Multi-Class Open Set Recognition Using Probability of Inclusion

Lalit P. Jain[1], Walter J. Scheirer[1,2], and Terrance E. Boult[1,3⋆]

[1]University of Colorado Colorado Springs
[2]Harvard University
[3]Securics, Inc.

Abstract. The perceived success of recent visual recognition approaches has largely been derived from their performance on classification tasks, where all possible classes are known at training time. But what about open set problems, where unknown classes appear at test time? Intuitively, if we could accurately model just the positive data for any known class without overfitting, we could reject the large set of unknown classes even under an assumption of incomplete class knowledge. In this paper, we formulate the problem as one of modeling positive training data at the decision boundary, where we can invoke the statistical extreme value theory. A new algorithm called the P_I-SVM is introduced for estimating the unnormalized posterior probability of class inclusion.

1 Introduction

Recent classification results reported for the ImageNet Large-Scale Visual Recognition Challenge [31,32] have captured the computer vision community's interest. With such low error rates (the top performing algorithm on the 2013 ImageNet challenge, a convolutional neural network, achieves an error rate of 11.1%), one might believe that we are closer to solving real-world visual object recognition than ever before. However, it is fair to ask if a scenario in which all classes are known during training time leads to an accurate assessment of the overall state of object recognition. Importantly, the *detection* results from the 2013 ImageNet challenge tell a different story. When unknown objects must be rejected in the process of detecting the location and label of a known object, no approach produces a result as impressive as what we see for classification: the best result is a mean average precision of just 22.6%. Detection falls under the general class of machine learning problems known as *open set recognition* [45], *i.e.* when the possibility of encountering novel classes not present in training exists during testing.

Emerging research that moves beyond typical binary models of positive/negative class association for open set recognition has examined the issues of detecting novel classes [16,5,4], rejecting outlier or unknown classes [24,57,2], and/or simultaneously detecting and recognizing known classes in the midst of unknown classes [45,11,14]. These approaches have been a good start, but they do not directly address the overarching problem: *multi-class open set recognition*, wherein models should account for multiple known classes as well as provide an option to detect novel classes or reject unknown

⋆ This work was supported in part by ONR MURI N00014-08-1-0638 and NSF IIS-1320956.

Pre-print of paper to appear in LNCS 8692 *ECCV 2014* (http://link.springer.com/chapter/10.1007/978-3-319-10578-9_26).
The original publication is available at http://www.springerlink.com.

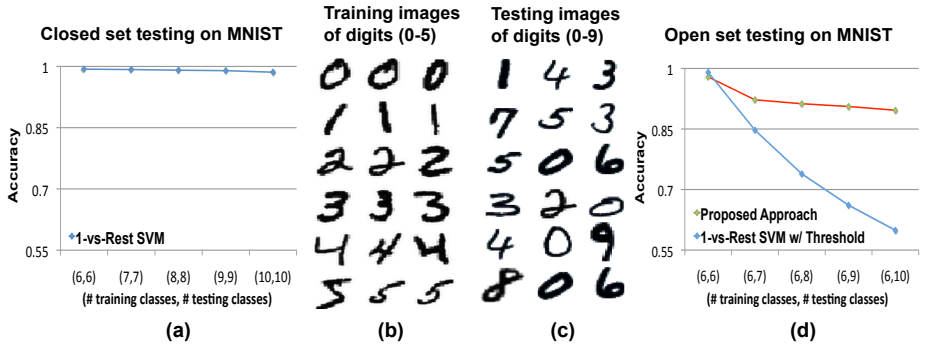


Fig. 1: What happens when the MNIST database of handwritten digits [33] is converted from a closed set classification task to an open set recognition task? (a) Standard supervised learning algorithms approach ceiling on the original MNIST classification problem. Here we show results for a 1-vs-Rest SVM with Platt [37] probability estimates using the same number of training and testing classes, where all classes are seen during training. (b) Training data consisting of six classes from MNIST. (c) Testing data consisting of all ten classes from MNIST, including four classes unseen during training. (d) By changing the testing regime to *cross-class-validation*, where some number of classes are held out during training (*e.g.* subfigure b) but included in testing (*e.g.* subfigure c), MNIST once again becomes a challenge. As soon as classes are withheld during training, the accuracy of the 1-vs-Rest SVM (with a rejection option provided by thresholding the Platt probability estimates) drops significantly. In this paper, we propose a new algorithm called the P_I-SVM, which retains higher levels of accuracy as the problem grows to be more open.

classes. In this paper, we introduce a new and effective algorithm for this task. Moreover, in contrast to popular detection challenges such as PASCAL VOC [19] where background classes have the same distribution in the training and test sets, the problem considered here assumes that completely new background classes can appear at test time.

Careful experimental design is necessary to evaluate multi-class open set recognition. Ideally, we would like to use well-known data sets. However, open set recognition requires an experimental regime that provides classes unseen during training – using all testing classes during training inflates performance on recognition problems. To address this, we can extend the familiar idea of cross-validation to *cross-class-validation*, wherein we simulate the unknown classes of an open set scenario by defining the number of training classes, target classes to recognize, classifiers, and validation classes, leaving some classes out during training while including them in testing. After moving to the open set multi-class recognition scenario, even what appear to be very simple "solved" pattern recognition tasks become quite difficult. To illustrate this point, Fig. 1 demonstrates how cross-class-validation transforms a classic closed set classification problem such as the MNIST database of handwritten digits [33] into a challenging multi-class open set recognition problem. In closed set classification (Fig. 1(a)), standard approaches such as multi-class 1-vs-Rest SVM achieve an average accuracy rate of approximately 98%. However, their accuracy drops significantly when using open set cross-class-validation testing (Fig. 1(d)).

