In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Rio de Janeiro, Brazil, October 2007.

Active Learning with Gaussian Processes for Object Categorization

Ashish Kapoor
Microsoft Research
Redmond, WA 98052, USA
akapoor@microsoft.com

Kristen Grauman
Univ. of Texas at Austin
Austin, TX 78712, USA
grauman@cs.utexas.edu

Raquel Urtasun and Trevor Darrell
MIT CSAIL
Cambridge, MA 02139, USA
{rurtasun, trevor}@csail.mit.edu

Abstract

Discriminative methods for visual object category recognition are typically non-probabilistic, predicting class labels but not directly providing an estimate of uncertainty. Gaussian Processes (GPs) are powerful regression techniques with explicit uncertainty models; we show here how Gaussian Processes with covariance functions defined based on a Pyramid Match Kernel (PMK) can be used for probabilistic object category recognition. The uncertainty model provided by GPs offers confidence estimates at test points, and naturally allows for an active learning paradigm in which points are optimally selected for interactive labeling. We derive a novel active category learning method based on our probabilistic regression model, and show that a significant boost in classification performance is possible, especially when the amount of training data for a category is ultimately very small.

1. Introduction

Collecting training data for large-scale image category models is a potentially expensive process. While certain categories may have a large number of training images available, many more will have relatively few. A number of ingenious schemes have been developed to obtain labeled data from people performing other tasks (e.g., [30, 29]), or directly labeling objects in images [1]. To make the most of scarce human labeling resources it is imperative to carefully select points for user labeling. The paradigm of active learning has been introduced in the machine learning community to address this issue [8, 26, 16, 18, 34]; with an active learning method, generally new test points are selected so as to minimize the model entropy.

To develop an active learning method for object category recognition, a probabilistic category estimation method is needed. Current results on benchmark category recognition tasks suggest that discriminative methods offer the best performance, but most such methods are not explicitly probabilistic and offer little guidance as to where estimates are accurate or where a model may generalize poorly. Probabilistic models are therefore desirable as they provide uncertainty estimates as part of the inference process.

We introduce a new Gaussian Process (GP) regression method for object category recognition using a local feature correspondence kernel. Local feature based object recognition has been shown to have several important advantages, including invariance to various translational, rotational, affine and photometric transformations and robustness to partial occlusions. Our method is based on a GP with a covariance function derived from a Pyramid Match Kernel [9], which offers an efficient approximation to a partial-match distance function and can therefore handle outliers and occlusions.

GPs have received limited attention in the computer vision literature to date perhaps due to the fact that they are conventionally limited to modest amounts of training data: the learning complexity is $O(n^3)$, cubic in the number of training examples. While recent advances in sparse GPs are promising (e.g., [12, 21, 24]), we focus here on the case of active learning with relatively small numbers of examples (10-100), which is feasible with existing implementations. In this realm, we show that active learning provides significantly more accurate estimates per labeled point than does a conventional random selection of training points.

The two main contributions of this paper are 1) a probabilistic discriminative category recognition scheme based on a Gaussian Process prior with a covariance function defined using the Pyramid Match Kernel, and 2) the introduction of an active learning paradigm for object category learning which optimally selects unlabeled test points for interactive labeling. With active learning very small amounts of interactively labeled data can provide very accurate category recognition performance.

2. Previous Work

Object category recognition has been a topic of active interest in the computer vision literature. Methods based on local feature descriptors (c.f. [14, 17]) have been shown to offer invariance across a range of geometric and photometric conditions. Early models captured appearance and shape variation in a generative probabilistic framework [7], but

more recent techniques have typically exploited methods based on SVMs or Nearest Neighbors in a bag-of-visual-words feature space [23, 19, 33, 5].

Several authors have explored correspondence-based kernels [33, 31], where the distance between a set of local feature descriptors—potentially including appearance and shape/position—is computed based on associating pairs of descriptors. However, the polynomial-time computational cost of correspondence-based distance measures makes them unsuitable for domains where there are large databases or large numbers of features per image. In [9] the authors introduced the Pyramid Match Kernel (PMK), an efficient linear-time approximation to a partial match correspondence, and in [13] it was demonstrated that a spatial variant—which efficiently represents the distinction between appearance and image location features—outperformed many competing methods.

Semi-supervised or unsupervised visual category learning methods are related to active learning. Generative models which model visual words as arising from a set of underlying objects or "topics" based on recently introduced methods for Latent Dirichlet Allocation have been developed [22, 25] but as yet have not been applied to active learning nor evaluated on purely supervised tasks. A semi-supervised method using normalized cuts to cluster a graph defined by Pyramid Match distances between examples was presented in [11], but this method was not probabilistic and did not provide for an active learning formalism.

In the machine learning literature active learning has been a topic of recent interest, and numerous schemes have been proposed for choosing unlabeled points for tagging. For example, Freund *et al*. [8] propose disagreement among the committee of classifiers as a criterion for active learning, and show an application to image classification [2]. Tong and Koller [26] explore the selection of unlabeled cases to query based on minimizing the version space within the SVM formulation. Chang *et al*. [3] use active learning with SVMs for the task of image retrieval using color and texture.

