Appears in European Conference on Computer Vision, Prague, 2004.

The Kullback-Leibler Kernel as a Framework for Discriminant and Localized Representations for Visual Recognition

Nuno Vasconcelos[1], Purdy Ho[2], and Pedro Moreno[2]

[1] Department of Electrical and Computer Engineering,
University of California San Diego,
9500 Gilman Drive, MC 0407, San Diego, CA 92093, USA
nuno@ece.ucsd.edu
[2] HP Cambridge Research Laboratory,
One Cambridge Center, Cambridge MA 02142
{purdy.ho,pedro.moreno}@hp.com

Abstract. The recognition accuracy of current discriminant architectures for visual recognition is hampered by the dependence on holistic image representations, where images are represented as vectors in a high-dimensional space. Such representations lead to complex classification problems due to the need to 1) restrict image resolution and 2) model complex manifolds due to variations in pose, lighting, and other imaging variables. Localized representations, where images are represented as bags of low-dimensional vectors, are significantly less affected by these problems but have traditionally been difficult to combine with discriminant classifiers such as the *support vector machine* (SVM). This limitation has recently been lifted by the introduction of probabilistic SVM kernels, such as the *Kullback-Leibler* (KL) kernel. In this work we investigate the advantages of using this kernel as a means to combine discriminant recognition with localized representations. We derive a taxonomy of kernels based on the combination of the KL-kernel with various probabilistic representation previously proposed in the recognition literature. Experimental evaluation shows that these kernels can significantly outperform traditional SVM solutions for recognition.

1 Introduction

The formulation of visual recognition as a problem of statistical classification has led to various solutions of unprecedented success in areas such as face detection, face, texture, object, and shape recognition, or image retrieval. There are, however, various fundamental questions in the design of classifiers for recognition that remain largely unanswered. One of the most significant is that of identifying the most suitable classification architecture. Broadly speaking, there are two major architecture classes: that of discriminant classifiers and that of classifiers based on generative models. On one hand, modern learning theory favors the use of discriminant solutions, namely the large-margin classifiers inspired by VC

theory [5], for which there is an appealing guiding principle ("do not model more than what is needed") and a more rigorous understanding of properties such as the generalization error than what is available for generative solutions. On the other hand, generative models have various properties of great appeal for the implementation of recognition systems. In particular they 1) have much better scalability in the number of classes and amount of data per class, 2) enable the encoding of knowledge about the classification problem in the choice of statistical models and, therefore, are significantly more flexible, and 3) allow modular solutions, where Bayesian inference is used to integrate the contributions of various modules into the optimal decision for a large classification problem.

For visual recognition, one of the fundamental differences between the two approaches is the set of constraints that are imposed on image representation. While generative models favor a representation of the image as a large collection of relatively low-dimensional features, discriminant solutions work best when images are represented as points in some high-dimensional space. Hence, while a *localized* image representation is usually adopted in the generative setting (e.g. by representing each image as a bag of 8×8 image blocks), on the discriminant setting the representation frequently consists of a *holistic* low-resolution replica, e.g. 20×20 pixels, of the original image. While this holistic representation has the clear advantage of capturing global attributes of the objects of interest, e.g. that eyes, nose, and mouth always appear in a given configuration in face images, it has various disadvantages over the localized representation. These include 1) a much higher susceptibility to invariance problems due to either image transformations, non-rigid objects, or occlusion and 2) a significant loss of information due to the need to downsample images severely in order to keep the dimensionality of the space tractable. Due to these problems, localized representations are frequently advocated or adopted for recognition tasks, leading to generative classifiers [4, 6]. While there is a sense that such classifiers imply some loss in recognition accuracy, the difficulty of combining discriminant techniques with the localized representation makes the discriminant alternative impractical.

