Nonparametric One-class Classification with Gaussian Processes

Michael Kemmler, Erik Rodner*, Esther-Sabrina Wacker, Joachim Denzler

Computer Vision Group, Friedrich Schiller University of Jena, Germany

http://www.inf-cv.uni-jena.de

Abstract

Detecting instances of unknown categories is an important task for a multitude of problems such as object recognition, event detection, and defect localization. This article investigates the use of Gaussian process (GP) priors for this area of research. Focusing on the task of one-class classification, we analyze different measures derived from GP regression and approximate GP classification. We also study important theoretical connections to other approaches and discuss their underlying assumptions.

Experiments are performed using a large number of datasets and different image kernel functions. Our findings show that our approaches can outperform the well-known support vector data description approach indicating the high potential of Gaussian processes for one-class classification. Furthermore, we show the suitability of our methods in the area of attribute prediction, defect localization, bacteria recognition, and background subtraction. These applications and experiments highlight the easy applicability of our method as well as its state-of-the-art performance compared to established methods.

Key words: one-class classification, novelty detection, kernel methods, Gaussian processes, visual object recognition

1. Introduction

Many machine learning tasks in real-world applications have to deal with a large set of examples from a single class (positive examples) and only few or zero learning examples from a counter class (negative examples). Learning a classifier in such situations is known as one-class classification (OCC), novelty detection, outlier detection and also strongly related to density estimation. These scenarios arise due to the difficulty of obtaining training examples for rare cases, such as images of defects in defect localization tasks [18] or data from non-healthy patients in medical applications [49]. In these cases, one-class classification (OCC) allows for describing the distribution of positive examples and to treat negative examples as outliers, which can be detected without explicitly learning their corresponding model. Another motivation to use OCC is the difficulty of describing a background or counter class. This problem of finding an appropriate unbiased set of representatives exists for example in the area of object detection or content based image retrieval [8, 27]. However, it is also a common problem in defect localization tasks, where using only a small number of defective examples likely leads to a strongly biased classifier.

Earlier work concentrates on density estimation with parametric generative models such as single normal distributions or Gaussian mixture models [51, 3, 36]. These methods often make assumptions about the nature of the underlying distribution. Kernel methods like one-class Support Vector Machines (1-SVM, Schölkopf et al. [43]) or the highly related Support Vector Data Description (SVDD, Tax and Duin [53]), offer to circumvent such assumptions in the original space of feature vectors by using the kernel trick. These methods inherit provable generalization properties from learning theory [43] and can handle even infinite dimensional feature spaces.

*Corresponding author

Email addresses: michael.kemmler@uni-jena.de (Michael Kemmler), erik.rodner@uni-jena.de (Erik Rodner), esther.platzer@uni-jena.de (Esther-Sabrina Wacker), joachim.denzler@uni-jena.de (Joachim Denzler)

In this article, we propose OCC approaches that are based on Gaussian process (GP) priors. Machine learning with GP priors allows for formulating kernel-based learning in a Bayesian framework [38] and has proved to be competitive with SVM-based classifiers for binary and multi-class categorization of images [22]. Nevertheless, their use for OCC scenarios has mostly been studied in the case of proper density estimation [1], which requires sophisticated Markov chain monte carlo (MCMC) techniques to obtain a properly normalized density.

We derive several new OCC methods from the GP framework and show their theoretical connections to existing approaches. Furthermore, we investigate the suitability of approximate GP inference methods for one-class classification, such as Laplace approximation (LA) or expectation propagation (EP) [38], and analyze the influence of kernel hyperparameters on the resulting classification performance. The proposed approaches achieve state-of-the-art performance and are easy to implement (only a few lines in MATLAB). Our experimental analysis not only covers an in-depth analysis and comparison to previous work but also experiments with a wide range of applications. We apply our method to visual object recognition, attribute prediction, defect localization, bacteria recognition, and background subtraction in video sequences, which clearly demonstrates the suitability of our method for different kinds of datasets and task characteristics.

This article is based on our previous work on OCC with Gaussian processes [23]. In addition to including experimental findings from large-scale object categorization and other challenging applications (wire rope analysis [40], bacteria recognition [24], background subtraction), several new theoretical connections to related methods are drawn.

This paper is structured as follows: First, we briefly review previous work in the area of one-class classification in Sect. 2. This is followed by introducing the reader to the GP framework and its use for regression and classification in Sect. 3. Building on these fundamentals, Sect. 4 explains our one-class classification methods as well as their theoretical connections to previous work. An experimental analysis in Sect. 5 with image categorization tasks provides further insights into the methods behavior compared to related approaches. In Sect. 6, we show the applicability of our methods to a wide range of possible applications. A summary of our findings and a discussion of future research directions conclude the paper.

2. Short Overview of Related Work

In the following, we briefly review previous work done in the area of one-class classification. A detailed comparison of our approach to other methods is given in Sect. 4.

Over the past years, several approaches have been proposed for novelty detection. Trivially, density estimation techniques such as Parzen windowing [2] can be used to achieve this goal. Apart from this, many machine learning techniques have been adapted to the task of one-class classification. One popular strategy is to enclose the provided training data by, e.g., a hypersphere [52] or the convex hull [5], and to measure the distance to the estimated boundary. Especially the support vector data description approach of Tax and Duin [52], which is equivalent to the 1-SVM of Schölkopf et al. [43] for stationary kernels, achieves in many cases state-of-the-art performance. Raetsch et al. [37] showed how to translate 1-SVM into a boosting approach with a comparable performance. In contrast, Smola et al. [46] estimate novelty relative to another reference set, which is assumed to be given.

Subspace projection methods such as principal component analysis [51] and its kernelized counterpart [19], where the negative re-projection error is employed as a membership score, are also successfully used for one-class classification. Other methods directly make use of the local neighborhood structure of the data. Tax and Duin [50] estimate a measure of local density by computing nearest neighbor ratios. Alternatively, Juszczak et al. [21] propose to compute a minimum spanning tree of the data and use the distance to this tree structure as a novelty score. Other approaches for one-class classification include the information theoretic method of Filippone and Sanguinetti [15] as well as the boundary adaptation technique of Guo et al. [17].

The approach of Roth [41] uses a kernel fisher discriminant (KFD) classifier to perform one-class classification. We show that their approach is to a large part a special case of our framework. Kim and Lee [26] presents a clustering approach indirectly using a GP prior and the predictive variance. In contrast to Kim and Lee [26], we present several new novelty scores derived from the GP framework and analyze their differences.

For further literature on existing one-class classification techniques and their applications, we refer the interested reader to the PhD thesis of Tax [51] and the comprehensive survey of Chandola et al. [7].

When speaking of one-class classification, we will always refer to the case where absolutely no novelties are available at training time. This is in contrast to other approaches, where a certain amount of information is available,

e.g., unlabeled data [58] for training or a separate validation set containing novelties [9]. This restriction is common in a lot of applications, like for wire-rope defect detection (Sect. 6.1), where even data with a few unlabeled defects is hard to obtain for realistic settings.

3. Classification with Gaussian Processes

This section gives a brief introduction to GP classification. Since classification is motivated from non-parametric Bayesian regression, we first briefly introduce the regression case with real-valued outputs $y \in \mathbb{R}$, before we discuss approximate methods for GP classification with binary labels $y \in \{-1, 1\}$.

3.1. The Regression Case

The regression problem aims at finding a mapping from input space $\mathcal{X}$ to output space $\mathbb{R}$ using labeled training data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathcal{X}^n$, $\mathbf{y} = [y_1, \ldots, y_n]^T \in \mathbb{R}^n$. In the following, it is assumed that an output y is generated by a latent function $f : \mathcal{X} \to \mathbb{R}$ and additive noise ε, i.e., $y = f(\mathbf{x}) + \varepsilon$. Rather than restricting f to a certain parametric family of functions, we only assume that the function is drawn from a specific probability distribution $p(\mathbf{f}|\mathbf{X})$. This allows for a Bayesian treatment of our problem, i.e., we infer the probability of output y_* given a new input x_* and old observations $[\mathbf{X}, \mathbf{y}]$ by integrating out the corresponding non-observed function values $f_* = f(\mathbf{x}_*)$ and $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^T$:

$$p(y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) = \int_{\mathbb{R}} p(f_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)\, p(y_*|f_*)\, \mathrm{d}f_* \ , \tag{1}$$

$$p(f_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) = \int_{\mathbb{R}^n} p(f_*|\mathbf{X}, \mathbf{f}, \mathbf{x}_*)\, p(\mathbf{f}|\mathbf{X}, \mathbf{y})\, \mathrm{d}\mathbf{f} \ . \tag{2}$$

The central assumption in GP regression is a Gaussian process prior over latent functions f, which we write by $f \sim \mathcal{GP}(m, \kappa)$. A Gaussian process can be thought of as a generalization of multivariate Gaussian distributions to infinite dimensionality. The latent function f is said to be distributed according to a Gaussian process, if and only if every finite subset of function values is jointly normally distributed. Therefore, the function values $\mathbf{f}$ obey the following model:

$$\mathbf{f}|\mathbf{X} \sim \mathcal{N}(m(\mathbf{X}), \kappa(\mathbf{X}, \mathbf{X})) \tag{3}$$

This distribution is solely specified by the mean function $m(\cdot)$ and covariance function $\kappa(\cdot, \cdot)$. If we further assume that the additive noise ε is modeled by a zero-mean Gaussian distribution, i.e.,

$$p(y|f(\mathbf{x})) = \mathcal{N}(y|f(\mathbf{x}), \sigma_n^2) \ , \tag{4}$$

we are able to solve the integrals in closed form. This is derived in detail in Rasmussen and Williams [38] and in this article we only present the result of this derivation. Using a zero-mean GP, the predictive distribution (2) is again Gaussian [38] with moments

$$\mu_* = \mathbf{k}_*^T \left(\mathbf{K} + \sigma_n^2 \cdot \mathbf{I}\right)^{-1} \mathbf{y} \tag{5}$$

$$\sigma_*^2 = k_{**} - \mathbf{k}_*^T \left(\mathbf{K} + \sigma_n^2 \cdot \mathbf{I}\right)^{-1} \mathbf{k}_* \tag{6}$$

using abbreviations $\mathbf{K} = \kappa(\mathbf{X}, \mathbf{X})$, $\mathbf{k}_* = \kappa(\mathbf{X}, \mathbf{x}_*)$, and $k_{**} = \kappa(\mathbf{x}_*, \mathbf{x}_*)$. Furthermore, we use $\mathbf{I}$ to denote the $n \times n$ identity matrix. Since we assume i.i.d. Gaussian noise, this also implies that $y_* = f_* + \varepsilon$ as a sum of independent Gaussian random variables is normally distributed with mean μ_* and variance $\sigma_*^2 + \sigma_n^2$.