Within the Gaussian Process framework, the method of choice has been to look at the expected informativeness of an unlabeled data point [12, 15]. Specifically, the idea is to choose to query cases that are expected to maximally influence the posterior distribution over the set of possible classifiers. Additional studies have sought to combine active learning with semi-supervised learning [16, 18, 34]. Our work is significantly different as we focus on local feature approaches for the task of object categorization. We explore the GP models, which provide estimates for uncertainty in prediction and can be easily extended to active learning.

Gaussian Processes have been recently introduced to the computer vision literature. While they have been used in [27, 28] for human motion modeling and in [32] for stereo segmentation, we are unaware of any prior work on visual object recognition in a Gaussian Process framework.

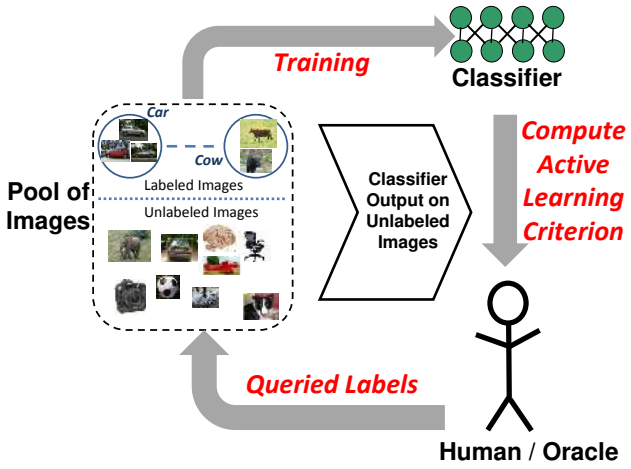

Figure 1. The active learning framework.

3. Approach

Active learning tackles the problem of finding the most crucial data in a set of unlabeled examples so that the classification system gains the most if it is given the label of that example. Figure 1 shows the proposed framework for the image categorization task. Given a pool of images of which few are labeled, the system aims to actively seek labels for unlabeled images by considering information from both the labeled and unlabeled sets of images.

At the core of this system is the classification framework, and in this work we explore classification using Gaussian Process priors with covariance functions defined by the Pyramid Match Kernel (GP-PMK). The next section reviews classification using GP priors. We then present our GP-PMK model which is directly suitable for supervised learning. Finally, we derive an active learning variant that can optimally select points for interactive labeling.

Note that in this paper we assume that there is one object of interest in an image. Handling multiple objects in the same image is an interesting and more challenging problem and will be the focus of future work.

4. Background: Classification with Gaussian Processes

Gaussian Process (GP) classification is related to kernel machines such as Support Vector Machines (SVMs) [4] and Regularized Least Square Classification (RLSC) and has been well explored in machine learning. In contrast to these methods GPs provide probabilistic prediction estimates and thus are well suited for active learning. In this section we briefly review regression and classification with Gaussian Process priors.

Given a set of labeled data points $\mathbf{X}_L = \{\mathbf{x}_1, .., \mathbf{x}_n\}$, with class labels $\mathbf{t}_L = \{t_1, .., t_n\}$, we are interested in

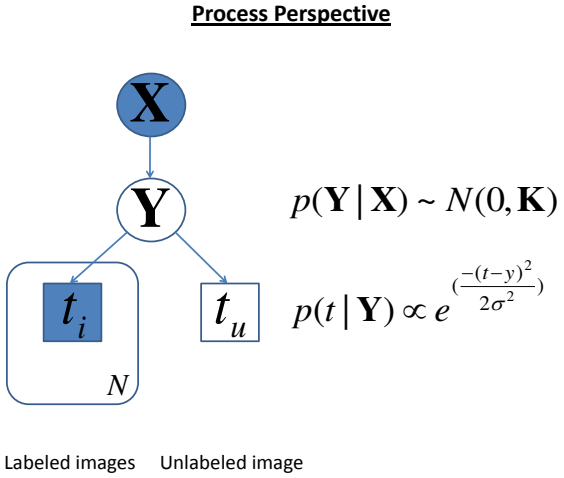

Figure 2. Graphical models in plate notation for classification via Gaussian Processes. The rounds and squares represent continuous and discrete random variables, respectively. A filled (unfilled) round/square denotes that the random variable is fully observed (unobserved). $\mathbf{X} = \{\mathbf{x}_1, .., \mathbf{x}_n, \mathbf{x}_u\}$ is the set of all images and is observed for both labeled and unlabeled data points. The corresponding $\mathbf{Y} = \{y_1, ..y_n, y_u\}$ is completely unobserved and the labels $\{t_1, .., t_n\}$ are observed only for the training images $\{\mathbf{x}_i, ..\mathbf{x}_n\}$ and unobserved for the unlabeled image $\mathbf{x}_u$.

classifying the unlabeled data $\mathbf{x}_u$. Under the Bayesian paradigm, we are interested in the distribution $p(\mathbf{t}_u|\mathbf{X}, \mathbf{t}_L)$. Here $\mathbf{X} = \{\mathbf{X}_L, \mathbf{x}_u\}$ and $\mathbf{t}_u$ is the random variable denoting the class label for the unlabeled point $\mathbf{x}_u$. For sake of simplicity in discussion we limit ourselves to two-way classification, hence, the labels are, $t_i \in \{-1, 1\}$.

