

# Towards Open Set Recognition

Walter J. Scheirer, *Member, IEEE*,  
 Anderson Rocha, *Member, IEEE*,  
 Archana Sapkota, *Student Member, IEEE*,  
 and Terrance E. Boult, *Member, IEEE*

**Abstract**—To date, almost all experimental evaluations of machine learning-based recognition algorithms in computer vision have taken the form of “closed set” recognition, whereby all testing classes are known at training time. A more realistic scenario for vision applications is “open set” recognition, where incomplete knowledge of the world is present at training time, and unknown classes can be submitted to an algorithm during testing. This article explores the nature of open set recognition, and formalizes its definition as a constrained minimization problem. The open set recognition problem is not well addressed by existing algorithms because it requires strong generalization. As a step towards a solution, we introduce a novel “1-vs-Set Machine,” which sculpts a decision space from the marginal distances of a 1-class or binary SVM with a linear kernel. This methodology applies to several different applications in computer vision where open set recognition is a challenging problem, including object recognition and face verification. We consider both in this work, with large scale experiments performed over data from the Caltech 256, ImageNet, and Labeled Faces in the Wild sets. The experiments highlight the effectiveness of machines adapted for open set evaluation compared to existing 1-class and binary SVMs for the same tasks.

**Index Terms**—Open Set Recognition, 1-vs-Set Machine, Machine Learning, Object Recognition, Face Verification, Support Vector Machines.

## 1 INTRODUCTION

Both recognition and classification are common terms in computer vision. What is the difference? In classification, one assumes there is a given set of classes between which we must discriminate. For recognition, we assume there are some classes we can recognize in a much larger space of things we do not recognize. A motivating question for our work here is: What is the general object recognition problem? This question, of course, is a central theme in vision. According to Duin and Pekalska [1], how one should approach multi-class recognition is still an open issue. Should it be performed as a series of binary classifications, or by detection, where a search is performed for each of the possible classes? What happens when some classes are ill-sampled, not sampled at all or undefined?

The general term recognition (and the specific terms object recognition and face verification that we consider in this article) suggests that the representation can handle different patterns often defined by discriminating features. It also suggests that the patterns to be recognized will be in general settings, visually mixed with many classes. For some problems, however, we do not need, and often cannot have, knowledge of the entire set of possible classes (Fig. 1). For instance, in a recognition application for biologists, a single species of fish might be of interest. However, the classifier

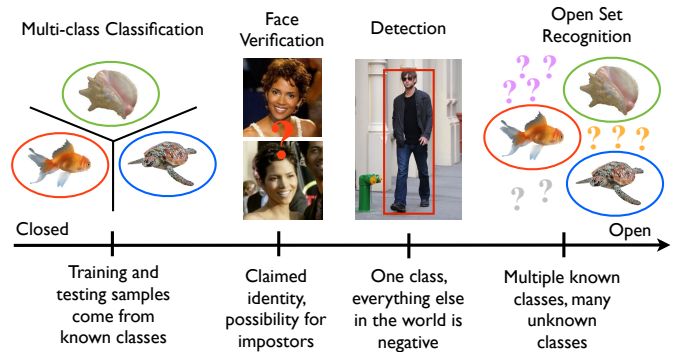


Fig. 1. Vision problems arranged in order of “openness”. For some problems, we do not have knowledge of the entire set of possible classes during training, and must account for unknowns during testing. In this article, we develop a deeper understanding of those open cases.

must consider the set of all other possible objects in relevant settings as potential negatives. Similarly, verification problems for security-oriented face matching constrain the target of interest to a single claimed identity, while considering the set of all other possible people as potential impostors. In addressing general object recognition, there is a finite set of known objects in myriad unknown objects, combinations and configurations – labeling something new, novel or unknown should always be a valid outcome. This leads to what is sometimes called “open set” recognition, in comparison to systems that make closed world assumptions or use “closed set” evaluation.

For many vision problems, researchers have assumed one

- Walter Scheirer, Archana Sapkota and Terrance Boult are with the University of Colorado at Colorado Springs, Colorado Springs, CO, 80918. E-mail: lastname@vast.uccs.edu
- Anderson Rocha is with the Institute of Computing, University of Campinas (Unicamp), Campinas, Brazil. Email: anderson.rocha@ic.unicamp.br

has examples from all classes, and have subsequently labeled the entire space in binary fashion as either positive (+1) or negative (−1). In contrast, an open set scenario has classes, not just instances, in testing that were not seen in training. It is somewhat reasonable to assume we can gather examples of the positive class, but the number and variety of “negatives” is not well modeled. The important difference is, in the words of Zhou and Huang [2] (with a bit of inspiration from Tolstoy), “All positive examples are alike; each negative example is negative in its own way”. Furthermore, even if all of the negative classes were known, from a pragmatic point of view, we generally cannot have sufficiently many positive examples to balance the required sampling of the negatives. In either case, we seek to generalize the problem from a closed world assumption to an open set.

Object detection is perhaps the most familiar vision problem that does not exist in a specific closed setting. The goal of detection is to locate an object of interest in an image. Since a negative detection is anything other than the class of interest, the problem is much more open than closed. Popular detection approaches train binary classifiers with a relatively modest sampling of positive examples and a very large sampling (often on the order of millions) of negatives from thousands of different classes. This is an appropriate strategy when a good sampling of the negative classes is possible, but with very incomplete knowledge of the possible negative classes it can lead to inaccuracies in many settings. In addition, we are generally left with a “negative set bias” [3] defined by the very large sampling of classes we do know about. In a sense, when we have very limited knowledge of the domain of possible classes, detection becomes a special case of open set recognition, with just one class of interest.

Fig. 1 depicts a few popular vision problems with varying qualitative degrees of openness. Intuitively, a problem with only a single class of interest is less open than one with many. However, the number of unknown classes we might encounter should also play a critical role. Let us formalize the “openness” of a particular problem or data universe by considering the number of target classes to be identified, the number of classes used in training, and the number of classes used in testing:

$$\text{openness} = 1 - \sqrt{\frac{2 \times |\text{training classes}|}{|\text{testing classes}| + |\text{target classes}|}} \quad (1)$$

The above formulation yields percent openness (values between 0% and 100%), where 0% represents a completely closed problem, and larger values more open problems. For a fixed number of training classes, increasing the number of testing classes increases openness, as does increasing the number of target classes to identify. Increasing the fraction of classes available during training decreases openness. Table 1 shows values of openness for different examples considered in our work and others from the spectrum of problems in Fig. 1. The number of training instances per class is important to the accuracy of a given classifier, but is not a property of the problem itself, and thus not part of this definition. For almost any unconstrained real world problem, the number of testing classes can grow rapidly with openness approaching 100%.

TABLE 1

Examples of openness values for the vision problems of Fig. 1 as a function of the number of target classes to be identified, training classes and testing classes, calculated using Eq. 1. Multi-class classification is always 0% open.

Problem	Targets	Training	Testing	Openness
Typical Multi-class [1]	$x$	$x$	$x$	0%
Our work: Face Verif.	12	12	50	38%
Typical Detection [4]	1	100,000	1,000,000	55%
Our work: Obj. Recog.	88	12	88	63%
Our work: Obj. Recog.	88	6	88	74%
Our work: Obj. Recog.	212	6	212	83%

Potential solutions to the open set recognition problem must optimize for unknown classes, as well as the known classes. An important difference from typical multi-class classification is that a general open set multi-class solution must be able to label the input as one of the known classes or as *unknown*. It is not sufficient to just return the most likely class: the classifier must also support rejection. The first insight we offer here is that Support Vector Machines (SVMs) define half-spaces, and will classify data that is very far from any training sample. While we need solutions that support strong generalization, there should be a limit on how far from known data a sample associated with a given label can be.

Empirical risk, measured on training data, is what is classically defined and optimized over. However, for open set recognition it is critical to consider how to extend the model to capture the risk of the unknown from insufficient generalization or specialization. This is different from the binary classifier approach that tries to maximize the margin, which is the gap between the positive and negative decision boundary. While maximum-margins can be very effective for closed set problems, the approach generally results in overgeneralization for open set problems. For example, in Fig. 2, the space containing unknowns (“?”) would likely be labeled “dog” as there is nothing to limit the positive label propagation if the decision boundary exists between birds & frogs and dogs. The SVM found a plane to separate positive and negative classes, but only by considering the known negatives. One might view the maximum-margin approach as assuming all unknown points are equally likely to be positive and negative based on what is nearest, even if that point is quite far away. For a sample coming from an unknown class, such as the raccoon, that is an incorrect assumption. We believe that good solutions to the open set recognition problem require minimizing the open space representing the learned recognition function  $f$ , outside the reasonable support of the training samples.

The primary goal of this work is to develop a thorough understanding of open set recognition in a supervised learning setting. We construct the first formalization of this problem, and provide an empirical case expanding existing 1-class and binary SVMs with linear kernels to address open set recognition. The resulting *1-vs-Set Machine* is a step towards a solution. Specifically, we revisit the ideas of the 1-class and binary SVM for open set recognition problems, and address the generalization/specialization issues through a novel learning

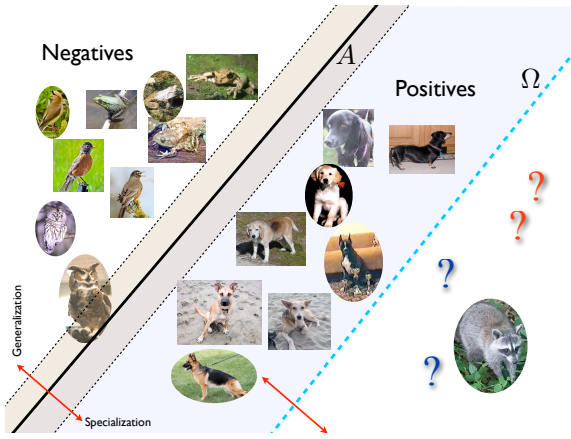


Fig. 2. The Open Set Recognition Problem explicitly assumes not all classes are known *a priori*. Square images are from training, oval images are from testing. The class of interest (“dog”) is surrounded by other classes, which can be known (“frog”, “birds”), or unknown (“owl”, “raccoon”, “?”). Plane  $A$  maximizes the SVM margin making “dog” a half-space – which is mostly open space. The 1-vs-Set machine adds a second plane  $\Omega$  and defines an optimization to adjust  $A$  and  $\Omega$  to balance empirical and open space risk.

technique. Instead of tackling the generalization/specialization problem as an error minimization of the training function of the SVM, we introduce a concept of open space risk and then minimize an error function combining empirical risk over training data with the risk model for the open space. The known class training data represents the “Set” of 1-vs-Set.

