A hybrid generative/discriminative classification framework based on free energy terms

A.Perina[1] M.Cristani[1,2] U.Castellani[1] V.Murino[1,2] N.Jojic[3]

[1] University of Verona, Strada le Grazie 15, 37134 Verona, Italy
[2] IIT - Istituto Italiano di Tecnologia, Via Morego 30, 16163 Genova, Italy
[3] Microsoft Research. One Microsoft Way. Redmond, WA 98052

Abstract

Hybrid generative-discriminative techniques and, in particular, generative score-space classification methods have proven to be valuable approaches in tackling difficult object or scene recognition problems. A generative model over the available data for each image class is first learned, providing a relatively comprehensive statistical representation. As a result, meaningful new image features at different levels of the model become available, encoding the degree of fitness of the data with respect to the model at different levels. Such features, defining a score space, are then fed into a discriminative classifier which can exploit the intrinsic data separability. In this paper, we present a generative score-space technique which encapsulates the uncertainty present in the generative learning phase usually disregarded by the state-of-the-art methods. In particular, we propose the use of variational free energy terms as feature vectors, so that the degree of fitness of the data and the uncertainty over the generative process are included explicitly in the data description. The proposed method is automatically superior to a pure generative classification, and we also experimentally illustrate it on a wide selection of generative models applied to challenging benchmarks in hard computer vision tasks such as scene, object and shape recognition. In several instances, the proposed approach beats the current state of the art in classification performance, while relying on computationally inexpensive models.

1. Introduction

The design of models for classification and recognition purposes is one of the fundamental issues in computer vision. Among the several possible taxonomies, two apparently orthogonal approaches can be found in the literature: the generative and the discriminative paradigms. Generative models encode intuitively data correlations and a-priori knowledge by means of a graph structure, establishing a "correspondence" between parts of the model and features in the image through a hierarchy of conditional distributions. These models allow for integration over hidden variables, and can thus deal elegantly with missing, unlabeled and varying-length data. Likelihoods under such parameterized class-specific models can then be used for classification using the Bayes rule. Discriminative methods, on the other hand, define the separation boundaries among classes, rather than the distribution over instances of a the class, and tend to maximize data separability. Depending on the features, this often results in classification performances higher than those achieved by the generative models, especially when large training sets are available [17]. Recently, hybrid generative-discriminative approaches have been proposed with the goal of benefiting from the best of both paradigms. Several lines of research have been pursued here. On one hand, there are the generative score-space methods, which represent a large family of two-step techniques, in which a generative phase is followed by a discriminative counterpart. For example, in [3], a classification framework is proposed which uses by-products of generative models, like the conditional distributions, as features for a discriminative machine. In the same class of methods, we find the similarity-based approaches [1, 20] which concatenate class conditional likelihoods of different generative models providing feature patterns which can be separated by discriminative techniques. Finally, "classical" score-space methods [7, 21] are characterized by their ability to grab deeper structural information from the generative models, evaluating how well the test samples "agree" with the generative process described by the models' parameters, under the form of fixed-length generative score. For example, the Fisher score [7] extracts features from the first-order derivative of the model likelihood (for a given test sample). The other big branch of the generative-discriminative family aims at blending the two orthogonal perspectives employing generative models which are discriminatively trained, exploiting hyperparameters which determine which

PARADIGM	PROS	CONS
Generative [8,12]	• Manages - unlabelled data - variable length data - need few data	• Does not exploits data separability
Discriminative [11,18,22]	• **TH:** Superior to generative when enough data	• Does not manage - unlabelled data - variable length data
Score based [1,3,6,7,20,21]	• Manages - unlabelled data - variable length data • **TH:** Superior to generative	• Does not take into account model uncertainty
Hybrid [10,14]	• Manages - unlabelled data - variable length data • Superior to generative	• Computationally demanding (only numerical optimizations)

Figure 1. Comparative table of pros and cons of different classification paradigms. **TH** stands for facts for which does exist a theoretical explanation.

part of the model (generative or discriminative) has to be favored [10, 14]. A comparison of the typical existing generative, discriminative and hybrid methods is given in Fig. 1. The prevailing wisdom is that generative score-based approaches can provide superior performances since features derived from a generative model capture both data and generative process; in this sense can be understood the good performances on scene recognition by [3], or the theoretical results of [10, 14].

In this paper, we present a generative score space method which considers the free energy of a generative model as a primary source of features for classification. Free energy is a popular function in statistical physics which, in this context, is usually to be minimized in variational model training, which represent a lower bound on the negative log-likelihood of the observations. " A valuable aspect of the free energy, exploited in our framework, is that it reflects both the uncertainty accumulated during the training process and the fit of a new data point to the learned model, as a summation of terms. Such terms can be partitioned in a *entropy* set and in a *cross-entropy* set. The latter encodes errors in the model's fit to the data, distributing such discrepancies across several terms, each one focusing on a particular factor of the generative joint distribution. The former encodes the uncertainty of this fit, again separated across the model components. The entropy and cross-entropy terms are then employed as features for discriminative classifiers. Asymptotically, the proposed framework can trivially be shown to perform at least as well as the generative model upon which is built. Moreover, as illustrated in the experimental section, our approach tends to outperform the performance of generative score-space based methods proposed in the literature. Finally, it has been employed to tackle a wide range of applications here, using challenging datasets and providing results superior to the current state-of-the-art classification performance.