3.2. From Regression to Classification

The goal in binary GP classification is to model a function predicting a confidence for each class $y \in \{-1, 1\}$, given a feature vector $\mathbf{x}$. In order to make probabilistic inference about the output given a training set, we can directly apply the Bayesian formalism from eq. (1) and (2). However, the key problem is that the assumption of Gaussian noise no

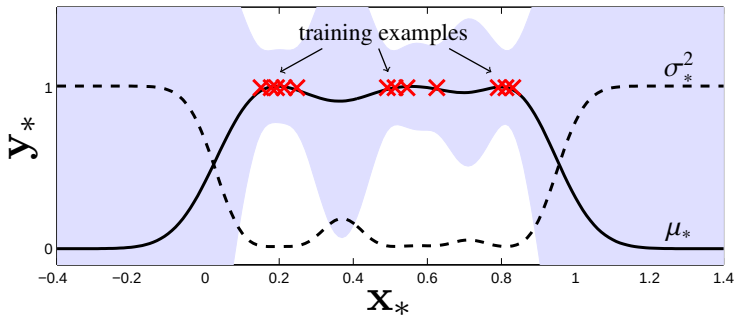

Figure 1: GP regression using a zero-mean GP prior in a one-dimensional OCC setting. The predictive distribution is visualized via its mean and corresponding confidence interval (scaled variances), where training points are marked as crosses.

longer holds, since the output space is discrete. We could either ignore this issue and perform regression on our labels, or we could use a more appropriate likelihood such as the cumulative Gaussian

$$p(y|f) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{yf} \exp\left(-\frac{1}{2}z^2\right) \mathrm{d}z \tag{7}$$

The disadvantage of the latter procedure is that our predictive distribution (2) is no longer a normal distribution. To overcome this issue, we follow the standard approach to approximate the posterior $p(\mathbf{f}|\mathbf{X}, \mathbf{y})$ with a normal distribution $\hat{p}(\mathbf{f}|\mathbf{X}, \mathbf{y})$. Two well-known approaches, which are also used in this work, are Laplace approximation (LA) and expectation propagation (EP). The interested reader is referred to Rasmussen and Williams [38].

For the final prediction step, approximations $\hat{p}(\mathbf{f}|\mathbf{X}, \mathbf{y})$ are used to solve (1). Using both Gaussian approximations to the posterior (2) and cumulative Gaussian likelihoods $p(y|f)$, it can be shown that the predictive distribution (1) is also equal to a cumulative Gaussian and can thus be evaluated in closed form [38].

4. One-class Classification with Gaussian Process Priors

In this section, we derive one-class scores from the predictive distribution of Gaussian process regression, show connections to existing machine learning methods and briefly discuss the issue of automatic hyperparameter tuning.

4.1. From the Predictive Distribution to One-class Scores

On a first glance when considering GP techniques for one-class classification, we are faced with two problems: First, GP approaches are discriminative techniques to tackle the problem of classification [38]. This follows from eq. (1) where the conditional density $p(y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)$ is modeled. Using discriminative classification techniques directly for one-class classification is a non-trivial task, due to the fact that the density of the input data is not taken into account. A second problem is that applying regression technique directly to labels $y = 1$, mostly results in a constant regression function, because the solution fits to the data and has a low model complexity.

Nevertheless, utilizing a properly chosen Gaussian process prior enables us to derive useful membership scores for OCC in a very intuitive manner. The main idea is to use a mean of the prior with a smaller value than our positive class labels (e.g., $y = 1$), such as a zero mean. This restricts the space of probable latent functions to functions with values gradually decreasing when being far away from observed points. In combination with choosing a smooth covariance function, an important subset of latent functions is obtained which can be employed for OCC (see Figure 1).

This highlights that the predictive probability $p(y_* = 1|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)$ (**Method P**) can be utilized, in spite of being a discriminative model. Due to the fact that the predictive probability is solely described by its first and second order moments, it is natural to also investigate the power of predictive mean (**Method M**) and variance as alternative membership scores. Their suitability is illustrated in Figure 1: The mean decreases for inputs distant from the training data and can be directly utilized as an OCC measure. Due to the constant labels $y = 1$, the formula simplifies to:

$$\mu_* = \mathbf{k}_*^T \left(\boldsymbol{K} + \sigma_n^2 \boldsymbol{I}\right)^{-1} \mathbf{1} \tag{8}$$

Table 1: Different measures derived from the predictive distribution, which are suitable for OCC membership scores as explained in Sect. 4.1. The letter enclosed in brackets is used for abbreviations in combination with the inference methods used (GP-Reg (label regression), Laplace approximation (GP-LA), expectation propagation (GP-EP)). For example, the predictive mean together with the GP regression method is denoted as GP-Reg-M.

	Method	Formula
M	Mean (M)	$\mu_* = \mathcal{E}(y_*\|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)$
V	neg. Variance (V)	$-\sigma_*^2 = -\mathcal{V}(y_*\|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)$
P	Probability (P)	$p(y_* = 1\|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)$
H	Heuristic (H)	$\mu_* \cdot \sigma_*^{-1}$

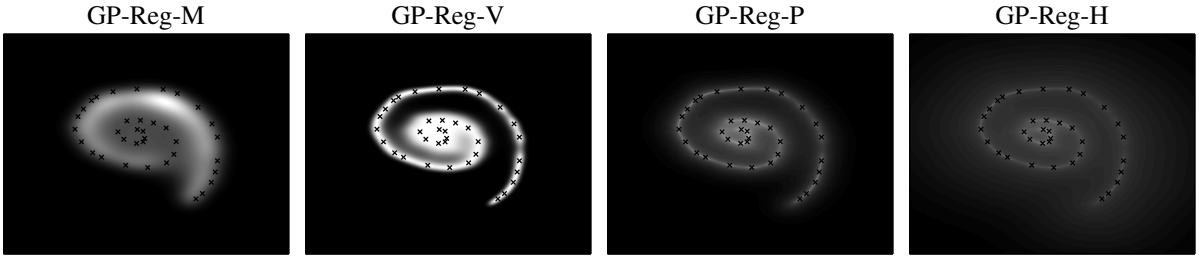

Figure 2: One-class classification using GP regression (GP-Reg) and measures listed in Table 1. All measures capture the distribution quite well.

where $\mathbf{1}$ denotes the n-dimensional vector containing only the value 1. In contrast to the predictive mean, the predictive variance σ_*^2 is increasing, which suggests that the negative variance value can serve as an alternative criterion for OCC (**Method V**). The latter concept is used in the context of clustering by Kim and Lee [26]. Additionally, Kapoor et al. [22] propose the predictive mean divided by the standard deviation as a combined measure for describing the uncertainty of the estimation (**Method H**). They applied this heuristic successfully in the field of active learning.

All variants, which are summarized in Table 1, are available for GP regression (GP-Reg) and approximate GP classification with Laplace (GP-LA) approximation or expectation propagation (GP-EP). Note that Table 1 also contains the abbreviation used in this paper for each of the methods presented. The different membership scores produced by the proposed measures are visualized in Figure 2 using an artificial two-dimensional example.

In the following sections, we additionally motivate the use of the mean and variance of GP regression by highlighting the strong relationship to Parzen estimation, normal density distributions, and several feature space interpretations.

4.2. Connections to other Methods

Relation to Unnormalized Density Estimation. The suitability of the predictive mean estimated by Gaussian process regression (GP-Reg-M) for OCC can be most easily demonstrated when using the exponential kernel. In the special case $\mathbf{X} = \mathbf{x}$, i.e., when the training set only contains a single example, the predictive mean score can be written as

$$\mu_* = \frac{1}{1+\sigma_n^2} \exp\left(-\|\mathbf{x}_* - \mathbf{x}\|^2/(2\sigma^2)\right) \propto \mathcal{N}(\mathbf{x}_* \mid \mathbf{x}, \sigma^2) \tag{9}$$

with noise variance σ_n^2 and hyperparameter σ, and can be thus interpreted as an unnormalized normal distribution centered on $\mathbf{x}$.

A more general connection, without the restriction to the exponential kernel and with an arbitrary number of learning examples, can be derived by rewriting the predictive mean in the following manner:

$$\mu_* = \mathbf{k}_*^T \left(\boldsymbol{K} + \sigma_n^2 \boldsymbol{I}\right)^{-1} \mathbf{1} = \sum_{i=1}^{n} \left[\sum_{j=1}^{n} \left(\boldsymbol{K} + \sigma_n^2 \boldsymbol{I}\right)_{ij}^{-1}\right] \kappa(\mathbf{x}_i, \mathbf{x}_*) \ , \tag{10}$$

which bears close resemblance with Parzen density estimation. In comparison to Parzen windowing, our predictive mean is unnormalized and additionally provides a scaling of similarities $\kappa(\mathbf{x}_i, \mathbf{x})$ between test and training examples

based on the kernel matrix $\boldsymbol{K}$ and the assumed noise level σ_n^2. For $\boldsymbol{K} = \boldsymbol{I}$ and $\sigma_n = 0$, unnormalized Parzen windowing is obtained. This shows that Parzen density estimation is a special case of our predictive mean approach, which assumes that there are no correlations within the training set. Therefore, our approach is more flexible and allows for easy incorporation of the existing correlations in the training set by modeling it with the same kernel function already used for the correlations with a new test example.

