Pre-print of article that will appear in T-PAMI.

# Probability Models for Open Set Recognition

Walter J. Scheirer*, *Member, IEEE,*
Lalit P. Jain*, *Student Member, IEEE,*
and Terrance E. Boult, *Senior Member, IEEE*

***Abstract*—Real-world tasks in computer vision often touch upon open set recognition: multi-class recognition with incomplete knowledge of the world and many unknown inputs. Recent work on this problem has proposed a model incorporating an open space risk term to account for the space beyond the reasonable support of known classes. This article extends the general idea of open space risk limiting classification to accommodate non-linear classifiers in a multi-class setting. We introduce a new open set recognition model called Compact Abating Probability (CAP), where the probability of class membership decreases in value (abates) as points move from known data toward open space. We show that CAP models improve open set recognition for multiple algorithms. Leveraging the CAP formulation, we go on to describe the novel Weibull-calibrated SVM (W-SVM) algorithm, which combines the useful properties of statistical extreme value theory for score calibration with one-class and binary support vector machines. Our experiments show that the W-SVM is significantly better for open set object detection and OCR problems when compared to the state-of-the-art for the same tasks.**

***Index Terms*—Machine Learning, Support Vector Machines, Open Set Recognition, Statistical Extreme Value Theory.**

## I. Introduction

In a recent article in this journal [27], we raised the issue of open set recognition for visual learning, where not all classes encountered during testing are known during training. This is a necessary and difficult problem to tackle. As an initial solution, we proposed an algorithm called the 1-vs-Set machine, which is suitable for single-class detection tasks in an open set scenario. In essence, the 1-vs-Set machine manages the risk of the unknown by solving a two-plane optimization problem that yields a linear classifier. Detection is a useful operation (almost every digital camera has an automatic face detector these days), but in many cases, we would like to recognize which known *classes*, if any, are associated with the input image. This can enable applications such as unconstrained optical character recognition (OCR), and photo or video tagging without constraints on the input. In this article we consider the multi-class open set recognition problem.

Multi-class open set recognition is a fundamental problem in computer vision. Intuitively, we classify objects with respect to a fixed set of known classes, but we recognize something we know among the set of all possible inputs that can include things for which we do not explicitly have class or training data. For example, when you look at a face in a photo, you might have a set of people you know and want to recognize in mind, but there are far more people who you don't know that may show up in the image. You must also ignore the presence of things that are not people: dogs, cars, buildings, trees, etc. Knowing that we do not recognize something is what sets multi-class recognition apart from multi-class classification.

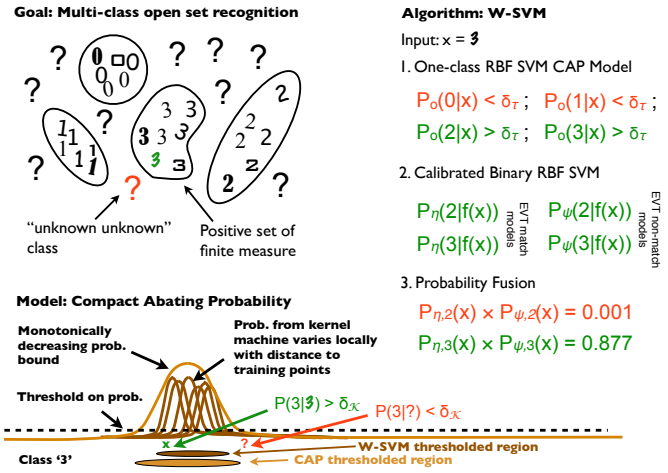


Fig. 1. Open set recognition must address both the known and unknown classes that might occur in the real world. For instance, considering OCR for a parcel delivery application, classifiers must recognize known characters (*e.g.* "3") and reject an unspecifiable variety of other input including symbols, marks, stickers, and photos that can appear on a package. Standard statistical learning, using any mixture of discriminative and generative models, does not address unknown classes. The goal of this work is to approach the problem of multi-class open set recognition by limiting open space risk using labeled training sets of finite measure. The Compact Abating Probability (CAP) model we introduce bounds probability estimates of feature space that decay away from the training data. By truncating the abating probability, CAP models provably reduce open space risk. In addition, probability estimates from a calibrated binary kernel machine can have even better estimates than the CAP bounds, further reducing the open set risk. By taking advantage of the CAP model and the statistical extreme value theory for probability estimation, the novel W-SVM provides solutions for non-linear multi-class classification in an open set scenario. In the multi-step example above, classes marked in red indicate rejection, and those marked in green indicate acceptance.

Adapting Donald Rumsfeld's famous "There are known knowns" statement [23], we assert that recognition must consider three basic categories of classes:

1) *known classes*, *i.e.* the classes with distinctly labeled positive training examples (also serving as negative examples for other known classes);
2) *known unknown classes*, *i.e.* labeled negative examples, not necessarily grouped into meaningful categories;
3) *unknown unknown classes*, *i.e.* classes unseen in training.

Traditional classification, which is the dominant model used for multi-class computer vision problems, considers only known classes. Including known unknown classes results in models with an explicit "other class," or a detector trained with unclassified negatives. Algorithms designed specifically to address *unknown unknown classes* are the focus of open set recognition – the subject of this article.

Our formal definition of open set recognition [27] introduced the concept of open space risk, and then combined it, via regularization, with empirical risk to formulate an open set risk minimization problem. Open space risk is the relative measure of positively labeled space, far from known samples

W.J. Scheirer is with the School of Engineering and Applied Sciences, Department of Molecular and Cellular Biology, and Center for Brain Science, Harvard University, Cambridge, MA, 02138.
Corresponding Author's E-mail: wscheirer@fas.harvard.edu
L.P. Jain and T.E. Boult are with the Department of Computer Science, University of Colorado Colorado Springs, Colorado Springs, CO, 80918.
* W.J. Scheirer and L.P. Jain contributed equally to this work.

over the overall measure of the space. However, we left the space "far from known samples" open to interpretation. Further, the 1-vs-Set machine is strictly a linear classifier. It reduces the open space risk by replacing the half-space of a binary linear classifier with a positive region bounded by two parallel planes. While the measure of the resulting positive region is smaller than a half-space, it still has infinite measure, and only reduces the risk because the definition of open space risk considers relative measure. This article seeks to incorporate non-linear kernels into a solution that further limits open space risk by positively labeling only sets with finite measure.