The rest of the paper is organized as follows. In Sec. 2, the proposed framework is described, with a mathematical analysis showing why our generative score space achieves better classification performances than the corresponding generative model alone. A mathematical comparative analysis with respect to state of the art techniques is also provided. Sec. 3 details experimental comparisons. Finally, the contributions of the work are summarized in Sec. 4.

2. The proposed score space framework

Our framework, as common for the generative score space methods, can be partitioned into two steps: first, features from a generative model are extracted, and second, the features forming the score space are analyzed by discriminative methods.

2.1. Extraction of free energy features

We denote a generative model as the joint distribution $P(y,x) = P(\theta)\prod_{t=1}^{T} P(y^{(t)}, x^{(t)}|\theta)$ where $x = \{x^{(t)}\}_{t=1}^{T}$ is a set of i.i.d. training observations, with $x^{(t)} \in \mathcal{R}^{d(t)}$, $d(t) \in \mathcal{N}\ \forall\ t = 1 \dots T$; $y = \{y^{(1)}, \dots, y^{(T)}, \theta\}$ is the set of hidden variables associated with the individual observations, and θ a set of parameters shared across observations.

Once a generative model has been learned, data analysis consists of probabilistic inference, i.e., computing estimates of all the hidden quantities by evaluating the posterior distribution $P(y|x)$. This is often intractable, and variational inference is used as a fast approximation keeping some uncertainty in the posterior distribution, but avoiding correlations which would lead to combinatorial explosion [4]. The main idea is to approximate $P(y|x)$ by a simpler variational distribution $Q(y) = \prod_{t=1}^{T} Q(y^{(t)})$, which can be easily evaluated.

Free energy [9] is a score function whose minimization ensures high similarity between exact posterior and variational distributions, and is commonly arranged as:

$$\mathcal{F} = \mathbb{KL}(Q,P) - \ln P(x) = \sum_{[y]} Q(y) \ln \frac{Q(y)}{P(y,x)} \qquad (1)$$

where $\mathbb{KL}$ stands for the Kullback-Leibler (KL) divergence, evaluated between the exact posterior and the Q distribution, and $\sum_{[y]}$ indicates the sum over all the possible values assumed by each of the hidden variables.

The KL divergence is always positive and zero only if $Q(y)$ equals the true posterior probability; therefore, minimizing $\mathcal{F}$ with respect to Q will always provide the negative log-likelihood, i.e., $-\ln P(x)$.

Minimization of $\mathcal{F}$ is usually achieved by variational inference using Expectation Maximization (EM) algorithm [9] which alternates between optimizing $\mathcal{F}$ with respect to Q and θ respectively, holding the other fixed. To make the inference tractable Q can be constrained to belong to a simplified family of distributions $\mathcal{Q}$ [4].

Once the parameters are estimated, namely $\hat{\theta}$, we can rearrange equation 1 to exploit the i.i.d. characteristic of the data and to highlight the contributions of each sample to the final free energy, i.e.,

$$\begin{aligned}\mathcal{F} &= \sum_t \left(\sum_{[y^{(t)}]} Q(y^{(t)}|\hat{\theta}) \cdot \ln Q(y^{(t)}|\hat{\theta}) - \right. \\ &- \left. \sum_{[y^{(t)}]} Q(y^{(t)}|\hat{\theta}) \cdot \ln P(y^{(t)}, x^{(t)}|\hat{\theta}) \right) \end{aligned} \quad (2)$$

which can be rewritten as $\mathcal{F} = \sum_t \mathcal{F}^t$. In equation 2, the first term in each $\mathcal{F}^t$, the entropy of the posterior, encodes the *ambiguity* in the data fit; the second term, the cross-entropy term, approximates the *divergence* between the sample $x^{(t)}$ and the model parameters $\hat{\theta}$. The factorization properties of the generative model and the particular choice of the family $\mathcal{Q}$, also allow for these ambiguity and divergence patterns in $\mathcal{F}^t$ to be further highlighted locally in a summation of terms, each corresponding to a portion of the generative process. For example, if the generative model is described by a Bayesian network, its joint distribution can be written as $P(\mathbf{v}) = \prod_i P(v_i|\mathbf{PA}_i)$, where $\mathbf{v}$ denotes the set of the variables (hidden or visible) and $\mathbf{PA}_i$ are the parents of v_i. We can compute the contributions for each of the hidden variables of the model in the cross-entropy term (assuming here for convenience that each variable has the same dimensionality d):

$$\begin{aligned}&\textstyle\sum_{v_1^{(t)}=1\ldots d} Q(v_1^{(t)} \cup \mathbf{PA}_1|\hat{\theta}) \cdot \ln P(v_1^{(t)}|\mathbf{PA}_1, \hat{\theta}) + \ldots \\ &+ \textstyle\sum_{v_N^{(t)}=1\ldots d} Q(v_N^{(t)} \cup \mathbf{PA}_N|\hat{\theta}) \cdot \ln P(v_N^{(t)}|\mathbf{PA}_N, \hat{\theta}) \end{aligned} \quad (3)$$

