created a new one consisting of five visual categories: cups (11 objects), planes (12 objects, obtained by the union of planes and fighters), lizards (9 objects), snakes (8 objects) and sport cars (9 objects) (see Figure 6.4). Then we created 6 different partitions for the training and test set as follows: training set consisted of all the objects of a category minus one, and for each object we took all the views relative to that object. Test set consisted, for each category, of all the views of the object not in the training set. Thus, training and test set were disjoint. For the plane category we followed the same procedure, except that we considered as test set two objects (a plane and a fighter).

We ran a first experiment on partition1 for learning the kernel parameters, resulting in $a = 0.7, b = 1.0$. Results are reported in Table 6.2, first raw. Then, for each of the other 5 partitions (partition2, partition3, ...partition6), we ran experiments using the learned kernel parameters. Classification results obtained for each experiment, for each category, are reported in Table 6.2. We report also the average recognition rate obtained in these experiments. The average is computed in two different ways. In the first way (Table 6.2, average1) we consider all the results obtained in the six experiments. In the second way (Table 6.2, average2) we don’t consider for each visual category the best and the worst result. This should exclude the effect of one outliers as well as the effect of an object almost identical to one contained into the training set. We see that considering all the experiments,
we obtain an average recognition rate of 91.20 %. The best average recognition rate for category is obtained for lizards (98.72 %), followed by cups (97.63 %). The worst performance was obtained for planes (87.15 %). Note that the category plane has the highest number of objects (12), but it is obtained by the fusion of two categories, planes and fighters. Thus, it could have been expected to be the most challenging. It is interesting to note that with the second kind of average, categories behave differently. Those who already had a high recognition rate have an increase in performance smaller than the others. This seems to indicate that more sample objects would be needed for those categories, in order to have a sufficient statistic.

6.3.4 Discussion

In this Section we presented a probabilistic method for the recognition of visual categories. We modeled each category on the basis of the sufficient statistic of each member of the category, and we classified every new object using a Maximum A Posteriori probability classifier. We modeled the parametric form of the probab-
Table 6.2: Classification results for recognition of visual categories.

<table>
<thead>
<tr>
<th># experiment</th>
<th>cups (%)</th>
<th>planes (%)</th>
<th>lizards (%)</th>
<th>snakes (%)</th>
<th>sportcars (%)</th>
<th>total (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>part1</td>
<td>100</td>
<td>56.77</td>
<td>98.72</td>
<td>100</td>
<td>58.97</td>
<td>76.11</td>
</tr>
<tr>
<td>part2</td>
<td>93.55</td>
<td>100</td>
<td>93.59</td>
<td>100</td>
<td>94.87</td>
<td>97.28</td>
</tr>
<tr>
<td>part3</td>
<td>95.48</td>
<td>91.61</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>95.28</td>
</tr>
<tr>
<td>part4</td>
<td>100</td>
<td>76.77</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>89.70</td>
</tr>
<tr>
<td>part5</td>
<td>97.42</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>97.44</td>
<td>99.14</td>
</tr>
<tr>
<td>part6</td>
<td>99.35</td>
<td>97.74</td>
<td>100</td>
<td>26.92</td>
<td>91.03</td>
<td>89.70</td>
</tr>
<tr>
<td>avg1</td>
<td>97.63</td>
<td>87.15</td>
<td>98.72</td>
<td>87.82</td>
<td>90.38</td>
<td>91.20</td>
</tr>
<tr>
<td>avg2</td>
<td>98.06</td>
<td>91.53</td>
<td>99.68</td>
<td>100</td>
<td>95.84</td>
<td>97.02</td>
</tr>
</tbody>
</table>

This work can be developed in many ways: first, it would be important to test the method on bigger databases and in realistic scenarios. Second, the approach should be benchmarked with other probabilistic and kernel methods presented in literature. Finally, we plan to test K-CSC for categorization; our expectation (based on results reported in the previous Section) is that using category-specific kernels will lead to performances competitive with state of the art algorithms for categorization.

### 6.4 Statistical Mechanics of Kernel Associative Memories

Learning and recognition in the context of neural networks is an intensively studied field. A lot of work has been done on networks in which learning is Hebbian, and recognition is represented by attractor dynamics of the network. Particularly, the Hopfield-Little model (H-L, (Hopfield, 1982; Little, 1974)) is a network exhibiting associative memory based on the Hamiltonian

\[ H = -\frac{1}{2N} \sum_{\mu=1}^{P} \sum_{i \neq j}^{N} s_i s_j \xi_i^{(\mu)} \xi_j^{(\mu)} , \]  

(6.15)

where the \( s_i \) are \( N \) dynamical variables taking on the values \( \pm 1 \) and the \( \xi_i^{(\mu)} \) (with \( \xi_i^{(\mu)} \) \( \pm 1 \)) are \( P \) fixed patterns which are the memories being stored. The storage capacity and thermodynamic properties of this model have been studied in detail within the context of spin glass theory (Mezard et al, 1987); it was found that at \( T = 0 \), for \( P < 0.14N \), memory patterns can be recovered with an accuracy better than 97%. At higher \( P \) or higher temperature there is a transition to a spin glass
6.4. STATISTICAL MECHANICS OF KERNEL ASSOCIATIVE MEMORIES

state in which memory patterns cannot be recovered.

Many authors studied generalizations of the H-L model which includes interactions between \( p \) (> 2) Ising spins ((Abbott et al, 1987; Gardner, 1987)), They consider a generalized Hamiltonian of the form

\[
H_{\text{Gen}} = - \sum_{i_1 < \ldots < i_p} J_{i_1 \ldots i_p} s_{i_1} \ldots s_{i_p},
\]

(6.16)

thus excluding all terms with at least two indexes equal (symmetric terms). As a general result, higher order Hamiltonians present an increase in the storage capacity compared to the H-L model; the price to be paid is an increase in the computer memory required to store the interaction matrix (Gardner, 1987).

As we pointed out through this thesis, if the symmetric term is included in the Hamiltonian (6.15), it can be written as a function of the scalar product between \( \xi^{(\mu)} \) and \( s \). The Euclidean scalar product can thus be substituted by a Mercer kernel (Vapnik, 1998), providing a new higher order generalization of the H-L energy. We call this new model Kernel Associative Memory (KAM). There are several reasons for considering this generalization. First, we will show that this model presents a storage capacity of \( P < \alpha_c(p)N^{p-1} \), thus much higher than the storage capacity of the H-L model. This increase is similar to that found for higher order Hamiltonians (6.16) (see (Abbott et al, 1987; Gardner, 1987; Baldi et al, 1987)), but, for the KAM model, kernel properties could be used in order to avoid the increase in the computer memory for storing the \( J_{i_1 \ldots i_p} \). Second, the higher order generalization of the H-L model via Mercer kernel gives the possibility to study a much richer class of models, due to the variety of possible Mercer kernels (Vapnik, 1998). Here we study the storage capacity and thermodynamic properties of KAM for polynomial and Gaussian kernels, in a replica symmetry ansatz (Parisi, 1980; Mezard et al, 1987).

To our knowledge, no previous works has considered the storage capacity of such a generalization of the H-L model. We find that the storage capacity increases with respect to the H-L one, but is lower compared to that of the higher order Hamiltonians (6.16) not including symmetric terms.

6.4.1 Kernel Associative Memories

The H-L energy (6.15) can be rewritten in the equivalent form

\[
H = -\frac{1}{2} \sum_{\mu=1}^{P} \left[ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} s_{i} \xi_{i}^{(\mu)} \right]^2 - 1.
\]

(6.17)

This energy can be generalized to higher order correlations via Mercer kernels (Vapnik, 1998):

\[
H_{\text{KAM}} = -\frac{1}{2} \sum_{\mu=1}^{P} K \left( \frac{1}{\sqrt{N}} s, \xi^{(\mu)} \right).
\]

(6.18)
The possibility to kernelize the H-L energy function was recognized first in (Caputo et al, 2001). We call this model Kernel Associative Memory (KAM). It is fully specified once the functional form of the kernel is given. In this paper we will consider polynomial and Gaussian kernels:

$$K_{\text{poly}}(x, y) = (x \cdot y)^p, \quad K_{\text{Gauss}}(x, y) = \exp(-\rho \|x - y\|^2). \quad (6.19)$$

Our goal is to study the storage capacity of energy (6.18) for kernels (6.19), using tools of statistical mechanics of spin glasses. To this purpose, we note that the study of energy (6.18) can be done for both kernels (6.19) considering the general case

$$H = -\frac{1}{N^{1-p/2}} \sum_{\mu=1}^P F \left( \frac{1}{\sqrt{N}} \sum_{i=1}^N s_i \xi_i^{(\mu)} \right), \quad F(x) = A_p x^p. \quad (6.20)$$

For $A_p = 1$, $p$ finite, equations (6.20) represent the KAM energy (6.18) for polynomial kernels (6.19). For $A_p = p^p / p!$, $p \to \infty$, equations (6.20) describe the behavior of the KAM energy (6.18) for Gaussian kernels (6.19). This can be shown as follows: note first that

$$\exp(-\rho \|s - \xi^{(\mu)}\|^2) = \exp(-\rho [s \cdot s + \xi^{(\mu)} : \xi^{(\mu)} - 2s \cdot \xi^{(\mu)}]) = \exp(-2\rho [N - s \cdot \xi^{(\mu)}]).$$

The multiplicative factor can be inglobated in the $\rho$, and the constant factor can be neglected. The Gaussian kernel function thus becomes

$$\exp(-\rho \|s - \xi^{(\mu)}\|^2) \rightarrow \exp(\rho s \cdot \xi^{(\mu)}) \approx 1 + \rho s \xi^{(\mu)} + \frac{\rho^2}{2!} (s \xi^{(\mu)})^2 + \frac{\rho^3}{3!} (s \xi^{(\mu)})^3 + \ldots$$

A generalization of the H-L energy function in the form (6.20) was proposed first by Abbott in (Abbott et al, 1987). In that paper, storage capacity and thermodynamic properties were derived for a simplified version of (6.20), where all terms with at least two equal indexes were excluded; a particular choice of $A_p$ was done. Other authors considered this kind of simplifications for higher order extensions of the H-L models (see for instance (Gardner, 1987)), which can be described in the most general way by equation (6.16) with an appropriate definition of $J_{i_1 \ldots i_p}$. The analysis we present here include all the terms in the energy and is not limited to a particular choice of the coefficient $A_p$; thus is more general. To the best of our knowledge, this is the first analysis on the storage capacity and thermodynamic properties of a generalization of the H-L model via kernel functions.

