Combining Color and Shape Information for Appearance-Based Object Recognition Using Ultrametric Spin Glass-Markov Random Fields

B. Caputo[1], Gy. Dorkó[2], and H. Niemann[2]

[1] Smith-Kettlewell Eye Research Institute
2318 Fillmore Street, San Francisco, 94115 California, USA
[2] Department of Computer Science, Chair for Pattern Recognition
Erlangen-Nuremberg University
Martenstrasse 3, 91058 Erlangen, Germany

Abstract. Shape and color information are important cues for object recognition. An ideal system should give the option to use both forms of information, as well as the option to use just one of the two. We present in this paper a kernel method that achieves this goal. It is based on results of statistical physics of disordered systems combined with Gibbs distributions via kernel functions. Experimental results on a database of 100 objects confirm the effectiveness of the proposed approach.

1 Introduction

Object recognition is a challenging topic of research in computer vision [8]. Many approaches use appearance-based methods, which consider the appearance of objects using two-dimensional image representations [9,15,23]. Although it is generally acknowledged that both color and geometric (shape) information are important for object recognition [11,22], 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 shape information [11,22,10]. Systems using this kind of approach show very good performances [11,22,10]. 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 geometrical information alone, depending on the task considered. A huge literature shows that color only, or shape only representations work very well for many applications (see for instance [8,9,21,23]). 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 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

S.-W. Lee and A. Verri (Eds.): SVM 2002, LNCS 2388, pp. 97–111, 2002.

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 shape information than single representations.

In this paper we propose a new strategy to this problem. Given a shape only and color only representation, we focus attention on how they can be combined together as they are, rather than define a new representation. At the end, we use a new kernel method: Spin Glass-Markov Random Fields (SG-MRF) [2]. They are a new class of MRF that integrates results of statistical physics of disordered systems with Gibbs probability distributions via non linear kernel mapping. The resulting model, using a Hopfield energy function [1], has shown to be very effective for appearance-based object recognition and to be remarkably robust to noise and occlusion. Here we extend SG-MRF to a new SG-like energy function, inspired by the ultrametric properties of the SG phase space. We will show that this energy can be kernelized as the Hopfield one, thus, it can be used in the SG-MRF framework. The structure of this energy provides a natural framework for combining shape and color representations together, without any need to define a new representation. There are several advantages to this approach:

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

To the best of our knowledge, there are no previous similar approaches to the problem of combining shape and color information for object recognition. Experimental results show the effectiveness of the new proposed kernel method.

The paper is organized as follows: after a review of existing literature (Section 2), we will define the general framework for appearance-based object recognition (Section 3) and Spin Glass-Markov Random Fields (Section 4). Section 5 will present the new ultrametric energy function, show how it can be used in a SG-MRF framework (Section 5.1) and how it can be used for combining together shape and color representation for appearance-based object recognition (Section 5.2). Experiments are presented in Section 6; the paper concludes with a summary discussion.

2 Related Work

Appearance-based object recognition is an alternative approach to the geometry-based methods [8]. In an appearance-based approach [17] the objects are modeled by a set of images, and recognition is performed by matching directly the input image to the model set. Swain and Ballard [23] proposed representing an object by its color histogram. 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. Therefore, color histograms have been combined with geometric information

(see for instance [22,10]). In particular, the SEEMORE system [11] uses 102 different feature channels which are each sub sampled and summed over a pre--segmented image region. The 102 channels comprise color, intensity, corner, contour shape and Gabor-derived texture features. Strikingly good experimental results are given on a database of 100 pre-segmented objects of various types. Most interestingly, a certain ability to generalize outside the database has been observed.

Schiele and Crowley [21] 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 small number of vectors.

Principal component analysis has been widely applied for appearance-based object recognition [24,14,7,19]. The attractiveness of this 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 have been investigated in order to handle complete parameterized models of objects [14], to cope with occlusion [7,19] and to be robust to outliers and noise [9].

Recently, Support Vector Machines (SVM) have gained in interest for appearance based object recognition [5,16]. Pontil [18] examined the robustness of SVM to noise, bias in the registration and moderate amount of partial occlusions, obtaining good results. Roobaert et al. [20] examined the generalization capability of SVM when just a few views per object are available.