In this work we consider one of the most popular discriminant architectures, the *support vector machine* (SVM). SVMs are large-margin classifiers obtained by solving a convex programming problem that depends on the training data through a kernel matrix that captures the distances between all pairs of examples. For a training set of size N, this results in a $O(N^2)$ complexity for any SVM learning algorithm, rendering localized representations (where each image can lead to a bag of thousands of examples) intractable. It has, however, been recently observed [7] that a natural extension of this formulation is to consider kernel matrices that capture distances between *the generative models associated with each bag of examples instead of the examples themselves*. This observation has motivated the introduction of various kernels based on probabilistic models, e.g. the *Fisher* [7], *Kullback-Leibler* [8], *TOP* [11], and *Battacharya* [12] kernels. In this paper, we investigate the benefits of the Kullback-Leibler (KL) kernel for visual recognition. In particular, we show that it subsumes many kernels based on localized representations that have been argued to be interesting, or shown

to work well, for recognition. We provide closed-form expressions for the kernel as a function of the parameters of the probabilistic models whenever they exist, and discuss alternatives for the construction of the kernel matrix when this is not the case. Finally, a detailed experimental evaluation is presented, illustrating the result of the various trade-offs associated with the various combinations of localized vs holistic representations and generative vs discriminant classifiers.

2 SVMs and kernel functions

In this section, we present a brief review of the SVM architecture and the constraints that it poses on image representation for recognition.

2.1 The SVM architecture

Consider a binary classification problem with a training set consisting of (input,output) pairs $(\mathbf{x}_i, y_i) \in \mathcal{X} \times Y$, $Y = \{-1, 1\}$. Assuming that the data is separable[1], the optimal (in the maximum margin sense) *linear* separating hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$ is the solution to the constrained optimization problem [5]

$$\max \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \quad \text{subject to} \quad \sum_i \alpha_i y_i = 0,\ \alpha_i \geq 0 \qquad (1)$$

where $\{\alpha_i\}$ is a set of Lagrange multipliers, and

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \qquad b = 1/|I| \sum_{i \in I} [y_i - \sum_j y_j \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j)] \qquad (2)$$

with $I = \{i | \alpha_i > 0\}$. One of the appealing properties of this formulation is that it depends only on the dot products of the training vectors. This allows the automatic extension of the optimal solution to the, seemingly much more complicated, problem of designing large-margin classifiers with non-planar boundaries.

This extension consists of introducing a non-linear mapping $\mathbf{\Phi} : \mathcal{X} \rightarrow \mathcal{Z}$ from the original input space $\mathcal{X}$ to a new feature space $\mathcal{Z}$. Typically $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Z} = \mathbb{R}^p$ where p is significantly larger than d and linear boundaries in $\mathcal{Z}$ are equivalent to non-linear boundaries in $\mathcal{X}$. It follows from the discussion above that the optimal solution in $\mathcal{Z}$ is given by (1),(2) with the inner products $(\mathbf{x}_i \cdot \mathbf{x}_j)$ replaced by $\mathbf{\Phi}(\mathbf{x}_i) \cdot \mathbf{\Phi}(\mathbf{x}_j)$. In $\mathcal{X}$, this is equivalent to simply introducing a *kernel function* $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$, under the constraint that this function must be an inner product in some space $\mathcal{Z}$, i.e.

$$\exists (\mathcal{Z}, \mathbf{\Phi}), \mathbf{\Phi} : \mathcal{X} \rightarrow \mathcal{Z} \ \text{ such that } \ \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{\Phi}(\mathbf{x}_i) \cdot \mathbf{\Phi}(\mathbf{x}_j). \qquad (3)$$

Mercer's theorem assures that this condition holds whenever $\mathcal{K}(\mathbf{x}, \mathbf{y})$ is a positive definite form [5]. Notice that an inner product is nothing but a measure of vector

[1] All the results in this paper apply equally well to the extension to the non-separable case [5]. We omit it here for simplicity.

similarity and, since $||\mathbf{x}-\mathbf{y}||^2 = (\mathbf{x}\cdot\mathbf{x}) - 2(\mathbf{x}\cdot\mathbf{y}) + (\mathbf{y}\cdot\mathbf{y})$, the standard dot product implies an Euclidean metric on $\mathcal{X}$. Under this interpretation, the role of the kernel is to enable extensions to non-Euclidean measures of similarity. Hence, the *kernel matrix* $\mathcal{K}(\mathbf{x}_i \cdot \mathbf{x}_j)$ can be seen as capturing the similarity between points $\mathbf{x}_i, \mathbf{x}_j$ under the similarity measure that is most suited for the problem at hand.