### 6.4.2 Free Energy and Order Parameters

We study the overlap of a configuration $s_i$ with one of the stored patterns, arbitrarily taken to be $\xi_i^{(1)}$, defined as (Amit et al, 1985)

$$m = \left\langle \frac{1}{N} \sum_{i=1}^N \langle s_i \rangle \xi_i^{(1)} \right\rangle, \quad (6.21)$$
6.4. STATISTICAL MECHANICS OF KERNEL ASSOCIATIVE MEMORIES

where the angle bracket $\langle \ldots \rangle$ represents a thermodynamic average while the double brackets $\{\ldots\}$ represents a quenched average over the stored patterns $\xi_i^{(\mu)}$. The quenched average over patterns is obtained using the replica methods (Parisi, 1980; Mezard et al, 1987). In the mean-field approximation, the free energy depends on $m$ and on the order parameters (Amit et al, 1985)

$$q_{ab} = \left\langle \left\langle \frac{1}{N} \sum_i^N \langle s_i^a \rangle \langle s_i^b \rangle \right\rangle \right\rangle, r_{ab} = \frac{1}{MN} \sum_{\mu=2}^M \left\langle \left\langle \sum_i^N \langle s_i^a \rangle \xi_i^{(1)} \sum_j^N \langle s_j^b \rangle \xi_j^{(1)} \right\rangle \right\rangle,$$

(6.22)

with $a,b$ replica indexes. The calculation we present below is analogous to the one obtained in (Abbott et al, 1987; Amit et al, 1985; Amit, 1989).

Considering a replica symmetry ansatz (Parisi, 1980; Mezard et al, 1987), $q_{ab} = q, r_{ab} = r$; we can derive the free energy at temperature $T = 1/\beta$:

$$f = (p - 1)A_p m^p + \frac{\alpha \beta}{2} [r(1 - q) - G(q)] - \frac{1}{\beta} \int Dz \ln[2 \cosh \beta(\sqrt{\alpha} rz) + pA_p m^{p-1}],$$

(6.23)

with $P = \alpha N^{p-1}$ and $Dz = \frac{dz}{\sqrt{2 \pi}} e^{-z^2/2}$. The function $G(q)$ is given by

$$G(q) = A_p^2 \sum_{c=1}^p \frac{1}{2} \left[ 1 + (-1)^{p+c} \right] \left( \frac{p!}{(p-c)!} \right)^2 (1 - q^c).$$

(6.24)

From the free energy we can compute the following order parameters:

$$m = \int Dz \tanh(\sqrt{\alpha} rz + pA_p m^{p-1}),$$

(6.25)

$$r = -\frac{\partial G(q)}{\partial q},$$

(6.26)

$$q = \int Dz \tanh^2(\sqrt{\alpha} rz + pA_p m^{p-1})$$

(6.27)

At all temperatures, these equations have a paramagnetic solution $r = q = m = 0$.

6.4.3 The Zero-temperature Limit: $\beta \to \infty$

In the zero-temperature limit $\beta \to \infty$, we see that equations (6.25) - (6.27) simplify to

$$q \to 1,$$

(6.28)

$$m \to \text{erf} \left( \frac{pA_p m^{p-1}}{\sqrt{\alpha}r} \right), \quad r \to \frac{1}{2} A_p^2 \sum_{c=1}^p R_c(p)$$

(6.29)
with

\[
R_c(p) = [1 + (-1)^{p+c}] \frac{c}{c!} \left[ \frac{p!}{(p-c)!!} \right]^2
\]  \hspace{1cm} (6.30)

Note that, in this approximation, \(m\) does not depend on \(A_p\):

\[
m = \text{erf} \left( \frac{pA_p m^{p-1}}{\sqrt{\alpha r}} \right) = \text{erf} \left( \frac{p m^{p-1}}{\sqrt{\alpha} \frac{A_p}{A_p \sqrt{\sum_{c=1}^{p} R_c(p)/2}}} \right).
\]  \hspace{1cm} (6.31)

Graphical solutions of equation (6.31) are shown in Figure 6.5-6.6, for several values of \(p\) (odd and even). From these graphics we can make the following considerations: first, higher order Hamiltonians (6.20) with \(p\) odd eliminate the symmetry between the stored memory and their complementary states, in which the state of each neuron is reversed (see Figure 6.6). Second, as \(p\) increases, the value of \(\alpha\) for
Figure 6.6: Graphical representation of the solutions of the retrieval equation (6.32) for $p = 4, 6, 8$ and $\alpha = 0.1$ (left), $\alpha = 0.01$ (top right), $\alpha = 0.001$ (bottom left), $\alpha = 0.0001$ (right).

which there is more than one intersection ($m = 0$, the spin-glass solution (Amit et al, 1985)), as to say the storage capacity, goes to zero (see Figure 6.5-6.6). Figure 6.7 shows the percentage of errors $(1 - m)/2$ made in recovering a particular memory configuration as a function of $\alpha$. The percentage of errors for the KAM model, for $p = 3, 5$, is compared with the same quantity obtained by Abbott (Abbott et al, 1987), considering a higher order Hamiltonian. In that case $m$ is given by (Abbott et al, 1987):

$$m^{Abbott} = \text{erf} \left[ \left( \frac{p}{2\alpha t!} \right)^2 m^{p-1} \right]$$

(6.32)

in the zero temperature limit. In both cases ($p=3,5$) and for both models (KAM and Abbott), the overlap with the input pattern $m$ remains quite close to 1 even for $\alpha \rightarrow \alpha_c$. This can be seen because the fraction of errors is always small. Thus
the quality of recall is good. Nevertheless, the fraction of errors is smaller for the Abbott model as \( p \) increase.

Even if \( \alpha_c \to 0 \) for large \( p \), the total number of memory states allowed is given by \( P = \alpha N^{p-1} \), thus it is expected to be large. In the limit \( p \to \infty \), \( \alpha_c \) and \( P \) can be calculated explicitly (Amit, 1989): for large values of the argument of the \( erf \) function, it holds

\[
m \approx 1 - \frac{\sqrt{\alpha/2} \sum_{c=1}^{p} R_c(p)}{\sqrt{\pi} pm^{p-1}} \exp \left[ - \frac{p^2 m^{p-1}}{\alpha/2 \sum_{c=1}^{p} R_c(p)} \right]. \tag{6.33}
\]

For stability considerations, \( m \approx 1 \), thus the second term in equation (6.33) must go to zero. This leads to the critical value for \( \alpha \)

\[
\alpha_c = \frac{2p^2}{\sum_{c=1}^{p} R_c(p) \ln p} \approx \frac{2p^2}{p! 2p \ln p [\sum_{c=1}^{p} c(1 + (-1)^{c+p})]} \tag{6.34}
\]

The total number of memory states that is possible to store is (using Stirling’s approximation):

\[
P = \alpha N^{p-1} \to \left( \frac{e N}{p} \right)^{p-1} \sqrt{\frac{2p}{\pi}} \frac{e}{2p \ln p \sum_{c=1}^{p} (1 + (-1)^{c+p})}. \tag{6.35}
\]

This result must be compared with the one obtained by Abbott (Abbott et al, 1987):

\[
\alpha_c^{Abbott} = \frac{p}{2p! \ln p}; \tag{6.36}
\]

it is easy to see that

\[
\alpha_c = \alpha_c^{Abbott} \cdot \frac{p}{2p \sum_{c=1}^{p} c[1 + (-1)^{c+p}]} \tag{6.37}
\]

Figure 6.7: The error rate \((1-m)/2\) as a function of \( \alpha \), for KAM and Abbott model, for \( p = 3 \) (left) and \( p = 5 \) (right). For \( p = 3 \) the two curves are indistinguishable, for \( p = 5 \) the Abbott approximation gives the lower error rate.
6.5. SUMMARY

thus, the introduction of terms in which more than one index is equal in the generalized Hamiltonian leads to a decrease in the storage capacity; the decrease will be higher the higher is the number of these terms introduced in the energy formulations, according with the retrieval behavior (see Figure 6.7).

6.4.4 Discussion

In this Section we considered kernel associative memories as a higher order generalization of the H-L model. Their storage capacity is studied in a replica symmetry ansatz. The main result is that the storage capacity is higher than the H-L’s one, but lower than the storage capacity obtained by other authors for different higher order generalizations. This work can be developed in many ways: first of all, the mean field calculation presented here is only sensitive to states that are stable in the thermodynamic limit. We expect to find in simulations correlations between spurious states and the memory patterns even for $\alpha > \alpha_c$. Simulations should also provide informations about the size of the basin of attraction. Second, replica symmetry breaking effects should be taken in account. Third, it should be explored how kernel properties can be used in order to reduce the memory required for storing the interaction matrix (Gardner, 1987). Finally, this study should be extended to other classes of kernels and to networks with continuous neurons. Future work will be concentrated in these directions.

6.5 Summary

In this Chapter we presented three research contributions for computer vision, machine learning and statistical mechanics of spin glasses. For all of them, Spin Glass-Markov Random Fields were the key element, and all of them open many perspectives for new research; new with respect to the topic, and new with respect to the approach. These results underline the versatility of the model and its potential; they also emphasize how the three above mentioned research fields, with which Spin Glass-Markov Random Fields are linked, can benefit from its development.
Chapter 7

Conclusion and Perspective

This Chapter summarizes the main results and contributions presented in this thesis, discusses the open issues and sketches possible future directions of research. Section 7.1 summarizes the main contributions to the field of object recognition, underlying theoretical as well experimental findings. Section 7.2 discusses controversial results and suggests future lines of research.