Moreover, for discrete hidden variables, each summation of Eq. 3 can be further organized as a summation of factors over the values that the related variable, e.g., for v_j,

$$\begin{aligned}&Q(v_j^{(t)} = 1, \cup \mathbf{PA}_j|\hat{\theta}) \cdot \ln P(v_j^{(t)} = 1|\mathbf{PA}_j, \hat{\theta}) + \ldots \\ &+ \; Q(v_j^{(t)} = d, \cup \mathbf{PA}_j|\hat{\theta}) \cdot \ln P(v_j^{(t)} = d|\mathbf{PA}_j, \hat{\theta}) \end{aligned} \quad (4)$$

In a similar fashion, the entropy term $\sum_{[y^{(t)}]} Q(y^{(t)}|\hat{\theta}) \cdot \ln Q(y^{(t)}|\hat{\theta})$ of $\mathcal{F}^t$ can be further decomposed into a sum of terms as dictated by the factorization in the particular family $\mathcal{Q}$ chosen for the posterior distributions.

At this point, we can formally define a feature extractor as a function ζ that maps a given a point (an image for example) into a score space composed by the pieces of free energy $\mathcal{F}^t$ under a model $\hat{\theta}$ as $\zeta(x^{(t)}, \hat{\theta}) = [f_1^{(t)}, \ldots, f_i^{(t)}, \ldots, f_M^{(t)}]$, where M is the number of factors involved in the summation $\mathcal{F}^t = \sum_i^M f_i^{(t)}$. By the above recipe, $\mathcal{F}^t$ can be broken down into a different number of terms, as either the coarse factorization of (3) or a further refined factorization of (4) can be used. Furthermore, contributions from various parts of the model can be summed together to control the dimensionality of the feature vector $\zeta(x^{(t)}, \hat{\theta})$. The experimental section will provide specific examples of this strategy.

Finally, given the learned models θ_c for different classes $c = 1 \ldots C$, we concatenate all the free energy pieces of all models together to form the vector of free energy components describing each data point:

$$\mathbf{FE} = \phi_\Theta^{FE}(x^{(t)}) = [\ \zeta(x^{(t)}, \hat{\theta}_1), \ldots, \zeta(x^{(t)}, \hat{\theta}_C)\ ] \quad (5)$$

2.2. Free energy decomposition/selection outperforms generative classification

By the terminology introduced in [21], ϕ_Θ^{FE} is a *model-dependent feature extractor* because different generative models θ_c lead to different feature vectors. As in [21] where TOP (and Fisher) kernels were shown to outperform the generative model on which they are based, we can directly show that a similar use of ϕ_Θ^{FE} will also outperform generative model alone.

Let $x \in \mathcal{X}$ be the input points, and $y \in \{-1, +1\}$ be the class label in a two-class classification problem. $\mathcal{X}$ may be a finite set or an infinite set like $\Re^d$. Because Fisher and TOP kernels are commonly used in combination with linear classifiers such as linear SVMs, [21] proposes as a reasonable performance measure the classification error of a linear classifier $w^T \cdot \phi_{\hat{\theta}}(x) + b$ in the feature space $\Re^d$, where $w \in \Re^d$, $b \in \Re$ and $\hat{\theta}$ is the ML estimate of the parameters. To make a general analysis, we assume that w and b are chosen by an optimal learning algorithm. In this case, the classification error $R(\phi_{\hat{\theta}})$ is

$$R(\phi_{\hat{\theta}}) = \min_{w,b} E_{x,y}\Phi[-y(w^T \cdot \phi_{\hat{\theta}}(x) + b)] \quad (6)$$

where $\Phi[a]$ is the indicator function which is 1 when $a > 0$, and otherwise 0, $E_{x,y}$ denotes the expectation with respect to the true distribution $P(x, y|\theta^*)$.

In [7], it has been demonstrated that the Fisher kernel (FK) classifier can perform at least as well as its generative counterpart if the parameters of a linear classifier are properly determined. Such property, revised more concisely within this framework can be written as [21]:

$$R(\phi_{\hat{\theta}}^{FK}) \leq E_{x,t}\Phi[-y(P(y = +1|x, \hat{\theta}) - \frac{1}{2})] = R(\theta) \quad (7)$$

where $R(\theta)$ represents the generative error.

It is straightforward to prove that a kernel classifier that uses **FE** features is asymptotically at least as good as the MAP labeling based on the generative models for the two classes. Let M be the length of the vector $\phi_{\hat{\theta}_i}^{FE}(x^{(t)})$ (see Eq. 5). Since the free energy can be written as $\mathcal{F}^t = w^T \cdot \phi_\Theta^{FE}(x)$, then:

$$\begin{aligned} R(\phi_\Theta^{FE}) &= \min_{w,b} E_{x,y}\Phi[-y(w^T \cdot \phi_\Theta^{FE}(x) + b)] \\ &\leq E_{x,y}\Phi[-y(w_G^T \cdot \phi_\Theta^{FE}(x) + b_G)] \\ &= R_{\mathcal{Q}}(\theta) \\ &for \quad w_G = [\overbrace{+1, \cdots, +1}^{M\ \ times}, \overbrace{-1, \cdots, -1}^{M\ \ times}], \ \ b_G = 0 \end{aligned} \quad (8)$$