Following the usual tenets of support vector machines [29], the 1-vs-Set machine simply assigns class labels to instances during testing. What we would like for a multi-class solution is a formulation that produces probabilistic decision scores. This allows us to assess the output of multiple classifiers to either accept the highest confidence label if the associated probability exceeds a threshold, or to reject as unknown if not. The formulation should be probabilistic because there is always some amount of uncertainty in any decision. For open set recognition in particular, there is a great deal of uncertainty when confronting the unknown. However, the derivation of a probabilistic learning formulation in an open set scenario is not as straightforward as it initially appears.

Assume the set of potential classes, known and unknown, are mutually exclusive, countable and hence can be labeled $y \in \mathbb{N}$. Let $x \in X \subseteq \mathbb{R}^d$ be a measured image from the set of all features $X$, where $x \in \mathcal{K}$ means it is from the feature space of known classes $\mathcal{K} \subset X$. While the overall joint probability $P(x, y)$ is well defined for open set recognition, the set of all $y$ is not (nor can be) known to the algorithm, thus the estimation of a generative model for $P(x, y)$ is not possible. A restricted generative model for $P(x, y)$ could be estimated for a known class $y$, $x \in \mathcal{K}$, but its use would be limited in a general setting. With unknown unknowns, many of the standard probabilistic and statistical learning tools cannot be directly applied.

Relating the joint distribution to the conditional distribution requires conditioning on the classes, and with unknown classes, one cannot properly normalize. Even with the assumption that all classes are mutually exclusive, the unknown unknowns prohibit the use of the law of total probability that underlies Bayes' theorem. Furthermore, open set recognition cannot just use the *maximum a posteriori probability* (MAP) estimate over the known classes as the optimal solution. MAP estimation needs the full posterior distribution, which again requires consideration of all classes. The consideration of just the known classes is insufficient.

To address these issues, we introduce a new formal model of probabilistic class association for open set recognition called *Compact Abating Probability* (CAP). In a CAP model, probability of class membership abates as points move from known data to open space, which accounts for the unknown unknowns without the need to explicitly model them. We also introduce a novel technique called the *Weibull-calibrated SVM* (W-SVM), which combines CAP with the statistical extreme value theory (EVT) for improved multi-class open set recognition. EVT statistics have been shown to yield well-grounded probability estimates for SVM applied to closed set recognition problems in computer vision [24], thus we revisit this approach in the context of open set recognition in this work. Fig. 1 provides a brief overview of the model and algorithm.

Our experiments show that incorporating CAP improves existing techniques and that the W-SVM is significantly better than existing approaches including: common binary and multi-class SVM formulations, multi-class SVM with a rejection option provided by thresholding Platt's sigmoid probability estimator [22], Multi-Attribute Spaces [24], the 1-vs-Set Machine [27], Logistic Regression, and Nearest Neighbor. For evaluation, we breathe new life into the classic data sets for multi-class classification, LETTER [12] and MNIST [21], by changing their testing protocols. Surprisingly, when recontextualized into open set problems, these once-solved data sets become significant challenges for recent algorithms. We also examine a difficult cross-data set object detection task with data from Caltech 256 [13] and ImageNet [7].

In summary, the contributions of this article are:

1) The theoretical formulation of a Compact Abating Probability (CAP) model for open set recognition.
2) A new algorithm called Weibull-calibrated SVM (W-SVM), which incorporates the CAP model and the statistical extreme value theory for probability estimates.
3) An experimental evaluation of CAP and the W-SVM in detection and multi-class open set scenarios.

## II. Background and Related Work

Consider the definition of open space risk that we introduced in [27], which the objective function of open set recognition, including multi-class formulations, must minimize. Let $f$ be a measurable recognition function where $f_y(x) > 0$ for recognition of the class $y$ of interest and $f_y(x) = 0$ when $y$ is not recognized, $\mathcal{O}$ be the "open space," and $S_o$ be a ball of radius $r_o$ that includes all of the known positive training examples $x \in \mathcal{K}$ as well as the open space $\mathcal{O}$. The probabilistic *Open Space Risk* $R_{\mathcal{O}}(f)$ for a class $y$ can be defined as

$$R_{\mathcal{O}}(f) = \frac{\int_{\mathcal{O}} f_y(x)dx}{\int_{S_o} f_y(x)dx} \tag{1}$$

where open space risk is considered to be the relative measure of positively labeled open space compared to the overall measure of positively labeled space (which includes the space near the positive examples). This definition, however, does not tell us how to define $\mathcal{O}$. In this article, we specifically look at a definition of $\mathcal{O}$ for kernels, including non-linear functions.

How to incorporate Eq. 1 into a model is an important question. There is an ongoing debate between the use of generative and discriminative models in statistical learning [2], [20], with arguments for the value of each. However, open set recognition introduces a new issue: neither discriminative nor generative models address the unknown unknowns that exist in open space; another constraint must be added. Moreover, strategies to learn discriminative class boundaries like hard negative mining [10], [14] are limited in open set recognition, since it is not possible to mine examples from the unknown classes. In response to these observations, we incorporate an *abating process*, a model enforced decay of probability away

from supporting evidence, into the CAP model we introduce in Sec. III. With missing classes preventing the use of MAP or a classical optimal Bayes estimator over known classes, we acknowledge that we must have what Lasserre et al. [20] would consider model mis-specification and hence expect a benefit from discriminative training.

The W-SVM, described in detail in Sec. IV, relies on a calibration process to transform raw scores to probabilities. Thresholding these probabilities provides a viable open set recognition algorithm. The idea of using thresholded probabilities for rejection is, unsurprisingly, not new [31], [1]. Chow [5] showed that the optimal rejection decision rule is a threshold over the *a postiori* probabilities. The raw SVM decision scores are uncalibrated values and not posterior probabilities, and thus rejection processes tend to calibrate/normalize them. Several different techniques [22], [8], [18], [3] have been proposed for converting uncalibrated SVM output to probabilistic calibrated output. For multi-class problems, the estimation is more complex because the calibrations across classes need to be related, highlighting the fact that the per-class models are insufficient. Multiple heuristic techniques have been developed [8], [18] for converting multi-class SVM output to an estimated posterior probability.