7.1 Summary

Recognizing objects on the basis of their visual appearance is one of the major goals in computer vision. This task has shown to be challenging mainly because of the large variability in objects appearance. This variability depends on many factors: noise, occlusions, cluttered background, changes in lighting condition, scale variations and so on. A vast literature in this field shows that both local and global approaches can achieve promising results, and can cope with some of the problems listed above (Barla et al, 2002; Leonardis et al, 2000; Matas et al, 1995; Nelson et al, 1998; Roobaert et al, 2001; Slater et al, 1995; Schmid et al, 1996). The goal of this thesis has been to develop a new probabilistic method for object recognition which allowed us to take into account local and global visual properties of objects. We assumed that objects were represented by their visual appearance, and that, for each object represented by a collection of views, we were able to extract meaningful features. The focus has been on how to develop an effective probabilistic model on the basis of given features, rather than defining new descriptors. Throughout the thesis, we explicitly addressed the issue of noise, occlusion and cluttered background.

The main contribution of this thesis consists of deriving a new class of probabilistic models which we have called Spin Glass-Markov Random Fields. This model has been developed on the basis of knowledge and results achieved in computer vision, machine learning and statistical physics of spin glasses (Mezard et al, 1987). This results in a probability distribution estimated via a fully connected Markov
random field. Full connectivity allows the model to take into account at the same time global and local characteristics of the visual pattern under consideration; hence the novelty of the approach. We have used an energy function derived from results of statistical physics of spin glasses, which has made it possible to benefit from theoretical analysis made within the physics research community on those classes of energies; the combination of Markov random fields and spin glass-like energies has been obtained via nonlinear kernel functions.

Object recognition has been the visual task central to this thesis, both as final application where to demonstrate the model and as motivation for the development of the model itself. Thus, we tested the effectiveness of Spin Glass-Markov Random Fields for object recognition on several series of experiments, increasingly challenging.

As a first proof of concept, we tested the suitability of Spin Glass-Markov Random Fields for appearance-based object recognition on a set of publicly available benchmark databases (Chapter 3). In all of them, objects views are taken in homogeneous background; while this makes the recognition task too easy and somehow unrealistic compared to real world conditions, still these experiments constitute a first and important way to evaluate a new object recognition algorithm with respect to those presented in the literature. We performed experiments on different databases with different kinds of features. Results were benchmarked with at least 2 other classification methods. In most cases, Spin Glass-Markov Random Fields obtained the best performance.

A key feature of any object recognition algorithm aiming to perform well in realistic scenarios is robustness. Objects appear in cluttered backgrounds, often occluded and under varying lighting conditions, and recording media can induce several types of noise in the image. We addressed the issue of robustness of Spin Glass-Markov Random Fields from a theoretical and experimental point of view (Chapter 4). We conducted a theoretical analysis on robustness of Spin Glass-Markov Random Fields to degraded visual information, using tools of statistical physics of spin glasses. We considered the case of degradation into the training and test set, and we found that adding degraded views in the training set leads to a higher robustness. In order to validate these findings from an experimental point of view, we performed an extensive series of experiments to test robustness of Spin Glass-Markov Random Fields with respect to noise, occlusion, decreasing size of training set and heterogeneous background. Experimental results have shown that Spin Glass-Markov Random Fields are highly robust to noise, occlusion and decreasing number of training views. Results obtained on heterogeneous background are weak, and we will discuss them in Section 7.2.

An important but somehow neglected issue in object recognition is how to use multiple cues. While the research community seems to generally agree that using multiple cues leads to a better performance (both from the point of view of higher recognition rate and of greater robustness), it is still unclear how to combine together several cues, given a specific task. Interestingly, a family of spin glass-like energy functions suggests a new and highly effective way to answer these questions
We derived theoretically an extension of Spin Glass-Markov Random Fields that presents a hierarchical structure. This new model is based on ultrametric properties of spin glass-like energy functions, thus we called it Ultrametric Spin Glass-Markov Random Field. This new energy function allows to use multiple cues separately at the modeling stage, but together at the classification stage. The result is a soft combination scheme which takes into account every cues contribution at every decision step, rather then relying on just one (as it would be the case by using a voting scheme (Duda et al., 2001)). As a consequence, the algorithm is capable to classify correctly views misclassified (possibly badly) by all the available cues. We performed experiments using color and textural/shape features, and compared the performance of Ultrametric Spin Glass-Markov Random Fields with that of Spin Glass-Markov Random Fields and other state-of-the-art algorithms. Results show that Ultrametric Spin Glass-Markov Random Field outperforms other methods.

Multi-cue integration is related to the issue of cue selection; the final contribution of this thesis addresses the problem of class specific feature selection. Different types of objects can be effectively represented and detected using a specific set of features; but when the recognition problem at hand involves several object classes, class specific features have to be abandoned, in favor of a common feature set. Here we have presented a possible solution to this problem: we derived a Kernel-Class Specific Classifier that allows to perform multi class object recognition using specific features for specific object classes (Chapter 6). The new classifier is derived by merging together Spin Glass-Markov Random Fields and Class Specific Classifiers. The class specific features are preserved by introducing a reference hypothesis class, common to all the other object classes. We performed experiments to compare the performance of the new classifier with Spin Glass-Markov Random Fields; results show a clear increase in performance for Kernel-Class Specific Classifier.

In conclusion, we can state that Spin Glass-Markov Random Fields (and its related algorithms) are a promising new probabilistic model for robust recognition of objects in real world scenes, and constitutes a valid alternative to state of the art algorithms for visual pattern recognition.

7.2 Open Issues

An open issue for Spin Glass-Markov Random Field is how to handle object recognition in heterogeneous background. We performed several sets of experiments, with different databases, in different setups and using different strategies (Chapter 3-5). On the basis of the results obtained, we can identify two different situations:

1. Prior Knowledge on the background, and/or view containing mostly the object to be recognized.
   In this case, Spin Glass-Markov Random Field can handle successfully heterogeneous background. If there is no prior knowledge on the background, but the object takes most of the view, then it is possible to consider heterogeneous background as a degradation of the visual information. Experiments
show that, in this case, the system still obtains good performances. When there is prior knowledge on the background, then it is possible to use it in the Ultrametric Spin Glass-Markov Random Field framework. Experiments show in this case a clear increase in performance with respect to Spin Glass-Markov Random Fields.

2. **No prior knowledge on the background, and/or views containing at least 50% of non-object information.**

In this case, Spin Glass-Markov Random Field gives low accuracy. When the object still takes most of the view, it is possible to use robustness properties of Spin Glass-Markov Random Fields, although results are not as good as in the previous case. But in the case of small objects into the database, the performance can decrease quite dramatically.

A possible explanation for this behavior is that Spin Glass-Markov Random Field is a global/local method: it is global because it is a fully connected Markov random field, and it is local because globality is achieved as the sum of all the learned significant localities. When there is a prior knowledge on the background surrounding the scene, it can be treated as context, and the algorithm can take advantage of it. When there is no prior knowledge, but the object views contain mostly the object to be recognized, then background is treated as some sort of noise, and the algorithm can handle it. But when the background is very textured, and it is the dominant visual information, the algorithm detects too many potentially significant localities, and performance degrades. A possible solution could be to do Markov random field modeling on local features rather than on global ones, as done in this thesis. In this case, the challenge would be to define suitable Mercer kernels for local features. In addition to this, we can expect anyway that, with very textured backgrounds, performance would decrease, even if not as dramatically as with global features.

A most promising strategy seems to decouple the problem, and use first attention strategies (Ramstrom et al., 2002; Bjorkman et al., 2002). This would facilitate getting rid of most of the background, and at the same time to zoom on the object of interest and obtain views containing mostly the object itself. As described above, with such a strategy Spin Glass-Markov Random Field has shown to perform successfully, and better than state-of-the-art algorithms.

Future work should include strategies for recognition in heterogeneous background, but not be limited to it. Some important theoretical points should be further investigated: first of all, the choice of prototypes. Selecting prototypes from the training set is equivalent to select key views for representing the considered object. It is important to study how the naive ansatz is related to existing work on this topic, and more generally how Spin Glass-Markov Random Fields can benefit from results achieved in this research area. Second, Kernel-Class Specific Classifier is a powerful algorithm, but still at a “proof of concept” level. It is still unclear how to learn efficiently the kernel parameters for each object class, and how to choose the reference hypothesis in a less heuristic manner. While a possible
solution for the reference hypothesis is to limit the choice to the object class to be recognized, the issue of learning efficiently the kernel parameters is still totally open.

From an experimental point of view, the Ultrametric Spin Glass-Markov Random Field approach to object recognition is extremely powerful; it should be tested against decision-tree algorithms, and should be extended so to include robustness properties of Spin Glass-Markov Random Field. All of this, with the final goal to merge together Kernel-Class Specific Classifiers with robust Ultrametric Spin Glass-Markov Random Fields, and use the new model in combination with attention strategies.
Appendix A

Connection Matrix and Associated Energy

In this Appendix we discuss in more details the H-L conditions on the connection matrix $J_{ij}$ and the properties of the associated energy (Amit, 1989).

In order to store $p$ prototypes as absolute minima of the energy (2.4), the $J_{ij}$ will be given by

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_{i}^{(\mu)} \xi_{j}^{(\mu)}$$

The configuration vector representing a prototype will be denoted by

$$\xi^{(\mu)} = (\xi_{1}^{(\mu)}, \xi_{2}^{(\mu)}, \ldots, \xi_{i}^{(\mu)}, \ldots, \xi_{N}^{(\mu)}),$$

where $\mu$ labels the different prototypes and $\xi_{i}^{(\mu)} = \pm 1$, chosen randomly, independently with equal probability, as to say

$$P(\xi_{i}^{(\mu)}) = \frac{1}{2} \delta(\xi_{i}^{(\mu)} - 1) + \frac{1}{2} \delta(\xi_{i}^{(\mu)} + 1).$$

This choice of the $\xi_{i}^{(\mu)}$’s implies that:

- the stored prototypes are unbiased, as to say, in a large network ($N \to \infty$)

$$\langle \langle \xi_{i}^{(\mu)} \rangle \rangle = \frac{1}{N} \sum_{i} \xi_{i}^{(\mu)} = 0; \quad (A.1)$$

- different stored prototypes are uncorrelated, as to say in a large network

$$\langle \langle \xi_{i}^{(\mu)} \xi_{i}^{(\nu)} \rangle \rangle = \frac{1}{N} \sum_{i} \xi_{i}^{(\mu)} \xi_{i}^{(\nu)} = 0, \quad \mu \neq \nu. \quad (A.2)$$
The double brackets indicate an average over the prototypes’ distribution; the equivalences in equations (A.1)-(A.2) are valid under the condition that \( p \) remains finite as \( N \to \infty \) (self averaging; see (Amit, 1989), chap 4, pag 187). Thus, these conditions can be summarized as

- \( a) \xi^{(\mu)} \perp \xi^{(\nu)}, \quad \mu \neq \nu, \quad \sum_i \xi^{(\mu)}_i = 0, \quad \forall \mu = 1, \ldots, p; \)
- \( aa) \ p \leq 0.138N, \quad N \to \infty, \)

where condition \( aa) \) integrates the self averaging condition with replica analysis results ((Amit, 1989), chap 6).