where $R_{\mathcal{Q}}(\theta)$ denotes the classification performance of the full free energy test, where the model with the lower free energy for the data point is chosen. When the family $\mathcal{Q}$ from which the posterior is chosen is expressive enough to capture the true posterior distribution, then free energy reduced to negative log likelihood, and the free energy test reduced to the likelihood ratio and $R_{\mathcal{Q}}(\theta) = R(\theta)$. In other cases, likelihood computation is intractable, and free energy test, and the corresponding $R_{\mathcal{Q}}(\theta)$ are used instead of the likelihood ratio test. At any rate, for the corresponding approximation family $\mathcal{Q}$ the optimal selection of the features of ϕ_{Θ}^{FE} must at least match the performance obtained by the purely generative approach.

2.3. Usage of the free energy features

Obviously, a number of discriminative methods can be utilized to design a classifier based on the features extracted from the free energies under a set of previously learned generative models Θ. As discussed above, if linear discriminant functions are adopted, the sum of the pieces of free energy $\{f_i\}$ will be re-weighted by a set of weights $\{w_i\}$. For example, if we employ logistic regression, we can estimate a set of weights β_i (one for each feature f_i), and classify using the sigmoid function $p(\mathbf{x}) = 1/(1 + exp(-(\beta_1 + \sum_i \beta_{i+1} f_i)))$. It is especially interesting to impose *sparsity* so that only few $\beta_i \neq 0$, i.e., only some free energy pieces will be taken into account for classification. This can be done efficiently by adding L_1 regularization term for the weights of the logistic regressor to the optimization criterion. At this point, we can better analyze the schema presented in Fig. 1, highlighting how our approach differs from other paradigms. For score-spaces [7, 21] classification error is theoretically lower than that of the generative methods: this trivially holds for our framework since generative classification rule is a special case of the proposed approach. Score-space methods and our approach perform better than standard discriminative methods since features extracted from a generative model encode both the data *and* the generative process. Moreover, unlike discriminative classifiers, they can handle missing and variable length data. As opposed to score space methods, our approach encodes ambiguities of data fit at different levels of the model, as dictated by the approximate posterior function rather than the derivatives with respect to the model parameters, making the approach both more tractable (no need for gradient computation as in [7, 21]), and qualitatively different.

3. Experiments

In this section, we focus on challenging vision applications such as object, scene and shape recognition, and also report comparative performances with respect to state-of-the-art generative models.

3.1. Object recognition: Probabilistic index map

The probabilistic index map (PIM) model has been introduced in [8]. Index maps are formally defined as ordered sets of indices $s_i \in 1, \ldots, S$, linked to spatially distinct areas $i \in 1, \ldots, K$ of images, where K is the number of such image areas (e.g., pixels). These indices point to a table of S possible local measurements (e.g., color), referred to as palette, modeled by a gaussian distribution $\mathcal{C}(s^{(t)}) = \{\mu_s^{(t)}, \psi_s^{(t)}\}$. Probabilistic index maps (PIMs) inject uncertainty into image indexing and into the nature of the palette: each image location i is associated with a learned prior distribution over indices $P(s_i = s)$, thus describing the image structure. We refer to an area of a $t-$th image with the same assigned index s as an image part which will be considered probabilistically as a (probabilistic) map of the segment $Q(s_i^{(t)} = s)$ of image locations $\{i\}$. In this way, the uncertainty of pixels of the $t-$th image belonging to the $s-$th image part can be modeled. Further, we infer the possible consistency in the palette across instances of the class through a palette prior $P(\mathcal{C}(s))$. Several maps, extracted from Caltech101 rhinoceros class, and the prior $P(s_i)$ and $P(\mathcal{C}(s))$ are shown in Fig. 2A, in which it is possible to note how the segment indexed by $s = 3$ represents the rhinoceros body for all the images with a palette prior peaked on the gray values. The free energy of the PIM model is defined as

$$\begin{aligned}\mathcal{F} &= \sum_t \sum_i \sum_s Q(s^{(t)}) \log Q(s^{(t)}) - Q(s^{(t)}) \log p(x^{(t)}|s^{(t)}) \\ &- Q(s^{(t)}) \cdot \log P(s) - p(\mathcal{C}(s^{(t)}))\end{aligned}$$

We carry out the sum over the pixels, keeping separated the other contributions. The resulting feature extractor ζ is therefore defined as follows:

$$\zeta(x^{(t)}, PIM) = [\sum_i Q(s_i^{(t)} = k) \log Q(s_i^{(t)} = k), \ldots \quad (9)$$

$$, -\sum Q(s_i^{(t)} = k) \log p(x_i|s_i^{(t)} = k), \quad (10)$$

$$\ldots, -\sum_i Q(s_i^{(t)} = k) \log P(s_i = k), \quad (11)$$

$$\ldots, -\log p(\mathcal{C}(s^{(t)} = k))]_{k=1}^{S} \quad (12)$$

In equations 9-12 four pieces of the free energy are evident: the first term, $Q(s^{(t)}) \log Q(s^{(t)})$, represents the entropy, the second one, $Q(s^{(t)}) \log P(x|s^{(t)})$, reflects how tight the color distribution within each image part is, the third term represents the agreement of the probabilistic image segmentation $Q(s^{(t)})$ with the prior $P(s)$, and the fourth term indicates the agreement of the palette with the palette prior (see Fig. 2B for the explanation of the third term, in which we have the agreement of the map $Q(s^{(t)} = 2)$ related to an image observation with respect to the prior $P(s = 2)$ of the squirrel class).