2.2 Constraints on image representation

Consider a binary recognition problem[2] with a training set consisting of I example images per class. The formulation of this problem as one of statistical classification can be based on two alternative image representations. The first, the holistic representation, makes $\mathcal{X}$ the space of all images and represents each image as a point in this space. Since images are high-dimensional, downsampling is always required to guarantee a space of manageable dimensionality. Typical image sizes after downsampling are on the order of 20×20 pixels, i.e. $dim(\mathcal{X}) \approx 400$. On the other hand, localized representations are based on a collection of local measurements (or *features*) extracted from the image. For example, the image can be broken down into a collection (*bag*) of small neighborhoods, e.g. 8×8 pixels, and $\mathcal{X}$ made the space of such neighborhoods, $dim(\mathcal{X}) = 64$. Each image no longer corresponds to a single point in $\mathcal{X}$, but to a collection of points.

While the dependence on the training examples only through their inner products is, theoretically, a very appealing feature of the SVM architecture, it also introduces a significant computational burden that places serious constraints on the image representations compatible with it. In particular, because SVM learning is an optimization problem with coefficients given by the entries of the kernel matrix, its complexity is quadratic in the size of the training set. Hence, if the localized representation originates K neighborhoods per image, this implies a $O(K^2I^2)$ complexity and a K^2-fold increase over the complexity associated with the holistic representation. As we will see in section 4, it is not difficult for the localized representation to originate on the order of $5,000$ neighborhoods per image, corresponding to a 25×10^6-fold increase in computation that is always undesirable and usually intractable. Furthermore, under the SVM formulation, there is no way to capture the natural grouping of image neighborhoods into images, i.e. the fact that the goal is to classify bags of examples instead of the examples independently. For these reasons, the localized representation is not suitable for traditional SVM-based recognition.

While the holistic representation has been successful for recognition [13, 3, 14] it should not be taken for granted that it is inherently better than its localized counterpart. On the contrary, it suffers from the following problems.

- **Resolution**: when images are severely downsampled a significant amount of information is lost. While this information may not be important for the classification of images far away from the classification boundary, it can be

[2] The discussion in this section generalizes to any number C of classes, since a C-way classifier can be implemented as a combination of C binary (one-vs-all) classifiers.

quite relevant to distinguish the ones that are close to it. Since the latter determine the classification error, low resolution can have an impact on recognition error. The best example of this phenomena are current state-of-the-art face detectors [13, 1]. While visual inspection of the errors committed by the classifier, at the low-resolution on which its decisions are based, reveals that it is quite hard to distinguish between faces and non-faces, a significant percentage of those errors becomes clear at full image resolution.

- **Invariance**: when images are represented as points in $\mathcal{X}$, a relatively simple image transformation can send the point associated with an image to another point that is significantly far away in the Euclidean sense. In fact, when subject to transformations, images span manifolds in $\mathcal{X}$ which can be quite convoluted and the correct distance for classification is the distance to these manifold. While the kernel function can, in principle, encode this, the traditional SVM formulation provides no hints on how to learn the kernel from examples. This can lead to significant invariance problems.
- **Occlusion**: since, for the holistic representation, occlusion originates a (possibly dramatic) change in some of the components of vector associated with the image to classify, an occluded pattern can, once again, be quite distant from the unoccluded counterpart. Unlike invariance, it is not even clear that occlusion leads to an image manifold (there could be creases, folds, or singularities in the space of occluded images) and it is therefore even less clear what metric, or kernel, would be appropriate to deal with occlusion.

Note that the localized representation does not place constraints on resolution (larger images simply generate more neighborhoods), and is significantly more invariant and robust to occlusion.