To improve the overall open set recognition error, our 1-vs-Set formulation balances the unknown classes by obtaining a core margin around the decision boundary  $A$  from the base SVM, specializing the resulting half-space by adding another plane  $\Omega$  and then generalizing or specializing the two planes (shown in Fig. 2) to optimize empirical and open space risk. This process uses the open set training data and the risk model to define a new “open set margin”. The second plane ( $\Omega$ ) allows the 1-v-Set machine to avoid the over generalization that would misclassify the raccoon in Fig. 2. The overall optimization can also adjust the original margin with respect to  $A$  to reduce open space risk, which can avoid negatives such as the owl.

We organize the rest of this article as follows. First, we formalize the open set recognition problem in Sec. 2. In Sec. 3, we take a look at the related work in open set recognition and machine learning across vision and pattern recognition. In Sec. 4, we formalize our theoretical model of margin generalization and specialization to develop the 1-vs-Set Machine. We apply this model, as well as common SVM models for comparison, to the problems of object recognition and face verification and present results in Sec. 5. We conclude and discuss some ideas for future work in Sec. 6.

## 2 FORMALIZATION OF THE PROBLEM

Assume images of classes are processed into  $d$  dimensional representations. Then we can define a class to be a partition  $\mathcal{P}$  over  $\mathbb{R}^d \times \{-1, 1\}$ , such that a sample value  $V \in \mathbb{R}^d$  is labeled  $L = +1$  if it belongs to the class to be recognized and  $L = -1$  if it is from any other class. For simplicity we assume the sample can be either positive or negative, but not both.

To define the open set problem, we assume that we are given valid training samples  $\hat{V} = \{v_1, \dots, v_m\}$  from  $\mathcal{P}$  as well as training samples  $\hat{K} = \{k_1, \dots, k_n\}$  from other known classes  $\mathcal{K}$ . Note also for simplicity we are assuming unique labeling (prohibiting nested or hierarchical classes) thus the samples from  $\mathcal{K}$  can be considered negative training examples with respect to  $\mathcal{P}$ . Let  $\mathcal{U}$  be the larger universe of allowed unknown (negative) classes. Note that  $\mathcal{U}$  does not need to contain every possible class (or image), though it still may be too large to explicitly specify or sample for training. For example,  $\mathcal{U}$  could contain any non-positive object one might observe in a zoo while looking for bears (elephants, horses, cages, strollers, popcorn, trees, etc.). Let  $\{t_1, \dots, t_z\}, t_i \in \mathcal{P} \cup \mathcal{K} \cup \mathcal{U}$  be our test data, which include unknown classes, and do not overlap with the training data.

Let  $f$  be a measurable recognition function for some class  $\mathcal{P}$  that can be applied over a set of test data. We can define the open space risk  $R(f)$  for the class of interest as the expected value of misclassification

$$R(f) = \mathbb{E}(\text{sign}(f(V)) \neq L) \quad (2)$$

where we have used expected risk. A critical element in defining risk for the open set problem is how to generalize from the known positives or specialize if there is already overlap between positive and negative training data. We seek models that address risk from the standpoint of generalization and specialization. This article will present a specific example of such a risk model in Sec. 4.1.

As discussed in Section 1.2.4 of [5], minimizing actual risk is ill posed, but the problem can be regularized by combining empirical risk with a smoothness constraint. Let  $\mathcal{H}$  be the space of “recognition functions” mapping  $\mathbb{R}^d \rightarrow \mathbb{R}$ , which should include some smoothness constraints. As a specific example here, assume  $\mathcal{H}$  is a Reproducing Kernel Hilbert Space (RKHS). If we minimize the risk only over functions in  $\mathcal{H}$ , we have a well posed regularized problem. While [5] discusses regularization using only empirical risk plus a convex penalty/smoothness term, we seek to expand our risk model to consider more than just the measured points. In particular, we seek to explicitly ensure that the risk of a false positive (overgeneralization) is proportional to the bounded space which is labeled positive, while also ensuring that overspecialization occurs if we define the region too narrowly around the training data.

One can view the maximum-margin concept as an explicit constraint, beyond the training data, that allows SVMs to better balance the empirical risk and open set risk. This leads to much better generalization when the problem appears “easy” and the negative training data is well separated. However, as the

number of unknown classes grows, the assumption of equally likely positives and negatives becomes unreasonable. Thus, we need to move beyond the basic maximum-margin concept. In the open set problem, only things that are reasonably close to the known positive points and not close to known negative points should be labeled positively. When possible explicit increased generalization via margin constraints is a useful way of addressing overspecialization risk. And minimization of the positive labeled space addresses overgeneralization risk. This is the overall model we develop in Sec. 4. Leaving the majority of space in an entire half-space labeled positive presents a significant risk of the unknown.

While we want to minimize risk of the unknown, we also need to balance the data error measure  $\mathcal{D}(f(\hat{V} \cup \hat{K}))$ , which incorporates performance metrics. The data error measure  $\mathcal{D}$  we consider most appropriate for the open set problem is the inverse of the F-measure (we look at this score in more detail in Sec. 5) over the classification results from the training data  $\hat{V} \cup \hat{K}$ :

$$\mathcal{D}_F = \frac{1}{\text{F-measure}} \quad (3)$$

Increasing F-measure is considered improving performance; we use an inverse measure so we can combine it with risk in a minimization framework. We can also include the specification of hard constraints, e.g., meeting at least a particular false accept or false reject rate. While it may be possible to fold the data error measure into the definition of risk or the space  $\mathcal{H}$  of recognition functions, we find it easier to consider two separate terms. We treat the data error measure as a fidelity term in a regularization formulation that minimizes the risk.

In summary, our goal is to balance the risk of the unknown (Eq. 2) with the empirical data error measure (Eq. 3). In this sense, we formally define the open set recognition problem as follows:

**Definition 1.** (The Open Set Recognition Problem) *Using training data with positive samples  $\hat{V}$ , and other known class samples  $\hat{K}$ , an open space risk function  $R$ , and a data error measure  $\mathcal{D}$ , find a measurable recognition function  $f \in \mathcal{H}$ , where  $f(x) > 0$  implies positive recognition, and  $f$  is defined by minimizing the open set risk:*

$$\operatorname{argmin}_{f \in \mathcal{H}} \left\{ R(f) + \lambda_r \mathcal{D}(f(\hat{V} \cup \hat{K})) \right\} \quad (4)$$

where  $\lambda_r$  specifies the regularization tradeoff between risk and data.

In Eq. 4, we have defined open set recognition as minimizing the open set risk, which combines the open space risk and empirical data error terms, over the space of all possible inputs. This definition balances what is known via  $\mathcal{D}$  while minimizing the open space risk  $R(f)$  associated with unknown classes.

We can also allow for explicit hard constraints on the training error (empirical risk) since in some applications one of these may be constrained for operational use (a false accept rate, for example) and that is not easily specified in the minimization equation. This is done by making Eq. 4 subject

to a constraint on the fraction of errors observed in the training set:

$$m\alpha \leq \sum_{i=1}^m \phi(f(v_i)) \quad \text{and} \quad n\beta \geq \sum_{j=1}^n \phi(f(k_j)) \quad (5)$$

where  $v_i \in \hat{V}$ ,  $m$  is the number of positive training samples,  $k_j \in \hat{K}$ ,  $n$  is the number of samples from other known classes,  $\alpha \geq 0$  and  $\beta \geq 0$  allow a prescribed limit on true positive and/or false positive rates, and  $\phi(z)$  is a given loss function, e.g., the classic soft margins hinge loss  $\phi(z) = \max(0, 1 - z)$  or squared hinge loss  $\phi(z) = \max(0, 1 - z)^2$  functions. For a prescribed  $\alpha$  and  $\beta$  it is possible there is no solution (100% classification of true positives with no false negatives is often not achievable). In a practical setting, we may fix either  $\alpha$  or  $\beta$  and then choose our data error term  $\mathcal{D}$  to require the system to optimize the other parameter. This can allow us to set precision or recall in an object recognition scenario, or bound the false accept rate in a face verification scenario.

### 3 RELATED WORK

Open set recognition has received only limited treatment in the literature, and almost all prior work focuses on evaluation. We are unaware of any prior formal definitions outside of evaluation protocols. In a study on evaluation methods for face recognition by Phillips et al. [6], a typical framework for open set identity recognition is described. The key to evaluation in open set recognition in the context described by Phillips et al. is the definition of an operating threshold  $\tau$ , which all classification scores must meet or exceed to be considered matches. An open set recognition system incorporating a threshold does not naïvely accept a top score as a match, allowing it to handle the cases where a sample does not correspond to a known class. Of course, the choice of  $\tau$  remains dependent on the requirements of the recognition system and its operating environment.