An obvious way [30,23,57] to approach the multi-class open set recognition problem is to incorporate a posterior probability estimator $P(y|x)$, where $y \in \mathbb{N}$ is a class label and $x \in \mathbb{R}^d$ is multi-dimensional feature data, and a decision threshold into an existing multi-class algorithm. Letting $\mathcal{C}$ be the set of known classes, testing is a two step process: 1) compute the maximum probability over known classes, and 2) label the data as "unknown" if that probability is below the threshold δ:

$$y^* = \begin{cases} \text{argmax}_{y_i \in \mathcal{C}} \, P(y_i|x) & \text{if } P(y^*|x) \geq \delta. \\ \text{"unknown"} & \text{Otherwise} \end{cases} \tag{1}$$

Such a thresholded probability model can support a multi-class classifier with a rejection option, *e.g.* the 1-vs-Rest SVM with a threshold applied over Platt calibrated [37] decision scores as shown in Fig. 1(d). A key question when applying any probability estimator is: how do we build a consistent probability model without over-fitting or over-generalizing? While better than a strict multi-class SVM, which always assigns a known class label, SVM with a rejection option is still not very good for open set recognition. It is weak because it implicitly makes closed set assumptions during the decision score calibration process. In open set recognition, a sample that is *not* from a known negative class does not imply that it is from the positive class. Furthermore, because we must consider the possibility of unknown classes, Bayes' theorem does not directly apply. The best we can do is produce an unnormalized posterior estimate. In essence, we need a good way, in an open class setting, to consistently estimate the unnormalized posterior *probability of inclusion* for each class.

In this paper, we introduce the novel idea of fitting *a robust single-class probability model over the positive class scores from a discriminative binary classifier*. The use of an underlying binary classification model helps to discriminate the positive class from the known negative classes, while the single-class probability model adjusts the decision boundary so unknown classes are not frequently misclassified as belonging to the positive class. For consistency with open set assumptions, this model does not directly use negative data in its probabilistic modeling. Our algorithm, the P_I-SVM, follows this approach by modeling the unnormalized posterior probability of inclusion for multiple classes using a multi-class SVM as a basis, and fitting probability distributions consistent with the Statistical Extreme Value Theory (EVT) [12] to decision scores from positive training samples. This paper extends the recent statistical learning work of Scheirer et al. [43,42], which is limited to closed set problems. Our extension directly models the probability of inclusion for open set problems.

2 Related Work

The related work spans one-class classifiers, open set recognition, decision score calibration and probability-based rejection techniques for multi-class recognition. Of these topics, one-class classifiers, which require only positive training data, are a natural starting place for a solution to open set recognition. Density Estimation, Support Vector Data Description (SVDD), and the One-class SVM are all prevalent techniques used for one-class classification. A simple way to obtain a one-class model is to fit a Gaussian

distribution to the positive training data for a class and set a threshold on the resulting density [50]. A more sophisticated approach to accomplish the same goal is SVDD [49], where a hypersphere with the minimum radius is estimated around the positive class data that encompasses almost all training points. Using a different strategy, the training procedure for a one-class SVM [46] treats the origin in feature space as the only member of the second class, and maximizes the margin with respect to it. One-class models are typically less effective than binary classifiers [5,45,54].

More powerful binary classification models have been proposed specifically for open set visual recognition tasks. Scheirer et al. [45,41] offer a formalization of the risk of the unknown in open set recognition that is used to develop the 1-vs-Set Machine (a dual-plane linear classifier) [45] and W-SVM (a calibrated non-linear classifier) [41] algorithms. An approach similar to the 1-vs-Set Machine was described by Cevikalp and Triggs [8] for object detection. Unlike the approach we introduce here, none of these algorithms leverages a robust probability estimator for a single class that is derived from a binary classifier. Also related to the idea of unknown data is the "universum" [55], which constructs a data-dependent structure on the set of admissible functions by using a set of unlabeled training examples. However, the resulting model is still a traditional closed set binary classifier.

To estimate probabilities, various researchers [26,53,37,56,17,28,6] have proposed different techniques for converting a raw decision score to calibrated output. In all of these techniques a parametric model is assumed for the underlying distribution; parameters are estimated from calibration data and the raw scores mapped based on the resulting model. In practice, Gaussian modeling is common. The most widely used technique for score calibration is Platt's approach [37], which was originally proposed for SVM calibration, but has since been extended and evaluated on many types of learning systems [36]. In a cross-validation style training regime, a sigmoid function is fit to the decision scores from each fold, which is then used as a probability estimator for the overall classification model. Zadrozny and Elkan [56] note that "Platt scaling is most effective when the size of training/calibration data is small," which is potentially useful for open set recognition, where known negative class data in training is always smaller than the full domain of negative class data encountered during testing. Hybrid classifiers such as Naive Bayes Nearest Neighbor (NBNN) [3,34] can also provide probability estimation, but do so under a closed set assumption. It may be possible to adapt their estimates for an open setting, but we have not found an efficient means to do so.

In this paper, we propose using the Extreme Value Theory [12] for calibrating SVM decision scores to unnormalized posterior probabilities reflecting class inclusion. For recognition problems in computer vision, EVT has been demonstrated to be a powerful explanatory theory [43] and an effective tool for statistical modeling [7,22,21], including fitting probability estimators [44,42,41]. The most relevant work in EVT modeling is the multi-attribute spaces approach of Scheirer et al. [42], which applies EVT calibration over binary classifiers for visual attribute assignment. In essence, the algorithm estimates the probability of exclusion from the negative class. For example, in a gender classifier, the probability of being female is $1 - P(\text{male})$. This may be viable in a binary closed set problem, but is not an option for open set recognition. Moreover, Scheirer et al. do not describe how to estimate the critical "tail size" parameter.

Finally, using probabilities it is possible to reject "unlikely" samples (see Eq. 1), which can often improve our ability to reject unknown inputs. For multi-class recognition problems in computer vision, posterior probabilities are widely used to make decisions in applications such as pedestrian classification and orientation estimation [18], image retrieval [15], attribute fusion [42], part-based human tracking [47], large-scale multi-class object categorization [4], and activity recognition [39], among others. To operate in open set scenarios, a threshold for these algorithms can be set at a certain confidence interval to reject unknown classes. Chow [10] showed that the optimal decision rule is always a threshold over the posterior probability. Thus various score thresholds have been studied as rejection techniques, *e.g.* [30,23,57]. Recent prior work on thresholding [2,24] extends the notion of rejection via a threshold to the loss function of SVM to increase the cost of confusing samples. However, we note that in open set recognition the derivation of optimality in [10] does not hold, bringing the closed set modeling of all of these approaches into question for the general problem.

3 Single-Class Probability Estimation from a Binary Model

Intuitively, a one-class classifier such as the one-class SVM [46] seems like it should help us solve the open set recognition problem by providing a per-class model using just the positive data for each class. One-class classifiers do not assume a closed world, nor do they make any assumptions about negative or unknown classes. Unfortunately, it is precisely because they do not use any negative data that one-class classifiers have trouble enforcing separation between known positive and negative classes. The example in Fig. 2 highlights this issue.