3 Probabilistic Kernels based on the KL-divergence

Since there are advantages to the localized representation, enabling the SVM architecture to support it is a relevant problem for visual recognition. This is the motivation behind the KL-kernel that we briefly review in this section. We then show that it 1) enables truly discriminant localized representations, and 2) can be naturally adapted to each classification problem. This allows the derivation of various kernels tailored for representations previously proposed for recognition.

3.1 The KL-kernel

The combination of SVMs and localized visual representations is related to that of SVMs and data sequences, a topic that has been addressed by various authors [7, 11, 8, 12, 15]. Since the role of the kernel is to capture similarities between examples, and sequences are naturally described by their probability densities, one idea that has recently received some attention is to replace the sequences by their probabilistic descriptions [7, 11, 8, 12]. This has various advantages, including the ability to 1) deal with sequences of variable lengths, 2) rely on a compact

sequence representation, and 3) exploit prior knowledge about the classification problem (through the selection of probability models) [7]. The KL-kernel is the extension of the standard Gaussian kernel to this family of probabilistic kernels: while the Gaussian is proportional to the negative exponent of the weighted Euclidean distance between two vectors, the KL-kernel is the negative exponent of the symmetric KL divergence [16]. This divergence is a measure of distance between two densities and has various interesting connections to the geometry of the manifold of probability distributions [17]. In particular, given densities $p(\mathbf{x})$ and $q(\mathbf{x})$, the KL-kernel is

$$KLK = \exp^{-a\mathcal{J}[p(\mathbf{x}),q(\mathbf{x})]+b}, \tag{4}$$

where $\mathcal{J}(p(\mathbf{x}), q(\mathbf{x})) = KL(p(\mathbf{x}), q(\mathbf{x})) + KL(q(\mathbf{x}), p(\mathbf{x}))$ is the symmetric KL divergence between $p(\mathbf{x})$ and $q(\mathbf{x})$,

$$KL(p(\mathbf{x}), q(\mathbf{x})) = \int_{-\infty}^{\infty} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} \tag{5}$$

the KL divergence between the two densities, and a and b constants [8].

3.2 A kernel taxonomy

One of the main attractives of probabilistic kernels is a significant enhancement of the flexibility of the SVM architecture. For example, the KL-kernel can be tailored to a classification problem by either 1) matching it to the statistics of the datasets under consideration, 2) taking advantage of approximations to the KL-divergence that have been shown to work well in certain domains, or even 3) combining feature and kernel design. In this section we give some examples of such tuning, but various other kernels could be derived in a similar fashion.

Parametric densities There are many problems where the class-conditional densities are known, or can be well-approximated, by parametric densities. In these cases (5) can usually be simplified. One common setting is for the densities to be members of a parametric family, such as the popular *exponential family*

$$p(\mathbf{x}|\theta) = \alpha(\mathbf{x}) \exp\left[a(\theta) + \mathbf{b}(\theta)\mathbf{c}(\mathbf{x})\right], \tag{6}$$

which includes densities such as Gaussian, Poisson, Binomial, Beta, among various others [18]. The KL-divergence between two such densities is

$$KL(p(\mathbf{x}|\theta_i), p(\mathbf{x}|\theta_j)) = a(\theta_i) - a(\theta_j) + \left[\mathbf{b}(\theta_i) - \mathbf{b}(\theta_j)\right]^T E_{\theta_i}[\mathbf{c}(\mathbf{x})] \tag{7}$$

where E_{θ_i} is the expectation with respect to $p(\mathbf{x}|\theta_i)$. One case of significant interest is that of the Gaussian density,

$$p(\mathbf{x}|\{\mu, \mathbf{\Sigma}\}) = \mathcal{G}(\mathbf{x}, \mu, \mathbf{\Sigma}) = \frac{1}{2\pi^{d/2}|\mathbf{\Sigma}|} exp\{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\} \tag{8}$$