In the first example, we used the Caltech101 dataset. After learning a PIM model (3 image parts, S=3) for each class, and calculating the free energies of each sample under every

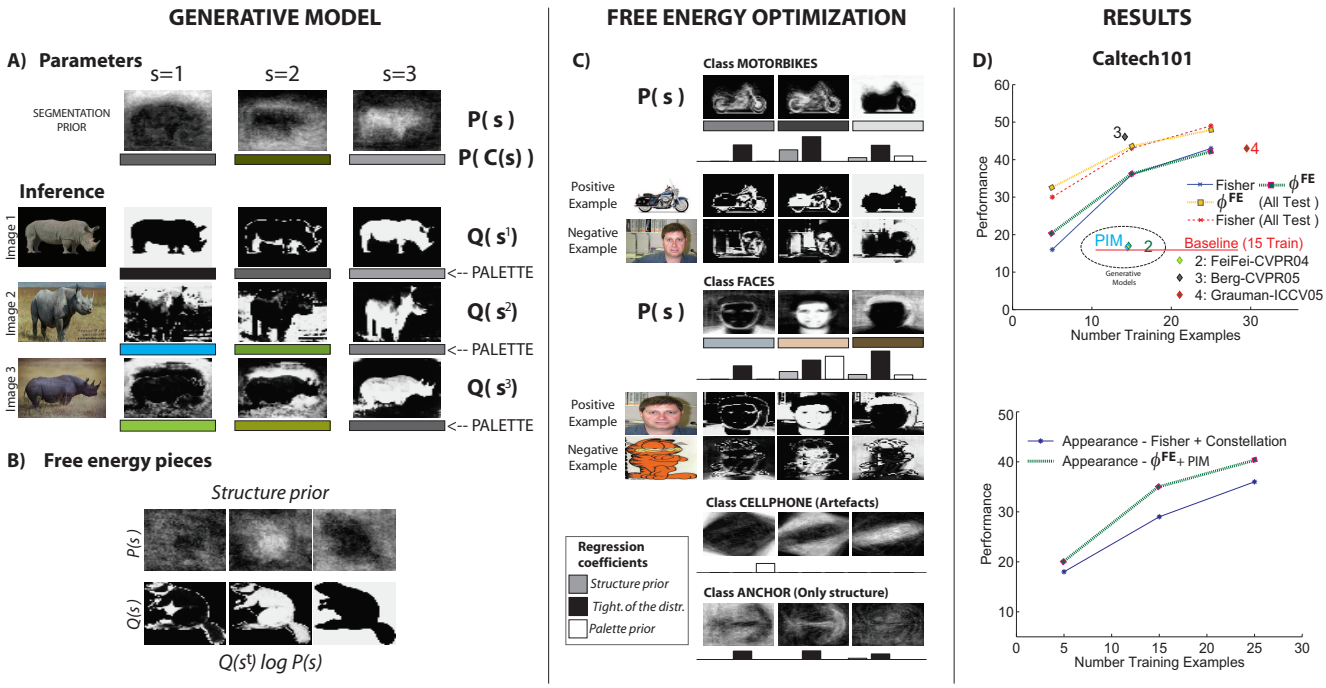

Figure 2. A) An illustration of the probabilistic index map generative model. Color bars below P(s) indicate the estimated means of the priors over the colors for different segment, while the color bars under Q(s) indicate the inferred color for the particular sample [8]. B) The illustration of the free energy terms as they relate to the image structure: brighter pixels indicate higher probabilities (the learned prior or estimated posterior, latter shown next to each image). C) The bars represent the discriminative contributions of the free energy parts for several data classes, obtained by normalizing the regression coefficients β_i (high bar = high contribution). The example indicates, for example, that for faces, the color prior is informative only for segment 2 encompassing the face, and leaving out the background and shoulders, whereas the uniformity of the color within a segment is important across all segments, followed in importance by the shape of the segment 2 (in terms of its agreement with the prior). D) Object recognition results (standard Caltech101 benchmarks): in the top chart, the dotted lines labeled a *all test* are obtained using all available testing samples for each class (see [6]). The lower chart, shows how starting from the same descriptor (appearance), our method outperforms [6]

model we describe each image using the feature vector:

$$\phi^{FE}(x^{(t)}) = [\zeta(x^{(t)}, PIM_1), \ldots, \zeta(x^{(t)}, PIM_{101})]$$

In Fig. 2C, we show the results of one-vs-all sparse logistic regression: the bars under the parameters $P(s)$ represent the estimated normalized regression coefficients associated with the three image parts $s = 1, 2, 3$, with respect to three free energy terms[1] under the model θ. In the picture, the higher the bar, the higher the importance of the free energy portion for the discrimination. For example, images of the class motorbikes are classified only considering their structural information as only positive examples would have high agreement with the segmentation prior inferred for that class (see segments 1 and 2 for the positive and negative example). Some classes can use a strong palette prior $p(\mathcal{C}(s))$ in some image parts as the source of discrimination power; this is the case of the faces class which can be correctly classified only looking at the agreement of the color entry with the color prior in the image part correspondent to the face (in the figure, s=2). The class cellphone can be correctly classified looking at the color in the background area (due to uniform clutter of the scene), while the most discriminant features for the anchor derive from the segmentation.

