

# To Each According to its Need: Kernel Class Specific Classifiers

B. Caputo and H. Niemann  
Computer Science Department, Chair for Pattern Recognition  
University of Erlangen-Nuremberg,  
Martenstrasse 3, D-91058, Erlangen, Germany,  
{ caputo, niemann }@informatik.uni-erlangen.de

## Abstract

We present in this paper a new multi-class Bayes classifier that permits using separate feature vectors, chosen specifically for each class. This technique extends previous work on feature Class Specific Classifier to kernel methods, using a new class of Gibbs probability distributions with nonlinear kernel mapping as energy function. The resulting method, that we call Kernel Class Specific Classifier, permits using a different kernel and a different feature set for each class. Moreover, the proper kernel for each class can be learned by the training data with a leave-one-out technique. This removes the ambiguity regarding the proper choice of the feature vectors for a given class. Experiments on appearance-based object recognition show the power of the proposed approach.

## 1 Introduction

This paper <sup>1</sup> presents a new multi-class Bayes classifier that permits using a different representation for each considered class. 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, zucchinis, 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 recognizable between others. This representa-

tion(s) can be different for different classes. Assuming it is possible to determine it (them), 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 [3].

Baggenstoss et al. [8, 2] 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 [8, 2]. 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 one, 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 done 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 [3, 7, 9]. 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 [10]. 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

---

<sup>1</sup>This work has been supported by the “Graduate Research Center of the University of Erlangen-Nuremberg for 3D Image Analysis and Synthesis”.

functions is known and limited [10]. 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 Spin Glass-Markov Random Fields (SG-MRFs). SG-MRFs are a new class of MRF [6] that use results of SG theory [1] via kernel methods [4]. The SG-MRF probability distribution uses a particular class of kernel functions -the Gaussian kernels -as energy function [4]. 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 function 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 paper is organized as follows: after the formulation of the general problem, we review CSC and SG-MRF methods (Section 2); Section 3 derives the K-CSC, and Section 4 presents experiments on appearance-based object recognitions. The paper concludes with a summary discussion.

## 2 A Few Landmarks

### 2.1 Problem Definition

Consider a given pattern recognition problem: let  $H_j, j = 1, \dots, M$  be  $M$  different classes or statistical hypothesis: given a data sample  $\mathbf{x}$ , produced by one of  $M$  possible classes, our goal is to classify  $\mathbf{x}$  as a sample from  $H_{j^*}$ , one of the  $H_j$  classes. Using the Maximum A Posteriori (MAP) classifier we get:

$$j^* = \operatorname{argmax}_j P(H_j|\mathbf{x}) = \operatorname{argmax}_j \{P(\mathbf{x}|H_j)P(H_j)\}; \quad (1)$$

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

$$j^* = \operatorname{argmax}_j P(\mathbf{x}|H_j).$$

### 2.2 The Feature-based Class Specific Classifier

The LFs have to be learned from data samples (training data). The usual approach extracts a small number of information-bearing statistics, called *features*. Let  $\mathbf{z} =$

$T(\mathbf{x})$  be such a set of features. The Bayesian classifier based on  $\mathbf{z}$  is

$$j^* = \operatorname{argmax}_j P(\mathbf{z}|H_j). \quad (2)$$

Thus, the features replace the raw data. The hidden implication here is that  $\mathbf{z}$  is a sufficient statistics for the classification problem:

$$P(\mathbf{x}|H_j) = g(T(\mathbf{x})|H_j)h(\mathbf{x}), j = 0, \dots, M.$$

This is the Neyman-Fisher factorization theorem [5]. The well known corollary of this theorem is that any likelihood ratio is invariant when written in terms of a sufficient statistic. Thus, if  $\mathbf{z}$  is a sufficient statistic,

$$\frac{P(\mathbf{x}|H_j)}{P(\mathbf{x}|H_k)} = \frac{P(\mathbf{z}|H_j)}{P(\mathbf{z}|H_k)} \quad (3)$$