The first question about the connection matrix (2.6) is whether, with the conditions \( a), aa) \) holding, the set of prototypes \( \{\xi^{(\mu)}\}_{\mu=1}^p \) actually are stable configurations. The stability condition for a single site is given by

\[
s_i h_i > 0, \quad \forall i = 1, \ldots, N, \quad \text{(A.3)}
\]
as to say, a site is in a stable configuration if the neighborhood interactions, represented by \( h_i \), do not make it change its state. Thus, in order to verify that the \( \{\xi^{(\mu)}\}_{\mu=1}^p \) are stable, it is sufficient to compute (A.3) on one of the prototypes of the set that we will call \( \xi^{(\mu_0)} \) (due to the symmetry of \( J \), what holds for one prototype will hold for any other):

\[
h_i s_i = \left( \frac{1}{N} \sum_{j=1}^{N} \sum_{\mu=1}^{p} \xi^{(\mu)}_i \xi^{(\mu)}_j \xi^{(\mu_0)}_j \right) \cdot \xi^{(\mu_0)}_i =
\]

\[
= \frac{1}{N} \left( \sum_{j=1}^{N} \xi^{(\mu_0)}_i \xi^{(\mu_0)}_j \xi^{(\mu_0)}_j \right) + \frac{1}{N} \left( \sum_{j=1}^{N} \sum_{\mu=1}^{p} \xi^{(\mu)}_i \xi^{(\mu)}_j \xi^{(\mu_0)}_j \xi^{(\mu_0)}_i \right) =
\]

\[
= 1 + \frac{1}{N} \left( \sum_{j=1}^{N} \sum_{\mu=1}^{p} \xi^{(\mu)}_i \xi^{(\mu)}_j \xi^{(\mu_0)}_j \xi^{(\mu_0)}_i \right) = 1 + R_i.
\]

The stability condition thus becomes

\[
1 + R_i > 0, \quad \forall i = 1, \ldots, N. \quad \text{(A.4)}
\]
The sum over \( \mu \) has been separated into two parts: a signal term, corresponding to the considered prototype, and a noise term, due to the contribution of all the other prototypes. The quantities

\[
R_i = \frac{1}{N} \left( \sum_{j=1}^{N} \sum_{\mu=1}^{p} \xi^{(\mu)}_i \xi^{(\mu)}_j \xi^{(\mu_0)}_j \xi^{(\mu_0)}_i \right) \quad \text{(A.5)}
\]
are random variables; in order to check whether the stability condition is satisfied, we must compute their mean value and variance. The mean value is

\[ E[R_i] = \frac{1}{N}E\left(\sum_{j=1}^{N} \sum_{\mu=1}^{N} \xi_i^{(\mu)} \xi_j^{(\mu)} \xi_j^{(\mu_0)} \xi_i^{(\mu_0)}\right) = \frac{1}{N} \sum_{j=1}^{N} \sum_{\mu=1}^{N} E[\xi_i^{(\mu)}] E[\xi_j^{(\mu)}] E[\xi_j^{(\mu_0)}] E[\xi_i^{(\mu_0)}] = 0 \]

The result is due to condition \(a\), which allows to factorize the expectation value of products in products of expectation values. The variance will be given by

\[ D[R_i] = E[(R_i - E[R_i])^2], \]

thus

\[ D[R_i] = E\left[\left(\frac{1}{N} \sum_{j=1}^{N} \sum_{\mu=1}^{N} \xi_i^{(\mu)} \xi_j^{(\mu)} \xi_j^{(\mu_0)} \xi_i^{(\mu_0)}\right)^2\right] = \frac{1}{N^2} E\left[\left(\sum_{j_1} \sum_{j_2} \sum_{\mu=1}^{N} \sum_{\mu \neq \mu_0} \xi_i^{(\mu_1)} \xi_j^{(\mu_1)} \xi_j^{(\mu_0)} \xi_i^{(\mu_0)} \xi_i^{(\mu_1)} \xi_j^{(\mu_1)} \xi_j^{(\mu_0)} \xi_i^{(\mu_0)}\right)^2\right] \quad (A.6) \]

The sums in equation \(A.6\) can be separated in different terms, some of which will be \(= 0\), again due to condition \(a\). The sum over \(\mu_1\) and \(\mu_2\) can thus be separated in two parts, one with all different indexes

\[ \xi_i^{(\mu_0)} \xi_j^{(\mu_0)} \sum_{\mu_1 \neq \mu_0} \sum_{\mu_2 \neq \mu_0} \sum_{\mu_2 \neq \mu_1} \xi_i^{(\mu_1)} \xi_j^{(\mu_1)} \xi_j^{(\mu_0)} \xi_i^{(\mu_0)} \xi_i^{(\mu_1)} \xi_j^{(\mu_1)} \xi_j^{(\mu_0)} \xi_i^{(\mu_0)} \quad (A.7) \]

and another which contains terms with \(\mu_1 = \mu_2\):

\[ \xi_i^{(\mu_0)} \xi_j^{(\mu_0)} \sum_{\mu \neq \mu_0} \xi_i^{(\mu)} \xi_j^{(\mu)} \xi_j^{(\mu_0)} \xi_i^{(\mu_0)} \xi_i^{(\mu)} \xi_j^{(\mu)} \xi_j^{(\mu_0)} \xi_i^{(\mu_0)} \quad (A.8) \]

The expectation value of the term \(A.7\) is \(= 0\), because all the quantities involved in the sum have different indexes, thus we can factorize the expectation value of the products as done before; the factor \(\xi_i^{(\mu_0)} \xi_j^{(\mu_0)}\) in front of the sum is \(= 1\) and doesn’t affect the argument. The variance thus becomes:

\[ D[R_i] = E[R_i^2] = \frac{1}{N^2} E\left(\sum_{j_1} \sum_{j_2} \sum_{\mu \neq \mu_0} \xi_j^{(\mu)} \xi_j^{(\mu_0)} \xi_j^{(\mu)} \xi_j^{(\mu_0)}\right) \]
Now we play with the sum over \( j_1, j_2 \) as we did for the sum (A.6); the final result is:

\[
D[R_i] = \frac{1}{N^2} \sum_{j_1} \sum_{j_2 \neq j_1} \sum_{\mu \neq \mu_0} E[\xi_{j_1}^{(\mu)}] E[\xi_{j_2}^{(\mu_0)}] E[\xi_{j_1}^{(\mu)}] E[\xi_{j_2}^{(\mu_0)}] + \frac{1}{N^2} E \left[ \sum_j \sum_{\mu \neq \mu_0} \xi_j^{(\mu)} \xi_j^{(\mu_0)} \right] \]

\[= \frac{N(p - 1)}{N^2} \approx \frac{p}{N}.\]

The mean value of \( R_i \)'s fluctuations is proportional to \( \sqrt{D[R_i]} \), as to say \( \sqrt{p/N} \). If condition \( aa) \) holds, it goes to zero, and the stability condition (A.4) is satisfied. This implies that every prototype is a fixed point. We underline that both conditions \( a), aa) \) are necessary in order to achieve this result.

Moreover, the prototypes are very stable fixed points. Let us change the values of a finite fraction \( d \) of sites of the prototype \( \xi^{(\mu_0)} \) (Amit, 1989); then (A.4) becomes

\[
(1 - 2d) + R_i > 0, \quad \forall i = 1, \ldots, N,
\]

and, in the condition in which the noise term vanishes, the deteriorated configuration will still lead to an equilibrium state: in other words, the prototypes have large basins of attraction. Finally, let us say a few words on the condition \( aa) \):

\[
p = \alpha N, \quad \alpha \leq \alpha_c = 0.138, \quad N \to \infty.
\]

Applying the signal to noise analysis, this condition becomes \( p \ll N \), as to say if the number of stored prototypes is finite and fixed, the noise term (A.5) will vanish and the condition stability (A.3) will be satisfied. Condition \( aa) \) is stronger, however: it tells us that the number of prototypes can also go to infinity as \( N \) does, just slower; the coefficient of proportionality has to be smaller than a threshold value \( \alpha_c \) (Amit, 1989). If this condition is not satisfied, the noise term becomes dominant and the energy landscape changes its properties very dramatically: we have a phase transition and the system loses its storage and retrieval properties (Amit, 1989).
Appendix B

Approximating the Partition function

In this Appendix we discuss the approximation of the partition function $Z$ (see Section 3.6). We present a sketch of the proof, and we send the interested reader to (Amit, 1989) for the complete discussion.