With GP models, a discrete label t for a data point $\mathbf{x}$ can be considered to be generated via a continuous hidden random variable y. The soft-hidden label arises due to a Gaussian Process, which in turn imposes a smoothness constraint on the possible solutions. A likelihood model $p(t|y)$ characterizes the relationship between the soft label y and the observed annotation t. Thus, when we infer the label t_u for the unlabeled data $\mathbf{x}_u$, we probabilistically combine the smoothness constraint and the information obtained by observing the annotations t.

The smoothness constraint is imposed using a Gaussian Process prior that defines the probabilistic relationship between the images $\mathbf{X}$ and the soft labels $\mathbf{Y}$. The distribution $p(\mathbf{Y}|\mathbf{X})$ gives higher probability to the labelings that respect the similarity between the data points. Intuitively, the assumption is that similar data points should have the same class assignments / regression values; the similarity between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined via a kernel $k(\mathbf{x}_i, \mathbf{x}_j)$. Below we derive a novel GP based on local feature correspondences.

Probabilistic constraints are imposed on the collection of soft labels $\mathbf{Y} = \{y_1, ..., y_n, y_u\}$. In particular, the soft labels are assumed to be jointly Gaussian and the covariance between two outputs y_i and y_j is specified using the kernel function applied to $\mathbf{x}_i$ and $\mathbf{x}_j$. Formally, $p(\mathbf{Y}|\mathbf{X}) \sim \mathcal{N}(0, \mathbf{K})$ where $\mathbf{K}$ is a $(n+1)$-by-$(n+1)$ kernel matrix with $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, and $n+1$ reflects the n labeled examples and 1 unlabeled example.

The likelihood models the probabilistic relation between the observed label t and the hidden label y. In this work we assume that t and y are related via a Gaussian noise model. Specifically,

$$p(t|y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-y)^2}{2\sigma^2}}. \quad (1)$$

The Gaussian noise model is commonly used in GP regression, as it leads to closed form inference. While a range of methods for GP classification have been proposed using additional latent "squashing" variables, inference with these methods is not possible in closed form [20]. For the experiments reported below we simply use regression to label variables.[1] Exploration of non-Gaussian noise models for our task such as the probit function or a sigmoid is a topic of interest for future work.

Given the labeled and unlabeled data points, our goal is then to infer $p(\mathbf{t}_u|\mathbf{X}, \mathbf{t}_L)$. Specifically:

$$p(t_u|\mathbf{X}, \mathbf{t}_L) \propto \int_{\mathbf{Y}} p(\mathbf{t}_u|\mathbf{Y}) p(\mathbf{Y}|\mathbf{X}, \mathbf{t}_L). \quad (2)$$

For a Gaussian noise model we can compute this integral using closed form expressions. Note that the key quantity to compute is the posterior $p(\mathbf{Y}|\mathbf{X}, \mathbf{t}_L)$, which can be written as:

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{t}_L) \propto p(\mathbf{Y}|\mathbf{X}) p(\mathbf{t}_L|\mathbf{Y}) = p(\mathbf{Y}|\mathbf{X}) \prod_{i=1}^{n} p(t_i|y_i). \quad (3)$$

This equation probabilistically combines the smoothness constraints $p(\mathbf{Y}|\mathbf{X})$ imposed via the GP prior and the information provided in the labels $(p(\mathbf{t}_L|\mathbf{Y}))$. The posterior as shown in Equation 3 is simply a product of Gaussians. We are interested in inferring the unknown label t_u. The Gaussian posterior over the soft label y_u has a particularly simple form. Specifically, $p(y_u|\mathbf{X}, \mathbf{t}_L) \sim \mathcal{N}(\bar{y}_u, \sigma_u^2)$, where:

$$\bar{y}_u = \mathbf{k}(\mathbf{x}_u)^T (\sigma^2 \mathbf{I} + \mathbf{K}_{LL})^{-1} \mathbf{t}_L \quad (4)$$

$$\Sigma_u = k(\mathbf{x}_u, \mathbf{x}_u) - \mathbf{k}(\mathbf{x}_u)^T (\sigma^2 \mathbf{I} + \mathbf{K}_{LL})^{-1} \mathbf{k}(\mathbf{x}_u). \quad (5)$$

Here, $\mathbf{k}(\mathbf{x}_u)$ is the vector of kernel function evaluations with n training points, and $\mathbf{K}_{LL}$ is the training covariance. Further, due to the Gaussian noise model that links t_u to y_u, the predictive distribution over the unknown label t_u is also a Gaussian: $p(t_u|\mathbf{X}, \mathbf{t}_L) \sim \mathcal{N}(\bar{y}_u, \Sigma_u + \sigma^2)$.

[1] This method is referred to as least-squares classification in the literature (see Section 6.5 of [20]) and often demonstrates performance competitive with more expensive GPC methods that require approximate inference.

Note that the posterior mean for both t_u and y_u is the same; thus, the unlabeled point $\mathbf{x}_u$ can be classified according to the sign of y_u. Unlike RLSC methods, we also get the variance in prediction. As we will show in the next section, we can exploit these measures of uncertainty to guide an active learning procedure. The computationally costly operation in GP inference is the inversion of $(\sigma^2\mathbf{I} + \mathbf{K}_{LL})$ which is $O(n^3)$. In addition to reducing manual labeling effort, an active learning formulation can help us reduce the computational overhead in inference by reducing the number of needed training points.