We performed classification over the entire dataset following the training-testing split of [6][2], where the authors used the Fisher kernel over the constellation model with two descriptors: shape and appearance. The purely generative application of the PIM model yields the classification accuracy of 17%, slightly outperforming the baseline. However, SVM classification based on the PIM's free energy features, yields a classification rate of 48% (see Fig. 2D), making the model comparable with similar hybrid approaches [6] not based on the PIM model. It is worth noting that we are proposing a novel hybrid feature that capture the generative process of the data which can be used with more complex discriminative classifiers in conjunction with other standard features, as intuitively **FE** captures rather different

[1]To ease the visualization and the understanding, we did not depict the entropy contributions which decode the ambiguity within the model

[2]Consisting in using part of the training set to learn the generative models and part to learn a support vector machine.

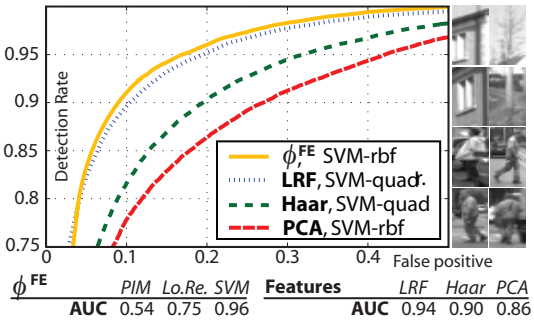

Figure 3. ROC curves for the pedestrian detection example. Bottom: results using ϕ^{FE} based on the PIM model, -PIM- refers to purely generative classification, -Lo.Re.- stands for logistic regression of **FE** features, while -SVM- for support vector machines with the RBF kernel.

aspects of the images.

Adopting a PIM model and the same evaluation procedure of [16], we also compared our approach on the pedestrian detection task to show a comparison between image-based features and hybrid features. We trained a probabilistic index map model using only half of the training images of pedestrians (see Fig. 3). Then, we used ϕ^{FE} to extract features for the remaining training images and for all the testing set. After ranking the features, we calculated the area under the ROC curve (AUC) for the generative classification, getting AUC$= 0.54$, and this quite poor result is due to the fact that, in general, positive examples have a low value of $Q(s^{(t)}) \log p(x^{(t)}|s^{(t)})$ (which portrays tightness with respect to a color distribution), hence it is hard to find a peaked color distribution for the pedestrians. On the other hand, we discover high values for $Q(s^{(t)}) \log P(s^{(t)})$, mirroring the fact that there is a structure prior which all the images are in agreement with (the structure portraying the human body). Summing the free energy pieces as the generative classification does, looses this difference. The more natural way to capture this difference is to employ kernel classifiers, which exploit the local differences exhibited by the free energy feature vectors. The SVM classification based on free energy features improved the performances to AUC$= 0.96$ (see Fig. 3) outperforming, while only using gray-level images and generative modeling, the previous classification rates on this task, including Haar wavelets, local receptive fields (LRF) and PCA coefficients which have been considered the best features for this task [16]. (LRF features could be considered as another example of a score space method, since they are extracted as by-product of a neural network. These features are only slightly inferior to free energies features.)

3.2. Shape recognition: hidden Markov models

Since we are using generative information in a discriminative framework, it is natural to compare our approach to well-known score spaces approaches like Fisher and Top kernels [7, 21]. Usually, these methods are used in conjunction with the hidden Markov model [19] as the generative starting point. First, an HMM model is fit to the data, then the tangent vector of the marginal log likelihood $\nabla_\theta \log p(x|\hat{\theta})$ is used as the feature vector for the Fisher Kernel [7], while the Top kernel [21] is derived from $\nabla_\theta(\log p(y=+1|x,\hat{\theta}) - \log p(y=-1|x,\hat{\theta}))$ where y is the label variable. In general, the main intuition of the generative score space approach is to distill the contribution of each parameter in the generation of a particular sample. This yields to a mapping of variable length sequences into a fixed-length feature vector (space). For HMMs, **FE** feature vectors will have a length dependent on the length of the particular sample t, making the usage of discriminative methods unfeasible. Details on the extraction of free energy for HMMs can be found in [13]. To solve this problem, we simply perform the sums over the sample length, normalizing the vectors over the length of the sequences to obtain a fixed number of features per data point.

We consider the Chicken Pieces Database[15] as the experimental data set, consisting of 446 binary images of five chicken piece classes: “wing”, “back”, “drumstick”, “thigh and back” and “breast”. Despite the limited number of classes, this is a challenging dataset where the best result does not go over the accuracy of 81%, to the best of our knowledge. In order to compute curvature sequences, we first extract the contours by using the Canny edge detector. The boundary is then approximated by segments of approximately fixed length. Finally, the curvature value at point x is computed as the angle between the two consecutive segments intersecting at x, resulting in real-valued sequences of different lengths.