3 Probabilistic Appearance-Based Object Recognition

Appearance-based object recognition methods consider images as feature vectors. Let $\mathbf{x} \equiv [x_{ij}], i = 1, \ldots \mathcal{N}, j = 1, \ldots \mathcal{M}$ be an $\mathcal{M} \times \mathcal{N}$ image. We will consider each image as a feature vector $\mathbf{x} \in G \equiv \Re^m, m = \mathcal{M}\mathcal{N}$. 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_j data samples, $d_j = \{\mathbf{x}_1^j, \mathbf{x}_2^j, \ldots, \mathbf{x}_{n_j}^j\}, j = 1, \ldots k$. We will assign each object to a pattern class $\Omega_1, \Omega_2, \ldots, \Omega_k$. The object classification procedure will be a discrete mapping that assigns a test image, showing one of the objects, to the pattern class the presented object corresponds to. How the object class Ω_j is represented, given a set of data samples d_j (relative to that object class), varies for different appearance-based approaches: it can consider shape information only, or color information only or both (see Section 2 for a review). Here we will concentrate on probabilistic appearance-based methods.

The probabilistic approach to appearance-based object recognition considers the image views of a given object Ω_j as random vectors. Thus, given the set

of data samples d_j and assuming they are a sufficient statistic for the pattern class Ω_j, the goal will be to estimate the probability distribution $P_{\Omega_j}(\mathbf{x})$ that has generated them. Then, given a test image $\mathbf{x}$, the decision will be made using a Maximum A Posteriori (MAP) classifier:

$$j^* = \underset{j}{\operatorname{argmax}} P_{\Omega_j}(\mathbf{x}) = \underset{j}{\operatorname{argmax}} P(\Omega_j|\mathbf{x}),$$

and, using Bayes rule,

$$j^* = \underset{j}{\operatorname{argmax}} P(\mathbf{x}|\Omega_j)P(\Omega_j). \tag{1}$$

where $P(f|\Omega_j)$ are the Likelihood Functions (LFs) and $P(\Omega_j)$ are the prior probabilities of the classes. In the rest of the paper we will assume that the prior $P(\Omega_j)$ is the same for all object classes; thus the Bayes classifier (1) simplifies to

$$j^* = \underset{j}{\operatorname{argmax}} P(\mathbf{x}|\Omega_j). \tag{2}$$

Many probabilistic appearance-based methods do not model the pdf on raw pixel data, but on features extracted from the original views. The extension of equation (2) to this case is straightforward: consider a set of features $\{\mathbf{h}_1^j, \mathbf{h}_2^j, \ldots, \mathbf{h}_{n_j}^j\}, j = 1, \ldots k$, where each feature vector $\mathbf{h}_{n_j}^j$ is computed from the image $\mathbf{x}_{n_j}^j, \mathbf{h}_{n_j}^j = T(\mathbf{x}_{n_j}^j), \mathbf{h}_{n_j}^j \in G \equiv \Re^m$. The Bayes classifier (2) will be in this case

$$j^* = \underset{j}{\operatorname{argmax}} P(\mathbf{h}|\Omega_j). \tag{3}$$

Probabilistic methods for appearance-based object recognition have the double advantage of being theoretically optimal from the point of view of classification, and to be robust to degradation of the data due to noise and occlusions [21]. A major drawback in these approaches is that the functional form of the probability distribution of an object class Ω_j is not known a priori. Assumptions have to be made regarding to 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 d_j. Thus, the performance will depend on the goodness of the assumption for the parametric form, and on whether the data set d_j is a sufficient statistic for the pattern class Ω_j and thus, permits us to estimate properly the distribution's parameters.

4 Spin Glass-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 provides a probabilistic foundation for modeling spatial interactions on lattice systems or, more generally, on interacting features. It considers each element of the random vector $\mathbf{h}$ as the result of a labeling of all the sites representing $\mathbf{h}$,

with respect to a given label set. The MRF joint probability distribution is given by

$$P(\mathbf{h}) = \frac{1}{Z}\exp\left(-E(\mathbf{h})\right), \qquad Z = \sum_{\{\mathbf{h}\}}\exp\left(-E(\mathbf{h})\right). \tag{4}$$

The normalizing constant Z is called the partition function, and $E(\mathbf{h})$ is the *energy function*. $P(\mathbf{h})$ measures the probability of the occurrence of a particular configurations $\mathbf{h}$; the more probable configurations are those with lower energies. Thus, using MRF modeling for appearance-based object recognition, eq (2) will become

$$j^* = \underset{j}{\operatorname{argmax}}\, P(\mathbf{h}|\Omega_j) = \underset{j}{\operatorname{argmin}}\, E(\mathbf{h}|\Omega_j) \tag{5}$$

Only a few MRF approaches have been proposed for high level vision problems such as object recognition [26,13], due to the modeling problem for MRF on irregular sites (for a detailed discussion about this point, we refer the reader to [2]). Spin Glass-Markov Random Fields overcome this limitation and can be effectively used for appearance-based object recognition [2]. To the best of our knowledge, SG-MRF is the first and only successful MRF-based approach to appearance-based object recognition.