5. Pyramid Match Kernel Gaussian Processes (GP-PMK)

To use GPs for object categorization, we need to define a suitable covariance function. We would like to exploit local feature methods for object and image representations. However, GP priors require covariance functions which are positive semi-definite (a Mercer kernel) and traditional covariance functions (e.g., RBF) are not suitable for representations that are comprised of sets of features.

We wish to define a GP with a covariance function based on a partial match distance function. The idea is to first represent an image as an unordered set of local features, and then use a matching over these sets of features to compute a smoothness prior between images. The optimal least-cost partial matching takes two sets of features, possibly of varying sizes, and pairs each point in the smaller set to a unique point in the larger one, such that the sum of the distances between the matched points is minimized. The cubic cost of the optimal matching makes it prohibitive for recognition with a large number of local image features, yet rich image descriptions comprised of densely sampled local features are known to yield better recognition accuracy.

Therefore, rather than adopt a full partial match kernel for the GP prior, we use the Pyramid Match [9]. The Pyramid Match is a linear-time kernel function over unordered feature sets that approximates the similarity measured by the optimal partial matching, and it forms a Mercer kernel. A multi-resolution partition (Pyramid) carves the feature space into increasingly larger regions. At the finest resolution level in the Pyramid, the partitions are very small; at successive levels they continue to grow in size until a single partition encompasses the entire feature space. The insight of the Pyramid Match algorithm is to treat points which share a bin in this Pyramid as being matched, and to use the histograms to read off the number of possible matches without explicitly searching for correspondences. Histogram intersection (the sum of the minimum number of points in a given histogram bin) is used to count the number of new matches that occur at each resolution level.

The input space S contains sets of feature vectors drawn from feature space $\mathcal{F}$: $S = \{\mathbf{F} | \mathbf{F} = \{\mathbf{f}_1, \ldots, \mathbf{f}_m\}\}$, where each feature $\mathbf{f}_i \in \mathcal{F} \subseteq \Re^d$, and $m = |\mathbf{F}|$. For example, $\mathcal{F}$ might be the space of SIFT [14] descriptors ($d = 128$), or image coordinate positions ($d = 2$), etc.; a set $\mathbf{F}$ contains a collection of these descriptors extracted from a single image or object. An L-level histogram Pyramid for input example $\mathbf{F} \in S$ is defined as: $\Psi(\mathbf{F}) = [H_0(\mathbf{F}), \ldots, H_{L-1}(\mathbf{F})]$, where $H_i(\mathbf{F})$ is a histogram vector formed over points in $\mathbf{F}$ using multi-dimensional bins. The partitions within each histogram H_i may be placed at uniform intervals to divide the feature space into equally sized grid cells, as in [9, 13], or they may be placed non-uniformly in a data-dependent manner, as in [10].

The Pyramid Match Kernel (PMK) value between two input sets $\mathbf{F}_1$, $\mathbf{F}_2 \in S$ is defined as the weighted sum of the number of feature matches found at each level of their Pyramids [9]:

$$K_\Delta\left(\Psi(\mathbf{F}_1), \Psi(\mathbf{F}_2)\right) = \sum_{i=0}^{L-1} w_i \Big(\mathcal{I}\left(H_i(\mathbf{F}_1), H_i(\mathbf{F}_2)\right) - \mathcal{I}(H_{i-1}(\mathbf{F}_1), H_{i-1}(\mathbf{F}_2)) \Big)$$

where $\mathcal{I}$ denotes histogram intersection, and the difference in intersections across levels ($\mathcal{I}\left(H_i(\mathbf{F}_1), H_i(\mathbf{F}_2)\right) - \mathcal{I}(H_{i-1}(\mathbf{F}_1), H_{i-1}(\mathbf{F}_2))$ serves to count the number of new matches formed at level i, which were not already counted at any finer resolution level. The weights are set to be inversely proportional to the size of the bins, in order to reflect the maximal distance two matched points could be from one another. As long as $w_i \geq w_{i+1}$, the kernel is Mercer. A variant of the PMK described in [13] first quantizes the appearance feature descriptors to form a bag-of-words representation, and then sums over the PMK values for each word in the space of image coordinates.

We thus define a Pyramid Match Gaussian Process model (GP-PMK) using

$$p(\mathbf{Y}|\mathbf{X}) \sim \mathcal{N}(0, K_\Delta). \tag{6}$$

In contrast to previous GP priors, this prior is well suited for visual category recognition as it naturally handles representations based on sets of local image features.

6. Active Learning for Object Categorization

In this section we assume that we have a pool of unlabeled data $\mathbf{X}_U = \{\mathbf{x}_{n+1}, .., \mathbf{x}_{n+m}\}$. The task in active learning is to seek the label for one of these examples and then update the classification model by incorporating it into the existing training set. The goal is to select the sample that would maximize the benefit in terms of the discriminatory capability of the system.

With non-probabilistic classification schemes a popular heuristic for establishing the confidence of estimates and identifying points for active learning is to simply use the distance from the classification boundary (margin). This approach can also be used with GP classification models, by inspecting the magnitude of the posterior mean $\bar{y}_u$: we would then choose the next point $\mathbf{x}^*$ as $\arg\max_{\mathbf{x}_u \in \mathbf{X}_U} \bar{y}_u$.