A series of thresholds can be considered to build a full performance curve (CMC, DET, PR, etc.), with parameters for matching instances selected by choosing one point on the curve. This idea is not just constrained to face recognition, and is familiar to many researchers working in recognition areas across vision. In [7], Fayin and Wechsler again view open set face recognition from just an evaluation perspective, describing it as a variation of the watch-list formulations in earlier face recognition testing at the National Institute of Standards and Technology (NIST). They state: “Open Set recognition operates under the assumption that not all the probes have mates in the gallery and it thus requires the reject option.”

Given our formal definition of open set recognition, we briefly discuss related work in recognition techniques that might satisfy that definition. A natural inclination towards solving the open set problem may be to consider binary and multi-class learning approaches with a representative sampling of negative training data to generalize the classifiers as much as possible. However, the nature of binary classification inhibits the controlled generalization needed for the open set problem. Consider the two examples in Fig. 3. 1-vs-1 classifiers [8] are

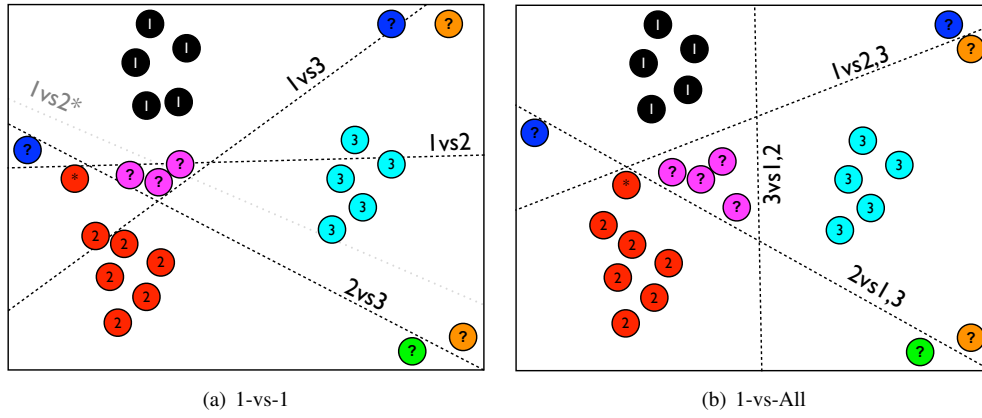


Fig. 3. The trouble with binary (1-vs-1) and multi-class (1-vs-All) classification for open set problems. In a 1-vs-1 scenario (a), good separation can be achieved between two classes during training, but this establishes margins that need not separate additional known or unknown classes. For instance, considering the margin between class 1 and class 2 above (labeled 1vs2), examples from class 3 and unknown classes fall indiscriminately across the margin. Similarly, in a 1-vs-All scenario (b), we see the same problem for unknown classes. In both cases, when considering just an additional training example (the red circle with a star in each figure), the results can be even worse, as the margins re-adjust for maximum separation. Far from the training classes this produces very significant margin plane movement, which can be seen in the light gray new margin separation plane 1vs2\* in (a).

trained by using positive examples from one class and negative examples from another. In a 1-vs-1 scenario, good separation can be achieved between the two classes during training, but this does not establish margins that separate additional known or unknown classes. 1-vs-All classifiers [8] are trained by using examples of a single class as the positive training set, and examples from all of the remaining (known) classes as the negative training set. In a 1-vs-All scenario, we can see the same problem that is present for 1-vs-1 for unknown classes. In both cases, when considering just an additional training example, the results can be even worse, as the margins re-adjust for maximum separation between the known data, while not taking other potential classes into account.

Another issue for any open set problem is that the training is both highly unbalanced and very incomplete (especially in the case of detection). Unbalanced data generally leads to overspecialization on the negative side. Resampling does not really solve the problem and the inherent imbalance in open set recognition presents issues that binary classifiers cannot easily overcome [9]. Thus, we turn to other methodologies that compensate for these deficiencies in our work.

One could think of the open set recognition problem as learning a minimum volume (MV) estimation that satisfies our data constraints. In this case, we might look for the minimal closed set covering a certain probability mass  $\mathcal{V}$  of the class of interest with respect to the unknown density (negative samples). If a new example  $t_i$  belongs to the minimal set, we classify it as being of the class of interest (+1) otherwise it is labeled (-1). According to Park et al. [10], we can directly estimate the MV-set by choosing the minimal set containing a portion of the positive sample points among a special class of measurable sets. Two approaches along this line are the 1-class Support Vector Machine method [11], and the Support Vector Domain Description or “Closed Balls” method [12]. These

methods, however, tend to have generalization problems given that they generate loose set coverings. The reason is that the MV-set in RKHS is not directly associated with the MV-set in the original space.

In this article, we consider the open set recognition problem using the 1-class and binary SVM as a basis and introduce a new formulation to solve the problem with respect to generalization/specialization. While it is possible that a density estimator (such as [10], [12], [13], [14]) could be used instead of the SVM, we restrict our focus in this article to linear kernel machines. SVM has a number of desirable traits for this work: its solutions are global and unique; it has a simple geometric interpretation; it does not depend on the dimensionality of the input space; and, very important given our definitions, uses structural risk minimization. And it has been considered for open set recognition before.

### 3.1 SVM Approches to Open Set Recognition

The 1-class SVM introduced by Schölkopf et al. [11] adapts the familiar SVM methodology to the open set recognition problem. With the absence of a second class in the training data, the origin defined by the kernel function serves as the only member of a “second class”. The goal then becomes to find the best margin with respect to the origin. The resulting function  $f$  after training takes the value +1 in a region capturing most of the training data points, and -1 elsewhere.

Let  $p(x)$  be the probability density function estimated from the training data  $\{x_1, x_2, \dots, x_m \mid x_i \in X\}$ , where  $X$  is a single class. A kernel function  $\Psi : X \rightarrow H$  transforms the training data into a different space. To separate the training data from the origin, the algorithm solves the following quadratic programming problem for  $w$  and  $\rho$  to learn  $f$ :

$$\min \frac{1}{2} \|w\|^2 + \frac{1}{\nu m} \sum_{i=1}^l \xi_i - \rho \quad (6)$$



subject to

$$(w \cdot \Psi(x_i)) \geq \rho - \xi_i \quad i = 1, 2, \dots, m \quad \xi_i \geq 0 \quad (7)$$

where  $\rho$  is an offset that parameterizes the hyperplane in the feature space defined by the kernel  $\Psi$ , and  $\xi_i$  are slack variables. The kernel function  $\Psi$  impacts density estimation and smoothness. The regularization parameter  $\nu \in (0, 1]$  controls the trade-off between training classification accuracy and the smoothness term  $\|w\|$ , and also impacts the choice and number of support vectors. In the 1-class SVM,  $p(x)$  is cut by the margin plane minimizing Eq. 6 and satisfying Eq. 7. Regions of  $p(x)$  above the margin plane define positive classification and capture most of the training data. As some researchers have pointed out in the literature [15], the 1-class SVM does not provide particularly good generalization or specialization ability, which has limited its use.

While not as much of an issue for binary SVMs, using Radial Basis Function (RBF) kernels, especially with a large  $\gamma$ , can also lead to over specialization. This can occur when “abusing” a 1-class SVM by performing grid search over the parameters and then testing with all available positive and negative examples for a given data set. While still formally a 1-class SVM, since only positive data is used for fitting, the optimization of class parameters to avoid negative training examples from the entire data set is not appropriate.

The 1-class SVM has received some (albeit limited) attention in the computer vision literature – mostly in the areas of image retrieval and face recognition. The application of 1-class SVMs to problems in computer vision was first made by Chen et al. [16] a decade ago. For binary classification, equal treatment is usually given to positive and negative training examples. However, Chen et al. argue that for image retrieval, while it is reasonable to assume that positive training examples cluster in a certain way, the same cannot be said about negative examples, since they can belong to *any* class. Thus, for an open set problem, it seems natural to consider a 1-class SVM, which is trained using only positive examples for a target class. The feasibility of this approach was shown by Chen et al. [16] (and in subsequent works [17] [18]), but with a caveat: kernel and parameter selection. Zhou and Huang note [15] that RBF and other Gaussian kernels are commonly used for 1-class SVMs, often leading to an “over-fitting” of the training data, with kernel parameters selected in an *ad hoc* manner, resulting in an overall lack of generalization to many classes. We believe the lack of generalization and specialization, combined with the common practice of closed set testing, are the primary reasons that the 1-class SVM did not gain much traction in vision.

Detection, as noted in the introduction, is an important open set problem, and several SVM techniques have been proposed to address it in that context. An interesting approach was presented by Hongliang et al. [19], where 1-class SVMs are used for face detection. By choosing to optimize the data used to train a 1-class SVM through subset selection and inclusion of *negative examples as positive*, they improved the generalization. This partially addresses the concerns of Zhou and Huang [15] at the training stage, but is logically inconsistent with no theoretical support. Using a 1-class SVM

trained with samples from the positive class and a few outlier cases, Wu and Ye [20] attempt to maximize the margin between the positive volume defined by a Gaussian kernel and the outliers for the task of novelty detection. This situation is similar to the approach proposed in this paper, with the following key differences:

- Our training data consists of a larger sampling of known data, instead of just a few outlier cases
- We consider a balanced risk formulation after SVM training
- We pursue a linear kernel approach that applies to both the 1-class and binary SVM

Revisiting binary SVMs for detection tasks, Malisiewicz et al. [21] note that a large ensemble of classifiers for a particular class trained with a single positive example and millions of negative examples yields surprisingly good generalization. In this article, we look at specific instances where more limited samplings of training data are assumed to be available, especially with respect to the known classes, where an approach like [21] can’t easily be applied.