The results are reported in Table 1, where we varied the fractions of the set of so obtained training curvature sequences used to train either the hidden Markov model or the SVM classifier (employing the same training images used by the database’s authors). Using the same classifier across different sequence features, we find that **FE** features outperform [7, 21]. We reached the same conclusions of [17], that is, generative models are more effective with less data,

% HMM	10%	30%	50%	70%
FE-lin	65,31%	79,99%	68,20 %	48,13%
FE-rbf	76,82%	**82,95%**	72,69 %	53,01%

Comparison with the state-of-the-art

[7] 79,12%	**FS**-rbf [7] 81,67%	**TK** [21] 80,80%

Table 1. Shape recognition results. **FE**-rbf outperforms [21] with a p-value = 0.00218 and **FS**-rbf with a p-value = 0.00967. The difference between **FE**-lin and [7] is statistically insignificant (p$>$0.05).

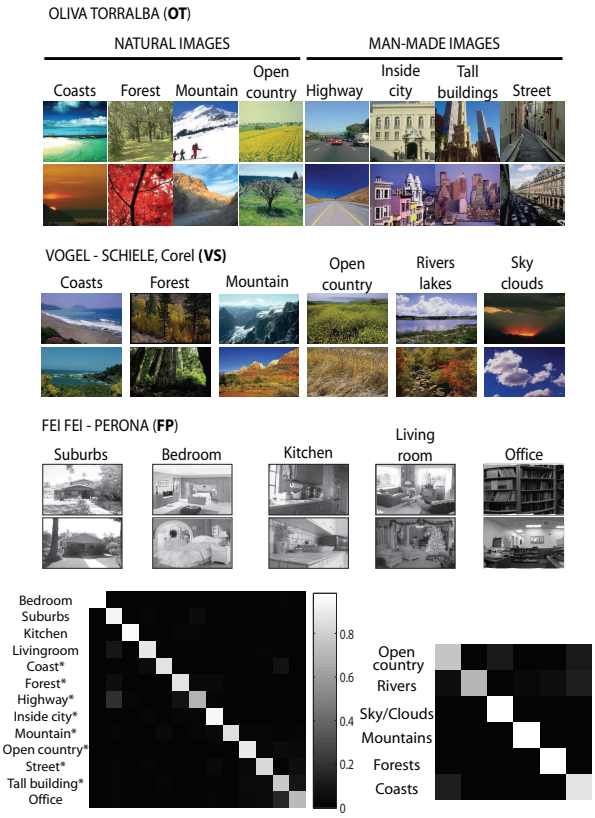

Figure 4. Some examples from each dataset used in the scene classification experiment. **FP** is composed of eight categories from **OT** and the additional five shown here. The last row shows the confusion matrices for **VS** and **FP** (**OT** is a subset of **FP**).

providing a generalizable class description which can be effectively isolated by the SVM, using the remaining training data.

3.3. Scene classification: pLSA/LDA

Topics models such as pLSA [5] and LDA [2] have been successfully employed in computer vision tasks such as scene classification [12, 3]. In this formulation, each image is represented as a collection of detected patches or visual words, taking word labels from a previously trained codebook of W words. The word occurrences are counted, resulting in a term-frequency vector for each image document, which ignores the words' spatial layout (as opposed to PIM model above). Given the term-frequency vector, the pLSA uses a finite number of hidden topics Z to model the co-occurrence of visual words inside and across images. Each image is explained as a mixture of hidden topics and these hidden topics refer to objects or object parts.

We used these models as the generative starting point and evaluated our classification algorithm on three different popular datasets: (1) Oliva and Torralba [18], (2) Vogel and Schiele [22], and (3) Fei Fei and Perona [12]. We will refer to these datasets as **OT**, **VS** and **FP**, respectively. We extracted SIFT features from 16x16 pixel patches computed over a grid with spacing of 8 pixels; we used 40 topics ($Z = 40$) and 175 codewords ($W = 175$).

pLSA: Comparison with the state-of-the-art

Dataset	*Best*	*Auth.*	[3]	ϕ^{FE}
VS (6)	85,7% [3]	75,1% [22]	85,7%	**90,30%**
OT^N (4)	90,2% [3]	89,0% [18]	90,2%	**95,21%**
OT^A (4)	92,5% [3]	89,0% [18]	92,5%	**94,38%**
OT (8)	86,5% [3]	n.a.	86,5%	**92,79%**
FP (13)	81,1% [11]	65,2% [12]	73,4%	**84,31%**

Table 2. Scene classification results. The numbers of classes in different datasets are shown in brackets. We compared ϕ^{FE} with the datasets' authors (*Auth.*), the best method (*Best*), and with [3]. For the **OT** dataset, **OT**^{N} refers to the restriction to four natural categories, and **OT**^{A} are the four man-made categories. SVM classifiers was found slightly superior to logistic regressors. The classification rate for the free energy decomposition-selection approach on the extended **FP** dataset used in [11] is **82,91%**

For each test we trained C pLSA models[3], one for each class, using half of the training set designated by the database authors. Afterward, we extracted the free energy features ϕ^{FE} from the rest of the dataset. We used the second half of the training set to learn the discriminative models. Finally, we calculated the classification accuracy over the test set, repeating the process 10 times and averaging the results. The free energy features are straightforward to extract from the description of pLSA/LDA in [5, 2].