for which (7) becomes

$$KL(\mathcal{G}(\mathbf{x},\mu_i,\mathbf{\Sigma}_i),\mathcal{G}(\mathbf{x},\mu_j,\mathbf{\Sigma}_j)) = \frac{1}{2}\log\frac{|\mathbf{\Sigma_j}|}{|\mathbf{\Sigma}_i|} - \frac{d}{2} + \frac{1}{2}tr(\mathbf{\Sigma}_j^{-1}\mathbf{\Sigma}_i) + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T\mathbf{\Sigma}_j^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \quad (9)$$

where d is the dimensionality of the $\mathbf{x}$. Since image data is not always well-approximated by densities in the exponential family, other probabilistic models are also used in the recognition literature. One popular model is the histogram, $\pi = \{\pi_1,\ldots,\pi_b\}$, where π_i are estimates for the distribution of the feature probability mass over a partition of the feature space $\mathcal{X}$ defined by a collection of non-overlapping cells, or bins, $\mathcal{C} = \{\mathcal{C}_1,\ldots,\mathcal{C}_b\}$. The KL-divergence between two histograms, $\pi^i = \{\pi_1^i,\ldots,\pi_b^i\}$ and $\pi^j = \{\pi_1^j,\ldots,\pi_b^j\}$, defined on $\mathcal{C}$ is

$$KL(\pi^i,\pi^j) = \sum_{k=1}^{b}\pi_k^i\log\frac{\pi_k^i}{\pi_k^j} \quad (10)$$

and extensions to the case where the two histograms are defined on different partitions, $\mathcal{C}_i$ and $\mathcal{C}_j$, are also available [9]. There are, nevertheless, models for which a closed-form solution to the KL-divergence does not exist. In these cases it is necessary to resort to approximations or sampling methods.

Approximations and sampling One popular approximation to the KL-divergence consists of linearizing the log around $x = 1$, i.e. $\log(x) \approx x - 1$. It is straightforward to show [10] that, under this approximation, the KL-divergence becomes the χ^2 statistic, a function that has been frequently proposed as a measure of histogram similarity [19]. For other models, the χ^2 approximation can still be quite difficult to compute in closed form. One such case is the popular Gaussian mixture and various approximations to the KL-divergence between Gaussian mixtures have been recently proposed in the literature, including 1) the *log-sum bound* [20], 2) the *asymptotic likelihood approximation* [9], and 3) approximations based on the *unscented transformation* [21]. Our experience is that, while these approximations tend to work rather well for ranking images by similarity, they do not always provide an approximation that is sufficiently tight for the purpose of evaluating the KL-kernel. An alternative that, at the cost of increase computation, eliminates this problem is a Monte-Carlo approximation

$$KL[p(\mathbf{x}|\theta_i),p(\mathbf{x}|\theta_j)] \approx \frac{1}{s}\sum_{m=1}^{s}\log\frac{p(\mathbf{x}_m|\theta_i)}{p(\mathbf{x}_m|\theta_j)} \quad (11)$$

where $\mathbf{x}_1,\ldots,\mathbf{x}_s$ is a sample drawn according to $p(\mathbf{x}|\theta_i)$.

4 Experiments and Results

We conducted a detailed experimental evaluation of the performance of the KL-kernel for recognition. The Columbia object database, COIL-100 [2], was the

source of data for these experiments. It consists of 100 classes, each containing 72 views of an object, obtained by rotating the object at 5° clockwise over 360°. All images have resolution of 128×128 pixels and 24-bit RGB color.