A variation on Platt's approach [22], included in LIBSVM [4], [18], is the most widely used probability estimator for a single SVM. Platt's technique fits a sigmoid function to uncalibrated SVM decision scores during training, and then computes calibrated values for novel instances using that model. While the data will generally be "roughly sigmoidal," there is no theoretical basis for the pure sigmoid: motivation for this model comes from specific empirical data instances [22]. In Sec. V we look at the performance of this estimator.

We consider an estimate to be *well-grounded* if there is a defensible theoretical justification for the choice of the underlying probability model, *e.g.* the central limit theorem justifies the use of a Gaussian distribution for some physical measurements. There is, however, no solid theoretical justification for SVM calibration using a Gaussian model computed over all of the data. As an alternative, we invoke the statistical extreme value theory [19] to develop grounded probability estimates.

The use of statistical extreme value theory in computer vision has been growing. Related work includes the use of EVT for "meta-recognition" introduced by Scheirer et al. in [26], [25], and expanded upon by Fragoso and Turk in [11]. However, none of these references considers open set recognition, nor do they suggest EVT for producing SVM probability estimates. The most related work is the Weibull-based normalization of SVM scores from visual attribute classifiers by Scheirer et al. [24], which considers data from the "other side" of the class of interest to estimate rejection probability, *i.e.* considering anything not in the negative class to be positive. We also compare with this model, which is rather weak for open set recognition. Rejecting association with known negatives is not as meaningful when there are unknown classes.

## III. The Compact Abating Probability model

This section defines our Compact Abating Probability (CAP) model for open set recognition. Intuitively, open space risk exists when a recognition model labels space far from any training data, *e.g.* if we are labeling location data using training data only from Colorado, it would be risky to apply that model to Boston. The idea of a CAP model is to ensure that the recognition function is decreasing away from the training data, so that thresholding it limits the labeled region.

The definition of open space risk, Eq. 1, requires a definition of open space $\mathcal{O}$. Open space is the space sufficiently far from any known positive training sample $x_i \in \mathcal{K}, i = 1 \dots N$. Thus, we offer a formal definition:

$$\mathcal{O} = S_o - \bigcup_{i \in N} B_r(x_i) \tag{2}$$

where $B_r(x_i)$ is a closed ball of radius $r$ centered around training sample $x_i$, and we consider all $N$ training samples. This defines open space as the space more than distance $r$ from any known training sample. Ideally, all samples from class $y$ will be outside $\mathcal{O}$. For a minimal radius $r^*$, any smaller radius $r'$ would include positive test samples of $y$: $r' < r^* \Rightarrow \exists x' \in \mathcal{O} \mid f_y(x') > 0$. Note that labeling the space within the balls as just positive may yield a poor recognition function; the balls likely contain a complex mixture of positive and negative regions. We consider $r$ to be a problem-specific parameter, which may be estimated via calibration, *e.g.* the maximum (or average) spacing between training samples.

We first describe the properties of an *Abating Bound*, and then expand that bound to a probabilistic formulation. An abating bound $A(r) : \mathbb{R} \rightarrow \mathbb{R}$ is a non-negative finite square integrable continuous decreasing function. This implies $\lim_{r \rightarrow \infty} A(r) = 0$. When $\forall x, \exists x^* \mid f(x) \leq A(\|x - x^*\|)$, we call the function $f$ "abating" because the spatial influence decreases with distance from $x^*$.

Next, let us assume features are transformed by the kernel trick into an inner product space with positive definite kernel $K(x, x_i) = \langle \Phi(x), \Phi(x_i) \rangle$, where $x_i \in \mathcal{K}$ indicates a specific positive training example, and $x \in X$ any example. The kernel $K$ defines a similarity measure over the feature space. We call kernel $K$ abating if there exists an abating bound $A$ such that

$$\forall x, x_i : \quad 0 < K(x, x_i) \leq A(\|x - x_i\|) \tag{3}$$

Standard RBF (Gaussian) kernels are decaying functions of radial distance and hence abating kernels.

To get probabilities, we process the kernels with a calibration technique, using training data to define a monotonically decreasing probability distribution $p_f(s; y)$ for the probability of the score $s$ from an RBF kernel $K$ being associated with the given class $y$. We call $p_f(s; y)$ an *Abating Probabilistic Point Model* because the probability of points associating abates (becomes less intense) as the spatial separation of any two points increases.

For recognition we need to combine multiple data points to build an effective class model. Consider fusing the abating models from the known training data $x_1 \dots x_m, x_i \in \mathcal{K}$, *i.e.* for any example $x \in X$ we define the model

$$M(x) = p_f(F(K(x, x_1) \dots K(x, x_m)); y) \tag{4}$$

where $F$ is the fusion operator. We typically consider $F$ to be either the canonical sum or canonical product rule, though

other fusion processes can be used. A *Fused Abating Property* is defined as any fusion function where $\exists x' \in X$ such that an abating bound function $A_{x'}$ exists $\forall x \in X$:

$$F(K(x,x_1)\dots K(x,x_m)) \leq A_{x'}(\|x'-x\|) \tag{5}$$

Eq. 5 states that after fusion there is an abating bound function centered at $x'$ such that the fused value $F$ is bounded from above by that abating function. According to Prop. 4.4 of [15], positive definite kernels are closed under canonical sums or products, so the result of fusion is still a proper kernel. Thus if $K(\cdot,\cdot)$ satisfies Eq. 3, then the canonical sum or product rules for fusion will satisfy Eq. 5.