Scaled Correlation in Kernel Feature Space. A different view on the Gaussian process mean can be derived from a geometrical view in the reproducing kernel Hilbert space associated to covariance function $\kappa(\cdot,\cdot)$. Let $\Phi(\cdot) : \mathcal{X} \to \mathcal{H}$ be the mapping from input space to the space $\mathcal{H}$ which is equipped with the inner product $\kappa(\mathbf{x},\mathbf{x}') = \Phi(\mathbf{x})^T\Phi(\mathbf{x})$. Let further $\boldsymbol{\Phi} = [\Phi(\mathbf{x}_1),\ldots,\Phi(\mathbf{x}_n)]$ denote the matrix of mapped input vectors and define $\boldsymbol{K} = \boldsymbol{\Phi}^T\boldsymbol{\Phi}$ and $\boldsymbol{C} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T$ as the inner product and the (scaled) second moment matrix of $\boldsymbol{\Phi}$, respectively. Note that $\boldsymbol{C}$ is defined over mapped *data points* and should not be confused with $\boldsymbol{K}$, the covariance matrix between *latent function values* $\boldsymbol{f} = [f(\mathbf{x}_1),\ldots,f(\mathbf{x}_n)]^T$. Let us further define the regularized matrices $\boldsymbol{K}_{reg} = \boldsymbol{K} + \sigma_n^2\boldsymbol{I}$ and $\boldsymbol{C}_{reg} = \boldsymbol{C} + \sigma_n^2\boldsymbol{I}$. If both $\boldsymbol{K}_{reg}$ and $\boldsymbol{C}_{reg}$ are invertible, the following equivalence pointed out in Pękalska and Haasdonk [34] holds:

$$\boldsymbol{\Phi}\boldsymbol{K}_{reg}^{-1} = \boldsymbol{C}_{reg}^{-1}\boldsymbol{\Phi} \tag{11}$$

This property follows from noticing

$$\boldsymbol{\Phi}\boldsymbol{K}_{reg} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \sigma_n^2\boldsymbol{\Phi} = \boldsymbol{C}\boldsymbol{\Phi} + \sigma_n^2\boldsymbol{\Phi} = \boldsymbol{C}_{reg}\boldsymbol{\Phi} \tag{12}$$

and multiplying with $\boldsymbol{C}_{reg}^{-1}$ from the left and with $\boldsymbol{K}_{reg}^{-1}$ from the right side. The predictive mean can hence be rewritten as

$$\mu_* = \mathbf{k}_*^T\boldsymbol{K}_{reg}^{-1}\mathbf{1} = \Phi(\mathbf{x}_*)^T\boldsymbol{\Phi}\,\boldsymbol{K}_{reg}^{-1}\,\mathbf{1} = \Phi(\mathbf{x}_*)^T\boldsymbol{C}_{reg}^{-1}\boldsymbol{\Phi}\,\mathbf{1} \tag{13}$$

By further realizing that the data mean $\boldsymbol{\mu}_{\boldsymbol{\Phi}}$ in $\mathcal{H}$ is given by $\boldsymbol{\mu}_{\boldsymbol{\Phi}} = n^{-1}\boldsymbol{\Phi}\,\mathbf{1} = n^{-1}\sum_{i=1}^n \Phi(\mathbf{x}_i)$, we can reformulate the predictive mean as

$$\mathbf{k}_*^T\boldsymbol{K}_{reg}^{-1}\mathbf{1} \propto \Phi(\mathbf{x}_*)^T\boldsymbol{C}_{reg}^{-1}\boldsymbol{\mu}_{\boldsymbol{\Phi}} \tag{14}$$

which is proportional (denoted with $\propto$) to a correlation of test data $\Phi(\mathbf{x}_*)$ and the data mean $\boldsymbol{\mu}_{\boldsymbol{\Phi}}$ in $\mathcal{H}$ scaled by the inverse regularized data second moment matrix $\boldsymbol{C}_{reg}^{-1}$.

What do we learn from this observation? The interesting fact is that although the predictive mean was derived from a probabilistic framework there is a clear geometrical interpretation when we go to feature space. The predictive mean indirectly uses the distance to the data mean in terms of normalized correlation. Whereas, this geometrical approach can be also applied directly in the input space, it is the kernel trick which turns these simple methods into flexible non-linear ones. A very similar geometrical interpretation can also be derived for the predictive variance.

Normal Distribution in Feature Space. One approach to describe the data is to estimate a normal distribution in feature space $\mathcal{H}$ induced via a mapping $\Phi : \mathcal{X} \to \mathcal{H}$. It has been shown by Pękalska and Haasdonk [34] that computing the variance term in GP regression is equal to the Mahalanobis distance (to the data mean in feature space) if the regularized ($\sigma_n > 0$) kernel-induced scaling matrix $\boldsymbol{\Sigma} = \widetilde{\Phi}(\mathbf{X})\widetilde{\Phi}(\mathbf{X})^T + \sigma_n^2\mathbf{I}$ is used:

$$\widetilde{\Phi}(\mathbf{x})^T\boldsymbol{\Sigma}^{-1}\widetilde{\Phi}(\mathbf{x}) \;\propto\; \widetilde{\kappa}(\mathbf{x},\mathbf{x}) - \widetilde{\kappa}(\mathbf{x},\mathbf{X})\left(\widetilde{\kappa}(\mathbf{X},\mathbf{X}) + \sigma_n^2\mathbf{I}\right)^{-1}\widetilde{\kappa}(\mathbf{X},\mathbf{x}) \tag{15}$$

where the tilde indicates operations on zero-mean normalized data, i.e., $\widetilde{\Phi}(\mathbf{x}) = \Phi(\mathbf{x}) - n^{-1}\sum_{i=1}^n \Phi(\mathbf{x}_i)$ and $\widetilde{\kappa}(\mathbf{x},\mathbf{x}') = \widetilde{\Phi}(\mathbf{x})^T\widetilde{\Phi}(\mathbf{x}')$. Since the GP variance argument from (6) does not utilize centered kernel matrices, we effectively use the logarithm of the unnormalized zero-mean Gaussian which best describes the data.

Relation to One-class KFD. Our approach is also related to Roth [41], where a Fisher discriminant classifier is used. In fact, his main underlying assumption is a normal distribution in feature space, which results in the use of the Mahalanobis distance. This is equivalent to our predictive variance method as shown in the previous section for centered kernels. Instead of directly deriving the necessary calculations in terms of kernel values, their derivation takes a diversion and is motivated from kernel discriminant analysis. In contrast, our methods are directly derived from the GP framework, which allows for developing several different novelty scores.

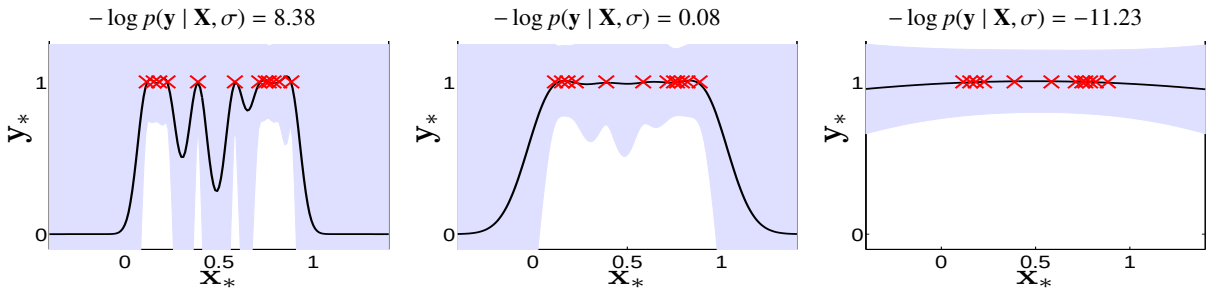

Figure 3: Gaussian process mean and variance scores using kernel $\kappa(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/(2\sigma^2))$, displayed along with negative log-likelihood values for a one-dimensional toy example. Kernel parameters which lead to smoother functions are preferred by the maximum marginal likelihood principle since both accurate and smooth functions are obtained. This is exemplified for hyperparameters $\log \sigma = \{-3, -2, +1\}$ from left to right in decreasing order of complexity.

Furthermore, Roth [41] uses the quantiles of the χ^2 distribution to select a suitable value for thresholding the novelty score. What we learn from the derived connection to Roth [41] is therefore that we can also apply this quantile technique in our case. Allowing for shifting the problem of adjusting the threshold for hard decisions to the more intuitive problem of tuning a confidence level.

Consistency Properties. In our experiments in Sect. 5.3, we observe that when using a fixed hyperparameter and an increasing number of training examples, the recognition performance can drop. At the first glance, this is counter-intuitive, because the one-class classification should be consistent and the estimated distribution should converge to the correct distribution with high probability for $n \to \infty$. Therefore, the question arises under which conditions, we can guarantee consistency.

Vert and Vert [56] show the consistency of a multitude of estimators with a similar underlying optimization problem as used in the one-class SVM formulation of Schölkopf et al. [43] equipped with a normalized exponential kernel:

$$\underset{f \in \mathcal{H}_\sigma}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^{n} \varphi(y_i \cdot f(\mathbf{x}_i)) + \lambda \|f\|^2_{\mathcal{H}_\sigma} \quad . \tag{16}$$

where φ is an arbitrary convex loss function. The feature space induced by a normalized exponential kernel with hyperparameter σ^2 [56, p. 2] is denoted by $\mathcal{H}_\sigma$ and plays an important role in their results. The consistency is only ensured if σ is decreasing for an increasing number of training examples. This condition is analogous to the consistency requirements of the Parzen estimator [44], where σ^2 is often referred to as bandwidth parameter. Optimization problem (16) is not only a generalization of the one-class SVM problem of Schölkopf et al. [43] but also a generalization of our one-class GP approach when using the predictive mean. This relationship can be seen by setting $\varphi(z) = (1 - z)^2$ and considering the alternative formulations of GP regression in [38, p. 144].

Therefore, the hyperparameter of the kernel function has to be set depending on the size of the training set. In case of the exponential kernel, the parameter has to decrease with increasing n to ensure consistency.