The  $M$ -ary classifier (1) can be constructed by knowing only the likelihood ratios; moreover, it is possible to use in the denominator an additional class,  $H_0$ :

$$j^* = \operatorname{argmax}_j \frac{P(\mathbf{x}|H_j)}{P(\mathbf{x}|H_0)}. \quad (4)$$

The direct consequence of equations (3)-(4) is that we may write the Bayes classifier as [8]

$$j^* = \operatorname{argmax}_j \frac{P(\mathbf{z}_j|H_j)}{P(\mathbf{z}_j|H_0)}, \quad (5)$$

where  $\mathbf{z}_j = T_j(\mathbf{x}), 1 \leq j \leq M$  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); CSCs major advantage is that allow to use different features for each class; a serious drawback is that they do not provide any criteria on how to chose the best possible set of features for each class.

### 2.3 Spin Glass-Markov Random Fields

Spin Glass-Markov Random Fields (SG-MRFs) [4] are a new class of fully connected MRFs [6] which use concepts and theoretical results developed within the framework of statistical physics of disordered systems [1]. They connect SG-like energy functions (mainly the Hopfield one, [1]) with Gibbs distributions via a non linear kernel mapping; the resulting model overcomes MRF modeling problems for irregular sites [6, 4], and enables to use the power of kernels in a probabilistic framework.

For a given data sample  $\mathbf{x}$ , generated by one of the statistical sources  $H_j, j = 0, \dots, M$ , the SG-MRF probability distribution is given by

$$P_{SG-MRF}(\mathbf{x}|H_j) = \frac{1}{Z} \exp[-E_{SG-MRF}(\mathbf{x}|H_j)], \quad (6)$$

$$Z = \sum_{\{\mathbf{x}\}} \exp[-E_{SG-MRF}(\mathbf{x}|H_j)],$$

with

$$E_{SG-MRF} = - \sum_{\mu=1}^{p_j} \left[ K(\mathbf{x}, \tilde{\mathbf{x}}^{(\mu)}) \right]^2, \quad (7)$$

where the function  $K(\mathbf{x}, \tilde{\mathbf{x}}^\mu)$  is a Generalized Gaussian kernel [10]:

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

$\{\tilde{\mathbf{x}}^\mu\}_{\mu=1}^{p_j}, j \in [0, M]$  are a set of vectors selected (according to a chosen ansatz, [4]) from the training data that we call *prototypes*. The number of prototypes per class must be finite, and they must satisfy the condition:

$$K(\tilde{\mathbf{x}}^i, \tilde{\mathbf{x}}^k) = 0, \quad (8)$$

for all  $i, k = 1, \dots, p_j, i \neq k$  and  $j = 0, \dots, M$ . The interested reader can find a detailed discussion regarding the derivation and properties of SG-MRF in [4].

### 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, a *kernel* is a function  $K$  such that, for all  $\mathbf{x}, \mathbf{y} \in X$ ,

$$K(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x})^T \Phi(\mathbf{y}),$$

where  $\Phi$  is a mapping from  $X$  to an (inner product) feature space  $F$  [10]. Thus, the SG-MRF energy function can be rewritten as:

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

where  $\tilde{K}$  represents a new kernel function [10]. So we can write the SG-MRF energy as

$$E_{SG-MRF} = - \sum_{\mu=1}^{p_j} \tilde{K}(\mathbf{x}, \tilde{\mathbf{x}}^\mu) = - \sum_{\mu=1}^{p_j} \Phi(\mathbf{x})^T \Phi(\tilde{\mathbf{x}}^\mu),$$

and the SG-MRF probability distribution becomes

$$P_{SG-MRF}(\mathbf{x}|H_j) = \frac{1}{Z} \exp \left[ \sum_{\mu=1}^{p_j} \Phi(\mathbf{x})^T \Phi(\tilde{\mathbf{x}}^\mu) \right]. \quad (9)$$