Beyond computer vision, 1-class SVMs have been considered in several other areas within pattern recognition, often implicitly to address open set recognition but without formal definitions of that problem. One of the earliest and best works is that of Manevitz and Yousef [22], which considers the problem of document classification. Using a 1-class SVM and a novel variation based on more strict outlier detection, the authors show high levels of classification accuracy on a standard document classification data set (Reuters). Manevitz and Yousef, like Zhou and Huang [15], point out that accuracy is quite sensitive to the choice of kernel and parameters, which they note is not well understood for this problem. Our own work [23] has considered 1-class SVMs for an open set analysis of literary style.

The field of speech processing has also considered 1-class SVMs for problems with unknown classes. In Shen and Yang’s work [24], a novel data description kernel based on the 1-class SVM is developed for text-dependent speaker verification. Kadri et al. [25] successfully apply 1-class SVMs to audio stream segmentation to overcome the problem of overlapping speech and very short speaker changes by maximizing the generalized likelihood ratio with respect to any probability distribution of the speech windows. Rossignol and Pietquin [26] use a 1-class SVM approach for audio segmentation in the context of overlapping speech. In a follow-up work to [25], Rabaoui et al. [27] move beyond stream segmentation to consider speech classification for recognition tasks.

### 3.2 Other Approaches to Open Set Recognition

Outside of the SVM framework, there are several other approaches that can apply to the open set problem, though they do not specifically address it. Recently, the vision community has produced some efforts to deal with the expressiveness and learnability of object models as well as the need for increasing amounts of training data [28]. Indeed, some work has been introduced to address the problem of object classification when training and test classes are disjoint (that is,

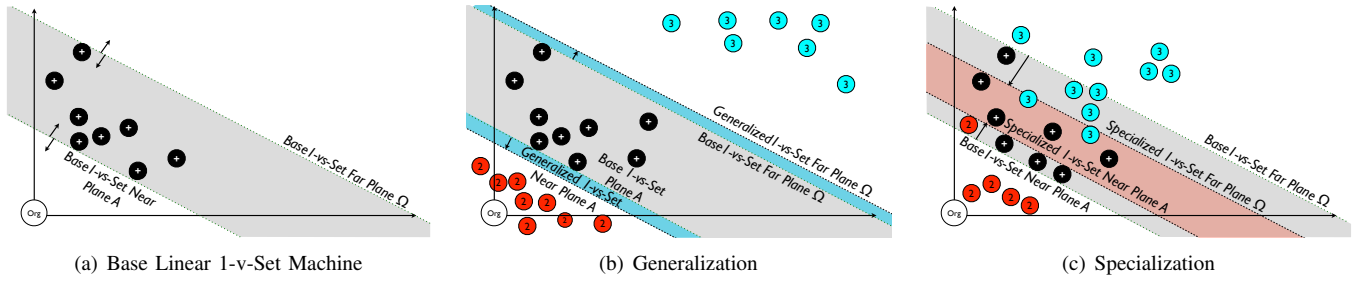


Fig. 4. Example of linear 1-vs-Set Machine showing the (a) base model, (b) the generalization, and (c) the specialization operators. Blue refers to generalization, red for specialization and gray for the base linear 1-vs-Set Machine.

no training examples of the target classes are available). In this direction, researchers have explored knowledge transfer for object class recognition such as: hierarchical structure of the object class space imposed by a general-to-specific ordering [29]; intermediate layer of descriptive attributes to represent object classes [30]; and direct similarity computations between known object classes [31]. In the machine learning literature there is also some work in this direction such as zero-shot [32] and one-shot [33] learning techniques. To deal with a classification problem for which no training data is available for some classes, all of these approaches need to introduce a coupling between known and unknown classes. According to Lampert et al. [30], given that training data for the unobserved classes is not available, this coupling cannot be learned from samples and often needs to be inserted into the system by human effort.

It is useful to point out the differences between these types of approaches and the one we discuss in this article. In the open set recognition problem, we have training samples for the class of interest and samples for only some of the negative classes. However, in our solution to this problem, there is no need for any coupling between known and unknown classes, neither is any human effort required. Some of these above approaches have formal definitions, but without constraints on smoothness or data accuracy. It should be possible to develop a formalization that combines the related definition for open set recognition with categories, which would formalize the open set variations of these problems.

Finally, the open set recognition problem we consider in this article is also different from general unsupervised and semi-supervised learning techniques (described in [34], [35]). Common unsupervised techniques (such as clustering) do not address the formal definition of the open set problem, which is a more precise labeling than just an identification of groups with similar appearance within a large collection of images [36]. We want to make full use of the available training examples that we have. In addition, semi-supervised learning, which aims at the development of techniques to take advantage of both labeled and unlabeled samples [36], also does not apply to our problem since we are not propagating labels from the known samples to the unknown ones. In fact, our objective is to minimize the total recognition error (Eq. 4) for the class of interest as we discussed in Sec. 2. Any solution to the open set recognition problem could be applied as a tool

in semi-supervised learning, but the criterion for evaluation might be significantly different.

## 4 INTRODUCING THE 1-VS-SET MACHINE

Our initial approach for the open set problem is based on a new variant of SVM that we call the 1-vs-Set machine. As we described in Sec. 2, the risk minimization inherent in solving the open set problem involves a minimization of the positive labeled region to address overgeneralization risk combined with margin constraints to minimize overspecialization risk. In this article, we introduce a formulation with a linear kernel that applies to both 1-class and binary SVMs. Since the open set recognition problem is directly related to human cognition, arguments can be made in favor of linear kernels as an idealized discriminator with biological grounding [37], [38]. Moreover, in our experience, linear kernels produce better results than non-linear kernels for the same open set data (we show this in Sec. 5).

The initial definitions of the 1-class SVM were based on RBF kernels, but multiple works can be found [22], [39] that use a 1-class SVM with linear kernels. Once the equations for the 1-class SVM are defined, as in Sec. 3.1, the minimization problem is still well defined for a linear kernel. Intuitively, 1-class linear SVMs can be viewed as taking all positive data, finding the plane that just touches the support vectors and has the origin on the opposite side of the plane from the training data. For binary SVMs, the linear kernel is a typical choice, and is often used for the open set problem of detection [3], [21]. Here we describe the details of the 1-vs-Set algorithm.

### 4.1 Solving the Optimization Problem

Our open set concept suggests that there is risk from including excess “volume” far from the positive data. This must be formalized before the risk can be optimized. Many features and machines have formally bounded data and so their volumes are well defined and finite (such is the case with an RBF kernel). For the infinite volumes that occur with linear SVMs in unbounded feature spaces, we can consider relative volume between two plane configurations during our optimization by using the volume of the positive subspace of a machine intersected with a sphere of large radius, taking the limit as the radius grows. When comparing the two plane configurations, one can take the ratio of the intersected volumes and then the

limit. To be precise, consider two volumes  $V_1$  and  $V_2$ . Let  $S$  be a sphere of radius  $r$ . We can consider relative volume risk as

$$R_v = \lim_{r \rightarrow \infty} \frac{(V_1 \cap S)}{(V_2 \cap S)} \quad (8)$$

If  $R_v < 1$  then risk from  $V_1$  is smaller; if  $R_v > 1$  then risk from  $v_2$  is smaller; and if  $R_v = 1$  the volume risks are essentially equivalent. Unfortunately, different half-spaces are all equivalent in relative volume risk, so to manage risk we must add more constraints.

We define the class of functions  $\mathcal{H}$  for the 1-vs-Set linear kernels in  $k$ -dimensions to be the volume between two parallel  $k$ -dimensional hyperplanes ( $A$  and  $\Omega$  introduced in Sec. 1). We initialize the planes to just contain all positive training data. We can generalize beyond the initial training data by further separating the planes, or, we can specialize by moving either of the planes, bringing them closer together. For first order approximation, the volume is given by separation distance between the planes, and the initial optimization just uses parameters based on plane separation. In particular, we define the overgeneralization risk as the expansion of plane distance:  $\frac{d_\Omega - d_A}{d^+}$ , where  $d_A$  is the marginal distance of the near plane,  $d_\Omega$  is the marginal distance of the far plane, and  $d^+$  is the separation needed to account for all positive data. In a similar manner, we define risk for overspecialization as  $\frac{d^+}{d_\Omega - d_A}$ . During the optimization, these two terms are balanced with the empirical risk (Eq. 3) determined by classifying the available training samples with respect to the original margin. The decision to limit the growth based on the spacing of the intra-class data is our initial solution to balance overgeneralization risk from false positives if we added large volumes, with the need to generalize to avoid future false negatives. In the margin spaces  $\omega_A$  around the near plane and  $\omega_\Omega$  around the far plane, we allow user specified control with parameters  $p_A$  and  $p_\Omega$ . We provide for these additional refinements (described in Sec. 4.2) because only the user can predict the openness of the problem. Combining the overgeneralization and overspecialization risk, along with any specified refinement, our open space risk is:

$$R_L = \frac{d_\Omega - d_A}{d^+} + \frac{d^+}{d_\Omega - d_A} + p_A \omega_A + p_\Omega \omega_\Omega \quad (9)$$

Given these definitions the exact optimization process for the 1-class and binary machines is detailed in Algs. 1 and 2.

Fig. 4 illustrates this process. The base linear 1-vs-Set machine, shown in Fig. 4(a), will just touch the extremes of the positive examples. We then turn to greedy optimization to move the planes simultaneously. If the negative examples overlap the base volume, the overspecialization risk will be 1, and the overgeneralization risk term and probably the data error term  $\mathcal{D}_F$  will require the planes to move inward as in Fig. 4(c). If all negative training classes are outside that volume, as in Fig. 4(b), the overspecialization risk terms will counteract the open space risk term and move the planes to generalize, as in Fig. 4(b).