The rest of this Section will review SG-MRFs (Section 4.1) and how they can be derived from results of statistical physics of disordered systems (Section 4.2). Section 5 will show how these results can be extended to a new class of energy function and how this extension makes it possible to use this approach for appearance-based object recognition using shape and color features combined together.

4.1 Spin Glass-Markov Random Fields: Model Definition

Spin Glass-Markov Random Fields (SG-MRFs) [2] are a new class of MRFs which connect SG-like energy functions (mainly the Hopfield one [1]) with Gibbs distributions via a non linear kernel mapping. The resulting model overcomes many difficulties related to the design of fully connected MRFs, and enables us to use the power of kernels in a probabilistic framework. Consider k object classes $\Omega_1, \Omega_2, \ldots, \Omega_k$, and for each object a set of n_j data samples, $d_j = \{\mathbf{x}_1^j, \ldots \mathbf{x}_{n_j}^j\}, j = 1, \ldots k$. We will suppose to extract, from each data sample d_j a set of features $\{\mathbf{h}_1^j, \ldots \mathbf{h}_{n_j}^j\}$. For instance, $\mathbf{h}_{n_j}^j$ can be a color histogram computed from $\mathbf{x}_{n_j}^j$. The SG-MRF probability distribution is given by

$$P_{SG-MRF}(\mathbf{h}|\Omega_j) = \frac{1}{Z}\exp\left[-E_{SG-MRF}(\mathbf{h}|\Omega_j)\right], \tag{6}$$

$$Z = \sum_{\{\mathbf{h}\}}\exp\left[-E_{SG-MRF}(\mathbf{h}|\Omega_j)\right],$$

with

$$E_{SG-MRF}(\mathbf{h}|\Omega_j) = -\sum_{\mu=1}^{p_j} \left[K(\mathbf{h}, \tilde{\mathbf{h}}^{(\mu_j)})\right]^2, \tag{7}$$

where the function $K(\mathbf{h}, \tilde{\mathbf{h}}^{(\mu_j)})$ is a Generalized Gaussian kernel [27]:

$$K(\mathbf{x}, \mathbf{y}) = \exp\{-\rho d_{a,b}(\mathbf{x}, \mathbf{y})\}, \qquad d_{a,b}(\mathbf{x}, \mathbf{y}) = \sum_i |x_i^a - y_i^a|^b. \tag{8}$$

$\{\tilde{\mathbf{h}}^{(\mu_j)}\}_{\mu=1}^{p_j}, j \in [1, k]$ are a set of vectors selected (according to a chosen ansatz, [2]) from the training data that we call *prototypes*. The number of prototypes per class must be finite, and they must satisfy the condition:

$$K(\tilde{\mathbf{h}}^{(i)}, \tilde{\mathbf{h}}^{(l)}) = 0, \tag{9}$$

for all $i, l = 1, \ldots p_j, i \neq l$ and $j = 1, \ldots k$. Note that SG-MRFs are defined on features rather than on raw pixels data. The sites are fully connected, which ends in learning the neighborhood system from the training data instead of choosing it heuristically. As we model the probability distribution on feature vectors and not on raw pixels, SG-MRF is not a generative model. Another key characteristic of the model is that in SG-MRF the functional form of the energy is given by construction. This is achieved using results for statistical physics of Spin Glasses. The next Section sketches the theoretical derivation of the model. The interested reader will find a more detailed discussion in [2].

4.2 Spin Glass-Markov Random Fields: Model Derivation

Consider the following energy function:

$$E = -\sum_{(i,j)} J_{ij}\ s_i\ s_j \qquad i, j = 1, \ldots N, \tag{10}$$

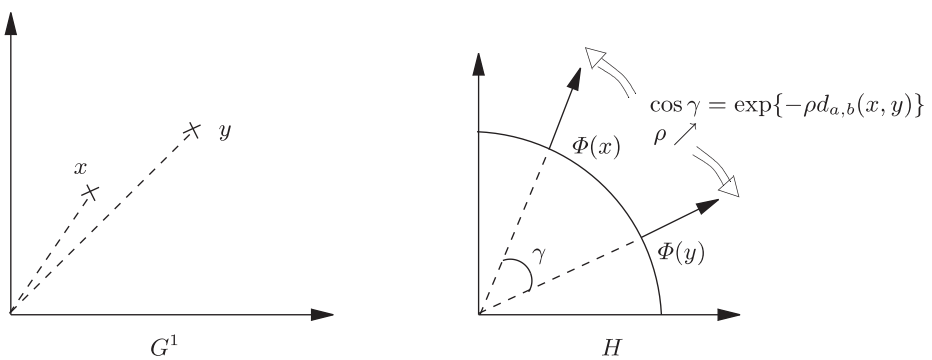

Fig. 1. Gaussian kernels map the data to an infinite dimension hyper-sphere of radius unity. Thus, with a proper choice of ρ, it is possible to orthogonalize all the training data in that space