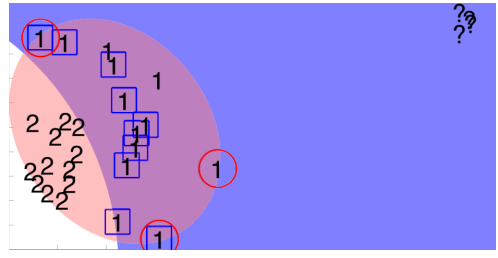


Fig. 2: Problems with existing models for two known classes ("1" and "2") when unknown classes ("?") are possible. If modeled by a one-class RBF SVM, the points with the red circles become support vectors defining the light red region as class 1. The model misclassifies most of class 2, but rejects unknowns. A binary RBF SVM separating 1 & 2 yields the blue region for class 1 with blue squares indicating positive support vectors. It can correctly classify class 1 and reject class 2, but it incorrectly classifies the unknowns with a cutting plane that over-extends rightward. SVM parameters in this example were optimized with 5-fold cross-validation grid-search.

To improve discrimination, binary classifiers such as RBF SVM use data from both positive and negative classes. But these models do not have an effective mechanism for rejecting classes; unknown classes must be classified as either positive or negative. In Fig. 2, a binary RBF SVM will misclassify all of the "?" points because there is no distinction between class 1 and the unknown classes in the region where they appear. Converting decision scores to probabilities using the estimation technique of Platt [37] could provide an indication of class membership for a test sample. For the example in Fig. 2, this means that an unknown test sample should receive a low probability score for association with either class 1 or class 2. However, this technique (like other

estimators [26,53,56,17,28,6]) assumes that all scores must be from a known positive or known negative class. Because of this, it is not very effective for open set recognition when combined with a threshold (see Sec. 6).

Thus we seek an approach for probability estimation that combines the ability to discriminate between known classes like a binary classifier, but with one-class-like rejection ability. To model this, the *probability of inclusion*, we only consider scores from the binary classifier that are associated with training data samples from a single class of interest in modeling. But what probabilty model should we use?

As shown by [1], for any $\zeta \in (1/2, 1)$ one can accurately estimate conditional probabilities in the interval $(1-\zeta, \zeta)$ only if support vectors are not sparse over that interval. For efficient classifiers we need some degree of sparsity, thus ζ should be close to $1/2$ and probability calibration is only well defined close to the decision boundary. And since that boundary is defined by the training samples that are effectively extremes, we conclude that proper models for efficient SVM calibration should be based on extreme value theory [25]. Different from previous calibration work for visual recognition [44,42,41] that has applied EVT via rejection of a hypothesis, we use EVT to directly model probability of inclusion P_I for a class of interest.

4 The P_I-SVM Algorithm

To begin, consider a kernelized SVM h that for any d-dimensional feature vector x will generate an uncalibrated hypothesis score s, which can be used to assign class membership:

$$h(x) = \sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + b \tag{2}$$

where α_i are support vectors, $K(x_i, x)$ is a radial basis function kernel, and b a bias term. A collection of such binary classifiers for each class $y \in \mathcal{C}$ forms a multi-class SVM. For an uncalibrated SVM hypothesis score $s = h_y(x), s > 0$ we assume larger scores imply more likely inclusion in class y. We also assume such inclusion scores are bounded from below, though it is straightforward to adapt the model to unbounded scores, depending on the desired EVT distribution for probability estimation.

We consider a multi-class problem for the known training classes $\mathcal{C}$ where we do not assume that the list of classes is exhaustive, *i.e.* at test time other classes may occur. The objective of the P_I-SVM is to compute a per class unnormalized posterior probability estimate for any input sample x. For P_I-SVM training, let $\{(x_1, y), (x_2, y), \ldots, (x_n, y)\}$ be a collection of training samples that will be used to fit a probability estimator for a single class. Let our overall match set be represented by $\mathcal{M}_y$, with $s_j \in \mathcal{M}_y$ if $h_y(x_j) > 0$. Let ℓ_y be the lower extremes of $\mathcal{M}_y$ (for SVM, the scores closest to 0).

After fitting, we can use the probability model of inclusion defined by a set of parameters θ_y as a robust probability estimator for a classifier. If we let $\rho(y)$ be the prior probability of class y, then we can estimate the posterior probability of inclusion P_I for the input x and class label y conditioned on the parameters θ_y as:

$$P_I(y|x, \theta_y) = \xi\rho(y)P_I(x|y, \theta_y) = \xi\rho(y)(1 - e^{-(\frac{x-\tau_y}{\lambda_y})^{\kappa_y}}) \tag{3}$$

Algorithm 1 Multi-class EVT-based Probability Modeling for P_I-SVM Training

Require: A set of class labels $\mathcal{C}$; a set of labeled training data points for each class $X_y, y \in \mathcal{C}$; a pre-trained 1-vs-Rest RBF SVM h_y for each class $y \in \mathcal{C}$, with positive support vectors α_y^+;

for $y = 1 \rightarrow |\mathcal{C}|$ **do**
  for $j = 1 \rightarrow |X_y|$ **do** ▷ Generate decision scores for fitting
    $s_{y,j} = h_y(x_{y,j})$
    if $s_{y,j} > 0$ **then**
      $\mathcal{M}_y = \mathcal{M}_y \cup \{s_{y,j}\}$
    end if
  end for
  $p_y = 1.5 * |\alpha_y^+|$ ▷ See Sec. 5 for an explanation of this step
  Sort $\mathcal{M}_y$ retaining p_y smallest items as vector ℓ_y
  $[\tau_y, \kappa_y, \lambda_y] = \text{wblfit}(\ell_y)$ ▷ Fit a Weibull distribution
  $\theta_y = [\tau_y, \kappa_y, \lambda_y]$
end for
return $\mathcal{W} = [\theta_1 \ldots \theta_{|\mathcal{C}|}]$ ▷ The result is a multi-class parameter set

for some constant ξ. If all classes and priors are known, then Bayes' Theorem yields

$$\xi = \frac{1}{\sum_{y \in \mathcal{C}} \rho(y) P_I(x|y, \theta_y)} \tag{4}$$

But we do not assume that all classes are known, so we let $\xi = 1$ and treat the posterior estimate as *unnormalized*. The use of unnormalized posterior estimation is well-known in computer vision [29,27,52,38,51], in part because as long as the missing normalization constant is consistent across all classes it still allows the use of maximum a posteriori estimation. Note that unnormalized posterior probabilities are always an approximation; the unknown constant ξ could be very large or very small which changes the true probability.