## 4.2 Refining with Near and Far Plane Pressures

Alg. 1 will result in an optimization where each plane is on a decision score from  $f$ . This is followed by a fine tuning

---

### Algorithm 1 Linear 1-vs-Set Machine Risk Optimization

---

**Require:** Parameter  $\lambda_r$ ; Optional parameters  $\alpha, \beta$

**Require:** Positive features  $\hat{V} = \{v_1, v_2, \dots, v_m \mid v_i \in \mathcal{P}\}$

**Require:** Negative features  $\hat{K} = \{k_1, k_2, \dots, k_n \mid k_i \in \mathcal{K}_j, 1 < j \leq c\}$ , for other known classes  $K_1, \dots, K_c$

```

1: procedure TRAIN( $\lambda_r, \alpha, \beta, \hat{V}, \hat{K}$ )
2:   if 1-class then Train a linear SVM  $f$  using  $\hat{V}$ 
3:   else Train a linear SVM  $f$  using  $\hat{V}, \hat{K}$ 
4:   end if
5:   for  $\forall u_i \in \hat{V}, \hat{K}$  do
6:     Classify  $\eta_i = f(u_i)$   $\triangleright$  Generate decision scores
7:   end for
8:    $\hat{s} = \text{sort}(\hat{\eta})$   $\triangleright$  Sort decision scores
9:    $s_k = \min(|s_1 \dots s_{m+n}|)$ 
10:   $s_j = \max(f(\hat{V}))$ 
11:   $A = \text{margin plane of } f$ 
12:   $\Omega = \text{plane parallel to } A \text{ at } s_j$ 
13:  Greedy Optimization iteratively move  $A$  to  $s_{k+1}$  or
     $s_{k-1}, \Omega$  to  $s_{j-1}$  or  $s_{j+1}$  to minimize  $R_L(f) + \lambda_r \mathcal{D}_F$ , while
    satisfying any constraints provided by  $\alpha, \beta$  in Eq. 5.
14: end procedure

```

---

to place each plane in between the point isolated during optimization and the next closest positive or negative point, with a special case when the plane is at an extreme of the data. We refine the plane positions, generalizing or specializing from the margin between the closest data and the plane based on parameterized “pressures”  $p_A$  and  $p_\Omega$  that control how far to move the plane between the decision scores. If a decision score is the extreme, then we cannot really define a margin-based refinement. This is a relatively common case for  $\Omega$ . When this occurs, we limit the generalization to be the user-specified pressure times the positive data width  $d^+$ . The procedure for using pressures to refine positions is detailed in Alg. 2.

The parameterized pressures impact how much specialization versus generalization to apply. When considering the risk from excess volume we note that the near plane is likely to have any unknown negative data impinge on or near the positive boundary. For the far plane, it is more likely that added positive data will be slightly beyond the existing data, while negatives may not be so close. Thus we provide separate pressures so users may specialize on the near plane while generalizing on the far plane. In our experiments, we typically had better results after applying Algs. 1 and 2 when the near plane specialized with respect to the normal SVM margin, while the far plane generalized from the initial optimization result. We note, however, that this is partially just semantics as any position of the far plane is really a specialization with respect to a standard SVM, which could be viewed as a “far plane” at infinity. *When addressing open set problems, the risk of the unknown is reduced by specializing and restricting the volume to be closer to the positive examples.*

Finally, from the learned model  $f$ , and the refined planes  $A$  and  $\Omega$ , any test vector  $t_x$  can be classified using Alg 3. In the software implementation, we sort the distances and search from the base position to optimize  $R_L(f) + \lambda_r \mathcal{D}_F$ . The code



**Algorithm 2** Linear 1-vs-Set Machine Plane Refinement

---

**Require:** Linear SVM  $f$  trained in Alg. 1  
**Require:** Planes  $A$  and  $\Omega$  from Alg. 1  
**Require:** Near and far plane pressures  $p_A$  and  $p_\Omega$   
**Require:** Counts of positive and negative features  $m, n$   
**Require:** Sorted decision scores  $\hat{s}$

```

1: procedure REFINE( $f, A, \Omega, p_A, p_\Omega, m, n, \hat{s}$ )
2:   Let  $i$  be the index of decision score  $s_i$  touching  $A$ 
3:   if  $i > 0$  then Shift  $A$  to  $s_i(\frac{1}{2} - p_A) + s_{i-1}(p_A - \frac{1}{2})$ 
4:   else Shift  $A$  to  $s_0 - p_A d^+$   $\triangleright$  No Margin, just generalize
5:   end if
6:   Let  $j$  be the index of decision score  $s_j$  touching  $\Omega$ 
7:   if  $j < (m+n)$  then Shift  $\Omega$  to  $s_j(p_\Omega - \frac{1}{2}) + s_{j+1}(\frac{1}{2} - p_\Omega)$ 
8:   else Shift  $\Omega$  to  $s_{m+n} + p_\Omega d^+$   $\triangleright$  No Margin, just generalize
9:   end if
10: end procedure

```

---

**Algorithm 3** 1-vs-Set Machine Prediction

---

**Require:** Test feature vector  $t_x$   
**Require:** Linear SVM  $f$  trained in Alg. 1  
**Require:** Planes  $A$  and  $\Omega$  from Alg. 2

```

1: function PREDICT( $t_x, f, A, \Omega$ )
2:   if ( $A \leq f(t_x)$  and  $f(t_x) \leq \Omega$ ) then Return +1
3:   else Return -1
4:   end if
5: end function

```

---

also supports setting fixed recall or precision, which is easy to implement given the explicit optimization process that satisfies both Def. 1 and Eq. 5. Since we are using an extension of the LIBSVM [40] library, and sorting, our implementation is non-optimized, but the overall complexity of the linear 1-vs-Set Machine can be made  $O(n)$  for  $n$  data items by using the ideas of Joachims in [41], and simple selection to find the points close to the near and far planes.

## 5 EXPERIMENTAL ANALYSIS

An important goal of our experiments is to highlight the radically different nature of data sets once they are recontextualized to reflect an open set problem. Torralba and Efros have recently noted that “Indeed, some datasets, that started out as data capture efforts aimed at representing the visual world, have become closed worlds unto themselves” [3]. They go on to analyze the various biases that exist in popular data sets, which are easily learned and leveraged to inflate recognition accuracies in a closed set scenario. By considering these same sets as open set problems in a cross-data context, we can directly address the problem of negative set bias (what the data set considers to be “the rest of the world” [3]). Here we propose testing scenarios that are more aligned with real world scenarios where we do not have knowledge of all classes.

For the object recognition experiments presented in Sec. 5.1, we make use of two different feature approaches. The first

approach is the popular Histogram of Oriented Gradients (HOG) [4] descriptor, which is commonly used for detection problems. Applying the standard procedure described by Dalal and Triggs, we produced a 3,780-dimension feature vector for each image considered in our experiments below. In the second approach, the underlying features used for classification are generated by extracting points of interest (PoIs) from the images using Difference of Gaussians as proposed in [42], and then computing an LBP-like [43] feature descriptor in a window around each detected PoI. Feature vectors are composed of 59-dimension histogram bins that summarize the feature descriptor information for each image.

For the face verification experiments presented in Sec. 5.2, we also make use of two different feature approaches. The first approach is the same LBP-like descriptor used for the object recognition experiments, but applied exactly as described for faces in Sapkota et al. [43]. This results in 3,776-dimension histogram bins that are used as feature vectors for learning. The second approach is the common Gabor feature, which has been shown to produce very good results for face verification [44]. Applying the feature process described by Pinto et al. [44], we generate 86,400-dimension feature vectors.

Open set recognition presents a couple of new challenges with respect to evaluation. Specifically, we need to address the choice of which statistic to evaluate classification accuracy, as well as the organization of the data sets. This leads us to a few procedures that are not commonly used in object recognition and face verification. Our experiments below consider several aspects of classification, including the statistical significance of the 1-vs-Set machine’s results, an assessment of the parameter space defined by  $p_A$  and  $p_\Omega$ , and the impact of problem openness on classification accuracy. All 1-vs-Set machines in these experiments follow the most general optimization of Eq. 4, and were not trained with explicit constraints (Eq. 5). The near and far plane pressures are set at default values of  $p_A = 1.6$  and  $p_\Omega = 4$  for all experiments to provide an extra measure of generalization.

Concerning statistics, accuracy is a natural choice for evaluating binary decision classifiers in machine learning. Simply defined, accuracy refers to the correctly classified samples (true positives and true negatives) out of all of the classification decisions. For open set recognition, this measure tends to underemphasize the distinction between correct positive and negative classifications. Remember – we are primarily interested in identifying a small number of positive samples out of a much larger pool of negatives. To highlight this point, consider a case where a classifier returns one true positive out of 100 positive test samples, and zero false positives out of 100,000 negative test samples. This classifier is 99.9% accurate on this test – even though it is essentially a “no” classifier. For this reason, recall and precision are a common alternative.

Recall refers to the amount of correctly classified positive examples with respect to all the available positive examples. Precision refers to the amount of correctly classified positive examples with respect to all of the false and true positives. If we consider precision and recall for the task of comparing different classifiers, we encounter the problem of an “apples to oranges” comparison, where a collection of statistics not fixed

TABLE 2

**Top half:** detailed comparison between all of the different classifiers for the HOG features over the open universe of 88 classes. **Bottom half:** summary comparisons between the 1-vs-Set Machines and all other classifiers for the HOG features over the open universe of 212 classes, and the LBP-like features for both open universes. The numbers are p-values for  $H_0$ : the F-measure of the Alg. on the row is lower than that of the Alg. on the column. p-values in red indicate the 1-vs-Set Machine is significantly better in F-measure, while those in blue indicate a baseline machine is significantly better (95% confidence interval). Dashes represent a worse F-measure, p-values in black indicate a better F-measure that is not significant, and a gray cell means no test was performed (a machine versus itself).