Relationship to Least Squares SVM. Other loss functions can also be integrated, e.g., quadratic loss instead of hinge loss, which leads to least-squares support vector machines (LS-SVM, [48]). It is interesting to note that LS-SVM with a zero bias term is equivalent to using the predictive mean estimated by GP regression. This can be seen when comparing the formulation in Rasmussen and Williams [38, p. 144] and the one in Suykens et al. [48]. Our extension to one-class classification problems, therefore, directly corresponds to the work of Choi [10] in this special case.

4.3. About the Difficulty of Tuning Hyperparameters

In principle, Gaussian process regression and classification enable an automatic way of tuning the kernel hyperparameters. By providing derivatives of the covariance function with respect to used hyperparameters, gradient-based optimization routines can be applied to maximize the marginal likelihood $p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta})$ or related objectives [38]. Although our one-class classification approach is based on Gaussian process regression, using this feature will not be

successful. This stems from the fact that only one class is available and the usual trade-off from regularized risk minimization practically no longer exists. In general, the overall goal in most machine learning approaches is to balance both complexity and accuracy on the training dataset. In the case of our Gaussian process based scores, a function $f(\mathbf{x}) = 1$ would give a perfect data fit while being extremely smooth. Solvers based on maximizing data likelihood hence run toward this extreme case (or more likely crash due to numerical instabilities of involved ill-posed kernel matrices). Figure 3 demonstrates this behavior for predictive mean and variance scores.

To summarize, tuning hyperparameters for one-classification tasks is a difficult task in a general setting without incorporating further model assumptions and should be done in an application-specific manner.

4.4. Computational Complexity

The main effort in computing our GP regression based one-class scores is associated to inverting kernel matrix $\boldsymbol{K}$. Instead of directly computing the inverse, a numerically more stable procedure is to solve the linear system $\boldsymbol{K}\boldsymbol{v} = \boldsymbol{w}$ with $\boldsymbol{w} = \mathbf{1}$ and $\boldsymbol{w} = \boldsymbol{k}_*$ for GP-Reg-M and GP-Reg-V, respectively. Since $\boldsymbol{K}$ is positive definite by definition, its Cholesky decomposition $\boldsymbol{K} = \boldsymbol{L}\boldsymbol{L}^T$ can be used to solve the easier problems $\boldsymbol{L}\boldsymbol{u} = \boldsymbol{w}$ and $\boldsymbol{L}^T\boldsymbol{v} = \boldsymbol{u}$ via forward and backward substitution. For both one-class scores, training time is hence bound by $O(n^3)$ due to calculating the Cholesky decomposition, where n denotes the number of training examples. For GP-Reg-M, the test time amounts to $O(n)$, given by the evaluation of inner product $\mu_* = \boldsymbol{k}_*^T\boldsymbol{v}$. On the other hand, GP-Reg-V scales quadratically in n, since the proxy $\boldsymbol{v}$ in $\sigma_*^2 = k_{**} - \boldsymbol{k}_*^T\boldsymbol{v}$ needs to be computed for each test example. Several speed-ups are possible but go beyond the scope of the paper. A comparison of standard methods is given in Chalupka et al. [6] and a method that exploits special kernel functions and can directly be used also for OCC has been very recently presented in Rodner et al. [39].

The computational complexity is hence similar to the support vector data description approach of Tax and Duin [52], which has an asymptotic of $O(m^2n)$ and $O(m)$ for training and testing, respectively, where m denotes the number of support vectors. The number of support vectors heavily depends on the hyperparameters of the kernel function as shown in Tax and Duin [52].

5. Experimental Analysis for Visual Object Recognition Tasks

In this section, we empirically analyze the proposed approach and its variants for visual object recognition tasks, which results in the following main outcomes:

1. For *Caltech 101*, a medium size object recognition database, OCC with the variance criterion estimated by GP regression (GP-Reg-V) is significantly better than all other methods using the color image kernels and it outperforms SVDD for various values of the outlier ratio ν (Sect. 5.2).
2. Approximate GP classification with LA and EP does not lead to a better OCC performance (Sect. 5.2).
3. The performance of the mean of GP regression (GP-Reg-M) varies for different categories and can even decrease with an increasing amount of training data (Sect. 5.3).
4. Parameterized image kernels offer additional performance boosts with the disadvantage of additional parameter tuning (Sect. 5.4).
5. Our OCC methods show comparable performance to state-of-the-art when evaluated on *ImageNet*, a large-scale dataset for object recognition (Sect. 5.5)

5.1. Experimental Setup and Image Kernel Functions

Our evaluation of one-class classification with GP priors is based on binary image categorization problems. To solve these problems, we utilize image-based kernel functions (image kernels) that rely on histogram representations of the image. The first image kernel used is the pyramid of oriented gradients (PHoG) kernel presented by Bosch et al. [4], which is based on gray-scale images only. The PHoG kernel computes histograms of gradient orientations in different parts of the image. The combination of the histograms is then done by utilizing a weighted exponential χ^2-kernel.

The other image kernel, which we refer to as color kernel, uses the bag-of-features approach [30]. Each image is represented as a set of local OpponentSIFT features (SIFT features for each of the channels in the opponent color

Table 2: Mean AUC Performance of OCC methods, averaged over all 101 classes. Bold font is used when all remaining measures are significantly outperformed. GP measures significantly superior to SVDD_ν (with optimal ν) are denoted in italic font.

	GP-Reg-P	GP-Reg-M	GP-Reg-V	GP-Reg-H	GP-LA-P	GP-EP-P
PHoG	*0.696*	0.693	0.692	*0.696*	0.684	0.683
Color	*0.761*	0.736	**0.766**	*0.755*	*0.748*	*0.747*
	GP-LA-M	GP-EP-M	GP-LA-V	GP-EP-V	$\text{SVDD}_{0.5}$	$\text{SVDD}_{0.9}$
PHoG	0.684	0.683	0.686	0.685	0.690	0.685
Color	0.745	0.744	*0.758*	*0.757*	0.739	0.746

space as proposed by van de Sande et al. [54]) and all features are clustered during learning. For clustering, we adapt the supervised algorithm of Moosmann et al. [33], which learns a random forest for all local features and uses the leaf nodes to obtain a clustering of the input space. We extend this technique to cluster local features of a single class by selecting a completely random feature and its median value in each inner node of the trees in the random forest. Afterwards, the clustering is utilized to compute histograms, which can be directly used as single global image descriptors.

Given global image descriptors, one could easily apply kernel functions such as an exponential kernel. The disadvantage of this method is that these kernels do not incorporate the position of local features as an additional cue for the presence of an object. Therefore, we use the spatial pyramid match kernel of Lazebnik et al. [30] to incorporate coarse, absolute spatial information with a 2×2 grid. Note that these kernel functions are not parameterized, i.e., they do not depend on hyperparameters, in their original formulation. This restriction is discussed in Sect. 5.4.

For all experiments, except the analysis conducted in Sect. 5.5, the Caltech 101 database [14] is used, considering all available 101 object categories. As performance measure we utilize the area under the ROC curve (AUC), which is estimated by 50 random splits in training and testing data. In each case, a specific number of images from a selected object category serves as training examples. Testing data consists of the remaining images from the category and all images of the Caltech background category.

5.2. Evaluation of One-class Classification Methods

To assess the OCC performance, we use 15 randomly chosen examples for training. First of all, we average the AUC over all classes and random repetitions to yield a final performance summary for each OCC method. Based on this performance assessment scheme, we compare predictive probability (-P), mean (-M) and variance (-V) of GP regression (GP-Reg) and GP classification using Laplace Approximation (GP-LA) or expectation propagation (GP-EP), respectively. We additionally analyze the heuristic $\mu_* \cdot \sigma_*^{-1}$ for GP regression (GP-Reg-H) and compare with SVDD using outlier fraction $\nu \in \{0.1, 0.2, \ldots, 0.9\}$ (SVDD_ν). The results for the PHoG and the color kernel are displayed in Table 2, which, for the sake of readability, lists only best performing SVDD measures. Note that SVDD is equivalent to the 1-SVM approach of Schölkopf et al. [43] in case of kernels with constant $\kappa(\boldsymbol{x}, \boldsymbol{x})$, which is the case in our experiments due to normalized features. Furthermore, we also skip a comparison with the approach of [41], due to the near equivalence with our GP-Reg-V measure as elaborated in Sect. 4.2.

It can be immediately seen that PHoG features are significantly inferior to color features. Therefore, experiments in subsequent sections only deal with color-based image kernels. Although the average performance of all measures are quite similar, SVDD is significantly outperformed for all tested ν (t-test, $p \leq 0.025$) by at least two GP measures. The method of choice for our task is GP regression variance (GP-Reg-V), which significantly outperforms all other methods using color features. Employing PHoG based image kernels, Reg-V also achieves at least comparable performance to SVDD for any tested parameter ν.

Our results also highlight that making inference with cumulative Gaussian likelihoods does not generally improve OCC, since LA and EP are consistently outperformed by GP regression measures GP-Reg-V and GP-Reg-P. Hence, the proposed OCC measures do not benefit from the noise model of (7) (and corresponding approximations) that are more suitable for classification from a theoretical perspective.

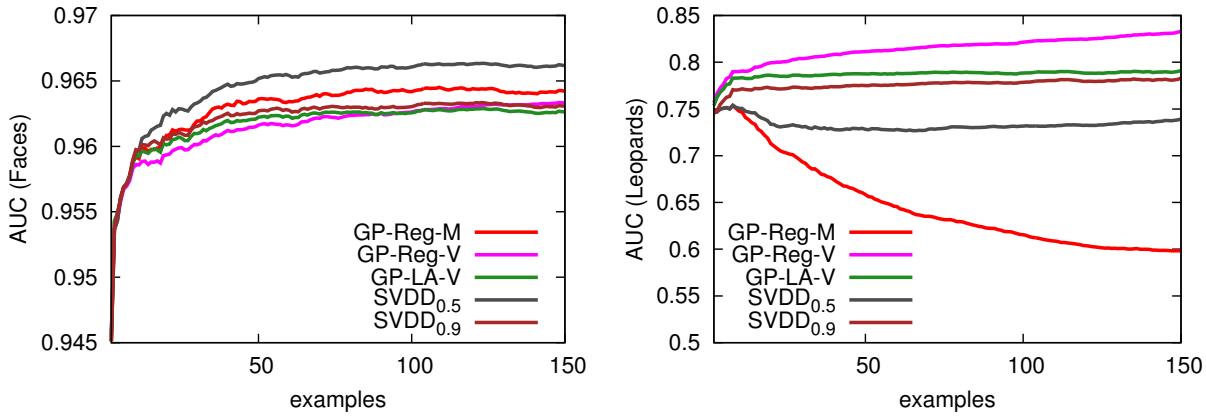

Figure 4: Results for color feature based image kernels regarding classes *Faces* (Left) and *Leopards* (Right) with varying number of training examples (using same legend).