Equation (9) tells that  $P_{SG-MRF}$  depends on  $\mathbf{x}$  via a mapping  $\Phi(\mathbf{x}) = \mathbf{z}$ . Thus, we can use this probability in the

CSC classifier (5), identifying the feature extraction operator  $T_j(\mathbf{x})$  with the mapping  $\Phi_j(\mathbf{x})$ , as to say using a different mapping, and thus a *different kernel* for each class. We get:

$$\begin{aligned} j^* &= \operatorname{argmax}_j \frac{P_{SG-MRF}(\Phi_j(\mathbf{x})^T | H_j)}{P_{SG-MRF}(\Phi_j(\mathbf{x})^T | H_0)} \\ &= \operatorname{argmax}_j \frac{\frac{1}{Z} \exp \left[ \sum_{\mu_j} \Phi_j(\mathbf{x})^T \Phi_j(\tilde{\mathbf{x}}^{\mu_j}) \right]}{\frac{1}{Z} \exp \left[ \sum_{\mu_0} \Phi_j(\mathbf{x})^T \Phi_j(\tilde{\mathbf{x}}^{\mu_0}) \right]} \end{aligned}$$

where  $\{\tilde{\mathbf{x}}^{\mu_j}\}, \mu_j = 1, \dots, p_j$  are the set of prototypes of class  $H_j$ ;  $\{\tilde{\mathbf{x}}^{\mu_0}\}, \mu_0 = 1, \dots, p_0$  are the set of prototypes of class  $H_0$ . As the mapping  $\Phi_j$  is the same for the numerator and denominator, the constant  $Z$  is the same for both terms, thus it simplifies. It follows:

$$\begin{aligned} &= \operatorname{argmax}_j \exp \left[ \sum_{\mu_j} \Phi_j(\mathbf{x})^T \Phi_j(\tilde{\mathbf{x}}^{\mu_j}) - \sum_{\mu_0} \Phi_j(\mathbf{x})^T \Phi_j(\tilde{\mathbf{x}}^{\mu_0}) \right] \\ &= \operatorname{argmin}_j \left[ - \sum_{\mu_j} \Phi_j(\mathbf{x})^T \Phi_j(\tilde{\mathbf{x}}^{\mu_j}) + \sum_{\mu_0} \Phi_j(\mathbf{x})^T \Phi_j(\tilde{\mathbf{x}}^{\mu_0}) \right]. \end{aligned}$$

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

$$j^* = \operatorname{argmin}_j \left[ - \sum_{\mu_j} [K_j(\mathbf{x}, \tilde{\mathbf{x}}^{\mu_j})]^2 + \sum_{\mu_0} [K_j(\mathbf{x}, \tilde{\mathbf{x}}^{\mu_0})]^2 \right]. \quad (10)$$

Given a training set, the kernel  $K_j$  can be *learned*, for each class  $H_j$ , with a leave-one-out technique. 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. The reader could wonder whether the  $\mathbf{z}_j = \Phi_j(\mathbf{x})$  are a sufficient statistics for the class  $H_j$ , as required by CSCs. It could be argued that it is not, as the mapping  $\Phi_j$  is a mapping in a higher dimensional space. The point is that, although  $\Phi_j$  maps the data into a higher dimensional space, it can be proved that the mapped data are embedded in a subspace of the mapped space  $F$ , which will be of dimension lower or equal to the dimension of the data set [10]. Thus, if  $\mathbf{x}$  is a sufficient statistic for the class  $H_j$ , so it will be  $\Phi_j(\mathbf{x})$ .

### 4 Experimental Results and Discussion

The application of the CSC method to vision problems is possible by principle, but has been very challenging until

now [8]; on the other side, many experiments show that SG-MRF are very effective for vision applications, particularly for appearance-based object recognition [4]. For these reasons, we decided to test the K-CSC performance on an appearance-based object recognition application. We ran all the experiments on a database of 59 objects [7]: 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 1. Each object is represented in the training set by a collection of views taken approximately 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 53 (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 53 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.