Results for each dataset are summarized in Table 2 where we compare the accuracy of our approach with the accuracy achieved by the datasets' authors, the current state of the art, and the results of [3]. The methods presented in [22, 18, 11] are purely discriminative: the features (SIFT or image patches) are directly used for SVM classification with well-suited kernels. In particular, for [22], the training requires manual annotation of 9 semantic concepts for 60000 patches making the pre-processing step rather expensive. The unsupervised approach of [3] trains a single pLSA model for all the classes, and then uses the marginal distribution $P(topic|document)$ as the input for a discriminative classifier, thus employing a hybrid technique. Finally, we also considered the semi-supervised generative approach of [12], which makes use of LDA likelihood to classify. In Fig. 4, we report the related confusion matrices.

Our method strongly outperforms the current best results on each dataset, which is also explained by its relationship with other methods. Our free energy decomposition/selection approach has the same discriminative power as [22, 18, 11] since it employs the same discriminative me-

[3] LDA performances were found to be slightly inferior.

thod. It has provably better performance in absence of overtraining than [12] since it uses the same free energy (see Eq. 8). Finally the information in $P(topic|document)$ used in [3] as features is contained in the free energy vector extracted by ϕ^{FE}.

4. Conclusions

In this paper, we present a novel generative score space approach exploiting variational free energy terms as features for discriminative classification task. Free energy terms are powerful descriptors as they encode ambiguity within the generative model at different levels mirroring the discrepancies of the test samples with respect to the generative data process. These discrepancies are distributed across several terms, each one focusing on a particular aspect of the generative joint distribution. Such sources of uncertainty are not directly separated in standard score operators that rely on derivatives of model parameter estimates. The family of possible mapping score operators resulting from free energy factorization has also been shown outperform generative classification both experimentally and theoretically. Finally, the experiments also show that the proposed framework is successful in a wide range of different applications, dealing with scene, object, and shape recognition tasks on public benchmark data sets. In these experiments, our method outperforms the state of the art, including approaches based on previously proposed score spaces for hybrid models. The proposed approach is highly flexible in the way it can deal with arbitrary generative models and the posterior factorization, and therefore, either fixed or variable-length data, missing data, and the data of different nature, while maintaining a relatively low computational cost.

Acknowledgements

We acknowledge financial support from the FET programme within the EU FP7, under the SIMBAD project (contract 213250).

References

[1] M. Bicego, V. Murino, and M. A. Figueiredo. Similarity-based classification of sequences using hidden markov models. *Pattern Recognition*, 37(12):2281 – 2291, 2004. 1

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation, 2003. 7

[3] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via plsa. 2006. 1, 2, 7, 8

[4] B. Frey and N. Jojic. A comparison of algorithms for inference and learning in probabilistic graphical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9):1392–1413, 2005. 2

[5] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, New York, NY, USA, 1999. ACM Press. 7

[6] A. D. Holub, M. Welling, and P. Perona. Combining generative models and fisher kernels for object recognition. *ICCV*, 1:136–143, 2005. 5

[7] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. *NIPS*, 1998. 1, 3, 4, 6

[8] N. Jojic and Y. Caspi. Capturing image structure with probabilistic index maps. In *CVPR (1)*, pages 212–219, 2004. 4, 5

[9] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999. 2

[10] J. A. Lasserre, C. M. Bishop, and T. P. Minka. Principled hybrids of generative and discriminative models. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 87–94, Washington, DC, USA, 2006. IEEE Computer Society. 2

[11] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages 2169–2178, 2006. 7

[12] F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, pages 524–531, 2005. 7, 8

[13] D. MacKay. Ensemble learning for Hidden Markov Models, 1997. Tech.Rep. Department of Physics, University of Cambridge. 6

[14] T. Minka. Discriminative models, not discriminative training. Technical Report TR-2005-144, Microsoft Research Cambridge, October 2005. 2

[15] R. A. Mollineda, E. Vidal, and F. Casacuberta. Cyclic sequence alignments: Approximate versus optimal techniques. *International Journal of Pattern Recognition and Artificial Intelligence*, 16:291–299, 2002. 6

[16] S. Munder and D. M. Gavrila. An experimental study on pedestrian classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1863–1868, 2006. 6

[17] A. Ng and M. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, 2002. 1, 6

[18] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175, 2001. 7

[19] L. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. of IEEE*, 77(2):257–286, 1989. 6

[20] N. Smith and M. Gales. Speech recognition using SVMs. In *NIPS*, pages 1197–1204. MIT Press, 2002. 1

[21] K. Tsuda, M. Kawanabe, G. Rätsch, S. Sonnenburg, and K.-R. Müller. A new discriminative kernel from probabilistic models. *Neural Comput.*, 14(10):2397–2414, 2002. 1, 3, 4, 6

[22] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision*. (in press). 7