5.3. Performance with an Increasing Number of Training Examples

To obtain an asymptotic performance behavior of all outlier detection methods, we repeat the experiments of Sect. 5.2 with a densely varying number of training examples. As can be seen in Figure 4, the performance behavior highly depends on the class. Classifying *Faces*, the performance increases with a higher number of training examples in almost all cases. A totally different behavior, however, is observed for *Leopards*, where the averaged AUCs of GP-Reg-M substantially decrease when more than 8 training examples are used. For small ν, SVDD also exhibits this behavior in our experiments.

Note that the large dependency of the number n of training examples upon the performance can be related to the property pointed out in Sect. 4.2. Convergence to a sensible one-class model is only ensured if the influence of an example on its surrounding decreases with increasing n. Despite the fact that we do not use a parameterized exponential kernel, oversmoothing effects can easily occur. This suggests to use a parameterized kernel alternative to account for this fact, as is investigated in the subsequent section.

5.4. Influence of an Additional Smoothness Parameter

Estimating the correct smoothness of the predicted distribution is one of the major problems in one-class classification and density estimation. This smoothness is often controlled by a parameterized kernel, such as an exponential kernel with a given variance. In contrast, our used kernel functions are not parameterized and the decreasing performance of the GP-Reg-M method in the last experiment might be due to this inflexibility.

For further investigation, we parameterize our image kernel function by transforming it into a distance:

$$d(\boldsymbol{x}, \boldsymbol{x}') = \kappa(\boldsymbol{x}, \boldsymbol{x}) - 2\kappa(\boldsymbol{x}, \boldsymbol{x}') + \kappa(\boldsymbol{x}', \boldsymbol{x}') \ , \tag{17}$$

which is then plugged into a distance substitution kernel κ_β [55]:

$$\kappa_\beta(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\beta \cdot d(\boldsymbol{x}, \boldsymbol{x}')\right) = \exp\left(-\beta\left(\kappa(\boldsymbol{x}, \boldsymbol{x}) - 2\kappa(\boldsymbol{x}, \boldsymbol{x}') + \kappa(\boldsymbol{x}', \boldsymbol{x}')\right)\right) \ . \tag{18}$$

This technique was inspired by the exponential kernel, where the Euclidean distance is used together with the exponential function. We perform experiments with κ_β utilizing 100 training examples and a varying value of β. The results for the categories *Faces* and *Leopards* are plotted in Figure 5.

Let us first have a look on the right plot and the results for *Leopards*. With a small value of β, the performance is comparable to the unparameterized version (cf. Figure 4, right side). However, by increasing the parameter we achieve a performance above 0.9 and superior to other methods, such as GP-Reg-V. This behavior differs significantly from the influence of β on the performance of the task *Faces*, which decreases after a small maximum. Right after the displayed points, we ran into severe numerical problems in both settings due to small kernel values below double precision. We expect a similar gain in performance by tuning the scale parameter of the cumulative Gaussian noise

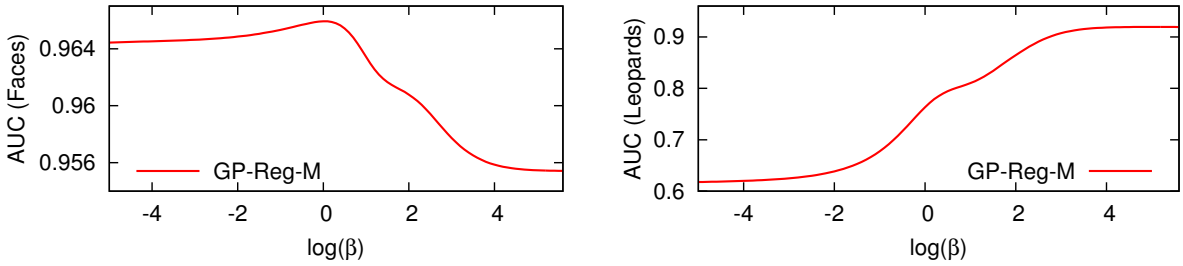

Figure 5: Influence of an additional smoothness parameter β of a re-parameterized image kernel on the OCC performance for the categories *Faces* (Left) and *Leopards* (Right).

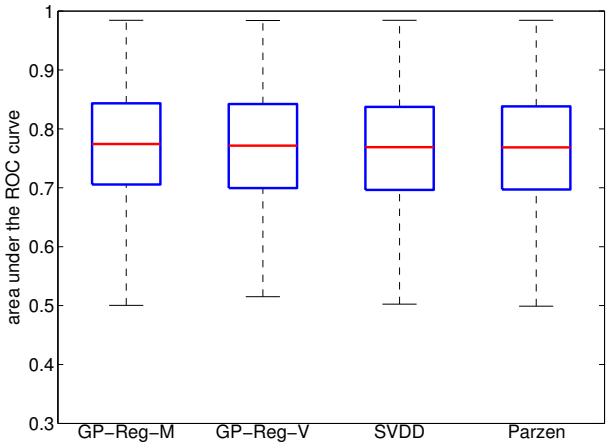

Figure 6: Results on the ImageNet dataset: averaged AUC values of 1 000 different one-class classification tasks.

model, but we skip this investigation to future research. This analysis shows that introducing an additional smoothness hyperparameter offers a great potential, though efficient optimization using the training set is yet unsolved.

5.5. Large-scale Evaluation on ImageNet

In the following, we evaluate our methods on the ImageNet database [11] and in particular the ILSRC'10 selection. This dataset contains 1 000 categories and 100 000 images for learning and 50 000 for testing[1]. Furthermore, we use the bag-of-features representation available for this dataset. One-class classification experiments are performed in the same manner as in previous experiments and the results of each of the 1 000 latent binary tasks are averaged. We use 100 examples and an exponential kernel $\kappa_{exp}(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/(2\sigma^2))$ with hyperparameter $\sigma^2 = \exp(-2)$ to learn with our OCC methods.

The final results are shown in Figure 6 for each method. As can be seen, our method achieves comparable performance to state-of-the-art methods. Using the Wilcoxon signed rank test, no significant difference was observed.

6. Further Applications beyond Visual Object Recognition

In the following section, we show the practical relevance of our work. There exists a great variety of OCC problems in many different application areas. Usually, there is not enough data available to build up a representation that covers all possible outcomes. There are two main problems, which result in OCC settings: (1) the collection of

[1] We use the public available validation dataset for testing.

training data lacks information for specific classes or (2) the full diversity of possible observations is not known in advance.

We investigate three typical but very different applications scenarios to prove the practical and broad applicability of our OCC methods: At first, we consider a visual inspection scenario and perform defect detection on wires ropes with help of our OCC techniques. In this case, only a small number of examples are available that show realistic defects, but a large set of non-defective data can be easily obtained. This is a very challenging application, as these defects are very small and inconspicuous. The application is one example for problem 1 mentioned above. Our second application, which corresponds to problem 2, is taken from microbiology. In this case, one often needs to distinguish between known microorganisms from the yet unknown without having any information about these novel observations. The last application we present is attribute prediction, which is useful for many object recognition problems and can be employed to be derive high-level features suitable for transfer learning [28] as well as object detection [13].

6.1. Wire-rope Defect Detection

We apply our OCC approach to wire rope defect detection. Given a lot of data of intact rope, the goal is to perform one-class classification to locate defects in the rope structure. We use two different rope datasets [35] acquired using a system of four line cameras. ROPE1 has a length of approximately 1.3km and ROPE2 is 400m long. The resolution of the line cameras is known to be 0.1 mm per camera line and each dataset was labeled by a human expert. Both rope datasets are different in nature with respect to the complexity of the surface defects. Whereas ROPE1 contains easily recognizable defects, defects contained in ROPE2 are often inconspicuous, small, and difficult to detect, even for a human expert. We compare the results obtained with GP-Reg-V, GP-Reg-M and SVDD to those previously published by Platzer et al. [36], which uses a Gaussian mixture model (GMM) with $m = 5$ latent Gaussians. All methods are trained on a defect-free rope region of 100 000 camera lines (10m rope, 5000 training examples). Evaluation is performed on the remaining rope sequence, which contains all labeled defects. As features we used gradient histograms, which were computed as described in Platzer et al. [36].

Note, that we do not utilize approximate inference techniques for GP classification, because they did not lead to a performance benefit for OCC tasks as observed in previous experiments. For our experiments, we use the exponential kernel κ_{exp} with a standard hyperparameter value of $\sigma^2 = \exp(-2.5)$. The outlier fraction of SVDD is set to 0.1, since other choices did not lead to significant changes of the result of SVDD.

Evaluation The results obtained for all methods are displayed in Figure 7 using ROC curves with the area under the ROC curve (AUC) given in the legend. Note, that the ROC curves are averaged over the results obtained for the four individual camera views.

It is obvious that all three kernel-based OCC approaches, GP-Reg-M, GP-Reg-V, and SVDD, clearly outperform the classical GMM strategy proposed by Platzer et al. [36]. Additionally, the AUC values suggest that the GP-based OCC approaches offer a slightly better performance than SVDD. GP-Reg-M achieves the best results and is also faster than GP-Reg-V during testing. Please note, that approaches which exploit the special structure of wire ropes achieve higher recognition results [35]. However, our OCC approach is not tailored to this special application scenario and can be applied to every defect localization task, as it does not depend on the features used.