With a set of scores bounded from below, the correct EVT distribution to model ℓ_y is the Weibull [12]. The Weibull distribution has three parameters: location τ, shape κ, and scale λ (for details of the Weibull distribution, see Eq. 4 of [43]). For this work, we used the libMR library provided by the authors of [43], which uses Maximum Likelihood Estimation (MLE) to find the $\tau_y, \kappa_y, \lambda_y$ that best fit ℓ_y. In a multi-class setting, these three parameters are defined for each known class y, and we let θ_y represent the vector of those parameters. Alg. 1 provides a precise description of the Weibull probability modeling of class inclusion for each of the classes present during P_I-SVM training.

For multi-class open set recognition using a set of Weibull models we set a minimum threshold δ on class probability and select

$$y^* = \underset{y \in \mathcal{C}}{\text{argmax}}\, P_I(y|x, \theta_y) \quad \text{subject to} \quad P_I(y^*|x, \theta_{y^*}) \geq \delta. \tag{5}$$

The formulation in Eq. 5 yields the most likely class, which is appropriate if the classes are exclusive (as in our testing). Alternatively, if the classes are overlapping, one might return all classes above a given probability threshold. Note that we are dealing with

Algorithm 2 Multi-class Probability Estimation for P_I-SVM Testing

Require: A set of class labels $\mathcal{C}$; a pre-trained 1-vs-Rest RBF SVM h_y for each class y; parameter set $\mathcal{W}$ from Alg. 1; probability threshold δ for rejection; class prior probability $\rho(y)$ for each class y; a test sample x

$y^* =$ "unknown"
$\omega = 0$ ▷ Maximum probability score
for $y = 1 \rightarrow |\mathcal{C}|$ **do**
  $P_I(x|y, \theta_y) = \text{wblcdf}(x, \theta_y)$ ▷ The Weibull CDF provides the probability of inclusion
  $P_I(y|x, \theta_y) = \rho(y)P_I(x|y, \theta_y)$ ▷ Unnormalized posterior probability
  if $P_I(y|x, \theta_y) > \delta$ **then**
    if $P_I(y|x, \theta_y) > \omega$ **then**
      $y^* = y$
      $\omega = P_I(y|x, \theta_y)$
    end if
  end if
end for
return $[y^*, \omega]$ ▷ The result is the label and unnormalized posterior probability

unnormalized posterior estimations so the priors $\rho(y)$ only need to be relatively scaled, *e.g.* they could sum to one even if there are unknown classes. It has been shown [10] that the optimal value for the threshold is a function of the risk associated with making a correct decision, making an error, or making a rejection respectively, as well as the prior probabilities of the known or unknown classes. In practice, these would come from the domain knowledge. In our experiments, we assume equal priors per class; accordingly, we set δ to be the prior probability of an unknown instance. Alg. 2 describes the P_I-SVM probability estimation process for a new test sample. The estimate for probability of inclusion $P_I(x|y, \theta_y)$ comes from the CDF of the Weibull model defined by the parameters θ_y (see the right-hand side of Eq. 3).

Algs. 1 & 2 can also be adapted to estimate the unnormalized posterior probability of class inclusion for a one-class SVM. We use this method as a performance baseline in Sec. 6. In the one-class variant P_I-OSVM, we fit a Weibull distribution to the lower extrema of the positive decision scores estimated from an RBF kernel machine trained over just the positive data for a single class. The multi-class EVT-based probability modeling and multi-class probability estimation for P_I-OSVM use the same steps of Algs. 1 and 2 – the only necessary change is the replacement of the 1-vs-Rest binary SVM with a one-class SVM for each class. To our knowledge, this is the first purely one-class kernel machine probability estimator.

5 A Principled Approach to Estimating Tail Size for EVT Fitting

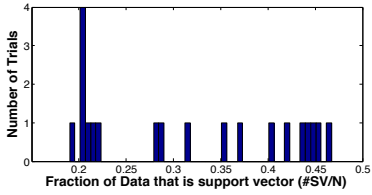
While EVT helps us model extrema, the theory tells us nothing about how many samples to use in fitting the EVT distribution. Prior work in EVT models for visual recognition [44,43,42,22,21] simply recommends choosing a tail size as an arbitrary percentage (not exceeding 50%) of the available data. How much data can be considered a proper

tail: 1%, 5%, 10%, 20%? We have found that the difference between a tail size of 5% and 20% of the data can produce a difference in recognition accuracy between 15-20%, with some models needing 5% and others needing 20% to achieve their best performance. Automatic estimation is a better strategy. One basic approach for estimating the tail size is to use cross-validation. However, our own experiments and those from others working in the area of financial modeling [40] have shown this approach to be unstable in practice. Why is this the case?

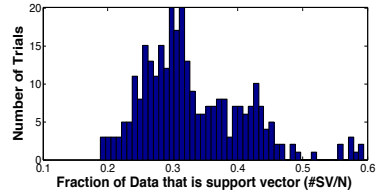
We believe the reason for this difficulty is that for visual learning, we apply EVT after mapping high-dimensional problems into one-dimensional scores. Inherent in our high-dimensional problems is a complex boundary where points can be near any part of it – there are many dimensions and directions in which points can appear as extrema. Reconsidering Fig. 2, what would the appropriate tail size be for modeling the score data from class 1? It depends not just on the training points but on the chosen classification model as well. When class 1 is modeled with a one-class SVM, the n-dimensional boundary is simpler (*i.e.* has fewer support vectors) than when it is modeled with a binary SVM. As the dimensionality of the data grows, the boundary can be far more complicated or it can be simple. With a complex dependency on dimensionality, sampling, and the problem, it should not be surprising that a tail defined by a fixed size or fixed percentage cannot easily predict how many points are on or near the boundary, or are extrema in general. We require a model that accounts for boundary complexity.