4.1 Holistic vs localized representation

To evaluate the impact of image resolution on the performance of the various classifiers, we created replicas of COIL-100 at three resolutions: 32×32, 64×64, and 128×128 pixels by downsampling and converting all images to grayscale. To test invariance, we created from each database 4 different combinations of train/test sets, following [3]: for each image class, I images were set aside as a training set, by sampling the view angle uniformly, the remaining ones being used for testing. As in [3], we considered $I \in \{4, 8, 18, 36\}$. We refer to the the dataset with $I = n$ as $\mathcal{D}_n$. In all experiments, the holistic representation was obtained by scan-converting each image into a vector. For the localized representation the image was transformed into a bag of 8×8 neighborhoods (obtained by shifting a 8×8 window by increments of two pixels horizontally and vertically). The discrete cosine transform (DCT) of each window was then computed and scanned into a vector of 64 features ordered by frequency of the associated DCT basis function. Only the 32 lowest frequency DCT coefficients were kept. This is a standard procedure that enables speeding-up the estimation of the density associated with each image without compromising classification performance.

Results The performance of the holistic representation was evaluated with traditional SVMs based on three different kernels: linear (L-SVM), polynomial of order 2 (P2-SVM), and Gaussian (G-SVM) [5]. The localized representation was evaluated with both a standard maximum-likelihood Gaussian mixture model (GMM) classifier and the KL-kernel using GMMs as probability models (KL-SVM). We used mixtures of 32 Gaussians in the former case and of 16 in the latter. Classification rates for all resolutions and datasets $\mathcal{D}_n$ are shown in Table 1. The best result for each combination of resolution/number of training images is shown in bold. These results support various interesting conclusions.

First, among the holistic kernels, G-SVM was consistently the best. Its performance is excellent when the number of training vectors is large, $I = 36$, achieving the best results of all classifiers tested. However, as the number of training examples decreases, the recognition rate drops significantly. In fact, for values of I other than 36, it is usually even inferior to that of the non-discriminant GMM classifier. While this may appear surprising, it underlines one of the points of section 2.2: *that the localized representation is inherently more invariant than the holistic one, therefore leading to simpler classification problems.* Due to this, the weaker classifier (GMM) outperforms the stronger one (SVM) when there are less views of each object in the training set and, therefore, the ability to generalize becomes more important. On the other hand, as expected, the combination of the localized representation with a discriminant classifier (KL-kernel SVM) outperforms that of the localized representation with a generative classifier (GMM). Overall, the *KL-kernel SVM achieves the best performance of all*

methods by combining the higher invariance of the localized representation with the better classification performance of discriminant methods.

Table 1. *Recognition rate (in %) for the various classifiers discussed in the text.*

	Resolution 32×32				Resolution 64×64				Resolution 128×128			
	$\mathcal{D}_4$	$\mathcal{D}_8$	$\mathcal{D}_{18}$	$\mathcal{D}_{36}$	$\mathcal{D}_4$	$\mathcal{D}_8$	$\mathcal{D}_{18}$	$\mathcal{D}_{36}$	$\mathcal{D}_4$	$\mathcal{D}_8$	$\mathcal{D}_{18}$	$\mathcal{D}_{36}$
L-SVM	67.24	82.67	92.98	97.31	67.54	82.84	92.85	97.39	67.85	82.80	92.89	97.50
P2-SVM	63.02	80.03	93.09	98.11	62.27	79.11	92.30	97.89	62.53	77.78	92.85	97.58
G-SVM	72.79	88.67	96.85	**99.78**	75.75	90.80	97.78	**99.68**	75.54	90.13	97.04	**99.17**
GMM	76.41	91.05	96.30	97.83	80.82	90.27	94.89	95.31	82.48	90.89	94.72	94.89
KL-SVM	**79.56**	**93.20**	**97.32**	98.28	**83.69**	**94.36**	**98.89**	98.83	**84.32**	**95.22**	**98.65**	98.67

Regarding the impact of resolution on classification rate, the tables also support some interesting observations. The first is that the performance of G-SVM is approximately constant across resolutions. This is remarkable since, for the holistic representation, 128×128 corresponds to a $16,384$ dimensional feature space. The fact is that, as resolution increases, the classification performance is subject to a tug-of-war between the nefast consequences of the curse of dimensionality and the benefits of added image information. For the holistic SVM these effects cancel out and performance is approximately constant. The localized representation, on the other hand, does not suffer from any increase in the dimensionality (only more vectors per image) and only has to benefit. Hence, *the gain in recognition rate of KL-SVM over G-SVM increases with image resolution.* For the hardest problems considered ($I = 4$) the decrease in error rate was as large as 36%. Once again, this underlines the points of section 2.2.