Just using an abating probabilistic point model does not, however, assure that open set recognition is being addressed because the model can have non-zero probability over all of $\mathbb{R}^n$ and yield a large risk when integrated over the open space. One way to to handle this is to define $M_\tau$ to be a *Compact Abating Probability Model* with distance threshold $\tau$ and an abating probabilistic point model $M$ satisfying the fused abating property such that for a given finite $\tau$ and $\forall x \in X$

$$\min_{x_i \in \mathcal{K}} \|x - x_i\| > \tau \;\Rightarrow\; M_\tau(x) = 0 \tag{6}$$

In a compact abating probability model, features beyond a given thresholded distance $\tau$ from the closest training point have zero probability. This is true even of very high dimensional sparse representations that are common in object recognition. The model controls for the possibility of representations that are close in feature space yet relatively far in label space via $\tau$. With these preliminaries, we can state our primary theorem:

**Theorem 1 (Open Space Risk of CAP models):** *Let $M_{\tau,y}(x)$ be a probabilistic recognition function that uses a CAP model over a known training set for class $y$, where $\exists x_i \in \mathcal{K} \mid M_{\tau,y}(x_i) > 0$. Let open space risk be $R_\mathcal{O}(f)$ and open space be $\mathcal{O}$, defined as in Eqs. 1 and 2 respectively. If $r$ in Eq. 2 satisfies $r > \tau$, then $R_\mathcal{O}(M_{\tau,y}) = 0$, i.e. when the CAP distance threshold is smaller than the open space radius, the CAP model has zero open space risk.*

Proof: Let $x$ be any point in $\mathcal{O}$. Since $x \in \mathcal{O}$ implies $x \notin \bigcup_{i \in N} B_r(x_i)$, we have $\forall x_i \in \mathcal{K}, \|x - x_i\| > r > \tau$. Therefore, by the compact abating property (Eq. 6) $M_{\tau,y}(x) = 0$. Placing this into the numerator of $R_\mathcal{O}(f)$ (Eq. 1) yields $\int_\mathcal{O} M_{\tau,y}(x)dx = 0$ and zero open space risk. □

**Corollary 1 (Thresholding CAP model probability manages Open Space Risk):** *For any CAP model, considering only points with sufficiently high probability will reduce open space risk. In particular, consider a canonical sum kernel-based CAP model with a probability threshold $0 \leq \delta_\tau \leq 1$ such that for the set of points $x_i \in \mathcal{K}$ and coefficients $\vartheta_i > 0$, $p_f(\sum_i \vartheta_i K(x,x_i); y) \geq \delta_\tau$. Increasing $\delta_\tau$ decreases open space risk, and there exists a $\delta_\tau^*$ such that any greater threshold produces zero open space risk.*

Thresholding probabilities provides a way to adjust the support of the CAP model because in the compact abating probability model $M$ the probabilities are bounded by the decreasing abating bound (Eqs. 4 and 5), therefore considering only points above a given probability threshold implicitly defines a $\tau$. Since we can adjust the CAP model's open space risk by thresholding the probabilities, it provides a powerful way to address open set recognition. While any model with sufficiently compact support could have zero open space risk, the abating property of the CAP model allows one to implicitly adjust $\tau$ to reduce the amount of open space that can be labeled positive. Note that the CAP property does not guarantee that the model assigns positive labels within the compact support region – it just ensures it that there is a zero probability of doing so outside the region. In general the quality of the CAP model will still depend on how well the probabilities model the actual underlying positive region of the class.

A simple CAP example, Nearest Neighbor + CAP (NN+CAP), is to let $d_x$ be the distance to the nearest neighbor of $x$, and to let $d_x > \tau \Rightarrow p_a(x) = 0$ and $p_a(x) = \frac{|\tau - d_x|}{\tau}$ otherwise. In a multi-class setting, this results in a thresholded nearest neighbor algorithm that can reject an input as unknown. Other vision algorithms have considered nearest neighbors within a distance threshold, *e.g.* [30]. With sufficently dense samples, NN+CAP reduces to nearest neighbor with all of its associated properties, *i.e.* having a limiting error of no more than twice the Bayes error rate. While NN+CAP provides more meaningful responses for open set recognition with finite samplings than nearest neighbor, it is still a weak probability model and hence not expected to perform very well on difficult problems (see Sec. V). Thus we seek alternative models. Our first step for non-linear kernels considers a one-class SVM-based model [28]:

**Theorem 2 (RBF One-Class SVM yields CAP model):** *Let $x_i \in \mathcal{K}, i = 1...m$ be the training data for class $y$. Let O-SVM be a one-class SVM with a square integrable monotonically decreasing RBF kernel $K$ defined over the training data, with associated Lagrangian multipliers $\alpha_i > 0$ [28], then $\sum_i \alpha_i y_i K(x,x_i)$ yields a CAP model.*

Proof: Since O-SVM has only positive data[1], we can view this function as providing a canonical sum over positive definite kernels $F(x) = \sum_i \vartheta_i(K(x,x_i))$, with coefficients $\vartheta_i > 0$ and training points $x_i$. To show this is a CAP model, let $g = \sum_i \vartheta_i = \sum_i \alpha_i y_i$. Let $i^* = \operatorname{argmin}_i \|x' - x_i\|$, then it is sufficient to let $A_{x'} = gK(x,x_{i^*})$, which by the theorem's kernel assumption is monotonically decreasing and in the space of square integrable functions. Hence $gK(x,x_{i^*})$ is an abating bound function for the sum, yielding a CAP model. □

## IV. Probability Estimation and the W-SVM

While the one-class model of Sec. III is intuitive, what about a binary SVM in this probabilistic mode? It is well known that one-class models are typically less effective than binary machines [27]. Unfortunately, the decision score of a binary SVM is not a canonical sum. It can, however, still be useful as improved probabilities will generally result in tighter bounds around the class of interest. Following [15], one can collect all the positive coefficients into one sum, and all of

[1]The decision function of the one-class SVM has a bias term $\rho$ that we ignore as it only shifts scores and hence is removed by probability normalization.

the negatives into a second sum, split the bias between them, and view the SVM as applying a decision rule on which is more similar. This effectively fuses both positive and negative evidence. Working with only the positive or negative data, we can get nicely bounded results from a binary SVM that can be used in conjunction with the one-class probabilities. We call this model the Weibull-calibrated SVM (W-SVM).

### A. Binary RBF SVM Incorporating a CAP Model

Discriminative trained classifiers such as binary SVMs can have very good closed set performance. However, a discriminative classifier estimating $P(y|x)$, trained on $x \in \mathcal{K}$, should be viewed as $P(y|x \in \mathcal{K})$, and has no basis for prediction when $x \notin \mathcal{K}$. Thus to improve the accuracy, we seek to combine probabilities computed for both one-class RBF SVMs and binary RBF SVMs. We use the one-class SVM CAP model as a conditioner: if the one-class SVM predicts $P_O(y|x) > \delta_\tau$, even with a very low threshold $\delta_\tau$, that a given input $x$ is a member of class $y$, then we will consider the binary classifier's estimates of $P(y|x)$.