A useful insight is that, by construction, support vectors are a type of extreme sampling that effectively describes the class boundary. It is natural to ask if there is a known parametric relationship between training data size, dimensionality, and the number of support vectors. Unfortunately, there is not. Vapnik has shown [53] that the number of support vectors can be relatively independent of the number of training samples and dimensions. Subsequently, Steinward [48] has developed asymptotic sharp upper and lower bounds on the number of support vectors. For an RBF SVM with L_1 regularization, the fraction of data that are support vectors tends to be twice the Bayes risk. For an RBF SVM with L_2 regularization, the fraction of support vectors tends to be the probability of noise. In both cases, the fraction of data that are support vectors depends on a problem specific property that is not known *a priori* and which is difficult to estimate. These results reinforce the difficulty of estimating tail size based on the percentage of training data and/or dimensionality.

We are not, however, at a dead end. The above insights suggest a different approach: consider extrema to be those points close to the boundary in the original feature space and count them. Using this strategy a new optimization problem, similar to soft-margin SVM optimization, could be defined to locate and minimize the number of extreme training samples on either side of the boundary while minimizing a loss function related to our goal of probability estimation. However, we have found that the exact size of the tail can vary moderately and have only minimal impact on final multi-class recognition system performance. Thus defining and solving a new optimization just to estimate the tail size parameter is not warranted. Since we are applying the EVT model to calibrate an SVM classifier, and that classifier already has a well defined boundary, a much simpler alternative is to consider points within some distance ϵ of the SVM decision boundary as the potential extrema. For problems not modeled by a binary SVM, *e.g.* those with only



(a) SV fractions for first class folds.



(b) SV fractions for all class folds.

Fig. 3: What is the trouble with assuming a fixed tail size? Consider 20 random trials for the different classifiers in our LETTER [35] experiments (Sec. 6). The above plots show how many trials have a given ratio of support vectors in the training data. (a) shows the variation for just a single class, and (b) shows the variation over 15 classes. The overall distribution is broad and asymmetric – it is not consistent with a constant model implicit in assuming tail size is a fixed fraction; our approach in Eqs. 6 & 7 is different from assumptions made in prior EVT modeling.

one class of data, a one-class SVM can still provide such an estimate. Given an SVM decision function, we define an indicator function B^+ and the positive tail size T_ϵ^+ via:

$$B^+(x;\epsilon) = \begin{cases} 1 & \text{if } h(x) \leq \epsilon \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad T_\epsilon^+ = \sum_{x \in \mathcal{M}_y} B^+(x;\epsilon) \quad (6)$$

For a soft margin RBF SVM, which we use in this paper, support vectors include all points on or outside the positive-class region boundary as defined by the SVM decision function in Eq. 2. Thus T_0^+ is just the number of support vectors that belong to the positive class. For $\epsilon > 0$, some points inside the positive boundary would be included. An approximation that we have found to be both stable and effective is to use a small multiple of the number of support vectors from the positive class, thereby allowing a few points inside but near the class boundary. Letting $|\alpha^+|$ represent the number of support vectors from the positive class, we approximate the tail size via:

$$\hat{T}_\epsilon^+ = \max(3, \psi \times \ |\alpha^+|) \quad (7)$$

where we need ≥ 3 distinct points to ensure a well-defined EVT fitting. One free parameter ψ must be estimated. Empirically, we have found that any $\psi \in [1.25 - 2.5]$ works well. This range has provided relatively stable multi-class recognition across multiple problems. For the experiments presented in the next section, we fix $\psi = 1.5$.

To help illustrate the significant difference between a fixed fraction and our approach, Fig. 3 shows the variation in fraction of data that are support vectors for the different classifiers in the LETTER data set [35] considered in Sec. 6 below. The data consists of 15 classes over 20 random trials with a mean of 0.33, standard deviation of 0.08, minimum of 0.19, and maximum of 0.59. The distribution is broad and choosing a fixed tail size across all classes results in a large measure of inconsistency. The number of support vectors is always a fraction of the data. Thus a post-hoc approach could choose any arbitrary fraction, but as a fixed size, would still be a poor approximation compared to Eqs. 6 & 7. The experiments in the next section show that in conjunction with the P_I-SVM, this principled approach to tail size estimation for SVM is quite effective.

6 Experimental Evaluation

Experiments are performed for two different open set scenarios: (1) the decision component of object detection, where individual classifiers are evaluated separately; and (2) multi-class open set recognition, where ensembles of classifiers are evaluated together. While our focus is on multi-class open set recognition, we chose to also evaluate a detection problem in order to compare the P_I-SVM[1] with recent published work in open set recognition, and to first establish viability in a more restrictive open set context.

Preliminaries. In all experiments we make use of the cross-class-validation evaluation methodology described in Sec. 1. Extending typical cross-validation for machine learning evaluation, cross-class-validation leaves out not only some training data on each fold to be used for validation purposes, but also some number of classes. Four parameters control how open the validation problem is: the number of training classes t, the target number of known classes $\eta \leq t$ that we would like to identify using m classifiers for the problem, and the number of validation classes $e \geq t$. The steps for the process are shown in Alg. 3 (for simplicity, we show 1-fold), with the final result being a set of validation statistics (*e.g.* accuracies or F-measures) that provides a realistic reflection of how well a particular classifier is performing in the midst of $e - t$ unknown classes during testing.

Cross-class-validation can be used for either detection-oriented problems or multi-class open set recognition problems. To evaluate a detection problem, the number of target classes η is set to 1 and the number of validation classes e is set to a value greater than t. To evaluate a multi-class open set recognition problem, η is set to a number greater than 1, and e is set to a value greater than η and t. A fully closed problem would set $e = t$, meaning the set of unknown classes $\mathcal{C}_u$ is empty. The parameters m, t, η, and e also allow us to quantify "openness" as a single number (where a larger value means a more open problem), providing a consistent frame of reference for plotting results. Like the prior work of Scheirer et al. [45], we plot "openness" vs. F-measure for the experiments in this section[2]. Adapting Eq. 1 from [45], openness is defined as:

$$\text{openness} = 1 - \sqrt{(2 \times t)/(m \times \eta + e)} \tag{8}$$

The primary question we seek to answer is how much improvement is achieved by the P_I-SVM over viable alternative approaches. To this end, we compare against a lengthy list of supervised learning algorithms including common classifiers and state-of-the-art algorithms for open set recognition. With respect to approaches that are suitable for detection[3], we consider standard SVM variants including the 1-vs-Rest binary RBF SVM, 1-vs-Rest binary linear SVM, and 1-vs-Rest binary RBF SVM with Platt Probability Estimation [37] and a threshold (all using LIBSVM implementations [9]). We also compare against the state-of-the-art EVT-based probability estimator Multi-Attribute Spaces (MAS) [42] with a threshold, and the 1-vs-Set Machine [45], a state-of-the-art

[1] Source code is available at https://github.com/ljain2/libsvm-openset.