4.2 The flexibility of the KL-kernel

Given that the most discriminant visual attributes for recognition depend on the recognition task (e.g. while shape might achieve the best results for digit recognition, texture is a better cue for discriminating outdoor scenes) a general-purpose classifier should support multiple image representations. As discussed in section 3.2, the flexibility of the KL-kernel makes it very attractive from this point of view. In this section, we evaluate the performance on COIL-100 of its combination with previously proposed representations for recognition, in particular, representations based on color, appearance, and joint color and appearance. Color-histograms were initially proposed for recognition in [22] and are today commonly used for object recognition and image retrieval [19]. Histogram similarity is frequently measured with the histogram intersection metric, which is equivalent to the L_1 distance between the histograms [22]. In the SVM context, this metric has been proposed as a kernel for visual recognition by [23], and denoted by *Laplacian kernel*. We compared its performance with that of the χ^2 approximation to the KL-divergence, a popular approximation for histogram-based recognition. For modeling local appearance we used the representation of the previous section (DCT coefficients of the luminance channel for appearance

alone, DCT coefficients of the three color channels for joint color and appearance). For global appearance we used the holistic representation. To jointly model color and global appearance we concatenated the vectors from the three color channels into a vector three times larger.

Results All experiments were based on 128×128 images and dataset $\mathcal{D}_4$. Color histograms were computed with $16 \times 16 \times 16$ bins, gray-level histograms with 16 bins. For joint color and local appearance, the DCT coefficients were interleaved into a 192 dimensional vector of which only the first 64 dimensions were used for density estimation. Table 2 presents a comparison of the recognition rates. The first interesting observation from this table is the importance of color as a cue for recognition on COIL, since all representations perform significantly better when color is used. Interestingly, in this case, the extremely localized histogram representation (features of pixel support) beats the less-localized (8×8 supported) appearance-based counterpart and both significantly outperform the holistic representation. This illustrates the trade-off between localization and invariance at an extreme level: *because color is so discriminant, even the invariance loss associated with the small* 8×8 *neighborhood is sufficient to degrade recognition performance. The invariance loss of the holistic representation is so large that its performance is not even close to those of the localized representations*. Note that the point is not to claim that the color histogram is the ultimate solution for object recognition. In fact, it would likely not perform as well if, for example, there were more objects with similar colors in the database. The point is that different visual attributes are most discriminant for different databases, and less discriminant attributes require representations of larger spatial support (which allow modeling *configurations* of features therefore increasing the discriminant power). However, larger support usually implies less invariance (since the manifolds spanned by the configurations are increasingly more complex) and the result is a trade-off between discriminant power and invariance. In table 2 the best value for this trade-off is achieved by the localized representation, for grayscale images, and by the histogram-based one, when color is used. The conclusion is that, even for a given classification problem, the optimal representation can vary depending on factors such as the composition of the database, its size, the visual features that can be reliably extracted, etc. In this sense, the ability of the KL-kernel to support a diversity of representations can be a great asset.

A second interesting observation is to compare the results in the table with those obtained by Roth et al. They used shape as the cue for recognition and proposed two representations. One based on explicit encoding of the position of pixels in the object contour, the second based on conjunctions of edges. The first achieved a rate of 81.46, i.e. superior only to the combination of the KL-kernel with the grayscale histogram, and the grayscale G-SVM. The second achieved a rate, 88.28, slightly superior to the grayscale KL-SVM kernel, and superior to the two holistic SVM representations, but clearly inferior to any of the KL-SVM kernels using color. Again, these results highlight the importance of different representations for different databases. While color does not produce a winner when combined with holistic appearance, it completely shatters the performance

of shape when combined with any of the localized representations. On the other hand, shape appears to be more discriminant than appearance in the absence of color. This suggests that it would be interesting to have a shape-based kernel for the KL-SVM, an area that we are now exploring.