6.2. Recognition of Novel Bacteria

In this application, we apply our OCC techniques to microbe identification based on Raman spectroscopy [42]. In this case, the large biodiversity of microorganisms prevents from building up a representative database, which covers all possible outcomes. The task of novelty detection can be easily cast as an OCC problem by identifying all known categories from the database as one positive super-class. Please also note that the negative class is always empty.

Our OCC approach is applied on a bacteria database containing 5 652 examples stemming from 50 different strains/classes. The microorganisms are described by a one-dimensional spectrum that covers the biochemical state of the sample at hand (see Figure 8). This response signal is sampled and quantized, which results in input vectors of fixed lengths.

For testing, an independent dataset comprising 130 spectra from 16 known and 169 spectra from 6 unknown strains is used. All spectra were pre-processed by using a median filter for cosmic spike elimination, baseline correction via iterative polynomial fitting [31], and unit length normalization.

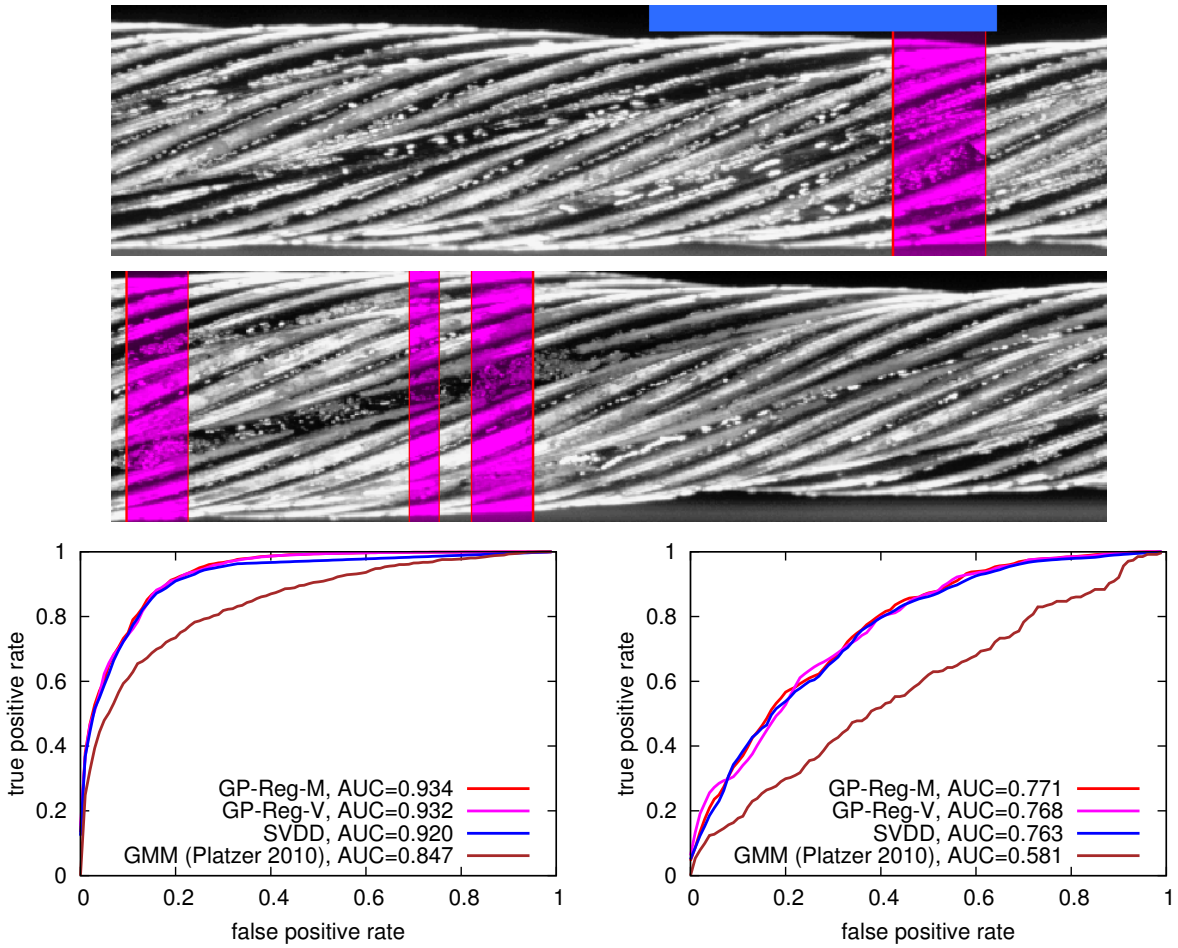

Figure 7: Wire rope defect detection: (left) example detections (ground-truth defect range is given in blue and our detections are highlighted with magenta), (center and right) average ROC curves rope datasets ROPE1 and ROPE2. The curves for all three kernel-based methods (GP-Reg-M, GP-Reg-V and SVDD) are very similar and best viewed in color.

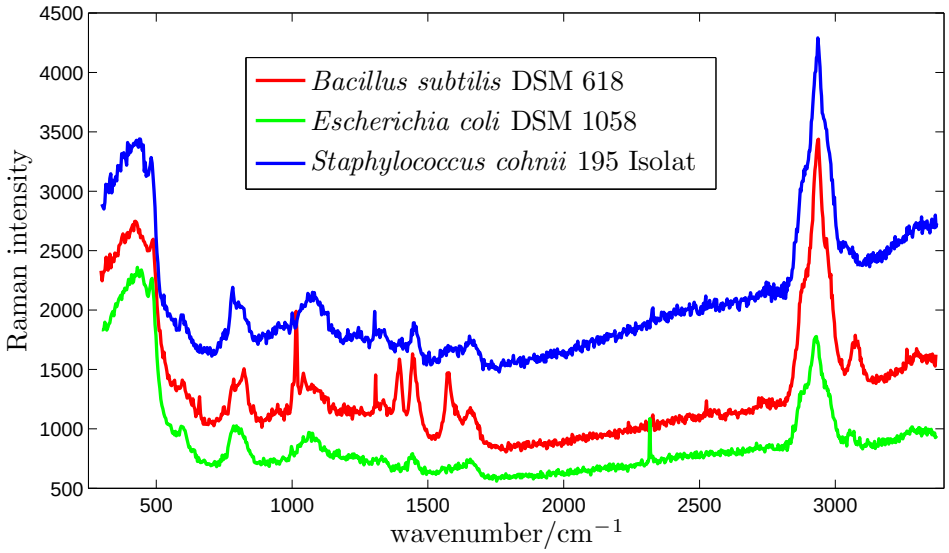

Figure 8: Example Raman spectra from three different bacterial strains. This figure is best viewed in color.

We compare our approach to GMMs [42], SVDD, and Parzen. For the kernel-based algorithms, the exponential kernel κ_{exp} was used. As pointed out in Sect. 4.3, hyperparameter tuning in an OCC settings is ill-posed. Therefore, we follow Lampert et al. [28] and set the length-scale parameter σ heuristically to the median of all pairwise distances between training points. In addition to Parzen density estimation using these settings, we also employ a normal distribution kernel with diagonal covariance matrix, i.e., $\kappa_{diag}(\mathbf{x}, \mathbf{x}') = \mathcal{N}(\mathbf{x}|\mathbf{x}', \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2))$, whose bandwidth parameters are tuned using *Silverman's rule of thumb* [45] separately in each dimension. For the latter approach and the GMM, the data is projected to the first $d = 30$ principal components for the sake of numerical stability.

Evaluation As is common practice in this domain, we use recognition rates based on sensitivity (true positive rate) and specificity (true negative rate) to assess all methods. To arrive at a crisp decision, threshold values were obtained using the 5-percentile of available training data scores. Note that this procedure artificially treats 5% of the training data as unknown, which is reminiscent of the outlier fraction $\nu = 0.05$ in SVDD.

Table 3 illustrates the performance of all tested methods. It is apparent that the decision boundary modeled by GP-Reg-V achieves the best trade-off between sensitivity and specificity with an average recognition rate of 76.5%. Both SVDD and GMM were tested for varying hyperparameters ($\nu \in \{0.1, \ldots, 0.9\}$, $m \in \{30, 100, 500, 1500\}$), where the best performances are displayed in Table 3. Both variants of Parzen density estimation failed to properly identify novel spectra, due to over- and underfitting effects.

A more closer look reveals that the low recognition rate of Parzen and SVDD is due to inappropriate hyperparameter σ estimated by the median heuristic. Except for GP-Reg-V, which is more robust to hyperparameter changes, the estimated hyperparameter generally seems to be unsuitable for the task of novelty detection. Figure 9 demonstrates this finding by varying the bandwidth parameter σ in a logarithmic domain, where additionally threshold-free AUC values are used for this analysis. It is clear that the estimated hyperparameter $\log \sigma = -1.8138$ is far away from an optimum for most of the tested OCC approaches. It becomes also apparent that a very good novelty detection performance for all methods can be achieved around $\log \sigma \approx -3.5$. However, this does not necessarily imply an improvement in terms of average recognition rate. This effect can be attributed to the interdependence between hyperparameter σ and the threshold-parameter used for realizing a hard novelty decision.