[2] For comparison, accuracy plots are provided in the supplemental material.

[3] We also tried reference code for the optimal Naive Bayes Nearest Neighbor algorithm [3], but at 72 hours per test, and with 2,640 tests (88 classes $\times$ 6 levels of openness $\times$ 5 folds) needed to add it to Fig. 4, including it was beyond the scope of this paper. See the longer note in the supplemental material.

Algorithm 3 Cross-Class-Validation (1-Fold)

Require: A set of class labels $\mathcal{C}$; a set of labeled data points for each class $X_y, y \in \mathcal{C}$; number of top-level classifiers to train m; number of training classes t; number of target classes $\eta \leq t$, number of validation classes $e \geq t$; a training objective ϕ, a fusion function F combining η bottom-level classifiers, and a ground-truth operator $\mathcal{Y}(x)$ returning label for x

for $i = 1 \rightarrow m$ **do**
  Randomly choose t classes for training label set $\mathcal{C}_t \subseteq \mathcal{C}$ ▷ Different on each iteration
  Randomly choose η classes for target label set $[y_1 \ldots y_\eta] \in \mathcal{C}_t$
  for $y = y_1 \rightarrow y_\eta$ **do**
    Randomly choose positive training set T_y^+ from X_y
    Randomly choose negative training set T_y^- sampling each $X_j; j \in \mathcal{C}_t, j \neq y$
    $f_y = \phi(T_y^+ \cup T_y^-)$ ▷ Train a decision model; one-class objectives ignore T_y^-
  end for
  Randomly choose $e - t$ additional class labels $\mathcal{C}_u \subset \mathcal{C}, \mathcal{C}_u \cap \mathcal{C}_t = \varnothing$
  Randomly choose known class validation set E^t sampling each $X_j, j \in \mathcal{C}_t; E_j^t \cap T_j^+ = \varnothing$
  Randomly choose unknown class validation set E^u sampling each $X_j, j \in \mathcal{C}_u$
  $v_i = \cup_{x \in (E^t \cup E^u)} \{\mathcal{Y}(x), F\left(f_{y_1}(x), \ldots, f_{y_\eta}(x)\right)\}$ ▷ Fuse classifiers; combine with label
  $\mathcal{V} = \mathcal{V} \cup \text{stats}(v_i)$ ▷ Accumulate overall evaluation statistics
end for
return $\mathcal{V}$ ▷ Return complete validation statistics for each classifier

algorithm for open set detection problems. For these latter two approaches, code was obtained from the public source repositories for the associated references.

With respect to approaches that are suitable for multi-class open set recognition, we consider standard multi-class SVM variants including the 1-vs-Rest Multi-class RBF SVM (LIBSVM ErrorCode implementation [28]), Pairwise Multi-class RBF SVM (LIBSVM implementation [9]), 1-vs-Rest Multi-class RBF SVM with Platt Probability Estimation and a threshold (LIBSVM ErrorCode implementation [28]), and Pairwise Multi-class RBF SVM with Platt Probability Estimation and a threshold (LIBSVM implementation [9]). As alternatives to standard SVM, we look at Logistic Regression analysis for multi-class probabilistic linear classification (LIBLINEAR implementation [20]), and MAS in a multi-class setting. Finally, the purely single-class P_I-OSVM is also used as a baseline in all experiments.

The first experiment uses a subset of Caltech 256 for training and images from Caltech 256 and ImageNet for separately testing open set "detection" for different classes. The setup is a replication of the experiment described in Sec. 5 of [45] (Fig. 7 of that article). 532,400 images are considered in total. Features are a 3,780-dimension vector of Histogram of Oriented Gradients (HOG) [13]. Using cross-class-validation (Alg. 3), we set $m = 88$, $\eta = 1$, and $e = 88$. The number of training classes t always includes one positive class and a varying number of negative classes to control the openness of the problem. Alg. 3 is invoked five times, always choosing a new set of 88 classes from the 256 we have available in $\mathcal{C}$. We report the average result over all trials.

The second experiment uses data from two classic visual learning benchmarks, LETTER [35] and MNIST [33], both of which are considered to be solved in their original closed set forms. To evaluate LETTER in a multi-class open set recognition

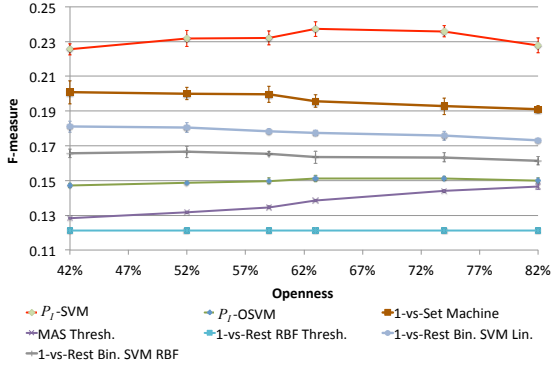
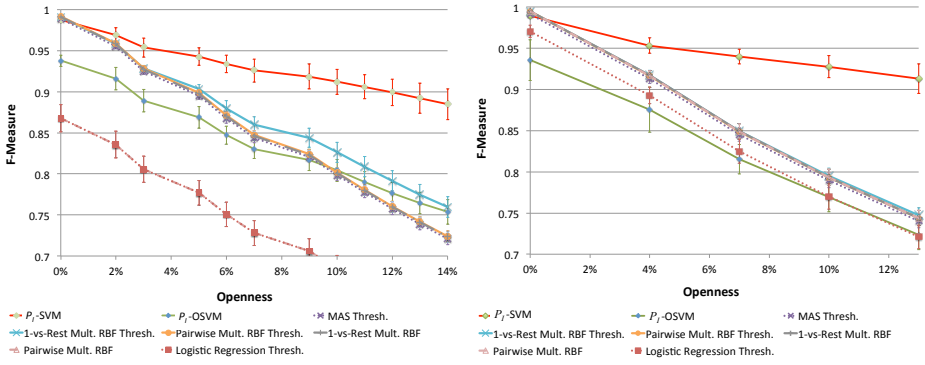