The partition function $Z$ is given by, using SG-MRF modeling:

$$Z = \sum_{\{f\}} \exp \left( \frac{1}{N} \sum_{\mu=1}^{p} \left[ K(f, \tilde{f}^{(\mu)}) \right]^2 \right), \quad (B.1)$$

as to say

$$Z = \sum_{\{s\}} \exp \left( \frac{1}{N} \sum_{\mu=1}^{p} (s \cdot \xi^{(\mu)})^2 \right), \quad (B.2)$$

For sake of simplicity, we limit again the discussion to the case of discrete variables, $s \in \{\pm 1\}$ (see (Shiino et al, 1990) for the continuous case). Using the identity

$$\exp\{AO^2\} = C \int_{-\infty}^{+\infty} dx \exp \left[ -\frac{x^2}{A} + xO \right],$$

we get:

$$Z \propto \int_{-\infty}^{+\infty} \prod_{\mu} dx^\mu \exp \left[ \sum_{\mu=1}^{p} \left( -N(x^\mu)^2 - x^\mu \sum_{i} s_i \xi^{(\mu)} \right) \right],$$

thus

$$Z \propto \int_{-\infty}^{+\infty} \prod_{\mu} dx^\mu \exp[-N f(\{x^\mu\})], f(\{x^\mu\}) = \sum_{\mu=1}^{p} (-N(x^\mu)^2 - \frac{1}{N} \sum_{i} \ln[2 \cosh(\sum_{\mu=1}^{p} x^\mu \xi^{(\mu)})]).$$
At this point we use a *saddle point* approximation, as to say

\[ I(a) = \int \exp[N\Psi(x, a)] dx \simeq \sum_j C \exp[N\Psi(\hat{x}_j(a), a)] \]

as \( N \to \infty \), where \( \hat{x}_j(a) \) are the points in the integration interval where the function \( \Psi(x, a) \) has its *absolute* maxima (minima if there is a negative sign in the exponent, of course). The absolute minima of \( f(x|\mu) \) are the absolute minima of the H-L energy function (Amit, 1989), hence

\[ Z \simeq \sum_{\{\tilde{f}(\mu)\}} \exp \left( \frac{1}{N} \sum_{\mu=1}^p \left[ K(f, \tilde{f}^{(\mu)}) \right]^2 \right) . \]  

(B.3)
Appendix C

Extra Experimental Results

C.1 Robustness of SG-MRF: a Detailed Record

In order to give an exact idea about the results obtained in the different series of experiments of Chapter 4, we present them in this Appendix in tabular form. In each table we summarize the results obtained with one approach, using different levels of degradation (noise or occlusions) and with respect to all mixtures of train and test sets. For sake of clarity, we remind the reader of all the parameter values used to perform each experiment. The term 2D-histograms denotes the combination $D_xD_y^\sigma$-16 with $\sigma = 1.0$, where 16 refers to the resolution of the histogram bin per axis; the term 3D-histograms refers to $D_xD_yLap^\sigma$-16 with $\sigma = 3.5$. In all the experiments the number of prototypes per object was 36 and the rest of the database was used for testing.

C.2 Robustness to Noise

These results were achieved using the following parameters:

- Kernel parameters: $a = 0.5, b = 1.0$.
- Feature representation: 2D-histograms.

Tables C.1, C.2 and C.3 show the results achieved on the COIL20 database, respectively with SG-MRF, $\chi^2$ and NNC. Tables C.4, C.5 and C.6 contain the results achieved on the COIL100 database, respectively with SG-MRF, $\chi^2$ and NNC.
### APPENDIX C. EXTRA EXPERIMENTAL RESULTS

#### Percentage of noise in testing set

<table>
<thead>
<tr>
<th>sigma</th>
<th>0%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>100</td>
<td>89.23</td>
<td>77.50</td>
<td>67.01</td>
<td>55.83</td>
</tr>
<tr>
<td>25%</td>
<td>99.93</td>
<td>98.19</td>
<td>97.50</td>
<td>95.90</td>
<td>94.58</td>
</tr>
<tr>
<td>50%</td>
<td>99.09</td>
<td>99.30</td>
<td>99.16</td>
<td>99.30</td>
<td>99.09</td>
</tr>
<tr>
<td>75%</td>
<td>97.22</td>
<td>97.91</td>
<td>98.54</td>
<td>99.09</td>
<td>99.79</td>
</tr>
<tr>
<td>100%</td>
<td>97.77</td>
<td>98.19</td>
<td>98.54</td>
<td>99.44</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sigma</th>
<th>0%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>100</td>
<td>76.45</td>
<td>52.84</td>
<td>29.23</td>
<td>5.62</td>
</tr>
<tr>
<td>25%</td>
<td>99.79</td>
<td>96.80</td>
<td>94.86</td>
<td>92.56</td>
<td>89.44</td>
</tr>
<tr>
<td>50%</td>
<td>98.19</td>
<td>97.91</td>
<td>97.91</td>
<td>97.01</td>
<td>97.01</td>
</tr>
<tr>
<td>75%</td>
<td>93.26</td>
<td>94.79</td>
<td>96.45</td>
<td>96.94</td>
<td>98.81</td>
</tr>
<tr>
<td>100%</td>
<td>22.56</td>
<td>41.45</td>
<td>61.31</td>
<td>80.34</td>
<td>99.86</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sigma</th>
<th>0%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>100</td>
<td>75.41</td>
<td>50.76</td>
<td>26.18</td>
<td>1.52</td>
</tr>
<tr>
<td>25%</td>
<td>99.79</td>
<td>95.83</td>
<td>92.22</td>
<td>87.36</td>
<td>82.91</td>
</tr>
<tr>
<td>50%</td>
<td>98.33</td>
<td>97.15</td>
<td>95.34</td>
<td>93.19</td>
<td>91.31</td>
</tr>
<tr>
<td>75%</td>
<td>93.40</td>
<td>93.95</td>
<td>94.86</td>
<td>94.58</td>
<td>95.20</td>
</tr>
<tr>
<td>100%</td>
<td>15.00</td>
<td>35.69</td>
<td>50.87</td>
<td>77.15</td>
<td>98.33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sigma</th>
<th>0%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>100</td>
<td>75.48</td>
<td>50.83</td>
<td>26.25</td>
<td>1.73</td>
</tr>
<tr>
<td>25%</td>
<td>99.79</td>
<td>92.63</td>
<td>86.87</td>
<td>80.06</td>
<td>73.05</td>
</tr>
<tr>
<td>50%</td>
<td>98.33</td>
<td>95.06</td>
<td>90.62</td>
<td>85.97</td>
<td>82.15</td>
</tr>
<tr>
<td>75%</td>
<td>93.47</td>
<td>92.84</td>
<td>92.77</td>
<td>90.06</td>
<td>88.81</td>
</tr>
<tr>
<td>100%</td>
<td>11.11</td>
<td>32.15</td>
<td>52.43</td>
<td>73.33</td>
<td>93.95</td>
</tr>
</tbody>
</table>

Table C.1: Robustness to noise on the COIL20 using SG-MRF

#### C.3 Robustness to the first Kind of Occlusion

These results were achieved using the following parameters:

- Kernel parameters: $a = 0.5, b = 1.0$.
- Feature representation: 2D-histograms.

Tables C.7, C.8 and C.9 show the results achieved on the COIL20 database, respectively with SG-MRF, $\chi^2$ and NNC. Tables C.10, C.11 and C.12 contain the results achieved on the COIL100 database, respectively with SG-MRF, $\chi^2$ and NNC.
### C.4 Robustness to the Second Kind of Occlusion

These results were achieved using the following parameters:

- Kernel parameters: $a = 0.5, b = 2.0$
- Feature representation: 3D-histograms.

Tables C.13, C.14 and C.15 show the results achieved on the COIL20 database respectively with SG-MRF $\chi^2$ and NNC. Tables C.16, C.17 and C.18 contain the results achieved on the COIL100 database, using respectively SG-MRF, $\chi^2$ and NNC, respectively with SG-MRF, $\chi^2$ and NNC.
### Table C.3: Robustness to noise on the COIL20 using NNC

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>0%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma = 10$</td>
<td>98.19</td>
<td>75.76</td>
<td>52.63</td>
<td>29.23</td>
<td>6.66</td>
</tr>
<tr>
<td></td>
<td>96.31</td>
<td>95.00</td>
<td>93.81</td>
<td>92.43</td>
<td>91.25</td>
</tr>
<tr>
<td></td>
<td>93.26</td>
<td>94.58</td>
<td>95.83</td>
<td>97.15</td>
<td>97.91</td>
</tr>
<tr>
<td></td>
<td>84.09</td>
<td>88.47</td>
<td>93.26</td>
<td>95.83</td>
<td>99.16</td>
</tr>
<tr>
<td></td>
<td>8.88</td>
<td>31.38</td>
<td>54.02</td>
<td>77.01</td>
<td>99.79</td>
</tr>
<tr>
<td>$\sigma = 50$</td>
<td>98.19</td>
<td>75.13</td>
<td>51.80</td>
<td>28.26</td>
<td>5.00</td>
</tr>
<tr>
<td></td>
<td>96.31</td>
<td>93.33</td>
<td>90.34</td>
<td>86.80</td>
<td>83.61</td>
</tr>
<tr>
<td></td>
<td>93.19</td>
<td>93.54</td>
<td>92.84</td>
<td>92.01</td>
<td>91.73</td>
</tr>
<tr>
<td></td>
<td>84.09</td>
<td>87.56</td>
<td>92.08</td>
<td>94.16</td>
<td>96.45</td>
</tr>
<tr>
<td></td>
<td>10.06</td>
<td>31.94</td>
<td>53.88</td>
<td>77.01</td>
<td>98.61</td>
</tr>
<tr>
<td>$\sigma = 80$</td>
<td>98.19</td>
<td>75.13</td>
<td>51.80</td>
<td>28.26</td>
<td>5.00</td>
</tr>
<tr>
<td></td>
<td>96.31</td>
<td>90.90</td>
<td>86.45</td>
<td>80.13</td>
<td>75.20</td>
</tr>
<tr>
<td></td>
<td>93.19</td>
<td>91.73</td>
<td>88.81</td>
<td>85.90</td>
<td>83.47</td>
</tr>
<tr>
<td></td>
<td>84.09</td>
<td>86.11</td>
<td>89.37</td>
<td>89.30</td>
<td>89.93</td>
</tr>
<tr>
<td></td>
<td>9.09</td>
<td>29.93</td>
<td>51.31</td>
<td>73.40</td>
<td>94.65</td>
</tr>
<tr>
<td>$\sigma = 120$</td>
<td>98.19</td>
<td>75.13</td>
<td>51.80</td>
<td>28.26</td>
<td>5.00</td>
</tr>
<tr>
<td></td>
<td>96.31</td>
<td>88.47</td>
<td>80.76</td>
<td>72.77</td>
<td>64.37</td>
</tr>
<tr>
<td></td>
<td>93.19</td>
<td>89.72</td>
<td>85.13</td>
<td>80.20</td>
<td>76.25</td>
</tr>
<tr>
<td></td>
<td>84.09</td>
<td>83.95</td>
<td>86.80</td>
<td>84.30</td>
<td>83.47</td>
</tr>
<tr>
<td></td>
<td>8.61</td>
<td>28.33</td>
<td>49.09</td>
<td>68.81</td>
<td>88.88</td>
</tr>
</tbody>
</table>