6.3. Attribute Prediction

In the last application, we apply our GP-Reg-V OCC method to the task of attribute prediction, which has shown to be useful for deriving high-level features for many visual recognition tasks [13, 28]. One possible scenario is the

Table 3: Results of novelty detection using a single positive class which comprises all training data.

method	specificity	sensitivity	ARR (%)
GPR-M	73 (43.2%)	108 (83.1%)	63.1
GPR-V	130 (76.9%)	99 (76.2%)	**76.5**
$\text{SVDD}_{0.1}$	28 (16.6%)	129 (99.2%)	57.9
Parzen	18 (10.7%)	127 (97.7%)	54.2
Parzen (κ_{diag})	169 (100.0%)	0 (0.0%)	50.0
GMM ($m = 100$)	59 (34.9%)	130 (100.0%)	67.5

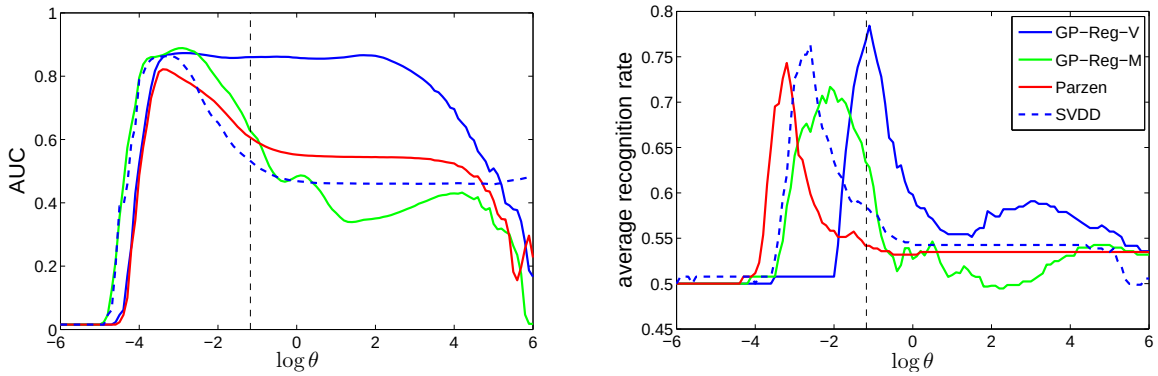

Figure 9: The effect of hyperparameter tuning upon kernel-based novelty detection methods using κ_{exp}. The estimated log-hyperparameter (vertical spaced line) $\log\sigma$ is generally unsuitable for most tested kernel-based techniques. Comparably good AUC values for all considered methods can be achieved at $\log\sigma \approx -3.5$. However, please note that this does not imply good performance in terms of average recognition results. This figure is best viewed in color.

estimation of a membership score for a specific (sub-)category. In our experiment, we used the Caltech database to estimate the membership score for the special sub-category *windsor chair*. We compared the results of GP-Reg-V against those obtained with SVDD. The results of GP-Reg-M are similar and therefore skipped in the following. We trained our GP methods and SVDD using 30 images of a type of chair called *windsor chair*, which has a characteristic wooden backrest. The performance is tested on all remaining *windsor chairs* and images of the category *chairs*. Results are illustrated in Figure 10 with the best and last ranked images. The qualitative results are similar, but the AUC values clearly show that GP-Reg-V is superior.

6.4. Background Subtraction

One-class classification can be also easily applied to background subtraction. To show this ability, we used image sequences provided by Elgammal et al. [12] and Monnet et al. [32] (similar to the setting used in Huang et al. [20]) and learned a GP-Reg-M classifier for each pixel by using normalized RGB values and the exponential kernel with fixed hyperparameters. For a given test image, each classifier gives us a novelty score, where a low value indicates that the pixel is likely not belonging to the background. To obtain a binary segmentation, which divides the image into foreground and background regions, we automatically determine a global threshold by analyzing the leave-one-out estimates [38, Section 5] in each pixel similar to the technique used in Sect. 6.2. Due to the large set of training examples, we use the 0.2%-percentile as a threshold.

Some example results can be seen in Figure 11. The results for Parzen and GMM have been obtained from Huang et al. [20]. Although we are not tuning anything to the task of background subtraction and only RGB values in each pixel are considered independently without any post-processing, the results are less noisy than the ones obtained with GMM and Parzen, as well as more accurate than the method of Monnet et al. [32]. For an efficient real-time application, the work of Rodner et al. [39] and Freytag et al. [16] can be considered, which allows a fast speed-up in the case of histogram intersection kernels and also fast incremental updates.

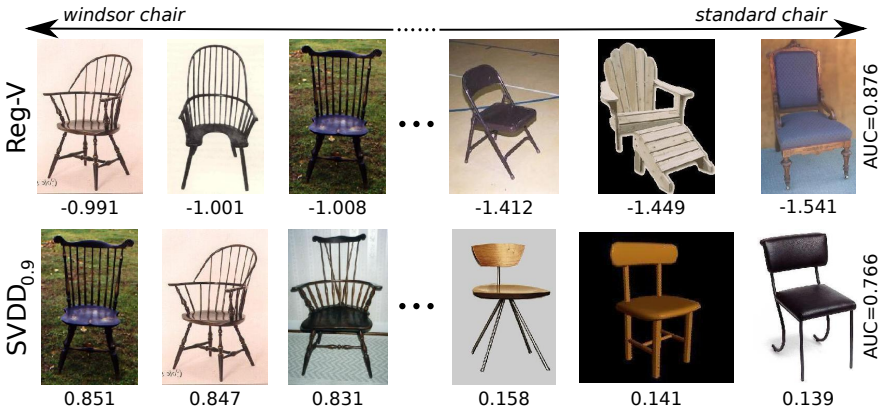

Figure 10: Results obtained by training different OCC methods for *windsor chair* and separating against *chair*: the three best ranked images (all of them are characteristic windsor chairs) and the three last ranked images with corresponding output values.

6.5. Summary of the Presented Applications

In order to properly evaluate and compare our proposed Gaussian process OCC scores, several challenging problems, ranging from object recognition, industrial computer vision, to bacteria recognition, and background subtraction, were studied. For all experiments, our GP based OCC methods GP-Reg-M and GP-Reg-V lead to results comparable or even superior to those obtained with state-of-the-art techniques such as SVDD or Parzen density estimation. This clearly shows that the derived OCC measures are applicable to a wide range of possible application scenarios. Our experiments also revealed that neither GP-Reg-M nor GP-Reg-V are to be preferred over the other in general, as their performances are highly application dependent. As becomes explicit in Figure 5 and Figure 9, this finding is tightly coupled with the inherent ill-posed problem of how hyperparameters are chosen.

7. Conclusions and Further Work

We have presented an approach for one-class classification (OCC) with Gaussian process (GP) priors and studied the suitability of different measures, such as mean and variance of the predictive distribution. The GP framework allows us to use different approximation methods to handle the underlying classification problem. From empirical evidence, Gaussian process label regression seems to be the method of choice. Moreover, based on the derived measures, relationships to several other well-known OCC methods are revealed. For example, it could be shown that the Parzen estimator is a specialization of one of our methods, which can be obtained when ignoring the correlations between training examples.

Our approach was tested on various novelty detection problems from challenging domains such as wire-rope failure detection and bacteria recognition. The obtained results clearly verified the suitability of Gaussian process based measures for the tasks at hand. While there is no clear winner for all experiments, Gaussian process regression based OCC scores always provided comparable or superior performances to other state-of-the-art techniques such as support vector data description or Parzen density estimation. Our work can be interesting for developers as well as machine learners. Whereas the latter group benefits from the connections we draw to other methods, developers can directly use our method without significant implementation overhead and use our analysis to see the effect of hyperparameters as well as the performance for a various set of applications.

Due to the ill-posed nature of the problem as a latent binary classification task (i.e., only training examples from a single class are provided), the choice of appropriate kernel hyperparameters still remains a problematic task. Automatic tuning based on maximizing the marginal training data likelihood, a standard model selection strategy for GP regression and classification, fails in the OCC setting due to the absence of a trade-off between functional fit and regularity. Future work should therefore concentrate on devising kernel hyperparameter tuning strategies that are applicable for one-class Gaussian process scores.

For convenience, we always considered homoscedastic Gaussian noise as the underlying noise process in our experiments. While heteroscedasticity is certainly more suited in many applications, existing work on heteroscedastic

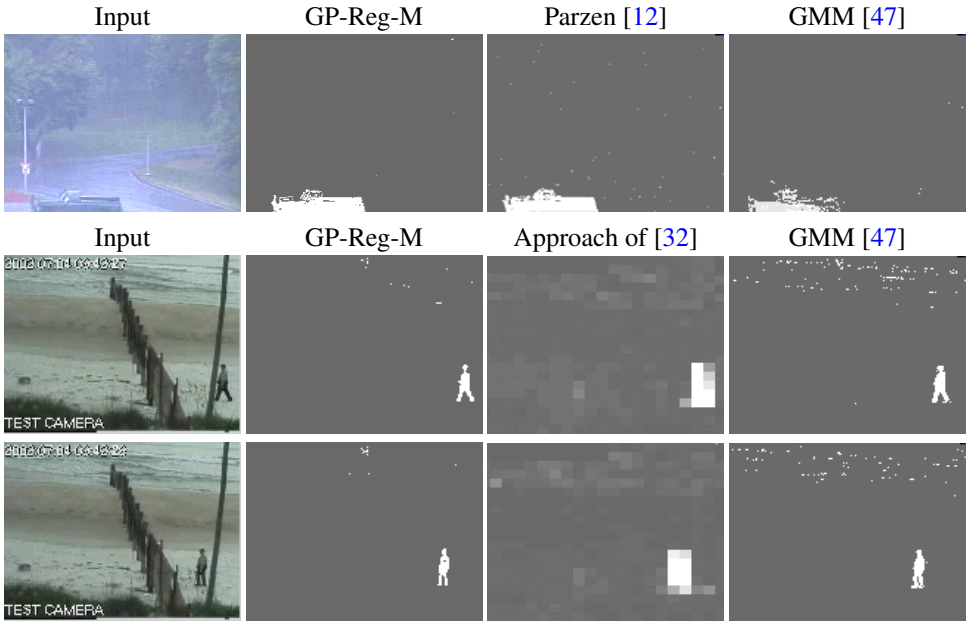

Figure 11: Background subtraction in video sequences using one-class classification: visual comparison of the obtained segmentations for the *RainCar* sequence of [12] and the *OceanPerson* sequence of [32].

Gaussian processes will not be applicable since they rely on automatic hyperparameter tuning strategies to infer the unknown variances [25, 29]. While we do not see a way out of this dilemma, we believe that domain-dependent knowledge may serve as a starting point to embed data-dependent uncertainty into the inference process. We also plan to apply our methods in the area of novelty detection for scene understanding [57].

Acknowledgements

We thank Mario Krause and Michaela Harz from the Institute of Physical Chemistry at the FSU Jena for capturing the microbe datasets. This work was partially funded by the TMBWK ProExzellenz project "MikroPlex" (PE113-1).