To estimate the binary classifier's probability, we start from the observation that separating positives and negatives is useful. Rather than computing the RBF SVM decision score and optimizing a sigmoid over all the scores, we seek to model the positive and negative scores separately. Assume a set of known classes $\mathcal{Y}$. Logically, for a class $y \in \mathcal{Y}$, we can use positive scores from $y$ to estimate $P^+(y|x)$. We can also use negative scores from other known classes to estimate $P^-(\mathcal{Y} \setminus y|x)$. In a closed set scenario, we could estimate $P^+(y|x) = 1 - P^-(\mathcal{Y} \setminus y|x)$, *i.e.* given input $x \in \mathcal{K}$ the probability of being a particular class label can be estimated as the probability of not being a negative example. In an open set scenario, we cannot, in general, make these estimations. Thus to minimize our open space risk, we only consider $P^-$ and $P^+$ when $P_O(y|x) > \delta_\tau$, *i.e.* when open space risk is small.

### B. Grounded Probability Estimation

The second issue for the W-SVM is the need for grounded probability estimation. Recent work [25], [24] has shown that the recognition problem itself is consistent with the assumptions of statistical extreme value theory (EVT) [19], which provides a way to determine probabilities, regardless of the overall distribution of the data. The extreme values of a score distribution produced by any recognition algorithm can always be modeled by an EVT distribution, which is a reverse Weibull if the data are bounded from above, and a Weibull if bounded from below. With respect to statistical modeling, the prior EVT approaches [25], [24] only fit on the "negative" side of the decision space and use a statistical hypothesis test for rejection, *i.e.* they indicate the probability of a positive by estimating the probability of "not being a negative."

We apply the EVT concept separately to the positive and the negative scores from the binary SVM. A reverse Weibull is justified for the largest scores from the negative examples because they are bounded from above. A Weibull is the expected distribution for the smallest scores from the positive examples because they are bounded from below. Based on this knowledge, we can formulate the calibration for the binary SVM component of the W-SVM. Let us separate our training examples into the match examples for a class $y$ as $x \in \mathcal{K}^+$ and all other non-match examples where the class $\neq y$ as $x \in \mathcal{K}^-$. Note that $\mathcal{K} = \mathcal{K}^+ \cup \mathcal{K}^-$. Letting $s_i = f(x_i)$ be the SVM decision score for $x_i$, we collect the scores into match and non-match sets where scores for matches are $s_j \in S^+$ if $x_j \in \mathcal{K}^+$ and scores for non-matches are $s_j \in S^-$ if $x_j \in \mathcal{K}^-$. Let $\psi$ be the upper extremes of the non-matches $S^-$, and let $\eta$ be the lower extremes of the matches $S^+$.

Returning to modeling, the reverse Weibull and Weibull have three parameters: location $\nu$, scale $\lambda$, and shape $\kappa$. We use the library provided by the authors of [24], applying Maximum Likelihood Estimation to estimate the $\nu_\eta, \lambda_\eta, \kappa_\eta$ that best fit $\eta$ and the $\nu_\psi, \lambda_\psi, \kappa_\psi$ that best fit $\psi$. To produce a probability score for a particular SVM decision $f(x)$, we use the CDF defined by the parameters. Given a test sample $x$, we have two independent estimates for $P(y|f(x))$: $P_\eta$ based on the Weibull CDF derived from the match data:

$$P_\eta(y|f(x)) = 1 - e^{-(\frac{-f(x)-\nu_\eta}{\lambda_\eta})^{\kappa_\eta}} \tag{7}$$

and $P_\psi$ based on the reverse Weibull CDF derived from the non-match data, which is equivalent to rejecting the Weibull fitting on the non-match data:

$$P_\psi(y|f(x)) = e^{-(\frac{f(x)-\nu_\psi}{\lambda_\psi})^{\kappa_\psi}} \tag{8}$$

$P_\eta$ is a novel direct probability estimation using only match data, while $P_\psi$ is similar to the normalization suggested by [24]. Our experiments in Sec. V show that, especially for open set testing, the use of $P_\eta$ is significantly better. This is intuitive since $P_\eta$, using only positive data, does not strongly depend on which classes are known. While $P_\eta$ is not formally related to any one-class estimation, its use of only positive data means it shares some of the characteristics of one-class SVMs, including CAP-like rejection of non-relevant classes. However, since the underlying classifier is a one-vs-all binary SVM, the resulting estimates are more discriminative. While the above description has been in the context of a binary SVM, Eq. 7 also serves as the calibration for the one-class SVM CAP model.

### C. The W-SVM Algorithm

Finally, we must say how to use these estimates. Since both estimators use independent points, the product $P_\eta \times P_\psi$ can be interpreted as "the probability that the input is from the positive class AND NOT from any of the known negative classes." Conversely, the sum $P_\eta + P_\psi$ would be interpreted as either a positive OR NOT a known negative, with the latter often being true for any unknown unknowns. In open set recognition, where there may be unknown classes, $P_\psi$ should generally be modulated by other supporting evidence of the sample being positive. Thus the product is the preferred combination.

We note the estimates are not completely conditionally independent since they share the underlying SVM structure, which is valid only when the input is from a known class. As described above, to further manage open set risk, we condition the use of the W-SVM with a thresholded CAP model. Letting $P_O(y|x)$ be the probability from Eq. 7 for the

RBF one-class SVM trained on positive examples of class $y$, we define an indicator variable: $\iota_y = 1$ if $P_O(y|x) > \delta_\tau$ and $\iota_y = 0$ otherwise. Multi-class W-SVM recognition for all known classes $\mathcal{Y}$ is then:

$$y^* = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} P_{\eta,y}(x) \times P_{\psi,y}(x) \times \iota_y \quad (9)$$
$$\text{subject to } P_{\eta,y^*}(x) \times P_{\psi,y^*}(x) \geq \delta_R.$$

Notice that the W-SVM has two parameters: $\delta_\tau$, which is generally very small (fixed to 0.001 for all experiments in this article) and is used to adjust what data the one-class SVM considers to be even remotely related, and $\delta_R$, which is the level of confidence needed in the W-SVM estimate itself [2].