C.5 Shape and Color Experiments: a Detailed Record

This Appendix reports all the results obtained using SVM with Gaussian kernel on shape only, color only and concatenated representation, as the value of the parameter $s$ changes. For details on the database and the chosen representations we refer the reader to Section 5.4.3.
Table C.4: Robustness to noise on the COIL100 using SG-MRF

<table>
<thead>
<tr>
<th>Sigma</th>
<th>Percentage of noise in testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>0%</td>
<td>99.80</td>
</tr>
<tr>
<td>25%</td>
<td>99.18</td>
</tr>
<tr>
<td>50%</td>
<td>97.75</td>
</tr>
<tr>
<td>75%</td>
<td>93.37</td>
</tr>
<tr>
<td>100%</td>
<td>94.34</td>
</tr>
<tr>
<td>0%</td>
<td>99.80</td>
</tr>
<tr>
<td>25%</td>
<td>98.77</td>
</tr>
<tr>
<td>50%</td>
<td>96.18</td>
</tr>
<tr>
<td>75%</td>
<td>88.84</td>
</tr>
<tr>
<td>100%</td>
<td>5.93</td>
</tr>
<tr>
<td>0%</td>
<td>99.80</td>
</tr>
<tr>
<td>25%</td>
<td>98.79</td>
</tr>
<tr>
<td>50%</td>
<td>96.20</td>
</tr>
<tr>
<td>75%</td>
<td>88.90</td>
</tr>
<tr>
<td>100%</td>
<td>2.31</td>
</tr>
<tr>
<td>0%</td>
<td>99.80</td>
</tr>
<tr>
<td>25%</td>
<td>98.79</td>
</tr>
<tr>
<td>50%</td>
<td>96.20</td>
</tr>
<tr>
<td>75%</td>
<td>88.90</td>
</tr>
<tr>
<td>100%</td>
<td>1.16</td>
</tr>
</tbody>
</table>
### Percentage of noise in testing set

<table>
<thead>
<tr>
<th>Sigma</th>
<th>0%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>σ = 10</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>99.52</td>
<td>79.19</td>
<td>58.88</td>
<td>38.38</td>
<td>18.04</td>
</tr>
<tr>
<td>25%</td>
<td>98.30</td>
<td>95.34</td>
<td>92.09</td>
<td>89.31</td>
<td>86.25</td>
</tr>
<tr>
<td>50%</td>
<td>95.66</td>
<td>95.76</td>
<td>95.76</td>
<td>95.45</td>
<td>95.36</td>
</tr>
<tr>
<td>75%</td>
<td>87.65</td>
<td>90.18</td>
<td>93.62</td>
<td>96.09</td>
<td>98.26</td>
</tr>
<tr>
<td>100%</td>
<td>33.13</td>
<td>49.51</td>
<td>66.36</td>
<td>82.73</td>
<td>99.45</td>
</tr>
<tr>
<td><strong>σ = 50</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>99.52</td>
<td>74.88</td>
<td>50.30</td>
<td>25.62</td>
<td>1.01</td>
</tr>
<tr>
<td>25%</td>
<td>98.22</td>
<td>92.36</td>
<td>86.12</td>
<td>80.19</td>
<td>74.19</td>
</tr>
<tr>
<td>50%</td>
<td>95.27</td>
<td>93.40</td>
<td>90.79</td>
<td>88.20</td>
<td>85.80</td>
</tr>
<tr>
<td>75%</td>
<td>86.55</td>
<td>88.01</td>
<td>90.30</td>
<td>91.43</td>
<td>92.08</td>
</tr>
<tr>
<td>100%</td>
<td>5.97</td>
<td>28.13</td>
<td>50.90</td>
<td>73.23</td>
<td>95.63</td>
</tr>
<tr>
<td><strong>σ = 80</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>99.52</td>
<td>74.90</td>
<td>50.26</td>
<td>25.51</td>
<td>0.93</td>
</tr>
<tr>
<td>25%</td>
<td>98.22</td>
<td>89.26</td>
<td>80.16</td>
<td>71.13</td>
<td>62.00</td>
</tr>
<tr>
<td>50%</td>
<td>95.27</td>
<td>90.22</td>
<td>85.26</td>
<td>79.69</td>
<td>74.12</td>
</tr>
<tr>
<td>75%</td>
<td>86.55</td>
<td>85.66</td>
<td>86.00</td>
<td>84.52</td>
<td>83.26</td>
</tr>
<tr>
<td>100%</td>
<td>3.11</td>
<td>24.30</td>
<td>46.13</td>
<td>67.65</td>
<td>88.91</td>
</tr>
<tr>
<td><strong>σ = 120</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>99.52</td>
<td>74.98</td>
<td>50.33</td>
<td>25.66</td>
<td>1.11</td>
</tr>
<tr>
<td>25%</td>
<td>98.22</td>
<td>86.51</td>
<td>74.76</td>
<td>62.66</td>
<td>50.66</td>
</tr>
<tr>
<td>50%</td>
<td>95.27</td>
<td>87.98</td>
<td>80.04</td>
<td>71.61</td>
<td>63.81</td>
</tr>
<tr>
<td>75%</td>
<td>86.55</td>
<td>83.45</td>
<td>81.70</td>
<td>77.91</td>
<td>74.05</td>
</tr>
<tr>
<td>100%</td>
<td>2.12</td>
<td>22.02</td>
<td>41.86</td>
<td>61.97</td>
<td>81.79</td>
</tr>
</tbody>
</table>

Table C.5: Robustness to noise on the COIL100 using $\chi^2$
C.5. SHAPE AND COLOR EXPERIMENTS: A DETAILED RECORD

<table>
<thead>
<tr>
<th>sigma</th>
<th></th>
<th>Percentage of noise in testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
<td>25%</td>
</tr>
<tr>
<td>σ = 10</td>
<td>0%</td>
<td>94.34</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>89.75</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>81.90</td>
</tr>
<tr>
<td></td>
<td>75%</td>
<td>70.38</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td>1.88</td>
</tr>
<tr>
<td>σ = 50</td>
<td>0%</td>
<td>94.34</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>89.75</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>81.91</td>
</tr>
<tr>
<td></td>
<td>75%</td>
<td>70.56</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td>1.27</td>
</tr>
<tr>
<td>σ = 80</td>
<td>0%</td>
<td>94.43</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>89.75</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>81.91</td>
</tr>
<tr>
<td></td>
<td>75%</td>
<td>70.56</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td>1.59</td>
</tr>
<tr>
<td>σ = 120</td>
<td>0%</td>
<td>94.43</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>89.75</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>81.91</td>
</tr>
<tr>
<td></td>
<td>75%</td>
<td>70.56</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td>1.18</td>
</tr>
</tbody>
</table>

Table C.6: Robustness to noise on the COIL100 using NNC
## APPENDIX C. EXTRA EXPERIMENTAL RESULTS

### Table C.7: Robustness to the first kind of occlusion on the COIL20 using SG-MRF

<table>
<thead>
<tr>
<th>occlusion</th>
<th>Percentage of noise in testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>c=1,r=1,w=8</td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>100</td>
</tr>
<tr>
<td>25%</td>
<td>100</td>
</tr>
<tr>
<td>50%</td>
<td>99.37</td>
</tr>
<tr>
<td>75%</td>
<td>99.16</td>
</tr>
<tr>
<td>100%</td>
<td>99.86</td>
</tr>
<tr>
<td>c=1,r=1,w=16</td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>100</td>
</tr>
<tr>
<td>25%</td>
<td>100</td>
</tr>
<tr>
<td>50%</td>
<td>99.30</td>
</tr>
<tr>
<td>75%</td>
<td>98.81</td>
</tr>
<tr>
<td>100%</td>
<td>99.37</td>
</tr>
<tr>
<td>c=1,r=1,w=32</td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>100</td>
</tr>
<tr>
<td>25%</td>
<td>100</td>
</tr>
<tr>
<td>50%</td>
<td>98.40</td>
</tr>
<tr>
<td>75%</td>
<td>94.93</td>
</tr>
<tr>
<td>100%</td>
<td>86.52</td>
</tr>
<tr>
<td>c=1,r=0,w=64</td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>100</td>
</tr>
<tr>
<td>25%</td>
<td>100</td>
</tr>
<tr>
<td>50%</td>
<td>98.33</td>
</tr>
<tr>
<td>75%</td>
<td>94.09</td>
</tr>
<tr>
<td>100%</td>
<td>55.48</td>
</tr>
</tbody>
</table>

Table C.7: Robustness to the first kind of occlusion on the COIL20 using SG-MRF
C.5. SHAPE AND COLOR EXPERIMENTS: A DETAILED RECORD

<table>
<thead>
<tr>
<th>occlusion</th>
<th>Percentage of noise in testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>c=1,r=1,w=8</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>c=1,r=1,w=16</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>c=1,r=1,w=32</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>c=1,r=0,w=64</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
</tbody>
</table>

Table C.8: Robustness to the first kind of occlusion on the COIL20 using $\chi^2$
### Table C.9: Robustness to first kind of occlusion on the COIL20 using NNC

<table>
<thead>
<tr>
<th>occlusion</th>
<th>Percentage of noise in testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>c=1, r=1, w=8</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>c=1, r=1, w=16</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>c=1, r=1, w=32</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>c=1, r=0, w=64</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>occlusion</td>
<td>Percentage of noise in testing set</td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>c=1,r=1,w=8</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>c=1,r=1,w=16</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>c=1,r=1,w=32</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>c=1,r=0,w=64</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
</tbody>
</table>