where the s_i are random variables taking values in $\{\pm 1\}$, $\mathbf{s} = (s_1, \ldots, s_N)$ is a configuration and $\mathbf{J} = [J_{ij}\ ], (i,j) = 1, \ldots, N$ is the connection matrix, $J_{ij} \in \{\pm 1\}$. Equation (10) is the most general Spin Glass (SG) energy function [1,12]; the study of the properties of this energy for different **J**s has been a lively area of research in the statistical physics community for the last 25 years.

An important branch in the research area of statistical physics of SG is represented by the application of this knowledge for modeling brain functions. The simplest and most famous SG model of an associative memory was proposed by Hopfield; it assumes J_{ij} to be given by

$$J_{ij} \ = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^{(\mu)}\ \xi_j^{(\mu)}\ , \tag{11}$$

where the p sets of $\{\xi^{(\mu)}\}_{\mu=1}^{p}$ are given configurations of the system (that we call *prototypes*) having the following properties: *(a)* $\xi^{(\mu)} \perp \xi^{(\nu)}, \forall \mu \neq \nu$; *(b)* $p = \alpha N, \alpha \leq 0.14, N \to \infty$. Under these assumptions it has been proved that the $\{\xi^{(\mu)}\}_{\mu=1}^{p}$ are the absolute minima of E [1]; for $\alpha > 0.14$ the system loses its storage capability [1]. These results can be extended from the discrete to the continuous case (i.e. $\mathbf{s} \in [-1, +1]^N$, see [6]); note that this extension is crucial in the construction of the SG-MRF model.

It is interesting to note that the energy (10), with the prescription (11), can be written as:

$$E = -\frac{1}{N} \sum_{i,j} \sum_{\mu} \xi_i^{(\mu)} \xi_j^{(\mu)} s_i s_j = -\frac{1}{N} \sum_{\mu} (\xi^{(\mu)} \cdot \mathbf{s})^2. \tag{12}$$

Equation (12) depends on the data through scalar products, thus it can be *kernelized*, as to say it can be written as

$$E_{KAM} = \frac{1}{N} \sum_{\mu} [K(\xi^{(\mu)}, \mathbf{s})]^2. \tag{13}$$

The idea to substitute a kernel function, representing the scalar product in a higher dimensional space, in algorithms depending on just the scalar products between data is the so called *kernel trick* [25], which was first used for Support Vector Machines (SVM); in the last few years theoretical and experimental results have increased the interest within the machine learning and computer vision community regarding the use of kernel functions in methods for classification, regression, clustering, density estimation and so on. We call the energy given by equation (13) Kernel Associative Memory (KAM). KAM energies are of interest in two different research fields: in the formulation given by equation (13) it is a non linear and higher order generalization of the Hopfield energy function [4]. The other research field is computer vision, on which we concentrate the attention here. Indeed, we can look at equation (13) as follows:

$$E = \frac{1}{N} \sum_{\mu} (\xi^{(\mu)} \cdot \mathbf{s})^2 = -\frac{1}{N} \sum_{\mu} [\Phi(\mathbf{h}^{\mu}) \cdot \Phi(\mathbf{h})]^2 = -\frac{1}{N} \sum_{\mu} [K(\mathbf{h}^{\mu}, \mathbf{h})]^2 \tag{14}$$

provided that Φ is a mapping such that (see Figure 1):

$$\Phi : G \equiv \Re^m \rightarrow H \equiv [-1,+1]^N, N \rightarrow \infty,$$

that in terms of kernel means

$$K(\mathbf{h},\mathbf{h}) = 1, \forall \quad \mathbf{h} \in \Re^m, dim(H) = N, N \rightarrow \infty. \tag{15}$$