To help readers see how this differs from Platt calibration, consider data from our object detection experiment below with widely separated training classes that have all match scores in the range $[0.99, 1]$ and all non-match scores in the range $[-1.0, -0.99]$. Platt's sigmoid will yield $P_s(0) \approx 0.5$. For W-SVM, $P_\eta(0) \approx 0.001$ and $P_\psi(0) \approx 0.999$ yielding a product of $\approx 0.001$. This is because the score of 0 is unlike anything seen in training; the sample producing it can be viewed as almost being in open space. W-SVM probabilities are qualitatively and quantitatively very different from prior calibration techniques.

## V. Experimental Evaluation

Our evaluation[3] of the W-SVM is focused on two challenging open set recognition scenarios. Like our previous work [27], we first examine binary object detection as a representative open set task in computer vision. Extending our experiments, we then look at the more difficult problem of multi-class open set recognition. The key question we seek to answer is how the W-SVM compares to the prior work in open set recognition, SVM probability estimation, binary and multi-class SVM formulations, and other common multi-class classifiers.

**Preliminaries.** Our comparison approaches include:

1) 1-vs-Set Machine [27]; a linear classifier for open set detection problems. Generalizes or specializes two planes to optimize empirical and open space risk (Eq. 1).
2) 1-vs-All Binary SVM; one positive class and all known negative classes are sampled for training a detector with a linear or RBF kernel. LIBSVM implementation [4].
3) 1-vs-All Binary RBF SVM with Platt Probability Estimation [22]; above detector training with empirical fit of training data to sigmoid, producing probabilistic decision scores. Rejection option is available by thresholding probability scores. LIBSVM implementation [4].
4) 1-vs-All Multi-class RBF SVM; all combinations of one positive class and all known negative classes. LIBSVM ErrorCode implementation [17].
5) 1-vs-all Multi-class RBF SVM with Platt Probability Estimation [22]; above SVM training with sigmoid fitting and rejection option. LIBSVM ErrorCode implementation.
6) Pairwise Multi-class RBF SVM [16]; a one-against-one multi-class formulation incorporating all known classes. LIBSVM implementation.

[2]Full pseudocode for the W-SVM and a visual walkthrough of the algorithm are provided as supplemental material.

[3]W-SVM code and all experimental data will be released after publication.

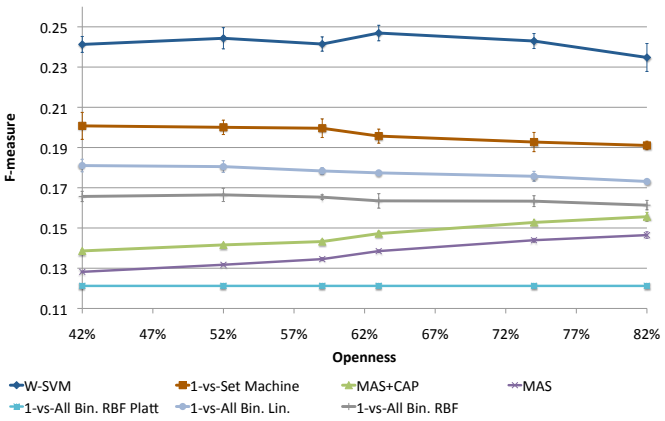


Fig. 2. Performance on an *open set binary object detection task* for an open universe of 88 classes [27]. Results are calculated over a five-fold cross-data set style test with images from Caltech 256 used for training and images from Caltech 256 and ImageNet for testing; error bars reflect standard deviation. The W-SVM significantly outperforms the prior state-of-the-art (1-vs-Set Machine), with a 20%–26% improvement in F-measure.

7) Pairwise Multi-class RBF SVM with Platt Probability Estimation; above SVM training with sigmoid fitting and a rejection option. LIBSVM implementation.
8) Multi-Attribute Spaces (MAS) [24]; binary RBF SVM calibration through EVT modeling of the decision scores from the non-match data (Eq. 8). We add a threshold over the probabilities for a rejection option.
9) Multi-Attribute Spaces + CAP Model (MAS+CAP); a novel extension to the MAS approach that provides a CAP model via one-class SVM to condition the decisions.
10) Nearest Neighbor (NN); simple non-parametric multi-class classification. Our own implementation.
11) Nearest Neighbor + CAP Model (NN+CAP); the algorithm described in Sec. III. $\tau$ is set via five-fold cross-validation on the training data.
12) Logistic Regression; regression analysis for multi-class probabilistic linear classification. LIBLINEAR implementation [9].

Following [27], we plot "openness" vs. F-measure, where we adapt standard data sets for open set cross-validation style analysis, holding out some classes in training and mixing them back in during testing. The definition we introduced in [27] quantifies "openness" as a function of the number of classes known in training and the number of classes observed during testing. Let $\mathcal{C}_R$ be the number of classes to be recognized, $\mathcal{C}_T$ be the number of classes used in training, and $\mathcal{C}_E$ be the number of classes used in evaluation (testing), then

$$openness = 1 - \sqrt{(2 \times \mathcal{C}_T / (\mathcal{C}_R + \mathcal{C}_E))}. \quad (10)$$

This provides a convenient way of assessing openness that varies from 0% and 100%. We chose F-measure instead of accuracy because it better emphasizes the distinction between correct positive and negative classifications[4]. For comparison, accuracy plots for all multi-class recognition experiments in this section are provided in this article's supplemental material.

[4]For an exposition of why, see Sec. 5 of [27]

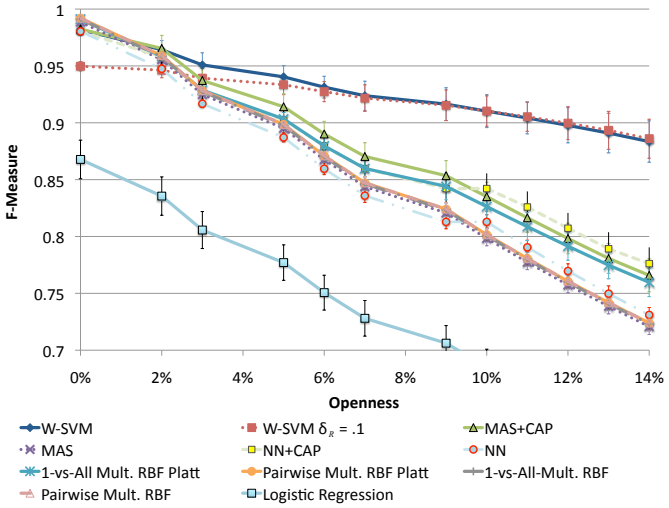


Fig. 3. F-measure for *multi-class open set recognition* on OLETTER. As openness increases, the W-SVM has the best performance. Common multi-class SVMs, probabilistic multi-class SVMs with a rejection option, logistic regression, and nearest neighbor degrade quickly (1-vs-All Mult. RBF, Pairwise Mult. RBF, and MAS are all comparable and visually overlap). The MAS and NN algorithms with a CAP model do a bit better than their baselines, but still degrade much more than the W-SVM. Error bars reflect standard deviation.