Fig. 4: Performance for the binary decision component of an *open set object detection task* for an open universe of 88 classes [45]. Results are calculated over a five-fold cross-data-set test with images from Caltech 256 used for training and images from Caltech 256 and ImageNet for testing; error bars reflect standard deviation. Approaches marked with "Thresh." have been augmented to support rejection. The P_I-SVM significantly outperforms the prior state-of-the-art (1-vs-Set Machine) with a 12%–22% improvement in F-measure, as well as the pure single-class P_I-OSVM model. We note that the P_I-OSVM is still measurably better than a standard one-class SVM (not plotted because its F-measures fall below the y-axis scale used in this figure), and is superior to other binary classifiers making use of probability estimation.

mode using Alg. 3, we set $m = 1$, $t = 15$, and $\eta = 15$. In this case, we vary the number of validation classes e by adding some number of additional class labels (not exceeding 11, the number of remaining letters outside of training) to the number of training classes t. To evaluate MNIST, we set $m = 1$, $t = 6$, and $\eta = 6$. We vary the number of validation classes e by adding some number of additional class labels (not exceeding 4, the number of remaining digits outside of training) labels to t. In both cases, we invoke Alg. 3 20 times and report the average, with standard deviation for error bars.

For multi-class open set recognition, the class with the maximum (depending on the operation of the algorithm) probability, decision score, or votes is the predicted class. Each approach producing a probability score has a rejection option via the threshold $\delta = 0.5 \times \text{openness}$. Each approach producing an uncalibrated decision score assigns a sample with a score less than zero as either a true negative if an unknown class, or a false negative if a known class. For multi-class algorithms with a rejection option we consider a rejected sample as either a true negative if an unknown class, or a false negative if a known class. Multi-class SVMs without a rejection option produce no negative decisions. RBF kernel parameters are tuned via 5-fold cross-validation on the training data, giving us ($C = 2$, $\gamma = 2$) for LETTER and ($C = 2$, $\gamma = 0.03125$) for MNIST.

Results. The results for the experiment evaluating the binary decision component of object detection are summarized in Fig. 4. The P_I-SVM significantly outperforms the prior state-of-the-art (1-vs-Set Machine) with a 12%–22% improvement in F-measure. An important effect in this experiment is the noticeable difference in F-measure between the P_I-SVM, which combines single-class probability estimation with a binary classification model, and the P_I-OSVM, which is purely single-class for probability estimation and classification. The extra discriminative power provided by the binary classifier



(a) Multi-Class Open Set Recog. for LETTER (b) Multi-Class Open Set Recog. for MNIST

Fig. 5: Two classic data sets evaluated in a *multi-class open set recognition scenario*. All existing algorithms we tested have significant trouble achieving good performance as the problem grows to be more open. The P_I-SVM is more stable than existing algorithms, and achieves high F-measure scores across all levels of openness. Note the large gap between the P_I-SVM and MAS [42] algorithms, indicating the EVT fitting strategy of the P_I-SVM is significantly better.

addresses the limitations inherent in the one-class model. The P_I-OSVM, however, is still measurably better than a standard one-class SVM (off the plot; see supplemental material), and shows improvement over other binary probability estimators.

The results for the multi-class open set recognition experiments are shown in Fig. 5. We expected low F-measure scores for all approaches in the first experiment, which examined a large number of classes for object detection. In contrast, the low scores present for basic OCR tasks with far fewer classes indicate that existing approaches are fundamentally constrained to the closed set classification tasks for which they were designed. The P_I-SVM, which does not possess this same limitation, achieves more stability and considerably better performance over all comparison approaches. We also tested thresholding 1-vs-Rest Multi-class linear SVM and RBF one-class SVM, both of which performed worse than the methods in Fig. 5.

7 Conclusion

A surprising finding of this work has been the impact of recasting basic visual benchmarks like LETTER and MNIST as multi-class open set recognition problems. This suggests that we are much farther away from solving very basic recognition tasks than the classification performance numbers initially led us to believe. As a solution, one-class models are appealing in that they do not suffer from any of the problems associated with negative class modeling for open set recognition, but they almost always overfit their training data. Hybrid models such as the P_I-SVM we introduced in this paper may be the key to achieving good generalization through some measure of discrimination with known negative classes and an estimate of probability of positive class inclusion. Future work includes incorporating objectness or object saliency as a pre-processing step, as well as extending novelty detection [5] to multi-class recognition.

References

1. Bartlett, P.L., Tewari, A.: Sparseness vs estimating conditional probabilities: Some asymptotic results. Journal of Machine Learning Research 8, 775–790 (2007)
2. Bartlett, P.L., Wegkamp, M.H.: Classification with a reject option using a hinge loss. Journal of Machine Learning Research 9, 1823–1840 (2008)
3. Behmo, R., Marcombes, P., , Dalalyan, A., Prinet, V.: Towards optimal naive Bayes nearest neighbor. In: ECCV (2010)
4. Bergamo, A., Torresani, L.: Meta-class features for large-scale object categorization on a budget. In: IEEE CVPR. pp. 3085–3092 (2012)
5. Bodesheim, P., Freytag, A., Rodner, E., Kemmler, M., Denzler, J.: Kernel null space methods for novelty detection. In: IEEE CVPR. pp. 3374–3381 (2013)
6. Bravo, C., Lobato, J.L., Weber, R., L'Huillier, G.: A hybrid system for probability estimation in multiclass problems combining svms and neural networks. In: Hybrid Intelligent Systems. pp. 649–654 (2008)
7. Broadwater, J., Chellappa, R.: Adaptive Threshold Estimation Via Extreme Value Theory. IEEE Transactions on Signal Processing 58(2), 490–500 (2010)
8. Cevikalp, H., Triggs, B.: Efficient object detection using cascades of nearest convex model classifiers. In: IEEE CVPR. pp. 886–893 (2012)
9. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011)
10. Chow, C.: On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory 16(1), 41–46 (1970)
11. Coates, A., Carpenter, B., Case, C., Satheesh, S., Suresh, B., Wang, T., Wu, D.J., Ng, A.Y.: Text detection and character recognition in scene images with unsupervised feature learning. In: Proceedings of the 2011 International Conference on Document Analysis and Recognition. pp. 440–445 (2011)
12. Coles, S.: An introduction to statistical modeling of extreme values. Springer Series in Statistics, Springer (2001)
13. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE CVPR. pp. 886–893 (2005)
14. Dardas, N.H., Georganas, N.D.: Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. IEEE Transactions on Instrumentation and Measurement 60, 3592–3607 (2011)
15. Deng, J., Berg, A.C., Li, F.F.: Hierarchical semantic indexing for large scale image retrieval. In: IEEE CVPR. pp. 785–792 (2011)
16. Ding, X., Li, Y., Belatreche, A., Maguire, L.P.: An experimental evaluation of novelty detection methods. Neurocomputing 135 (2014)
17. Duan, K.B., Keerthi, S.S.: Which is the best multiclass SVM method? An empirical study. In: Proceedings of the 6th International Conference on Multiple Classifier Systems. pp. 278–285 (2005)
18. Enzweiler, M., Gavrila, D.M.: Integrated pedestrian classification and orientation estimation. In: IEEE CVPR. pp. 982–989 (2010)
19. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html
20. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research 9, 1871–1874 (2008)
21. Fragoso, V., Sen, P., Rodriguez, S., Turk, M.: EVSAC: accelerating hypotheses generation by modeling matching scores with extreme value theory. In: IEEE ICCV. pp. 2472–2479 (2013)