Table 1. Active Learning Criteria

Method	Criteria
Distance from Boundary (SVM)	$\mathbf{x}^* = \arg\min_{\mathbf{x}_u \in \mathbf{X}_U} \lvert\bar{y}_u\rvert$
Variance	$\mathbf{x}^* = \arg\max_{\mathbf{x}_u \in \mathbf{X}_U} \Sigma_u$
Uncertainty (GP)	$\mathbf{x}^* = \arg\min_{\mathbf{x}_u \in \mathbf{X}_U} \frac{\lvert\bar{y}_u\rvert}{\sqrt{\Sigma_u + \sigma^2}}$

However, GP classification provides us with both the posterior mean as well as the posterior variance for the unknown label t_u. An alternative criteria could be to look at the variances and select the point that has the maximum variance, i.e. $\mathbf{x}^* = \arg\max_{\mathbf{x}_u \in \mathbf{X}_U} \Sigma_u$. However such an approach does not consider the mean $\bar{y}_u$ at all! Further, the expression for Σ_u does not consider labels from the annotated training data; this scheme uses only a very limited amount of information.

We therefore propose an approach which considers both the posterior mean as well as the posterior variance. Specifically, we select the next point according to:

$$\mathbf{x}^* = \arg \min_{\mathbf{x}_u \in \mathbf{X}_U} \frac{|\bar{y}_u|}{\sqrt{\Sigma_u + \sigma^2}}. \quad (7)$$

This formulation considers uncertainty in the labeling $\mathbf{x}_u$ as ± 1. Note that the predictive distribution for t_u is a Gaussian; however, we are interested in the binary label decided according to the sign of t_u. To this end we should consider the $p(t_u \geq 0) = \phi(\frac{\bar{y}_u}{\sqrt{\Sigma_u + \sigma^2}})$, where $\phi(\cdot)$ denotes the cdf of a standard normal distribution, to provide the hard label ± 1. Further, we are interested in selecting those samples where the uncertainty is maximum. The points where the classification model is most uncertain should have $p(t_u \geq 0)$ close to 0.5 - equivalently, $\frac{|\bar{y}_u|}{\sqrt{\Sigma_u + \sigma^2}}$ lies very close to zero. Thus, the criterion in Equation 7 chooses the unlabeled point where the classification is the most uncertain.

We summarize the methods for identifying points to be labeled in Table 1. Our active learning approach looks at all the points before choosing the active points; thus it considers the whole dataset instead of just looking at individual points. Further, this scheme considers both the distance from the boundary as well as the variance in selecting the points; this is only possible due to the availability of the predictive distribution in GP regression.

Other active learning criteria such as information gain score or differential entropy [12] are possible, and these have been demonstrated to have advantages in online learning and sparsifying methods. We plan to investigate the utility of these approaches for our active learning scheme in future work.

7. Experiments and Results

We performed experiments to 1) demonstrate the effectiveness of the GP-PMK classification framework, 2) compare different discriminative models, and 3) demonstrate

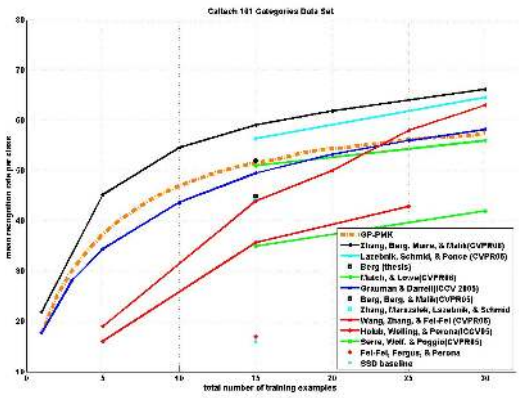

Figure 3. Performance comparison of GP-PMK classification.

that active learning can guide the learning procedure to select critical examples. We show how active learning with a GP-PMK yields classifiers which can learn object categories from relatively few examples.

We performed supervised and active learning experiments on two different datasets that are considered standards for the object categorization task: the Caltech-4 dataset and the Caltech-101 dataset (which is a superset of Caltech-4). We compute the similarity between all pairs of images in each database using the PMK. LIBSVM was used for SVM baseline tests. In our experiments we set the noise model variance $\sigma = 10^{-5}$ for the Gaussian process models and fix $C = 10000$ for SVM models. These parameter values worked well; we experimented with other values but found that both SVM and GP classification schemes were fairly insensitive to the choice of these parameters.

The object categorization task is a multi-class problem ($n_{class} = 101$ and $n_{class} = 4$ for the Caltech-101 and the Caltech-4). To handle multiple classes we use one-vs-all formulation, where we choose the label corresponding to the class with maximum value of the soft label y. For multi-class active learning in every round we select one example from each of the one-vs-all classifiers, thus adding n_{class} examples every time.

The Caltech-4 database contains 3188 images with four object classes. There are 1155 rear views of cars, 800 images of airplanes, 435 images of frontal faces, and 798 images of motorcycles. The second data base is the Caltech-101 database of 101 object categories [6]; there are 8677 images in this data set, with between 31 to 800 images for each of the 101 categories.