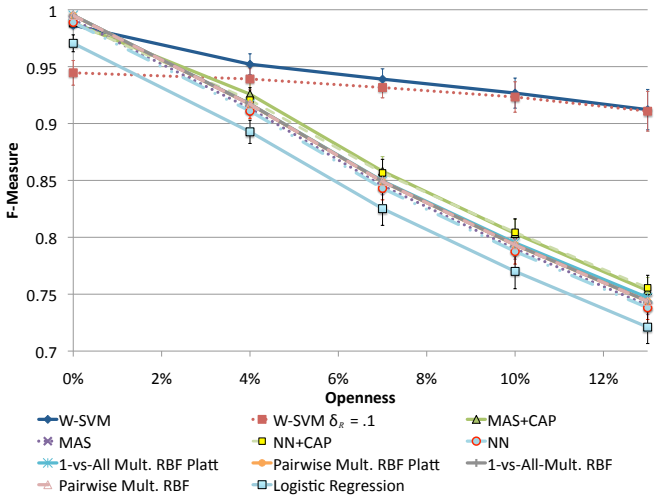


Fig. 4. F-measure for *multi-class open set recognition* on OMNIST. W-SVM again maintains high F-measure scores as the problem grows to be more open, but common multi-class SVMs and existing thresholded probability estimators degrade quickly. NN+CAP and MAS+CAP are again better than their baselines. All algorithms except the W-SVM, NN+CAP, MAS+CAP, and Logistic Regression are comparable and visually overlap. Error bars reflect standard deviation.

An openness statistic also helps us set the level of confidence needed for the W-SVM to make a positive decision. Recognizing that the more open the problem the higher the confidence we require to reject unknown unknowns, one can use expected openness (based on prior observations of similar problems outside of the testing regime at hand) to set this threshold as

$$\delta_R = 0.5 \times openness \tag{11}$$

Eq. 11 sets the threshold at 0 for closed problems; thresholds approach 0.5 (*i.e.* random chance) as the problem becomes more open. For a fair comparison, we also set the rejection threshold over the probability scores for all SVM algorithms with a rejection option and any other algorithm with a CAP model according to this formula.

For the open set binary detection experiment, we use a subset of Caltech 256 for training and images from Caltech 256 and ImageNet for testing. 532,400 images are considered in total. The setup is a replication of the open universe of 88 classes experiment described in [27] (Fig. 7 of that article), with five-fold cross-validation style testing. The 88 classes are selected at random, where one class is chosen to be positive, $n$ classes are chosen as negative training data (where $n$ varies with openness), and 87 known and unknown classes are used as negatives for testing. Each class is treated positively once per fold. Features are a 3,780-dimension vector of Histogram of Oriented Gradients (HOG) [6]. Following [27], all SVM parameters are set to the defaults specified by LIBSVM [4].

The second experiment examines a true open set multi-class recognition problem based on extensions of the standard visual learning benchmarks LETTER [12] and MNIST [21], which are commonly used to evaluate multi-class classification. To recast LETTER as an open set problem that we call OLETTER, we randomly choose 15 distinct labels as the known classes, and vary openness by adding a subset of the remaining 11 labels, repeating over 20-folds to get error bars. For open set testing on MNIST, we define OMNIST to use 6 labels as the known classes, and vary openness with the other 4 labels, again with 20 folds. We emphasize that while LETTER and MNIST are essentially solved problems in the literature, they are substantially more difficult in this open set configuration.

For multi-class recognition the class with the maximum (as a function of algorithm) score, probability, or votes is the predicted class. In multi-class algorithms with a rejection option, we consider rejected samples as either true "negatives" if an unknown class, or false negatives if a known class. SVMs without a reject option can produce no negative decisions, and thus have very poor precision as problems grow to be more open. In multi-class open set recognition, precision is critical for a high F-measure statistic. RBF parameters were tuned via cross validation on the training data, yielding ($C = 2$, $\gamma = 2$) for OLETTER and ($C = 2$, $\gamma = 0.03125$) for OMNIST. For the W-SVM, we also added an additional comparison case where instead of using Eq. 11 to set $\delta_R$, it is fixed to 0.1. We chose this constant to help illustrate the role of $\delta_R$ and the sensitivity of performance to the threshold. For problems that are more closed, a large rejection threshold degrades performance.

**Results.** The results of the detection experiment are summarized in Fig. 2. The W-SVM is significantly better than the prior state-of-the-art detection approach (1-vs-Set Machine) for open set scenarios. With respect to other approaches, rejection by thresholding sigmoid-based Platt probabilities is worse than standard SVM decision scores. A likely explanation for this is that the training classes are well separated, making the calibration model weak for unknown classes falling in between. Similarly, the performance of the MAS algorithm, which was designed for closed set classification, is also worse than binary SVM. This demonstrates a positive effect for adding a Weibull model of the match data (Eq. 7) to compliment the non-match model for increased generalization in the W-SVM. Another

important question is whether the CAP model provides any discernible benefit for open set recognition. The MAS+CAP curve shows adding a CAP model offers a large improvement in F-measure over the base MAS algorithm.

Figs. 3 and 4 highlight open set multi-class recognition F-measure performance for OLETTER and OMNIST respectively. On these tests F-measure degrades very rapidly for the most common multi-class SVM algorithms because they do not have a rejection option to handle the growing number of negatives. On OLETTER we see that the one-vs-all SVM with Platt probability estimation, which has a rejection option, is better than the standard algorithms – but the W-SVM is much better at tolerating increasing openness. On OMNIST, which has a larger number of testing images, the advantage of one-vs-all SVM with Platt probability estimation over typical SVMs disappears, but W-SVM retains most of its performance. Comparing the W-SVM using Eq. 11 and the W-SVM with the fixed $\delta_R$, Eq. 11 is the better strategy, producing higher F-measure statistics at lower levels of openness – especially for OMNIST. Again we see improvement for the MAS+CAP algorithm over its baseline (+3.3%), and a similar effect is observed for NN+CAP (+3.3%), but both are weaker than the W-SVM.