If we can find such a kernel, then we can use the KAM energy, with all its properties, for MRF modeling. As the energy is fully connected and the minima of the energy are built by construction, the usage of this energy overcomes all the modeling problems relative to irregular sites for MRF [2]. Conditions (15) are satisfied by generalized Gaussian kernels (8). Regarding the choice of prototypes, given a set of n_k training examples $\{\mathbf{x}_1^\kappa, \mathbf{x}_2^\kappa, \ldots, \mathbf{x}_{n_\kappa}^\kappa\}$ relative to class Ω_κ, the condition to be satisfied by the prototypes is $\xi^{(\mu)} \perp \xi^{(\nu)}, \forall(\mu \neq \nu)$ in the mapped space H, that becomes $\Phi(\tilde{\mathbf{h}}^{(\mu)}) \perp \Phi(\tilde{\mathbf{h}}^{(\nu)}), \forall \mu \neq \nu$ in the data space G. The measure of the orthogonality of the mapped patterns is the kernel function (8) that, due to the particular properties of Gaussian Kernels, has the effect of orthogonalize the patterns in the space H (see Figure 2). Thus, the orthogonality condition is satisfied by default: if we do not want to introduce further criteria for the choice of prototypes, the natural conclusion is to take all the training samples as prototypes. This approximation is called the *naive ansatz*. Note that when a single feature vector is computed from each view, the naive ansatz approximation becomes exact.

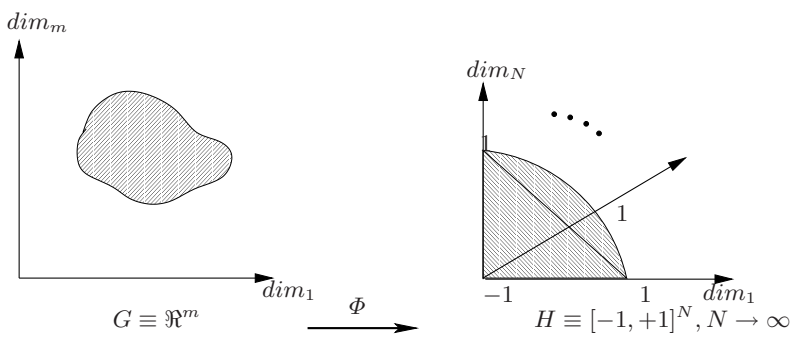

Fig. 2. The kernel trick maps the data from a lower dimension space $G \equiv \Re^m$ to a higher dimension space $H \equiv [-1,+1]^N, N \rightarrow \infty$. This permits to use the H-L energy in a MRF framework

5 Ultrametric Spin Glass-Markov Random Fields

SG-MRF, with the Hopfield energy function (10)-(11), have been successfully applied to appearance-based object recognition. The modeling has been done on

raw pixels [3], on shape representations [2] and on color representations [4]. In all cases, it has shown to be very effective, and has shown remarkable robustness properties. A major drawback of the Hopfield energy function is the condition of orthogonality on the set of prototypes. When the modeling is done on the raw pixel data (as in [3]), or when a single feature vector is computed from a single image (as in [2,4]), then the naive ansatz approximation becomes exact. But there are many applications in which the number of prototypes (as to say the number of features extracted from a single image) can be > 1. This is the case for example in most texture classification problems; it is also the case if we want to combine together shape and color features, as it is the purpose here. Two problems arises in this case: first, whether it is possible or not to combine together these representations. Second, assuming the answer is yes, whether the property of generalized Gaussian kernels is sufficient to ensure the orthogonality of prototypes. In other words, the naive ansatz can turn out to be in some cases too rough an approximation.

The solution we propose consists in kernelizing a new SG energy function, that allows us to store non mutually orthogonal prototypes. 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*. The interested reader will find a complete description of ultrametricity and of the ultrametric energy in [1,12]. In the rest of the Section we will present the ultrametric energy and we will show how it can be kernelized (Section 5.1); we will also show how it can be used for appearance-based object recognition using shape and color information contained in different representations.

5.1 Ultrametric Spin Glass-Markov Random Fields: Model Derivation

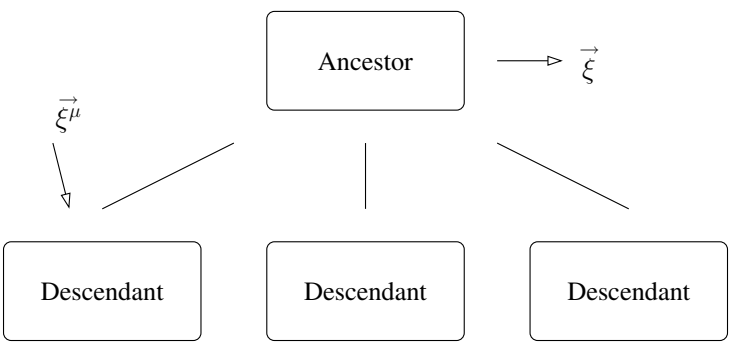

Fig. 3. Hierarchical structure induced by the ultrametric energy function

Consider the energy function (10)

$$E = -\sum_{ij} J_{ij} s_i s_j$$

with the following connection matrix:

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^{(\mu)} \xi_j^{(\mu)} \left(1 + \frac{1}{\Delta(a_\mu)} \sum_{\nu=1}^{q_\mu} (\eta_i^{(\mu\nu)} - a_\mu)(\eta_j^{(\mu\nu)} - a_\mu) \right) \tag{16}$$

with

$$\xi_i^{(\mu\nu)} = \xi_i^{(\mu)} \eta_i^{(\mu\nu)}, \qquad a_\mu^2 = \frac{1}{N} \sum_{i=1}^{N} \eta_i^{(\mu\nu)} \eta_i^{(\mu\lambda)}.$$

This energy induces a hierarchical organization of stored prototypes ([1], see Figure 3). The set of prototypes $\{\xi^{(\mu)}\}_{\mu=1}^{p}$ are stored at the first level of the hierarchy and are usually called the *ancestor*. Each of them will have q *descendants* $\{\xi^{(\mu\nu)}\}_{\nu=1}^{q_\mu}$. The parameter $\eta_i^{(\mu\nu)}$ measures the similarity between ancestors and descendants; the parameter a_μ 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 paper we will limit the discussion to the case[1]

$$a_\mu^2 = a^2.$$

The connection matrix thus becomes:

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^{(\mu)} \xi_j^{(\mu)} \left(1 + \frac{1}{1-a^2} \sum_{\nu=1}^{q_\mu} (\eta_i^{(\mu\nu)} - a)(\eta_j^{(\mu\nu)} - a) \right)$$

$$= \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^{(\mu)} \xi_j^{(\mu)} + \frac{1}{N(1-a^2)} \sum_{\mu=1}^{p} \xi_i^{(\mu)} \xi_j^{(\mu)} \sum_{\nu=1}^{q_\mu} (\eta_i^{(\mu\nu)} - a)(\eta_j^{(\mu\nu)} - a)$$

$$= \text{Term1} + \text{Term2}.$$

Term1 is the Hopfield energy (10)-(11); Term2 is a new term that allows us to store as prototypes patterns correlated with the $\{\xi^{(\mu)}\}_{\mu=1}^{p}$, and correlated between each other. This energy will have $p + \sum_{\mu=1}^{p} q^\mu$ minima, of which p absolute (ancestor level) and $(\sum_{\mu=1}^{p} q^\mu)$ local (descendant level).

When $a \to 0$, the ultrametric energy reduces to a hierarchical organization of Hopfield energies; it is remarkable to note that 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 [1].

[1] Considering the general case would not add anything from the conceptual point of view and would make the notation even heavier.

Here we are interested in using this energy in the SG-MRF framework shown in Section 4. To this purpose, we show that the energy (10), with the connection matrix (16), can be written as a function of scalar product between configurations:

$$E = -\frac{1}{N}\sum_{ij}\left[\frac{1}{N}\sum_{\mu=1}^{p}\xi_i^{(\mu)}\xi_j^{(\mu)}\left(1+\frac{1}{1-a^2}\sum_{\nu=1}^{q_\mu}(\eta_i^{(\mu\nu)}-a)(\eta_j^{(\mu\nu)}-a)\right)\right]s_i s_j$$

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

$$\frac{2a}{N(1-a^2)}\sum_{\mu=1}^{p}\sum_{\nu=1}^{q_\mu}(\xi^{(\mu)}\cdot\mathbf{s})(\xi^{(\mu)}\cdot\mathbf{s})+\frac{a^2}{N(1-a^2)}\sum_{\mu=1}^{p}\sum_{\nu=1}^{q_\mu}(\xi^{(\mu)}\cdot\mathbf{s})^2. \tag{17}$$

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

$$E = -\left[\frac{1}{N^2}\left[\sum_{\mu=1}^{p}(\xi^{(\mu)}\cdot\mathbf{s})^2+\sum_{\mu=1}^{p}\sum_{\nu=1}^{q_\mu}(\xi^{(\mu\nu)}\cdot\mathbf{s})^2\right]\right]. \tag{18}$$

The *ultrametric energy*, in the general form (17) or in the simplified form (18) 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 Random Fields (USG-MRF).