22. Fragoso, V., Turk, M.: SWIGS: a swift guided sampling method. In: IEEE CVPR. pp. 2770–2777 (2013)
23. Fumera, G., Roli, F.: Support vector machines with embedded reject option. In: International Workshop on Pattern Recognition with Support Vector Machines (SVM2002). pp. 68–82 (2002)
24. Grandvalet, Y., Rakotomamonjy, A., Keshet, J., Canu, S.: Support vector machines with a reject option. In: NIPS. pp. 537–544 (2008)
25. Gumbel, E.: Statistical Theory of Extreme Values and Some Practical Applications. US Govt. Printing Office (1954)
26. Hastie, T., Tibshirani, R.: Classification by pairwise coupling. In: Annals of Statistics. pp. 507–513. MIT Press (1996)
27. Hinton, G.E., Ghahramani, Z., Teh, Y.W.: Learning to parse images. In: NIPS. pp. 463–469 (1999)
28. Huang, T.K., Weng, R.C., Lin, C.J.: Generalized bradley-terry models and multi-class probability estimates. Journal of Machine Learning Research 7, 85–115 (2006)
29. Jepson, A., Mann, R.: Qualitative probabilities for image interpretation. In: IEEE ICCV. pp. 1123–1130 (1999)
30. Kwok, J.T.Y.: Moderating the outputs of support vector machine classifiers. IEEE Transactions on Neural Networks 10(5), 1018–1031 (1999)
31. ImagetNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012): http://image-net.org/challenges/lsvrc/2012/index (Accessed: 2014-02-18)
32. ImagetNet Large Scale Visual Recognition Challenge 2013 (ILSVRC2013): http://image-net.org/challenges/lsvrc/2013/index (Accessed: 2014-02-18)
33. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
34. McCann, S., Lowe, D.: Local naive Bayes nearest neighbor for image classification. In: CVPR (2012)
35. Michie, D., Spiegelhalter, D.J., Taylor, C.C., Campbell, J. (eds.): Machine Learning, Neural and Statistical Classification. Ellis Horwood (1994)
36. Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: ICML. pp. 625–632 (2005)
37. Platt, J.: Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In: Smola, A., Bartlett, P., Schölkopf, B. (eds.) Advances in Large Margin Classifiers. MIT Press (2000)
38. Ramanan, D., Sminchisescu, C.: Training deformable models for localization. In: IEEE CVPR. pp. 206–213 (2006)
39. Ryoo, M., Matthies, L.: First-person activity recognition: What are they doing to me? In: IEEE CVPR. pp. 2730–2737 (2013)
40. Samanta, R., LeBaron, B.: Extreme Value Theory and Fat Tails in Equity Markets. Computing in Economics and Finance 140, Society for Computational Economics (2005)
41. Scheirer, W., Jain, L., Boult, T.: Probability models for open set recognition. To appear in IEEE Transactions on Pattern Analysis and Machine Intelligence (2014)
42. Scheirer, W., Kumar, N., Belhumeur, P.N., Boult, T.E.: Multi-attribute spaces: Calibration for attribute fusion and similarity search. In: IEEE CVPR. pp. 2933–2940 (2012)
43. Scheirer, W., Rocha, A., Michaels, R., Boult, T.E.: Meta-recognition: The theory and practice of recognition score analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(8), 1689–1695 (2011)
44. Scheirer, W., Rocha, A., Micheals, R., Boult, T.: Robust fusion: extreme value theory for recognition score normalization. In: ECCV. pp. 481–495 (2010)
45. Scheirer, W., Rocha, A., Sapkota, A., Boult, T.: Toward open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(7), 1757–1772 (2013)

46. Schölkopf, B., Platt, J.C., Shawe-Taylor, J.C., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Computation 13, 1443–1471 (2001)
47. Shu, G., Dehghan, A., Oreifej, O., Hand, E., Shah, M.: Part-based multiple-person tracking with partial occlusion handling. In: IEEE CVPR. pp. 1815–1821 (2012)
48. Steinwart, I.: Sparseness of support vector machines–some asymptotically sharp bounds. In: NIPS. pp. 1069–1076 (2003)
49. Tax, D.M.J., Duin, R.P.W.: Support vector data description. Machine Learning 54, 45–66 (2004)
50. Tax, D.M.J.: One-class classification: Concept learning in the absence of counter-examples. Ph.D. thesis, Technische Universiteit Delft (2001)
51. Toronto, N., Morse, B.S., Ventura, D., Seppi, K.: The hough transform's implicit Bayesian foundation. In: IEEE ICIP. pp. 377–380 (2007)
52. Torr, P.H., Szeliski, R., Anandan, P.: An integrated Bayesian approach to layer extraction from image sequences. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(3), 297–303 (2001)
53. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience (1998)
54. Wah, C., Belongie, S.: Attribute-based detection of unfamiliar classes with humans in the loop. In: IEEE CVPR. pp. 779–786 (2013)
55. Weston, J., Collobert, R., Sinz, F., Bottou, L., Vapnik, V.: Inference with universum. In: ICML. pp. 1009–1016 (2006)
56. Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 694–699 (2002)
57. Zhang, R., Metaxas, D.: RO-SVM: Support vector machine with reject option for image categorization. In: BMVC. pp. 1209–1218 (2006)