For experiments described in this paper, we used the PMK with SIFT descriptors extracted densely from the images (dense PMK), where we compute features at every 8th pixel in the image, and concatenate each with their normalized image positions. We used PCA to reduce the dimensionality of the SIFT descriptors to 10. We also performed

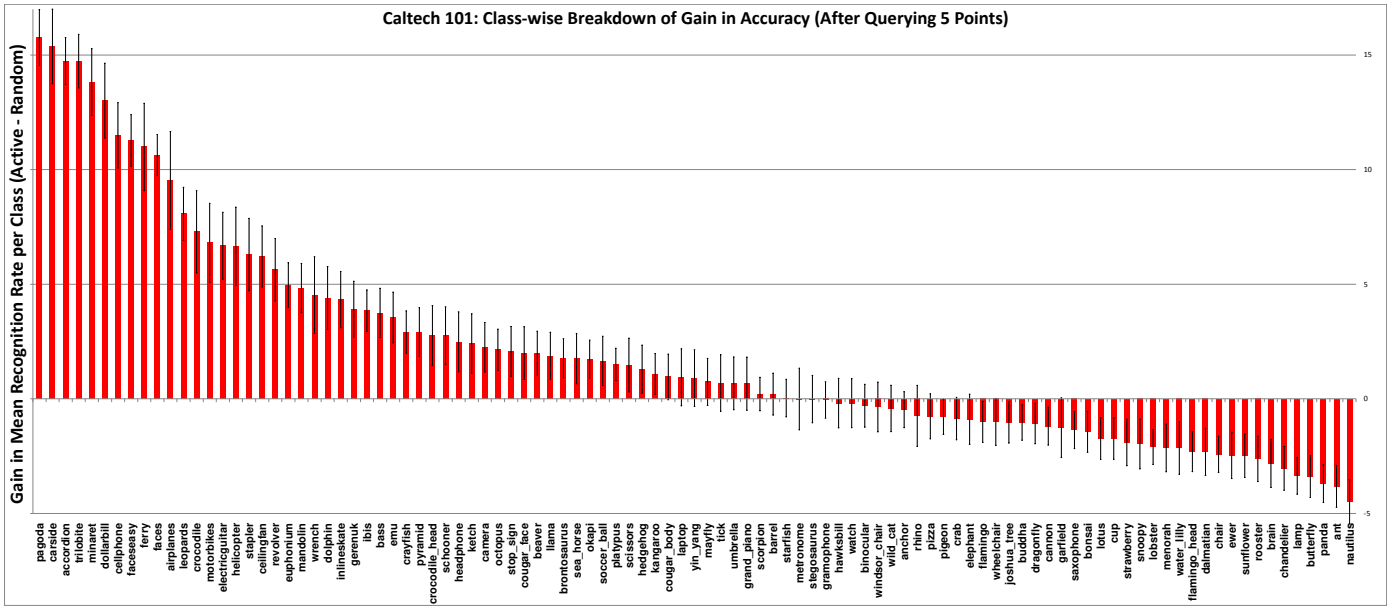

Figure 4. Average gain in performance over random selection when choosing 5 points using active learning. The graph shows the mean and the standard error for 100 runs for all the object classes in Caltech-101 database using GP-PMK.

experiments on two other flavors of PMK that included one with data-dependent partitions [10] and the other where the SIFT descriptors are extracted only at salient points detected with a Harris-Affine interest operator [17]. We observed similar trends in all the cases and due to space limitations report results only on uniform-bin PMK with dense features.

First, we compare GP-PMK classification with state-of-the-art supervised visual category learning methods. We follow the standard testing protocol, where a given number (say 15) of training images are taken from each class at random, and the rest of the data is used for testing. The mean recognition rate per class is used as a metric of performance. Note that this metric makes sure that the recognition accuracies are such that classes with large numbers of examples are not favored. This process is repeated 10 times and the average correctness rate is reported. Figure 3 shows the performance of an SVM and the classification with GP priors using the PMK along with the other state-of-the-art methods using the same evaluation methodology. The PMK was also earlier used by Grauman and Darrell [9] with SVMs. We show in figure 3 that classification with GP outperforms the SVM; thus, demonstrating the value in the proposed approach. The other approaches by Zhang *et al.* [33] and Lazebnik *et al.* [13] perform better; however, note that those approaches have different feature representations that may provide a significant advantage in the task. The point we wish to make here is that GP classification can often provide comparable or slightly improved classification performance when compared to SVMs; we do not have to lose accuracy to gain the predictive uncertainty offered by probabilistic recognition models.

Next, we show the value of active learning in selecting examples to annotate. For these experiments, we test the classification performance on a validation set that includes 10 examples from each class. We first consider the *binary* problem of detecting an object class. Starting with one labeled example per class, the procedure chooses the next image to query from the set of images not in the validation set. We compare the active version of the GP classification with a version that selects the points to query randomly. We again use the mean classification rate per class to compare the methods. We repeat this procedure for 100 different validation sets. Figure 4 shows the gain in performance on all the 101 binary problems, averaged over the 100 runs, made by the active learning scheme on the validation set after 5 examples are chosen. We can clearly see that for most of the categories there is a significant positive gain showing the benefit of the active learning scheme. Further, figure 5 shows the performance on various binary problems as we increase the size of the training set. The figure depicts that the active learning scheme quickly exploits the uncertainty in its estimates to select appropriate examples to seek the annotation for. The random policy on the other hand performs poorly. The fact that the Caltech-101 dataset has unbalanced numbers of examples per category affects the random sampling policy; it does not work well in these unbalanced scenarios because the training set will usually be skewed towards one class, resulting in poor accuracy. However, selecting points via active learning focuses on points with maximum uncertainty, irrespective of their label, making the procedure highly effective.