5.2 Ultrametric Spin Glass-Markov Random Fields: Model Application

Consider the probabilistic appearance-based framework described in Section 3. Given a view $\mathbf{x}_{n_j}^j$, we will suppose to extract two feature vectors from it, $\mathbf{hs}_{n_j}^j$ containing shape information and $\mathbf{hc}_{n_j}^j$ containing color information. USG-MRF provides a straightforward manner to use the Bayes classifier (3) using both these two representations separately. We will consider $\mathbf{hc}_{n_j}^j$ as the ancestor and $\mathbf{hs}_{n_j}^j$ as the descendant; for each level there will be a single prototype, thus the naive ansatz approximation will be exact. The USG-MRF energy function will be in this case:

$$E_{USG-MRF} = -\left\{[K_c(\mathbf{hc}_{n_j}^j,\mathbf{hc})]^2+[K_s(\mathbf{hs}_{n_j}^j,\mathbf{hs})]^2\right\}, \tag{19}$$

that leads to the Bayes classifier

$$j^* = \underset{j}{\operatorname{argmin}}\left\{-\left\{[K_c(\mathbf{hc}_{n_j}^j,\mathbf{hc})]^2+[K_s(\mathbf{hs}_{n_j}^j,\mathbf{hs})]^2\right\}\right\}. \tag{20}$$

The indexes c and s relative to the two kernels acting on the two different representations indicates that it is possible to use, at different levels of the hierarchy, different kernels. Here, we would like to remark, that this newly introduced model is adaptable for the combination of any kind of features similary to the demonstarted color and shape.

6 Experiments

In order to show the effectiveness of USG-MRF for appearance-based object recognition, we perform several sets of experiments. All of them were ran on the COIL database [15], 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 × 128 pixels. The images were obtained by placing the objects on a turntable and taking a view every 5°. In all the experiments we performed, 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 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 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 [23]. The CH was normalized to 1. As shape only representation, we chose Multidimensional receptive Field Histograms (MFH). This method was proposed by Schiele in order to extend the color histogram approach of Swain and Ballard; the main idea is to calculate multidimensional histograms of the response of a vector of receptive fields. An MFH is determined once we chose the local property measurements (i.e., the receptive fields functions), which determine the dimensions of the histogram, and the resolution of each axis. SG-MRF has been successfully used many times combined with MFH for appearance-based object recognition [2,4]. Here we chose for all the experiments we performed two local characteristics based on Gaussian derivatives:

$$D_x = -\frac{x}{\sigma^2}G(x,y); \qquad D_y = -\frac{y}{\sigma^2}G(x,y)$$

where

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

is the Gaussian distribution. Thus, our shape only representation consisted of two dimensional MFH, D_xD_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. **Shape experiments**: we ran the experiments using the shape features only. Classification was performed using SG-MRF with the kernelized Hopfield energy (10)-(11). The kernel parameters (a, b, ρ) were learned using a leave--one-out strategy. The results were benchmarked with those obtained with a χ^2 and $\cap$ similarity measures, which proved to be very effective for this representation.

Fig. 4. Examples of different views

2. **Color experiments**: we ran the experiments using the color features only. Classification and benchmarking were performed as in the shape experiment.
3. **Color-shape experiments**: we ran the experiments using the color and shape features combined together to form a unique feature vector. Again, classification and benchmarking were performed as in the shape experiment.
4. **Ultrametric experiment**: we ran a single experiment using the shape and color representation disjoint in the USG-MRF framework. The kernel parameters relative to each level (a_s, b_s, ρ_s and a_c, b_c, ρ_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 1 reports the error rates obtained for the 4 sets of experiments.

Results presented in Table 1 show that for all series of experiments, for all representations, SG-MRF always gave 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 χ^2 (best result obtained with a non SG-MRF technique). The fact that the error rates for the color experiments are all above 20% is an

Table 1. Classification results; we report for each set of experiments the obtained error rates. The kernel parameters learned for SG-MRF, for the color experiment were $a_c = 0.5, b_c = 0.4, \rho_c = 0.1$. For the shape experiment were $a_s = 0.4, b_s = 1.3, \rho_s = 0.1$. For the color-shape experiment were $a_{cs} = 0.3, b_{cs} = 0.6, \rho_{cs} = 0.1$ and finally for the ultrametric experiment were $a_c = 0.5, b_c = 0.4, \rho_c = 0.016589, a_s = 0.4, b_s = 1.3, \rho_s = 2.46943$

	Color (%)	Shape (%)	Color-Shape (%)	Ultrametric (%)
χ^2	23.47	9.47	19.17	
$\cap$	25.68	24.94	21.72	
SG-MRF	**20.10**	**6.28**	**8.43**	**3.55**

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 shape appearance-based object recognition.

7 Summary

In this paper we presented a kernel method that permits us to combine color and shape information for appearance-based object recognition. It does not require us to define a new common representation, but use the power of kernels to combine different representations together in an effective manner. This result is achieved using results of statistical mechanics of Spin Glasses combined with Markov Random Fields via kernel functions. Experiments confirm the effectiveness of the proposed approach. Future work will explore the possibility to use different representations for color and shape and will benchmark this approach with others, presented in literature.