References

[1] Adams, R. P., Murray, I., MacKay, D., 2009. The gaussian process density sampler. In: Advances in Neural Information Processing Systems (NIPS). pp. 9–16.
[2] Bishop, C. M., 1994. Novelty detection and neural network validation. IEEE Proceedings on Vision, Image and Signal Processing. Special Issue on Applications of Neural Networks 141 (4), 217–222.
[3] Bishop, C. M., 2007. Pattern Recognition and Machine Learning (Information Science and Statistics), 1st Edition. Springer.
[4] Bosch, A., Zisserman, A., Munoz, X., 2007. Representing shape with a spatial pyramid kernel. In: Proceedings of the 6th ACM international conference on Image and video retrieval. pp. 401–408.
[5] Casale, P., Pujol, O., Radeva, P., 2011. Approximate convex hulls family for one-class classification. In: 10th International Workshop on Multiple Classifier Systems. pp. 106–115.
[6] Chalupka, K., Williams, C. K., Murray, I., 2013. A framework for evaluating approximation methods for gaussian process regression. Journal of Machine Learning Research 14, 333–350.
[7] Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection: A survey. ACM Computing Surveys 41, 15:1–15:58.
[8] Chen, Y., Zhou, X., Huang, T. S., 2001. One-class svm for learning in image retrieval. In: Proceedings of the IEEE Conference on Image Processing. pp. 34–37.
[9] Cheplygina, V., Tax, D., 2011. Pruned random subspace method for one-class classifiers. In: Sansone, C., Kittler, J., Roli, F. (Eds.), 10th International Workshop on Multiple Classifier Systems. Vol. 6713 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 96–105.
[10] Choi, Y.-S., 2009. Least squares one-class support vector machine. Pattern Recognition Letters 30 (13), 1236–1240.
[11] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09). pp. 248 – 255.

[12] Elgammal, A., Harwood, D., Davis, L., 2000. Non-parametric model for background subtraction. In: European Conference on Computer Vision (ECCV'00).
[13] Farhadi, A., Endres, I., Hoiem, D., 2010. Attribute-centric recognition for cross-category generalization. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10). pp. 2352 – 2359.
[14] Fei-Fei, L., Fergus, R., Perona, P., 2006. One-shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (4), 594–611.
[15] Filippone, M., Sanguinetti, G., 2010. Information theoretic novelty detection. Pattern Recognition 43 (3), 805 – 814.
[16] Freytag, A., Rodner, E., Bodesheim, P., Denzler, J., 2012. Rapid uncertainty computation with gaussian processes and histogram intersection kernels. In: Asian Conference on Computer Vision (ACCV). pp. 511–524.
[17] Guo, S., Chen, L., Tsai, J., 2009. A boundary method for outlier detection based on support vector domain description. Pattern Recognition 42 (1), 77 – 83.
[18] Hodge, V., Austin, J., 2004. A survey of outlier detection methodologies. Artificial Intelligence Review 22 (2), 85–126.
[19] Hoffmann, H., 2007. Kernel pca for novelty detection. Pattern Recognition 40 (3), 863 – 874.
[20] Huang, J., Huang, X., Metaxas, D., 2009. Learning with dynamic group sparsity. In: International Conference on Computer Vision (ICCV'09).
[21] Juszczak, P., Tax, D. M., Pękalska, E., Duin, R. P., 2009. Minimum spanning tree based one-class classifier. Neurocomputing 72 (7-9), 1859 – 1869.
[22] Kapoor, A., Grauman, K., Urtasun, R., Darrell, T., 2010. Gaussian processes for object categorization. International Journal of Computer Vision 88 (2), 169–188.
[23] Kemmler, M., Rodner, E., Denzler, J., 2010. One-class classification with gaussian processes. In: Proceedings of the 10th Asian Conference on Computer Vision. pp. 489–500.
[24] Kemmler, M., Rodner, E., Rösch, P., Popp, J., Denzler, J., 2013. Automatic identification of novel bacteria using raman spectroscopy and gaussian processes. Analytica Chminica Acta (submitted).
[25] Kersting, K., Plagemann, C., Pfaff, P., Burgard, W., 2007. Most likely heteroscedastic gaussian process regression. In: Proceedings of the 24th International Conference on Machine Learning. pp. 393–400.
[26] Kim, H.-C., Lee, J., 2006. Pseudo-density estimation for clustering with gaussian processes. In: Advances in Neural Networks - ISNN. Vol. 3971. pp. 1238–1243.
[27] Lai, C., Tax, D. M. J., Duin, R. P. W., Pękalska, E., Paclík, P., 2002. On combining one-class classifiers for image database retrieval. In: Proceedings of the Third International Workshop on Multiple Classifier Systems. pp. 212–221.
[28] Lampert, C. H., Nickisch, H., Harmeling, S., 2009. Learning to detect unseen object classes by between-class attribute transfer. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09). pp. 951–958.
[29] Lázaro-Gredilla, M., Titsias, M. K., 2011. Variational heteroscedastic gaussian process regression. In: Proceedings of the 28th International Conference on Machine Learning. pp. 841–848.
[30] Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the 2006 IEEE Conf. on Computer Vision and Pattern Recognition. pp. 2169–2178.
[31] Lieber, C. A., Mahadevan-Jansen, A., 2003. Automated method for subtraction of fluorescence from biological raman spectra. Applied Spectroscopy 57, 1363–1367.
[32] Monnet, A., Mittal, A., Paragios, N., Ramesh, Y., 2003. Background modeling and subtraction of dynamic scenes. In: International Conference on Computer Vision.
[33] Moosmann, F., Triggs, B., Jurie, F., 2006. Fast discriminative visual codebooks using randomized clustering forests. In: NIPS. pp. 985–992.
[34] Pękalska, E., Haasdonk, B., 2009. Kernel discriminant analysis for positive definite and indefinite kernels. IEEE Trans. Pattern Anal. Mach. Intell. 31 (6), 1017–1032.
[35] Platzer, E.-S., Denzler, J., 2011. Combining structure and appearance for anomaly detection in wire ropes. In: Proceedings of the International Conference on Computer Analysis of Images and Patterns. pp. 163–170.
[36] Platzer, E.-S., Se, H., Ngele, J., Wehking, K.-H., Denzler, J., 2010. On the suitability of different features for anomaly detection in wire ropes. In: Computer Vision, Imaging and Computer Graphics: Theory and Applications. Vol. 68. pp. 296–308.
[37] Raetsch, G., Mika, S., Scholkopf, B., Muller, K., 2002. Constructing boosting algorithms from svms: an application to one-class classification. Pattern Analysis and Machine Intelligence, IEEE Transactions on 24 (9), 1184–1199.
[38] Rasmussen, C. E., Williams, C. K. I., 2005. Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press.
[39] Rodner, E., Freytag, A., Bodesheim, P., Denzler, J., 2012. Large-scale gaussian process classification with flexible adaptive histogram kernels. In: European Conference on Computer Vision (ECCV). Vol. 4. pp. 85–98.
[40] Rodner, E., Wacker, E.-S., Kemmler, M., Denzler, J., 2011. One-class classification for anomaly detection in wire ropes with gaussian processes in a few lines of code. In: Proceedings of the 12th IAPR Conference on Machine Vision Applications.
[41] Roth, V., April 2006. Kernel fisher discriminants for outlier detection. Neural Computation 18, 942–960.
[42] Schmid, U., Rösch, P., Krause, M., Harz, M., Popp, J., Baumann, K., 2009. Gaussian mixture discriminant analysis for the single-cell differentiation of bacteria using micro-raman spectroscopy. Chemometrics and Intelligent Laboratory Systems 96 (2), 159 – 171, chimiometrie 2007, Lyon, France, 29-30 November 2007.
[43] Schölkopf, B., Platt, J. C., Shawe-Taylor, J. C., Smola, A. J., Williamson, R. C., 2001. Estimating the support of a high-dimensional distribution. Neural Comput. 13 (7), 1443–1471.
[44] Scott, D. W., Sain, S. R., 2004. Multi-Dimensional Density Estimation. Elsevier, Amsterdam, pp. 229–263.
[45] Silverman, B., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall/CRC.
[46] Smola, A., Song, L., Teo, C. H., 2009. Relative novelty detection. In: Twelfth international conference on artificial intelligence and statistics. Vol. 5. Citeseer, pp. 536–543.
[47] Stauffer, C., Grimson, W. E. L., 1999. Adaptive background mixture models for real-time tracking. In: Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on. Vol. 2. p. 252.

[48] Suykens, J. A. K., Gestel, T. V., Brabanter, J. D., Moor, B. D., Vandewalle, J., 2002. Least Squares Support Vector Machines. World Scientific Pub. Co.
[49] Tarassenko, L., Hayton, P., Cerneaz, N., Brady, M., 1995. Novelty detection for the identification of masses in mammograms. In: Fourth International Conference on Artificial Neural Networks. pp. 442–447.
[50] Tax, D. M., Duin, R. P., 2000. Data description in subspaces. In: Proceedings of the 15th International Conference on Pattern Recognition. Vol. 2. pp. 672–675.
[51] Tax, D. M. J., June 2001. One-class classification. Ph.D. thesis, Delft University of Technology.
[52] Tax, D. M. J., Duin, R. P. W., 1999. Data domain description using support vectors. In: European Symposium on Artificial Neural Networks. pp. 251–256.
[53] Tax, D. M. J., Duin, R. P. W., 2004. Support vector data description. Mach. Learn. 54 (1), 45–66.
[54] van de Sande, K. E. A., Gevers, T., Snoek, C. G. M., 2010. Evaluating color descriptors for object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. (in press).
[55] Vedaldi, A., Soatto, S., 2008. Relaxed matching kernels for robust image comparison. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition. pp. 1–8.
[56] Vert, R., Vert, J.-P., 2006. Consistency and convergence rates of one-class svms and related algorithms. Journal for Machine Learning Research 7, 817–854.
[57] Yong, S.-P., Deng, J. D., Purvis, M. K., 2012. Novelty detection in wildlife scenes through semantic context modelling. Pattern Recognition 45 (9), 3439 – 3450.
[58] Zhang, B., Zuo, W., 2008. Learning from positive and unlabeled examples: A survey. In: Proceedings of the International Symposiums on Information Processing. pp. 650–654.