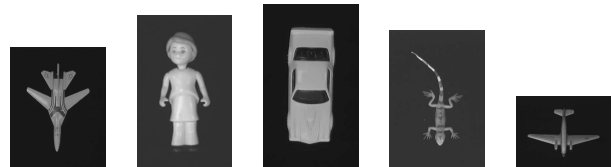
As the views in the database are of different sizes, we decided to use a Multidimensional receptive Field Histogram (MFH) representation for all classes [9], that has been already used successfully combined with SG-MRF [4]. Thus, the power of K-CSC will rely entirely on the choice of kernels. We used 2D MFH, with filters given by Gaussian derivatives along  $x$  and  $y$  directions and with  $\sigma = 1.0$ ; resolution for histogram axis 16 bins. We ran 3 sets of experiments: in the first we used 8 kernels ( $a = 1, 0.5; b = 0.5, 1, 1.5, 2$ ) and we performed the classification with K-CSC and SG-MRF used in a MAP-Bayes classifier [4]. As common reference hypothesis  $H_0$  we used a portion of the background of the database views. For the choice of prototypes, we made a naive ansatz [4], which means that all training views are taken as prototypes, and the  $\rho$  in the Gaussian kernel is learned so to satisfy condition (8). For SG-MRF we ran 8 experiments -one for each kernel -and we report just the best result. For K-CSC, for each of the 59 object classes, a kernel is selected out of the possible 8, using a leave-one-out technique. The other 2 sets of experiments were ran with the same procedure, but increasing each time the number of possible kernels: 20 ( $a = 1, 0.8, 0.6, 0.4, 0.2; b = 0.5, 1, 1.5, 2$ ) and 40 ( $a = 1, 0.9, \dots, 0.1$ , stepwise reduced;  $b = 0.5, 1, 1.5, 2$ ). The obtained recognition rates are reported in Table 1. K-CSC performs always better than SG-MRF; note also that, as the number of kernels increases, the performance of SG-MRF remains basically the same, while the K-CSC's increases, with respect to that of SG-MRF, up to + 3.87 %.

## 5 Conclusions

We presented in this paper a new kernel classifier that permits to use different kernels for different classes; this results in a remarkable increase in the recognition rate with respect to a standard Bayes classifier using the same kernel method.

	SG-MRF	K-CSC
8 kernels	<b>10.78</b> ; a=0.5, b=1	<b>8.93</b>
20 kernels	<b>10.51</b> ; a=0.4, b=1	<b>6.71</b>
40 kernels	<b>10.51</b> ; a=0.4, b=1	<b>6.59</b>

**Table 1. Classification results for SG-MRF and K-CSC for different choices of kernels parameters. We report the error rates.**



**Figure 1. An example of 5 objects of the 59 contained into the used database. Views have different sizes for different objects and for different pose parameters.**

In the future we plan to explore the robustness of the K-CSC and to compare its performance to that of other kernel methods, mainly Support Vector Machines.

## References

- [1] D. Amit. *Modeling Brain Function: The World of Attractor Neural Networks*. Cambridge University Press., 1989.
- [2] P. M. Bagginstoss. Class-specific features in classification. *IEEE Trans. SP*, pages 3428-3432, 1999.
- [3] C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.
- [4] B. Caputo and H. Niemann. From markov random fields to associative memories and back: Spin-glass markov random fields. *SCTV*, 2001.
- [5] D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman and Hall, London, 1974.
- [6] S. Z. Li. *Markov Random Field Modeling in Computer Vision*. Springer-Verlag, 1995.
- [7] R. C. Nelson and A. Selinger. A cubist approach to object recognition. *ICCV98*, pages 614-621, 1998.
- [8] H. N. P. M. Bagginstoss. A theoretically optimal probabilistic classifier using class-specific features. *ICPR00*, pages 767-772, 2000.
- [9] B. Schiele and J. L. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *IJCV*, 36(1):31-52, 2000.
- [10] B. Schölkopf and A. J. Smola. *Learning with kernels*. MIT Press, 2001.