Acknowledgments This work has been supported by the "Graduate Research Center of the University of Erlangen-Nuremberg for 3D Image Analysis and Synthesis", and by the Foundation BLANCEFLOR Boncompagni-Ludovisi.

References

1. D. J. Amit, *"Modeling Brain Function"*, Cambridge University Press, 1989. 98, 101, 103, 105, 106
2. B. Caputo, H. Niemann, "From Markov Random Fields to Associative Memories and Back: Spin Glass Markov Random Fields", SCTV2001. 98, 101, 102, 104, 105, 108
3. B. Caputo, J. Hornegger, D. Paulus, H. Niemann, "A Spin Glass-Markov Random Field", *Proc ICANN01 workshop on Kernel Methods*, Vienna, 2001. 105
4. B. Caputo, "A new kernel method for object recognition and scene modeling: Spin Glass-Markov Random Fields", *PhD thesis*, to appear. 103, 105, 108
5. T. Evgeniou, M. Pontil, C. Papageorgiou, T. Poggio, "Image representations for object detection using kernel classifiers" ACCV, 2000. to appear. 99
6. J. J. Hopfield, "Neurons with graded response have collective computational properties like those of two-state neurons", *Proc. Natl. Acad. Sci. USA*, Vol. 81, pp. 3088- 3092, 1984. 103
7. C.-Y. Huang, O. I. Camps, "Object recognition using appearance-based parts and relations",CVPR'97:877-883, 1997. 99
8. A.Jain,P. J. Flynn, editors. "Three–Dimensional Object Recognition Systems", Amsterdam, Elsevier, 1993. 97, 98
9. A. Leonardis, H. Bischof, "Robust recognition using eigenimages", CVIU,78:99-118, 2000. 97, 99
10. J. Matas, R. Marik, J. Kittler, "On representation and matching of multi-coloured objects", *Proc ICCV95*, 726-732, 1995. 97, 99
11. B. W. Mel, "SEEMORE: combining color, shape and texture histogramming in a neurally-inspired approach to visual object recognition", *NC*, 9: 777-804, 1997 97, 99

12. M. Mezard, G. Parisi, M. Virasoro, *" Spin Glass Theory and Beyond"*, World Scientific, Singapore, 1987. 103, 105
13. J. W. Modestino, J. Zhang. "A Markov random field model–based approach to image interpretation". *PAMI*, 14(6),606–615,1992. 101
14. H. Murase, S. K. Nayar, "Visual Learning and Recognition of 3D Objects from Appearance", *IJCV*,14(1):5-24, 1995. 99
15. Nene, S. A., Nayar, S. K., Murase, H., "Columbia Object Image Library (COIL-100)", *TR* CUCS-006-96, Dept. Comp. Sc., Columbia University, 1996. 97, 108
16. E. Osuna, R. Freund, F. Girosi, "Training support vector machines: An application to face detection", CVPR'97: 130-136, 1997. 99
17. J. Ponce, A. Zisserman, M. Hebert, "Object Representation in Computer Vision—II", Nr. 1144 in LNCS. Springer, 1996. 98
18. Pontil, M., Verri, A. "Support Vector Machines for 3D Object Recognition", *PAMI*, 20(6):637-646, 1998. 99
19. R. P. N. Rao, D. H. Ballard, "An active vision architecture based on iconic representations",AI:461– 505, 1995. 99
20. D. Roobaert, M. M. Van Hulle, "View-based 3D object recognition with support vector machines", Proc. IEEE Workshop. on NNSP, 1999. 99
21. B. Schiele, J. L. Crowley, "Recognition without correspondence using multidimensional receptive field histograms", *IJCV*, 36(1),:31- 52, 2000. 97, 99, 100
22. 97, 99
D. Slater, G. Healey, "Combining color and geometric information for the illuminantion invariant recognition of 3-D objects", *Proc ICCV95*, 563-568, 1995.
23. M. Swain, D. Ballard, "Color indexing",*IJCV*, 7(1):11-32, 1991. 97, 98, 108
24. M. Turk, A. Pentland, "Eigenfaces for recognition", Journal of Cognitive Neuroscience,3(1):71–86, 1991. 99
25. V. Vapnik, *Statistical learning theory*, J. Wiley, New York, 1998. 103
26. M. D. Wheeler, K. Ikeuchi. *Sensor modeling, probabilistic hypothesis generation, and robust localization for object recognition. PAMI*, 17(3):252–265, 1995. 101
27. B. Schölkopf, A. J. Smola, *Learning with kernels*, 2002, the MIT Press, Cambridge, MA. 102