Table C.10: Robustness to the first kind of occlusion on the COIL100 using SG-MRF
<table>
<thead>
<tr>
<th>occlusion</th>
<th>Percentage of noise in testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>c=1,r=1,w=8</td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>99.52</td>
</tr>
<tr>
<td>25%</td>
<td>98.50</td>
</tr>
<tr>
<td>50%</td>
<td>96.06</td>
</tr>
<tr>
<td>75%</td>
<td>89.41</td>
</tr>
<tr>
<td>100%</td>
<td>68.22</td>
</tr>
<tr>
<td>c=1,r=1,w=16</td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>99.52</td>
</tr>
<tr>
<td>25%</td>
<td>98.33</td>
</tr>
<tr>
<td>50%</td>
<td>95.50</td>
</tr>
<tr>
<td>75%</td>
<td>87.63</td>
</tr>
<tr>
<td>100%</td>
<td>43.97</td>
</tr>
<tr>
<td>c=1,r=1,w=32</td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>99.52</td>
</tr>
<tr>
<td>25%</td>
<td>98.22</td>
</tr>
<tr>
<td>50%</td>
<td>95.27</td>
</tr>
<tr>
<td>75%</td>
<td>86.54</td>
</tr>
<tr>
<td>100%</td>
<td>9.81</td>
</tr>
<tr>
<td>c=1,r=0,w=64</td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>99.52</td>
</tr>
<tr>
<td>25%</td>
<td>98.22</td>
</tr>
<tr>
<td>50%</td>
<td>95.26</td>
</tr>
<tr>
<td>75%</td>
<td>86.59</td>
</tr>
<tr>
<td>100%</td>
<td>7.41</td>
</tr>
</tbody>
</table>

Table C.11: Robustness to the first kind of occlusion on the COIL100 using $\chi^2$. 
### Percentage of noise in testing set

<table>
<thead>
<tr>
<th>occlusion</th>
<th>0%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>c=1, r=1, w=8</td>
<td>94.43</td>
<td>75.73</td>
<td>57.72</td>
<td>38.86</td>
<td>20.04</td>
</tr>
<tr>
<td></td>
<td>89.79</td>
<td>79.27</td>
<td>68.79</td>
<td>58.83</td>
<td>48.19</td>
</tr>
<tr>
<td></td>
<td>82.31</td>
<td>77.51</td>
<td>72.22</td>
<td>67.09</td>
<td>61.18</td>
</tr>
<tr>
<td></td>
<td>71.81</td>
<td>72.08</td>
<td>72.63</td>
<td>71.44</td>
<td>70.01</td>
</tr>
<tr>
<td></td>
<td>32.73</td>
<td>44.47</td>
<td>55.90</td>
<td>66.81</td>
<td>77.97</td>
</tr>
<tr>
<td>c=1, r=1, w=16</td>
<td>94.43</td>
<td>73.26</td>
<td>52.44</td>
<td>31.15</td>
<td>10.04</td>
</tr>
<tr>
<td></td>
<td>89.51</td>
<td>76.48</td>
<td>63.62</td>
<td>51.12</td>
<td>37.97</td>
</tr>
<tr>
<td></td>
<td>81.34</td>
<td>74.29</td>
<td>66.98</td>
<td>59.86</td>
<td>51.80</td>
</tr>
<tr>
<td></td>
<td>69.12</td>
<td>68.08</td>
<td>68.01</td>
<td>65.50</td>
<td>63.01</td>
</tr>
<tr>
<td></td>
<td>11.52</td>
<td>27.02</td>
<td>42.63</td>
<td>57.91</td>
<td>72.90</td>
</tr>
<tr>
<td>c=1, r=1, w=32</td>
<td>94.43</td>
<td>71.65</td>
<td>48.95</td>
<td>25.94</td>
<td>3.02</td>
</tr>
<tr>
<td></td>
<td>89.52</td>
<td>73.91</td>
<td>58.65</td>
<td>43.34</td>
<td>28.09</td>
</tr>
<tr>
<td></td>
<td>81.01</td>
<td>71.58</td>
<td>62.41</td>
<td>53.13</td>
<td>43.19</td>
</tr>
<tr>
<td></td>
<td>68.72</td>
<td>65.55</td>
<td>63.80</td>
<td>59.73</td>
<td>55.61</td>
</tr>
<tr>
<td></td>
<td>3.00</td>
<td>18.26</td>
<td>34.76</td>
<td>51.04</td>
<td>66.70</td>
</tr>
<tr>
<td>c=1, r=0, w=64</td>
<td>94.43</td>
<td>71.16</td>
<td>48.08</td>
<td>24.75</td>
<td>1.41</td>
</tr>
<tr>
<td></td>
<td>89.55</td>
<td>76.11</td>
<td>62.95</td>
<td>49.41</td>
<td>36.12</td>
</tr>
<tr>
<td></td>
<td>81.51</td>
<td>74.02</td>
<td>66.54</td>
<td>59.26</td>
<td>51.40</td>
</tr>
<tr>
<td></td>
<td>69.75</td>
<td>68.45</td>
<td>68.15</td>
<td>65.95</td>
<td>63.51</td>
</tr>
<tr>
<td></td>
<td>2.13</td>
<td>19.55</td>
<td>37.25</td>
<td>55.66</td>
<td>73.13</td>
</tr>
</tbody>
</table>

Table C.12: Robustness to the first kind of occlusion on the COIL100 using NNC
APPENDIX C. EXTRA EXPERIMENTAL RESULTS

<table>
<thead>
<tr>
<th>visible object portion</th>
<th>Percentage of noise in testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>20%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>30%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>40%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>50%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>60%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>70%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>80%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
</tbody>
</table>

Table C.13: Robustness to the second kind of occlusion on the COIL20 using SG-MRF
### C.5. SHAPE AND COLOR EXPERIMENTS: A DETAILED RECORD

<table>
<thead>
<tr>
<th>visible object portion</th>
<th>0%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>20%</td>
<td>100</td>
<td>79.86</td>
<td>59.44</td>
<td>39.38</td>
<td>19.03</td>
</tr>
<tr>
<td></td>
<td>99.79</td>
<td>88.68</td>
<td>77.08</td>
<td>64.24</td>
<td>53.19</td>
</tr>
<tr>
<td></td>
<td>99.03</td>
<td>91.18</td>
<td>81.88</td>
<td>74.17</td>
<td>65.62</td>
</tr>
<tr>
<td></td>
<td>95.07</td>
<td>90.56</td>
<td>86.11</td>
<td>80.90</td>
<td>74.86</td>
</tr>
<tr>
<td></td>
<td>3.61</td>
<td>22.29</td>
<td>41.67</td>
<td>61.18</td>
<td>80.42</td>
</tr>
<tr>
<td>30%</td>
<td>100</td>
<td>83.75</td>
<td>67.71</td>
<td>51.67</td>
<td>35.97</td>
</tr>
<tr>
<td></td>
<td>99.79</td>
<td>95.69</td>
<td>91.46</td>
<td>87.29</td>
<td>83.47</td>
</tr>
<tr>
<td></td>
<td>99.03</td>
<td>98.19</td>
<td>96.81</td>
<td>95.28</td>
<td>94.10</td>
</tr>
<tr>
<td></td>
<td>95.07</td>
<td>96.39</td>
<td>97.64</td>
<td>97.78</td>
<td>98.54</td>
</tr>
<tr>
<td></td>
<td>11.67</td>
<td>33.47</td>
<td>55.69</td>
<td>77.50</td>
<td>99.65</td>
</tr>
<tr>
<td>40%</td>
<td>100</td>
<td>87.57</td>
<td>74.51</td>
<td>61.60</td>
<td>49.10</td>
</tr>
<tr>
<td></td>
<td>99.79</td>
<td>97.57</td>
<td>95.69</td>
<td>93.61</td>
<td>91.88</td>
</tr>
<tr>
<td></td>
<td>99.03</td>
<td>99.17</td>
<td>98.06</td>
<td>98.40</td>
<td>98.12</td>
</tr>
<tr>
<td></td>
<td>95.07</td>
<td>96.60</td>
<td>97.77</td>
<td>98.47</td>
<td>99.72</td>
</tr>
<tr>
<td></td>
<td>29.79</td>
<td>47.57</td>
<td>64.79</td>
<td>82.22</td>
<td>100</td>
</tr>
<tr>
<td>50%</td>
<td>100</td>
<td>91.32</td>
<td>82.29</td>
<td>74.72</td>
<td>66.53</td>
</tr>
<tr>
<td></td>
<td>99.79</td>
<td>98.82</td>
<td>98.19</td>
<td>97.08</td>
<td>96.25</td>
</tr>
<tr>
<td></td>
<td>99.03</td>
<td>99.17</td>
<td>99.17</td>
<td>99.44</td>
<td>99.44</td>
</tr>
<tr>
<td></td>
<td>95.69</td>
<td>97.08</td>
<td>98.12</td>
<td>98.96</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>61.94</td>
<td>71.04</td>
<td>80.97</td>
<td>90.56</td>
<td>100</td>
</tr>
<tr>
<td>60%</td>
<td>100</td>
<td>95.97</td>
<td>91.46</td>
<td>86.60</td>
<td>83.33</td>
</tr>
<tr>
<td></td>
<td>99.79</td>
<td>99.44</td>
<td>99.51</td>
<td>98.96</td>
<td>98.96</td>
</tr>
<tr>
<td></td>
<td>99.03</td>
<td>99.38</td>
<td>99.44</td>
<td>99.58</td>
<td>99.72</td>
</tr>
<tr>
<td></td>
<td>96.53</td>
<td>97.71</td>
<td>98.26</td>
<td>99.24</td>
<td>99.93</td>
</tr>
<tr>
<td></td>
<td>92.22</td>
<td>93.82</td>
<td>96.32</td>
<td>97.85</td>
<td>99.93</td>
</tr>
<tr>
<td>70%</td>
<td>100</td>
<td>99.44</td>
<td>98.89</td>
<td>98.54</td>
<td>98.12</td>
</tr>
<tr>
<td></td>
<td>99.79</td>
<td>99.72</td>
<td>99.86</td>
<td>99.72</td>
<td>99.72</td>
</tr>
<tr>
<td></td>
<td>99.51</td>
<td>99.79</td>
<td>99.58</td>
<td>99.79</td>
<td>99.93</td>
</tr>
<tr>
<td></td>
<td>98.19</td>
<td>98.61</td>
<td>98.96</td>
<td>99.65</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>99.10</td>
<td>99.24</td>
<td>99.72</td>
<td>99.72</td>
<td>100</td>
</tr>
<tr>
<td>80%</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Table C.14: Robustness to the second kind of occlusion on the COIL20 using $\chi^2$
APPENDIX C. EXTRA EXPERIMENTAL RESULTS