Table 2. *Recognition rate (in %) of classifiers based on different visual cues: color, appearance, and joint color and appearance.*

	histogram-based	local appearance	global appearance
grayscale	χ^2 kernel: 71.72 Laplacian kernel: 69.90	KL-SVM: 84.32	G-SVM: 75.54
color	χ^2 kernel: 98.12 Laplacian kernel: 97.81	KL-SVM: 96.74	G-SVM: 84.90

References

1. P. Viola and M. Jones. Robust Real-Time Object Detection. In 2^{nd} *International Workshop on Statistical and Computational Theories of Vision*, 2001.
2. H. Murase and S. Nayar. Visual Learning and Recognition of 3-D Objects from Appearence. *International Journal of Computer Vision*, 14:5–24, 1995.
3. D. Roth, M. Yang, and N. Ahuja. Learning to Recognize Three-Dimensional Objects. *Neural Computation*, 14:1071–1103, 2002.
4. M. Weber, M. Welling, and P. Perona. Unsupervised Learning of Models for Recognition. In *European Conf. on Computer Vision*, pages 18–32, Dublin, Ireland, 2000.
5. V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, 1995.
6. H. Schneiderman and T. Kanade. A Statistical Method for 3D Object Detection Applied to Faces and Cars. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Hilton Head, South Carolina*, 2000.
7. T. Jaakkola, M. Diekhans, and D. Haussler. Using the fisher kernel method to detect remote protein homologies. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, August 1999.
8. P. J. Moreno, P. P. Ho, and N. Vasconcelos. A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications in *Proc. of NIPS 2003*
9. N. Vasconcelos. On the Efficient Evaluation of Probabilistic Similarity Functions for Image Retrieval, *IEEE Transactions on Information Theory*, to appear
10. N. Vasconcelos. A Unified View of Image Similarity. In *Proc. Int. Conf. Pattern Recognition*, Barcelona, Spain, 2000.
11. K. Tsuda, M. Kawanabe, G Ratsch, S. Sonnenburg, and K. Muller. A New Discriminative Kernel from Probabilistic Models. *Neural Computation*, 14(10):2397–2414, 2002.
12. R. Kondor and T. Jebara. A kernel between sets of vectors. In *International Conference on Machine Learning*, 2003.
13. H. Rowley, S. Baluja, and T. Kanade. Neural Network-Based Face Detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(1):23–38, January 1998.
14. B. Moghaddam and M. Yang. Gender Classification with Support Vector Machines. In 4^{th} *IEEE Int'l Conference on Automatic Face & Gesture Recognition*, 2000.

15. L. Wolf and A. Shashua. Kernel Principal Angles for Classification Machines with Applications to Image Sequence Interpretation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2003, Madison, Wisconsin.
16. S. Kullback. *Information Theory and Statistics*. Dover, New York, 1968.
17. D. Johnson and S. Sinanovic. Symmetrizing the Kullback-Leibler Distance. *IEEE Transactions on Information Theory*, March 2001. Submitted.
18. R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.
19. J. Puzicha, Y. Rubner, C. Tomasi, and J. Buhmann. Empirical Evaluation of Dissimilarity Measures for Color and Texture. In *International Conference on Computer Vision, Korfu, Greece*, pages 1165–1173, 1999.
20. Y. Singer and M. Warmuth. Batch and On-line Parameter Estimation of Gaussian Mixtures Based on Joint Entropy. In *Neural Information Processing Systems, Denver, Colorado*, 1998.
21. J. Goldberger, S. Gordon, and H. Greenspan. An Efficient Image Similarity Measure based on Approximations of the KL-Divergence Between Two Gaussian Mixtures. In *International Conference on Computer Vision*, 2003.
22. M. Swain and D. Ballard. Color Indexing. *International Journal of Computer Vision*, Vol. 7(1):11–32, 1991.
23. O. Chapelle, P. Haffner, and V. Vapnik. Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10(5):1055–1064, September 1999.