We also ran active learning experiments on the Caltech-4 dataset and figure 6 compares different classification ap-

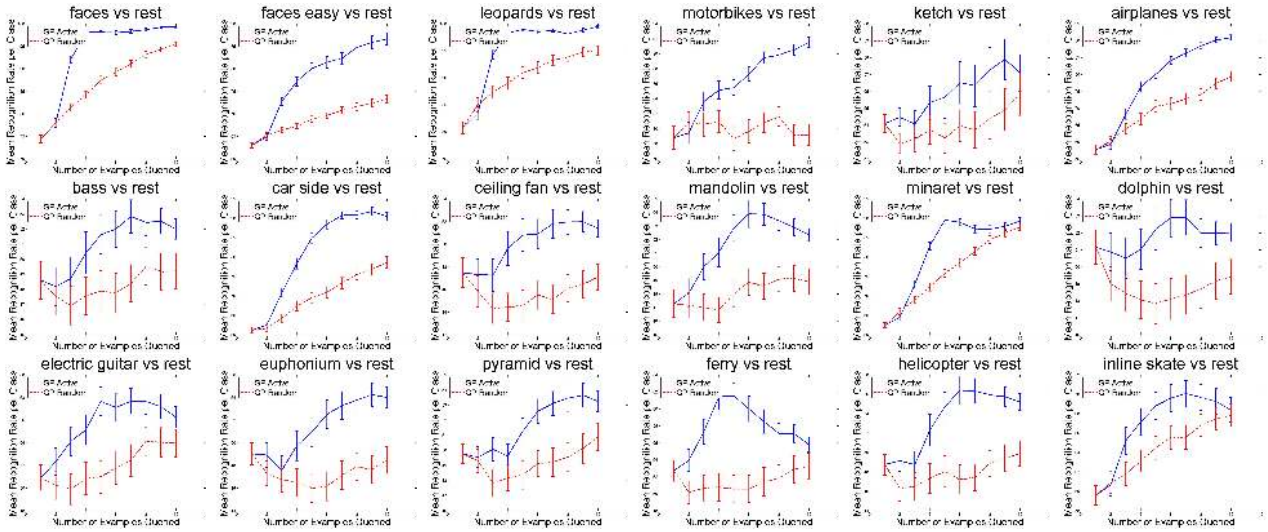

Figure 5. Performance comparison of GP classification with active learning and GP with random supervision for various object detection problems (binary) in Caltech-101 Database.

proaches. Essentially, the plot shows mean classification accuracy per class as we vary the total number of examples in the training data. The images not in the training set are considered as test set to compute the classification performance. We plot the performance of SVM and the GP classification with and without active learning. We start with zero labeled points and for SVM and supervised GP without the active learning, we randomly select points as we increase the size of the training set. The active learning GP classification uses uncertainty to guide its sample selection process. This process was repeated 40 times and figure 6 shows the mean performance. The errorbars denote the standard error and non-overlapping errorbars signify difference in performance levels with 95% confidence.

Figure 6 shows that GP classification performs competitively with the SVM and using active learning further improves the performance. In fact we can see that a mean accuracy per class close to a 90% can be obtained with just 20 labeled examples, whereas the non-active learners achieve around 85% accuracy for the same amount of labeled data. This demonstrates that active learning can provide a significant boost in accuracy, and makes it possible for the learning algorithm to learn the object classes even with very few labeled examples.

Table 2 shows the confusion matrix resulting after incorporating only 120 examples in the training set using the active learning methodology. We obtain an accuracy of 98.48%, which demonstrates the effectiveness of the framework. The completely supervised GP classification and SVM achieved a mean classification accuracy per class of 95.6% and 95.19% respectively. This shows that our active learning strategy allows us to learn object categories much more effectively than plain supervised classification.

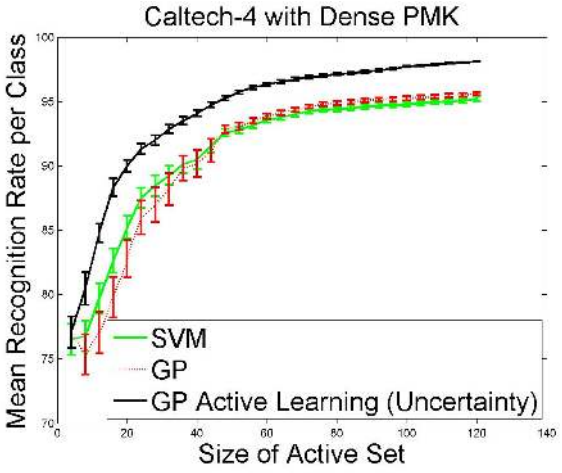

Figure 6. Active learning on Caltech-4 database using the Pyramid Match Kernel.

8. Discussion

The experiments in this paper indicate that classification using GP priors outperforms SVM on the Caltech-101 and performs competitively on the Caltech-4 database. However, we would like to point out that these experiments are not conclusive proof that classification using a GP prior is inherently superior than other classification techniques. The superior performance might be due to several reasons. For instance, one of the key requirements for any classification strategy to work well is that the underlying data supports the assumptions made by the model. In this case fortunately the underlying data density in the object categorization task for the Caltech-4 and the Caltech-101 databases are favorable to the assumptions of the classification model we are using.