## VI. Discussion

As one considers open set recognition, the assumptions of traditional statistical learning, Bayesian models, and generative and discriminative models often do not hold. However, they can be adapted to provide probabilities for thresholding decisions that pave a way forward – where decisions depend on the validity and shape of those probabilities. While this article has focused on SVMs, and our experiments show the strong impact of openness on SVMs, the improvements from using a CAP model applied to SVM and other techniques leads us to conjecture that *addressing open set recognition requires thresholding on estimates that are robust to unknown classes and decay away from training data.*

We used the statistical extreme value theory to develop a novel approach to probability estimation for SVMs. Leveraging prior work in EVT, we made use of the observation that the distribution of scores near the SVM boundary of the extreme values of both matching and non-matching class data follow EVT. However, this observation comes with a caveat: with very limited sampling in training for a class with large variation in its feature space, it may not always be possible to fit a good Weibull model to the data.

Small training sets are just one open issue. Humans have a remarkable ability to reason through more complicated open set scenarios where machines currently cannot. Overlapping classes and hierarchical classes, when mixed with unknown unknowns, are representative of these cases. Perhaps most interesting is the significant challenge posed by the new partitions of LETTER and MNIST for open set multi-class recognition testing. These data sets are considered solved for closed set classification, but become quite relevant again for open set recognition. Both demonstrate that state-of-the-art learning approaches are far more brittle than originally assumed.

## Acknowledgment

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

## References

[1] P. L. Bartlett and M. H. Wegkamp. Classification with a reject option using a hinge loss. *JMLR*, 9:1823–1840, 2008. 3
[2] G. Bouchard, B. Triggs, et al. The tradeoff between generative and discriminative classifiers. In *COMPSTAT*, 2004. 2
[3] C. Bravo, J. L. Lobato, R. Weber, and G. L'Huillier. A hybrid system for probability estimation in multiclass problems combining SVMs and neural networks. In *Int. Conf. Hybrid Intel. Sys.*, 2008. 3
[4] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Tran. on Int. Sys. & Tech.*, 2:27:1–27:27, 2011. 3, 6, 7
[5] C. Chow. On optimum recognition error and reject tradeoff. *IEEE Trans. Info. Theory*, 16(1):41–46, 1970. 3
[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE CVPR*, 2005. 7
[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, 2009. 2
[8] K. Duan and S. Keerthi. Which is the best multiclass SVM method? An empirical study. *Multiple Classifier Systems*, pp 732–760, 2005. 3
[9] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 9:1871–1874, 2008. 6
[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE T-PAMI*, 32(9):1627–1645, 2010. 2
[11] V. Fragoso and M. Turk. SWIGS: A swift guided sampling method. In *IEEE CVPR*, June 2013. 3
[12] P. W. Frey and D. J. Slate. Letter recognition using holland-style adaptive classifiers. *Machine Learning*, 6:161–182, 1991. 2, 7
[13] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Tech. Rep. 7694, California Institute of Technology, 2007. 2
[14] J. F. Henriques, J. Carreira, R. Caseiro, and J. Batista. Beyond hard negative mining: Efficient detector learning via block-circulant decomposition. In *IEEE ICCV*, December 2013. 2
[15] T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *Annals of Stat.*, pp 1171–1220, 2008. 4
[16] C.-W. Hsu and C.-J. Lin. A comparison of methods for multiclass support vector machines. *IEEE T-NN*, 13(2):415–425, 2002. 6
[17] T.-K. Huang. LIBSVM ErrorCode. http://goo.gl/cOcgDN. 6
[18] T.-K. Huang, R. C. Weng, and C.-J. Lin. Generalized Bradley-Terry models and multi-class probability estimates. *JMLR*, 7:85–115, Dec. 2006. 3
[19] S. Kotz and S. Nadarajah. *Extreme Value Distributions: Theory and Applications*. World Sci. Pub. Co., 2001. 3, 5
[20] J. A. Lasserre, C. M. Bishop, and T. P. Minka. Principled hybrids of generative and discriminative models. In *IEEE CVPR*, 2006. 2, 3
[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. of the IEEE*, 86(11):2278–2324, 1998. 2, 7
[22] J. Platt. Probabilistic outputs for support vector machines and comparison to regularize likelihood methods. In *Advances in Large Margin Classifiers*, pp 61–74, 2000. 2, 3, 6
[23] D. Rumsfeld. DoD News Briefing addressing *unknown unknowns.* http://www.defense.gov/transcripts/transcript.aspx?transcriptid=2636, 2002. [Last accessed 15-Mar-2014]. 1
[24] W. Scheirer, N. Kumar, P. N. Belhumeur, and T. E. Boult. Multi-attribute spaces: Calibration for attribute fusion and similarity search. In *IEEE CVPR*, June 2012. 2, 3, 5, 6
[25] W. Scheirer, A. Rocha, R. Michaels, and T. E. Boult. Meta-recognition: The theory and practice of recognition score analysis. *IEEE T-PAMI*, 33(8):1689–1695, Aug. 2011. 3, 5
[26] W. Scheirer, A. Rocha, R. Micheals, and T. Boult. Robust fusion: extreme value theory for recognition score normalization. In *ECCV*. Springer, 2010. 3
[27] W. Scheirer, A. Rocha, A. Sapkota, and T. E. Boult. Towards open set recognition. *IEEE T-PAMI*, 36(7):1757–1772, July 2013. 1, 2, 4, 6, 7
[28] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, July 2001. 4
[29] V. Vapnik. *The Nature of Statistical Learning Theory, 2nd Edition*. Springer, 1998. 2
[30] Y. Weiss, R. Fergus, and A. Torralba. Multidimensional spectral hashing. In *ECCV 2012*, 2012. 4
[31] R. Zhang and D. Metaxas. RO-SVM: Support vector machine with reject option for image categorization. In *BMVC*, 2006. 3