2-tailed paired t-test	binary 1-vs-Set	binary linear	binary RBF	1-class 1-vs-Set	1-class linear	1-class RBF
binary 1-vs-Set (HOG 88)		2.19E-16	2.85E-3	1.97E-33	2.33E-35	1.73E-33
binary linear (HOG 88)	—		—	1.45E-31	7.12E-34	6.31E-32
binary RBF (HOG 88)	—	1.81E-3		8.96E-23	2.33E-24	7.58E-24
1-class 1-vs-Set (HOG 88)	—	—	—		1.43E-11	—
1-class linear (HOG 88)	—	—	—	—		—
1-class RBF (HOG 88)	—	—	—	0.527	2.38E-3	
binary 1-vs-Set (HOG 212)		8.18E-7	0.021	9.56E-16	2.06E-16	4.50E-16
1-class 1-vs-Set (HOG 212)	—	—	—		0.645	0.047
binary 1-vs-Set (LBP-like 88)		9.37E-4	1.17E-14	4.15E-45	4.88E-37	8.99E-18
1-class 1-vs-Set (LBP-like 88)	—	—	—		2.76E-7	—
binary 1-vs-Set (LBP-like 212)		0.03	0.758	8.24E-11	9.54E-34	9.23E-10
1-class 1-vs-Set (LBP-like 212)	—	—	—		5.38E-19	0.186

to a specific precision or recall are present. For example, for the same training and testing data, the 1-vs-Set machine might produce a recall of 75% at a precision of 32%, while a binary SVM produces a recall of 62% at a precision of 25%. While *ad hoc* thresholding could be applied to the decision scores to produce a precision-recall curve, a better way to resolve this issue is to use F-measure, which provides us with a consistent point of comparison across inconsistent precision and recall numbers. In information retrieval and machine learning, F-measure is applied as a combination of precision and recall given by their harmonic mean:

$$\text{F-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

### 5.1 An Evaluation of Object Recognition

The data we consider for open set object recognition follows a cross-data set methodology adapted from [3]. For training, we choose all classes from the Caltech 256 set. For testing, we choose images for the positive class from Caltech 256, but for the negative classes, choose images from ImageNet [45]. Despite the bias within Caltech 256, we wanted to ensure some consistency between training and testing samples for the positive class, while attempting to generalize or specialize to the negatives from ImageNet based on a limited sampling of negatives from Caltech 256. While we have a sense of what our positives are during training, there is no way to know *a priori* if positive classes across data sets are consistent. However, negatives are definitely negative, thus we should be able to handle any examples from any data set based on our optimization. From this data, we construct two “open universes” of different sizes, allowing us to vary the training and testing data, which is somewhat constrained by the number of images provided by both data sets for the same classes. The first open universe consists of 88 classes, where we choose one class as positive,  $n$  classes as open set training data or binary negatives (where  $n$  varies depending on the experiment), and

87 classes as negatives for testing. The second, more open universe, consists of 212 classes, where we choose one class as positive,  $n$  classes as open set training or binary negatives, and 211 classes as negatives for testing.

We follow a multiple trial randomized testing procedure that selects different training, open set training, and testing sets for each experimental trial. This is done to verify consistency in the reported results across numerous trials, thus limiting any misleading impressions outliers might give for a single test. We cycle through all of the classes five times, treating each class positively once per iteration. The individual training and testing sample breakdowns vary as a function of experiment, and are noted below as we describe the individual tests. To ensure a fair comparison, the 1-class One-vs-Set Machine and all of the binary classifiers are trained with the exact same positive and negative examples. The 1-class machines use a default  $\nu$  parameter of 0.5; machines with an RBF kernel use a default  $\gamma$  parameter of one divided by the number of features (LIBSVM default settings). These tests represent a total of 532,400 images for the open universe of 88 classes, and 13,610,400 images for the open universe of 212 classes, in different combinations across all of the randomized tests.

Our primary goal is to establish, in a rigorous statistical manner, the advantage of the 1-class and binary 1-vs-Set machines over typical SVM classifiers for open set recognition problems. We do this by applying a 2-tailed paired t-test [46] over all of the results for the classes from each of the open worlds and for each of our two feature sets to generate summary statistics. The t-test allows us to determine if two sets of classification results differ from each other in a significant way. The resulting p-values are assessed at the 0.05 confidence level (95% confidence). Our null hypothesis  $H_0$  states the F-measures from the first set of classification results are lower than that of the second set. We reject  $H_0$  when the p-value is less than 0.05.

For the open universe test of 88 classes with both the HOG

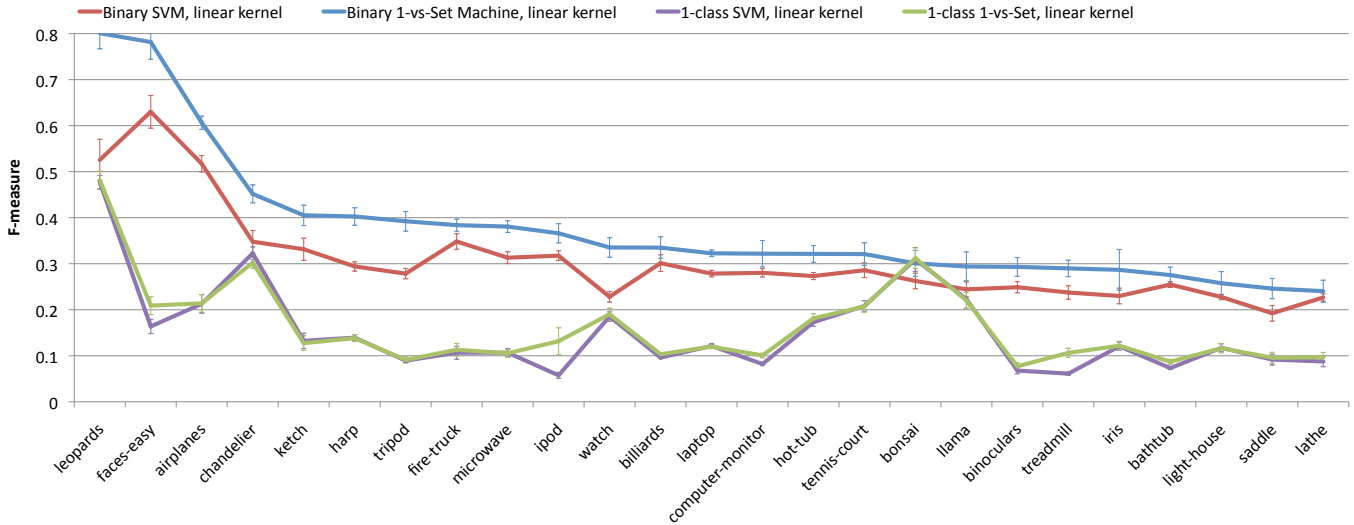


Fig. 5. A comparison between the 1-vs-Set Machine and typical SVMs with a linear kernel. This plot represents detail from the open universe of 88 classes with HOG features test found in Table 2. The classes shown here correspond to the top 25 for the binary 1-vs-Set Machine ranked by F-measure. Error bars reflect standard error. In every case shown, the binary 1-vs-Set Machine produces a higher F-measure compared to a binary SVM. The 1-class 1-vs-Set Machine also shows a gain in F-measure, albeit at far more modest intervals for these particular classes.

and LBP-like features, we train on 70 positive images and 14 negative images each from five other classes (approximately 5% of the available classes), and test on 30 positive images and 435 negative images across all the negative classes. For the more difficult open world of 212 classes, we train on 30 positive images and 5 negative images each from six other classes (approximately 3% of the available classes), and test on 30 positive images and 6,330 negative images across all the negative classes.

The results of these statistical tests are presented in Table 2. In a direct comparison with other machines utilizing a linear kernel, the results for the 1-class and binary 1-vs-Set machines are statistically significant: the null hypothesis is rejected in all but a single case. Even when we move to a cross-kernel comparison with typical machines utilizing RBF kernels, the binary 1-vs-Set machine produces better results that are statistically significant in all but a single case. Thus, we can conclude that for open set recognition problems, the 1-vs-Set machine is a suitable alternative to typical SVM classifiers.

In the next series of experiments, we break-out detail from the 88-class open universe set with HOG features to look at specific aspects affecting classification. First, the reader might be interested in the actual F-measures outside of the statistical summary presented above. Fig. 5 shows the F-measures for the top 25 classes ranked by binary 1-vs-Set machine accuracy. The binary SVM with a linear kernel is also plotted as a baseline comparison for the same classes. In every case shown, the binary 1-vs-Set machine is able to achieve a higher F-measure than the typical binary SVM by solving the constrained minimization problem over the exact same training data. The 1-class 1-vs-Set machine also shows a gain in F-measure, albeit at far more modest intervals for these particular classes. However, the 1-class 1-vs-Set machine should not

be neglected: for 27 of the 88 classes, it produces better F-measures than the binary 1-vs-Set Machine. Compared to closed set problems, these F-measures might seem low, but one must remember that this experiment is far more difficult: both machines saw only 5% of the total classes during training, and all of the classes during testing.

Next we turn to an assessment of the parameter space defined by the near pressure  $p_A$  and far pressure  $p_\Omega$  described in Sec. 4.2. To analyze our results in a broader context, F-measure for this experiment considers all of the true positives, false negatives and false positives across all classes from a series of randomized tests, as opposed to just those from a single class as we did for the experiments above. Using the 88-class open universe set with HOG features, we searched the parameter space using binary 1-vs-Set machines to gain a better understanding of the impact of plane movement. Plotting portions of this data around our default parameters of  $p_A = 1.6$  and  $p_\Omega = 4$  (Fig. 6), we can see that movement on the near and far planes during training is indeed affecting the results achieved during testing. Of particular interest is the impact of moving  $\Omega$  (blue curve in Fig. 6), which limits false positives in what was a positive half-space. A bit too much generalization from the plane established by Alg. 1 (point 7 on the x-axis of the blue curve in Fig. 6) can cause a dip in F-measure.