Table 2. Confusion matrix obtained for Caltech-4 database using active learning with the log information score. (120 labeled images, mean accuracy over all the classes = 98.48%).

	Recognized Class			
True Class	**Cars**	**Faces**	**Airplanes**	**Motorbikes**
Cars	1121	0	0	1
Faces	0	416	0	2
Airplanes	0	2	753	20
Motorbikes	10	0	10	733

The experiments in this paper strongly suggest that there is a value in looking at GP classification models for object categorization.

Another important aspect of our framework lies in its seamless extension to active learning. The probabilistic paradigm allows us to incorporate measures such as uncertainty, variance, and expected information gain that could be highly valuable in guiding a supervised learning procedure. One of the challenges in computer vision is the ability to learn object categories with a low number of examples. Humans are able to learn object categories and generalize from a very small number of examples. However, current machine vision systems are far from achieving performance akin to humans. One of the principal differences among humans and existing object classification systems is that humans have the ability to actively seek supervision from the environment and other sources of information. We believe that active learning might enable us to move towards vision systems that require few examples to learn successfully.

9. Conclusion and Future Work

We have presented a discriminative probabilistic framework based on Gaussian Process priors and the Pyramid Match Kernel, and shown its utility for visual category recognition. The GP-PMK provides direct estimates of prediction uncertainty using a smoothness prior that captures a correspondence-based notion of similarity between sets of local image features. Further, we introduced an active learning method for visual category recognition based on the uncertainty estimates provided by the GP-PMK, and showed empirically that active learning can be used to achieve recognition results using fewer training images than standard supervised learning approaches.

We plan to extend the framework to adopt non-Gaussian noise models, and investigate other active learning formulations such as value of information and/or criteria previously developed for sparsifying GPs [12]. By incorporating decision-theoretic formulations we should be able to learn object categories within a given budget. We also plan to extend the model to handle multiple objects in the same image, incorporate semi-supervised learning, and explore sparse GP techniques for large training sets.

References

[1] *http://labelme.csail.mit.edu/*. 1

[2] Y. Abramson and Y. Freund. Active learning for visual object recognition. Technical report, UCSD, 2004. 2

[3] E. Y. Chang, S. Tong, K. Goh, and C. Chang. Support vector machine concept-dependent active learning for image retrieval. *IEEE Transactions on Multimedia*, 2005. 2

[4] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1), 2000. 2

[5] B. T. F. Moosmann and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *NIPS*, 2007. 2

[6] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transaction on Pattern Recognition and Machine Intelligence*, 2006. 5

[7] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003. 1

[8] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3), 1997. 1, 2

[9] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, 2005. 1, 2, 4, 6

[10] K. Grauman and T. Darrell. Approximate correspondences in high dimensions. In *NIPS*, 2006. 4, 6

[11] K. Grauman and T. Darrell. Unsupervised learning of categories from sets of partially matching image features. In *CVPR*, 2006. 2

[12] N. Lawrence, M. Seeger, and R. Herbrich. Fast sparse Gaussian process method: Informative vector machines. *NIPS*, 2002. 1, 2, 5, 8

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 2, 4, 6

[14] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2), 2004. 1, 4

[15] D. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4), 1992. 2

[16] A. K. McCallum and K. Nigam. Employing EM in pool-based active learning for text classification. In *ICML*, 1998. 1, 2

[17] K. Mikolajczyk and C. Schmid. Indexing Based on Scale Invariant Interest Points. In *ICCV*, 2001. 1, 6

[18] I. Muslea, S. Minton, and C. A. Knoblock. Active + semi-supervised learning = robust multi-view learning. In *ICML*, 2002. 1, 2

[19] D. Nister and H. Stewenius. Scalable Recognition with a Vocabulary Tree. In *CVPR*, 2006. 2

[20] C. E. Rasmusen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. 3

[21] Y. Shen, A. Ng, and M. Seeger. Fast gaussian process regression using kd-trees. In *NIPS*, 2006. 1

[22] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering object categories in image collections. In *ICCV*, 2005. 2

[23] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *ICCV*, 2003. 2

[24] E. Snelson and Z. Ghahramani. Sparse gaussian processes using pseudo-inputs. In *NIPS*, 2006. 1

[25] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky. Describing visual scenes using transformed dirichlet processes. In *NIPS*, 2005. 2

[26] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In *ICML*, 2000. 1, 2

[27] R. Urtasun, D. J. Fleet, and P. Fua. Gaussian process dynamical models for 3d people tracking. In *CVPR*, 2006. 2

[28] R. Urtasun, D. J. Fleet, A. Hertzman, and P. Fua. Priors for people tracking from small training sets. In *ICCV*, 2005. 2

[29] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *ACM CHI*, 2004. 1

[30] L. von Ahn, R. Liu, and M. Blum. Peekaboom: A game for locating objects in images. In *ACM CHI*, 2006. 1

[31] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *ICCV*, 2003. 2

[32] O. Williams. A switched gaussian process for estimating disparity and segmentation in binocular stereo. In *NIPS*, 2006. 2

[33] H. Zhang, A. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *CVPR*, 2006. 2, 6

[34] X. Zhu, J. Lafferty, and Z. Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *Workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining at ICML*, 2003. 1, 2