<table>
<thead>
<tr>
<th>visible object portion</th>
<th>Percentage of noise in testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>20%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>30%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>40%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>50%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>60%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>70%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>80%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
</tbody>
</table>

Table C.15: Robustness to the second kind of occlusion on the COIL20 using NNC
<table>
<thead>
<tr>
<th>visible object portion</th>
<th>Percentage of noise in testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>99.97 76.31 52.68 28.99 5.29</td>
</tr>
<tr>
<td>25%</td>
<td>99.58 82.36 65.04 47.92 30.78</td>
</tr>
<tr>
<td>50%</td>
<td>97.96 83.19 68.33 53.82 38.90</td>
</tr>
<tr>
<td>75%</td>
<td>91.96 80.75 70.87 59.03 47.49</td>
</tr>
<tr>
<td>100%</td>
<td>1.32  14.36 27.79 41.53 54.62</td>
</tr>
</tbody>
</table>

| 0%                     | 99.97 78.10 56.67 34.92 13.19     |
| 25%                    | 99.58 91.36 83.06 75.36 67.15     |
| 50%                    | 97.96 94.60 90.97 87.11 83.07     |
| 75%                    | 91.96 92.08 92.82 92.10 91.64     |
| 100%                   | 0.75  24.65 48.40 72.16 96.00     |

| 0%                     | 99.97 80.96 62.17 43.43 24.42     |
| 25%                    | 99.58 94.54 89.56 85.22 80.42     |
| 50%                    | 97.96 97.11 95.94 94.68 93.68     |
| 75%                    | 91.96 93.62 95.54 96.74 98.04     |
| 100%                   | 6.35  29.69 52.83 76.46 99.54     |

| 0%                     | 99.97 84.53 69.44 54.51 39.19     |
| 25%                    | 99.58 96.53 93.57 70.76 87.96     |
| 50%                    | 97.96 97.94 97.72 97.49 97.32     |
| 75%                    | 91.98 94.03 96.22 97.85 99.56     |
| 100%                   | 21.97 41.00 60.92 80.08 99.62     |

| 0%                     | 99.97 89.79 79.67 69.87 59.40     |
| 25%                    | 99.60 97.68 96.07 94.79 92.92     |
| 50%                    | 98.17 98.33 98.42 98.49 98.56     |
| 75%                    | 92.53 94.49 96.58 98.19 99.76     |
| 100%                   | 52.01 64.06 75.67 87.89 99.97     |

| 0%                     | 99.97 96.71 93.75 90.72 87.31     |
| 25%                    | 99.71 98.86 98.57 98.26 97.43     |
| 50%                    | 98.69 98.90 99.14 99.37 99.56     |
| 75%                    | 95.36 96.60 97.96 98.72 99.86     |
| 100%                   | 86.75 90.10 93.28 96.78 100       |

Table C.16: Robustness to the second kind of occlusion on the COIL100 using SG-MRF
<table>
<thead>
<tr>
<th>visible object portion</th>
<th>Percentage of noise in testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>20%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>30%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>40%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>50%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>60%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>70%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>80%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
</tbody>
</table>

Table C.17: Robustness to the second kind of occlusion on the COIL100 using $\chi^2$
<table>
<thead>
<tr>
<th>visible object portion</th>
<th>Percentage of noise in testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>0%</td>
<td>98.35</td>
</tr>
<tr>
<td>25%</td>
<td>95.79</td>
</tr>
<tr>
<td>50%</td>
<td>90.62</td>
</tr>
<tr>
<td>75%</td>
<td>80.92</td>
</tr>
<tr>
<td>100%</td>
<td>1.53</td>
</tr>
<tr>
<td>30%</td>
<td>98.35</td>
</tr>
<tr>
<td>25%</td>
<td>95.79</td>
</tr>
<tr>
<td>50%</td>
<td>90.62</td>
</tr>
<tr>
<td>75%</td>
<td>80.92</td>
</tr>
<tr>
<td>100%</td>
<td>2.28</td>
</tr>
<tr>
<td>40%</td>
<td>98.35</td>
</tr>
<tr>
<td>25%</td>
<td>96.79</td>
</tr>
<tr>
<td>50%</td>
<td>90.62</td>
</tr>
<tr>
<td>75%</td>
<td>80.92</td>
</tr>
<tr>
<td>100%</td>
<td>3.53</td>
</tr>
<tr>
<td>50%</td>
<td>98.35</td>
</tr>
<tr>
<td>25%</td>
<td>95.81</td>
</tr>
<tr>
<td>50%</td>
<td>90.64</td>
</tr>
<tr>
<td>75%</td>
<td>80.99</td>
</tr>
<tr>
<td>100%</td>
<td>6.90</td>
</tr>
<tr>
<td>60%</td>
<td>98.35</td>
</tr>
<tr>
<td>25%</td>
<td>95.83</td>
</tr>
<tr>
<td>50%</td>
<td>90.76</td>
</tr>
<tr>
<td>75%</td>
<td>81.24</td>
</tr>
<tr>
<td>100%</td>
<td>12.57</td>
</tr>
<tr>
<td>70%</td>
<td>98.35</td>
</tr>
<tr>
<td>25%</td>
<td>95.86</td>
</tr>
<tr>
<td>50%</td>
<td>90.97</td>
</tr>
<tr>
<td>75%</td>
<td>82.07</td>
</tr>
<tr>
<td>100%</td>
<td>20.74</td>
</tr>
<tr>
<td>80%</td>
<td>98.35</td>
</tr>
<tr>
<td>25%</td>
<td>96.00</td>
</tr>
<tr>
<td>50%</td>
<td>91.39</td>
</tr>
<tr>
<td>75%</td>
<td>84.28</td>
</tr>
<tr>
<td>100%</td>
<td>36.33</td>
</tr>
</tbody>
</table>

Table C.18: Robustness to the second kind of occlusion on the COIL100 using NNC
### APPENDIX C. EXTRA EXPERIMENTAL RESULTS

<table>
<thead>
<tr>
<th>Value of $\sigma$</th>
<th>Error rate (%)</th>
<th>Value of $\sigma$</th>
<th>Error rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma = 0.001$</td>
<td>94.92</td>
<td>$\sigma = 0.01$</td>
<td>49.27</td>
</tr>
<tr>
<td>$\sigma = 0.03$</td>
<td>31.08</td>
<td>$\sigma = 0.04$</td>
<td>29.30</td>
</tr>
<tr>
<td>$\sigma = 0.05$</td>
<td>28.60</td>
<td>$\sigma = 0.06$</td>
<td>28.20</td>
</tr>
<tr>
<td>$\sigma = 0.07$</td>
<td>27.87</td>
<td>$\sigma = 0.08$</td>
<td>27.37</td>
</tr>
<tr>
<td>$\sigma = 0.09$</td>
<td>26.90</td>
<td>$\sigma = 0.1$</td>
<td>26.77</td>
</tr>
<tr>
<td>$\sigma = 0.2$</td>
<td><strong>25.30</strong></td>
<td>$\sigma = 0.3$</td>
<td>25.47</td>
</tr>
<tr>
<td>$\sigma = 0.4$</td>
<td>26.05</td>
<td>$\sigma = 0.5$</td>
<td>27.38</td>
</tr>
<tr>
<td>$\sigma = 0.6$</td>
<td>28.40</td>
<td>$\sigma = 0.7$</td>
<td>29.53</td>
</tr>
<tr>
<td>$\sigma = 0.9$</td>
<td>31.20</td>
<td>$\sigma = 1$</td>
<td>32.03</td>
</tr>
</tbody>
</table>

Table C.19: Classification results for SVM, Gaussian kernels, different sigmas, color representation

<table>
<thead>
<tr>
<th>Value of $\sigma$</th>
<th>Error rate (%)</th>
<th>Value of $\sigma$</th>
<th>Error rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma = 0.001$</td>
<td>91.97</td>
<td>$\sigma = 0.01$</td>
<td>34.00</td>
</tr>
<tr>
<td>$\sigma = 0.02$</td>
<td>31.08</td>
<td>$\sigma = 0.03$</td>
<td>28.17</td>
</tr>
<tr>
<td>$\sigma = 0.04$</td>
<td>25.77</td>
<td>$\sigma = 0.05$</td>
<td>24.05</td>
</tr>
<tr>
<td>$\sigma = 0.06$</td>
<td>22.43</td>
<td>$\sigma = 0.07$</td>
<td>21.35</td>
</tr>
<tr>
<td>$\sigma = 0.08$</td>
<td>20.50</td>
<td>$\sigma = 0.09$</td>
<td>19.87</td>
</tr>
<tr>
<td>$\sigma = 0.1$</td>
<td><strong>19.78</strong></td>
<td>$\sigma = 0.2$</td>
<td>24.22</td>
</tr>
<tr>
<td>$\sigma = 0.3$</td>
<td>27.88</td>
<td>$\sigma = 0.4$</td>
<td>28.7</td>
</tr>
<tr>
<td>$\sigma = 0.5$</td>
<td>28.73</td>
<td>$\sigma = 1$</td>
<td>32.70</td>
</tr>
</tbody>
</table>

Table C.20: Classification results for SVM, Gaussian kernels, different sigmas, shape representation
Bibliography


R. Fergus, P. Perona, A. Zisserman, ”Object class recognition by unsupervised scale-invariant learningÃ’Â´, *International conference on Computer Vision and Pattern Recognition (CVPR03).*


J.J. Hopfield, “Neurons with a graded response have collective computational properties like those of two-state neurons”, *Proceedings of the National Academy of Science USA*, 81, 1984, pp 3088-3092.


B. Leibe, A. Leonardis, B. Schiele, “Combined object categorization and segmentation with an implicit shape model”, *Workshop on Statitical Learning in Computer Vision (in conj. with ECCV04)*.


P. Viola, M. Jones, “Rapid object detection using a boosted cascade of simple features”, *Proc International Conference on Computer Vision and Pattern Recognition (CVPR01)*.