Finally, we consider the impact of openness on F-measure. Intuitively, when more classes are available during training, we expect that the resulting classifiers should be more accurate. And this is exactly what we observe in practice. For the curves shown in Fig. 7, we chose 60 images as positive testing data for each class, and varied the openness of the test from 42% (30 negative classes seen during training) to 82% (3 negative classes seen during training). Testing remained constant at 30 positive images and 435 negative images. Again, F-measure

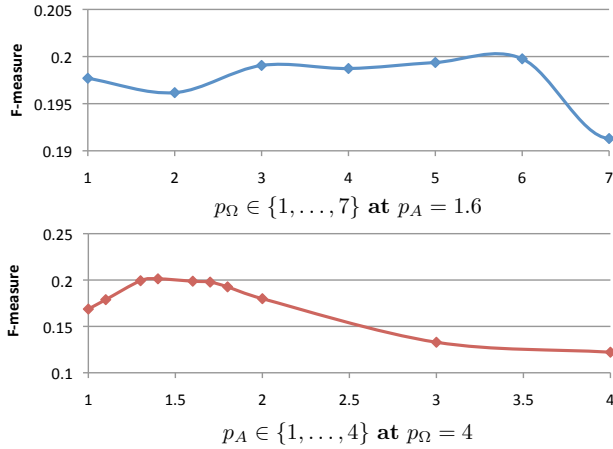


Fig. 6. Examples of the near and far plane pressure parameter space and corresponding F-measures when one of the two parameters is fixed at our selected default. F-measure in this plot is calculated over all of the classes in the open world of 88 classes with HOG features. Notice how movement on the near and far planes during Algs. 1 and 2 makes a difference in the resulting F-measures over the test data.

for this experiment considers all of the true positives, false negatives and false positives across all classes from a series of tests. All three binary classifiers decrease in accuracy as the world grows to be more open. However, even when just a small number of classes are available during training, the 1-vs-Set machine is able to drastically reduce the number of false positives compared to the other machines.

## 5.2 An Evaluation of Face Verification

We also analyze the task of face verification, where people are the classes, with less obvious inter-class variations. We chose to evaluate classes from another well-known and challenging data set: Labeled Faces in the Wild (LFW) [43]. LFW (like many verification sets) is traditionally used for image-pair matching, which is really neither an open nor closed problem based on the *learning* criteria of this article. However, we can still define training and testing sets from it. The twelve people with at least 50 images (providing sufficient training data) were selected as gallery classes. For open set testing, we selected 82 “impostors” from other people in the LFW set, yielding a total of 1,316 test images across all classes.

The impact of problem “openness” is also a very important factor for face verification, where a benchmark test might not reflect the performance of an algorithm over time as more people attempt to verify. Our experiments evaluated this scenario, starting with a completely closed world of twelve people, and adding more impostors in each individual experiment. Once again, we follow a five trial randomized testing procedure that selects different training (35 positive and negative samples per person) and testing sets (14 different test samples per person) for each experiment. Galleries are represented by binary 1-vs-Set machines and binary SVMs with linear kernels, which are

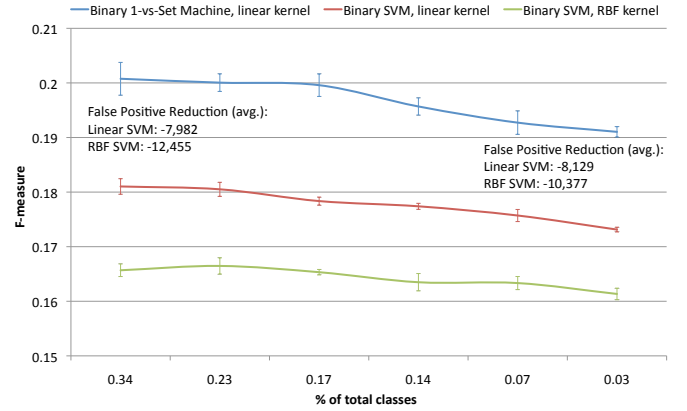


Fig. 7. An assessment of F-measure as a function of openness (growing from left to right) for a collection of binary classifiers. F-measure in this plot is calculated over all of the classes in the open universe of 88 classes with HOG features. As expected, all three machines decrease in accuracy as the universe grows to be more open. Even in the most open setting (82%), the 1-vs-Set Machine yields 8,129 fewer false positives compared to the binary SVM with a linear kernel, and 10,377 fewer false positives compared to the binary SVM with an RBF kernel. All 1-vs-Set Machine results are significantly better at a 95% confidence interval.

trained with the exact same positive and negative examples. The closed set scenario for verification considers only the twelve known people for both the gallery and the probes (the test samples). The four subsequent experiments consider different numbers of *probe* classes to vary openness from 24% (30 probe classes) to 52% (94 probe classes). The gallery remains fixed at twelve people. This is slightly different from the experiments described for object recognition, where we varied the number of training classes, but more consistent with a critical analysis of typical face verification testing.

Fig. 8 shows a comparison between the binary 1-vs-Set machine and binary SVM with a linear kernel across the LBP-like and Gabor features. When the experiment is completely closed, the problem appears easy, with all machines producing high F-measures that are similar. However, as the problem grows to be more open, a large gap in F-measure appears between the 1-vs-Set machine and the binary SVM. Once experiments move beyond closed set testing protocols, it quickly becomes clear that typical machines with strong features are not always sufficient to address the open set aspect of the problem. This is particularly important for face verification, which is used in real world authentication applications. We also evaluated the statistical significance of the 1-vs-Set machine’s results using the 2-tailed paired t-test. In all but a single case (the Gabor features at 0% openness), the results are significantly better at a 95% confidence interval.

## 6 DISCUSSION AND FUTURE WORK

By revisiting the ideas of the general recognition problem and SVM-based recognition systems, we have gained a better



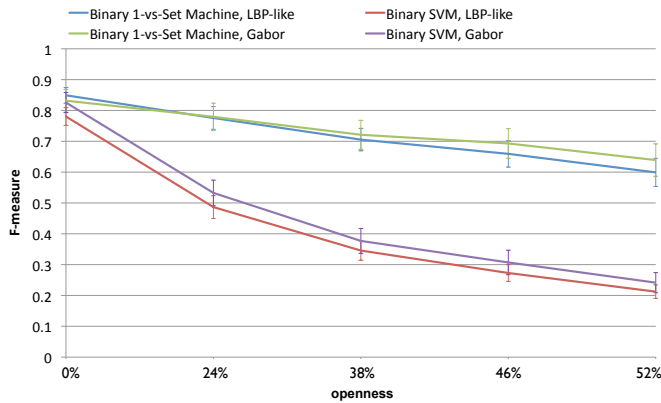


Fig. 8. Face verification results as a function of openness (growing from left to right) for a collection of binary classifiers and LBP-like and Gabor features. F-measure in this plot is calculated over all of the classes in the subset of LFW we consider. Notice that in closed set testing (0%), there is not much difference between the 1-vs-Set Machine and the typical binary classifiers. In all but a single case (the Gabor features at 0%), the 1-vs-Set machine results are significantly better at a 95% confidence interval.

understanding of the challenges of the problem and the shortcomings of the most frequently used solutions. In an open world, we cannot have knowledge of all classes, and it is impossible for us to sample and train on every possible image configuration for a class. Even if we could, with negatives greatly outnumbering positives, choosing representative negative training examples for a binary or multi-class classifier is problematic. The open set assumption changes how we must evaluate what is a “solution” – the risk of unknown classes must be accounted for without causing unforeseen errors.

With this in mind, we formalized the open set recognition problem as a risk-minimizing constrained functional optimization problem. As a first step towards a solution, we introduced a novel “1-vs-Set Machine” as an extension of the 1-class and binary Support Vector Machines to better support generalization and specialization in a manner that is consistent with the open set problem definition. The experiments for object recognition and face verification show that the 1-vs-Set machine is highly effective at improving accuracy when compared to 1-class and binary SVMs for the same problems. Upon this article’s acceptance for publication, we will release our source code for the 1-vs-Set machine, as well the computed features for all experiments.

Torralba and Efros [3] point out the effect of the closed-world assumption: a focus on beating the latest benchmark reports on the newest data set. Many researchers in the vision community have lost the sight of the original purpose of these data sets: *recognizing the visual world*. By reformulating the recognition problem as open set recognition, we naturally avoid over-training biases. An open set testing methodology reduces the chance of data set bias because one cannot train on most of the data. While leave-one-out cross-validation is

popular, the open set formulation suggests a leave-most-out validation. By restructuring tests over existing data sets we hope to encourage researchers to begin to address the more natural open set form of the recognition problem. For instance, our results on an open set reformulation of LFW may not be as impressive as the closed-set testing, but they highlight the actual difficulty of unconstrained face verification.

The next step for this work is to extend our 1-vs-Set machine model to RBF kernels, which have a bounded volume that can also be adapted via generalization or specialization. This includes exploring alternative kernel density estimators outside of an SVM framework. We emphasize that the 1-vs-Set machine is only a first step towards an algorithm that is a truly suitable solution for open set recognition. Specific learning approaches that incorporate open set recognition into their fundamental design (especially at the initial density estimation stage) are of great interest.

## ACKNOWLEDGMENTS

The authors thank the financial support of ONR MURI N00014-08-1-0638, FAPESP 2010/05647-4, Army SBIR W15P7T-12-C-A210, and Microsoft.

## REFERENCES

- [1] R. P. Duin and E. Pekalska, “Open Issues in Pattern Recognition,” in *Computer Recognition Systems*, M. Kurzynski, E. Puchala, M. Wozniak, and A. Zolnierrek, Eds. Springer, 2005, pp. 27–42.
- [2] X. Zhou and T. Huang, “Small Sample Learning during Multimedia Retrieval using BiasMap,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2001.
- [3] A. Torralba and A. A. Efros, “Unbiased Look at Dataset Bias,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2011, pp. 1521–1528.
- [4] N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2005.
- [5] B. Schölkopf and A. J. Smola, *Learning with Kernels*. The MIT Press, 2002.
- [6] P. Phillips, P. Grother, and R. Micheals, “Evaluation Methods on Face Recognition,” in *Handbook of Face Recognition*, A. Jain and S. Li, Eds. Springer, 2005, pp. 329–348.
- [7] L. Fayin and H. Wechsler, “Open Set Face Recognition Using Transduction,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1686–1697, Nov. 2005.
- [8] A. Rocha and S. Goldenstein, “From Binary to Multi-class: Divide to Conquer,” in *Intl. Conference on Computer Vision Theory and Applications*, Lisbon, Portugal, February 2009, pp. 323–330.
- [9] B. Raskutti and A. Kowalczyk, “Extreme Re-balancing for SVMs: a Case Study,” *ACM SIGKDD Explorations*, vol. 6, no. 1, pp. 60–69, 2004.
- [10] C. Park, J. Z. Huang, and Y. Ding, “A Computable Plug-In Estimator of Minimum Volume Sets for Novelty Detection,” *Operations Research*, vol. 58, pp. 1469–1480, September 2010.
- [11] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, “Estimating the Support of a High-dimensional Distribution,” Microsoft Research, Tech. Rep. MSR-TR-99-87, 1999, available at: <http://ftp.research.microsoft.com/pub/tr/tr-99-87.pdf>.
- [12] D. M. Tax and R. P. Duin, “Support Vector Domain Description,” *Pattern Recognition Letters*, vol. 20, pp. 1191–1199, November 1999.
- [13] X. Wang, P. Tino, M. A. Fardal, S. Raychaudhury, and A. Babul, “Fast Parzen Window Density Estimator,” in *IEEE Intl. Joint Conference on Neural Networks*, 2009, pp. 3267–3274.
- [14] E. Parzen, “On Estimation of a Probability Density Function and Mode,” *The Annals of Mathematical Statistics*, vol. 33, pp. 1065–1076, September 1962.
- [15] X. Zhou and T. Huang, “Relevance Feedback in Image Retrieval: A Comprehensive Review,” *Multimedia Systems*, vol. 8, no. 6, pp. 536–544, 2003.



- [16] Y. Chen, X. Zhou, and T. Huang, "One-class SVM For Learning in Image Retrieval," in *IEEE Conf. on Image Processing*, 2001, pp. 34–37.
- [17] C. Zhang, X. Chen, M. Chen, and S. Ching, "A Multiple Instance Learning Approach for Content Based Image Retrieval Using One-Class Support Vector Machine," in *IEEE Intl. Conf. on Multimedia and Expo.*, 2005, pp. 1142–1145.
- [18] K. Goh, E. Change, and B. Li, "Using One-Class and Two-Class SVMs for Multiclass Image Annotation," *IEEE Trans. on Knowledge and Data Engineering*, vol. 17, no. 10, pp. 1333–1346, Oct. 2010.
- [19] J. Hongliang, L. Qingshan, and L. Hanqing, "Face Detection Using One-Class-Based Support Vectors," in *IEEE Automated Face and Gesture Recognition*, 2004, pp. 457–462.
- [20] M. Wu and J. Ye, "A Small Sphere and Large Margin Approach for Novelty Detection Using Training Data with Outliers," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 2088–2092, Nov. 2009.
- [21] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of Exemplar-SVMs for Object Detection and Beyond," in *Intl. Conf. on Computer Vision*, 2011.
- [22] M. Manevitz and M. Yousef, "One-class SVMs for Document Classification," *Journal of Machine Learning Research*, vol. 2, pp. 139–154, March 2002.
- [23] C. Forstall, S. Jacobson, and W. Scheirer, "Evidence of Intertextuality: Investigating Paul the Deacon's *Angustae Vitae*," *Literary & Linguistic Computing (LLC)*, vol. 26, no. 3, pp. 285–296, Sept. 2011.
- [24] Y. Shen and Y. Yang, "A Novel Data Description Kernel Based on One-Class SVM for Speaker Verification," in *Intl. Conf. on Acoustics, Speech and Signal Processing*, 2007, pp. 489–492.
- [25] H. Kadri, M. Davy, A. Rabaoui, Z. Lachiri, and N. Ellouze, "Robust Audio Speaker Segmentation Using One Class SVMs," in *EURASIP European Signal Processing Conf.*, 2008.
- [26] S. Rossignol and O. Pietquin, "Single-Speaker/Multi-Speaker Co-Channel Speech Classification," in *Conf. of the International Speech Communication Association*, 2010.
- [27] A. Rabaoui, H. Kadri, Z. Lachiri, and N. Ellouze, "One-Class SVMs Challenges in Audio Detection and Classification Applications," *EURASIP Journal on Advances in Signal Processing*, 2008.
- [28] M. Rohrbach, M. Stark, and B. Schiele, "Evaluating Knowledge Transfer and Zero-Shot Learning in a Large-Scale Setting," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2011, pp. 1641–1648.
- [29] A. Zweig and D. Weinshall, "Exploiting Object Hierarchy: Combining Models from Different Category Levels," in *Intl. Conf. on Computer Vision*, 2007, pp. 1–8.
- [30] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2009, pp. 951–958.
- [31] E. Bart and S. Ullman, "Single-example Learning of Novel Classes Using Representation by Similarity," in *British Machine Vision Conference*, 2005, pp. 951–958.
- [32] M. Palatucci, D. Pomerleau, G. Hinton, and T. M. Mitchell, "Zero-Shot Learning with Semantic Output Codes," in *Neural Information Processing Systems*, 2009, pp. 1–9.
- [33] L. Wolf, T. Hassner, and Y. Taigman, "The One-shot Similarity Kernel," in *Intl. Conf. on Computer Vision*, 2009.
- [34] M. G. Quiles, Z. Liang, F. Breve, and A. Rocha, "Label Propagation Through Neuronal Synchrony," in *IEEE Intl. Joint Conference on Neural Networks*, Barcelona, Spain, July 2010, pp. 2517–2524.
- [35] G. Heidemann, "Unsupervised Image Categorization," *Image and Vision Computing*, vol. 23, no. 10, pp. 861–876, October 2004.
- [36] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. Springer, 2006.
- [37] T. Serre and T. Poggio, "A Neuromorphic Approach to Computer Vision," *Communications of the ACM*, vol. 53, no. 10, pp. 54–61, 2010.
- [38] J. DiCarlo and D. Cox, "Untangling Invariant Object Recognition," *Trends in Cognitive Science*, vol. 11, no. 8, pp. 333–341, Aug. 2007.
- [39] T. Onoda, H. Murata, and S. Yamada, "Non-relevance Feedback Document Retrieval Based on One Class SVM and SVDD," in *IEEE Intl. Joint Conference on Neural Networks*, 2006, pp. 1212–1219.
- [40] C. Chang and C. J. Lin, "LIBSVM: a Library for Support Vector Machines," 2001, available at: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [41] T. Joachims, "Training Linear SVMs in Linear Time," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2006, pp. 217–226.
- [42] D. Lowe, "Distinctive Image Features From Scale-Invariant Keypoints," *Intl. Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [43] A. Sapkota, B. Parks, W. Scheirer, and T. Boulton, "FACE-GRAB: Face Recognition with General Region Assigned to Binary Operator," in *IEEE Intl. Workshop on Biometrics*, 2010.
- [44] N. Pinto, J. J. DiCarlo, and D. D. Cox, "How Far Can You Get with a Modern Face Recognition Test Set Using Only Simple Features?" in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.
- [45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.
- [46] NIST, *NIST/SEMATECH e-Handbook of Statistical Methods*, ser. 33. U.S. GPO, 2008.



**Walter J. Scheirer** received his M.S. degree in computer science from Lehigh University (2006), and his Ph.D. in Engineering, from the University of Colorado at Colorado Springs (2009). He is currently the director of research and development at Securics, Inc., and an assistant professor adjunct at the University of Colorado at Colorado Springs. His primary research interests include computer vision, machine learning, biometrics, and digital humanities.



**Anderson de Rezende Rocha** received his B.Sc (Computer Science) degree from Federal Univ. of Lavras (UFLA), Brazil in 2003. He received his M.S. and Ph.D. (Computer Science) from Univ. of Campinas (Unicamp), Brazil, in 2006 and 2009, respectively. Currently, he is an assistant professor in the Institute of Computing, Unicamp, Brazil. He is a Microsoft Research Faculty Fellow. His interests include digital image and video forensics, pattern analysis, machine learning, and general computer vision.



worked in various industry roles for 3 years.

**Archana Sapkota** is a doctoral candidate in the Department of Computer Science at University of Colorado at Colorado Springs. She works as a Research Assistant at the Vision and Security Technology Lab. (VAST) under the supervision of Dr. T. E. Boulton. Her research interests include: Face Recognition, Computer vision, Pattern Analysis, and Biometrics. Prior to joining UCCS, she received a bachelor's degree in computer science and engineering from Indian Institute of Technology, Roorkee, India and



**Terrance E. Boulton**, El Pomar Professor of Innovation and Security at the University of Colorado at Colorado Springs had published over 180 Papers and holds 9 patents (8 pending). Prior to joining UCCS, Dr. Boulton held professorships at Lehigh and Columbia universities. He is also CEO/CTO of Securics Inc., a company in the biometrics and security space. Dr. Boulton has served as an Assoc. Editor for TPAMI, has been the PAMI-TC chair and is a member of the IEEE Golden Core Society.