Towards Open World Recognition

Abhijit Bendale, Terrance Boult
University of Colorado at Colorado Springs
{abendale,tboult}@vast.uccs.edu

Abstract

With the of advent rich classification models and high computational power visual recognition systems have found many operational applications. Recognition in the real world poses multiple challenges that are not apparent in controlled lab environments. The datasets are dynamic and novel categories must be continuously detected and then added. At prediction time, a trained system has to deal with myriad unseen categories. Operational systems require minimal downtime, even to learn. To handle these operational issues, we present the problem of ***Open World Recognition*** *and formally define it. We prove that thresholding sums of monotonically decreasing functions of distances in linearly transformed feature space can balance "open space risk" and empirical risk. Our theory extends existing algorithms for open world recognition. We present a protocol for evaluation of open world recognition systems. We present the Nearest Non-Outlier (NNO) algorithm that evolves model efficiently, adding object categories incrementally while detecting outliers and managing open space risk. We perform experiments on the ImageNet dataset with 1.2M+ images to validate the effectiveness of our method on large scale visual recognition tasks. NNO consistently yields superior results on open world recognition.*

1 Introduction

Over the past decade, datasets for building and evaluating visual recognition systems have increased both in size and variation. The size of datasets has increased from a few hundred images to millions of images, and the number of categories within the datasets has increased from tens of categories to more than a thousand categories. Co-evolution of rich classification models along with advances in datasets have resulted in many commercial applications [8, 40, 29]. A multitude of operational challenges are posed while porting recognition systems from controlled lab environments to the real world. A recognition system in the "open world" has to continuously update with additional object categories and be robust to unseen categories and have minimum downtime. Despite the obvious dynamic and open nature of the world, a vast majority of recognition systems assume a *static* and closed world model of the problem where all categories are known a priori. To address these operational issues, this paper formalizes and presents steps towards the problem of open world recognition. The key steps of the problem are summarized in Fig. 1.

As noted by [34], "when a recognition system is trained and is operational, there are finite set of known objects in scenes with myriad unknown objects, combinations and configurations – labeling something new, novel or unknown should always be a valid outcome." One reason for the domination of "closed world" assumption of today's vision systems is that matching, learning and classification tools have been formalized as selecting the most likely class from a closed set. Recent research [34, 33, 15], has re-formalized learning for recognition as open set recognition. However, this approach does not explicitly require that inputs be as known or unknown. In contrast, for open world recognition, we propose the system explicitly label novel inputs as unknown and then incrementally incorporate them into the classifier. Furthermore, open set recognition as formulated by [34] is designed for traditional one-vs-all batch learning scenario. Thus, it is open set but not incremental and does not scale gracefully with the number of categories.

While there is a significant body of work on incremental learning algorithms that handle new instances of known classes [3, 4, 45], open world requires two more general and difficult steps: continuously detecting novel classes and when novel inputs are found updating the system to include these new classes in its multi-class open set recognition algorithm. Novelty detection and outlier detection are complex issues in their own right with long histories [26, 14] and they are still active vision research topics [2, 25]. After detecting a novel class, the requirement to add new classes leaves the system designer with the choice of re-training the entire system. When the number of categories are small, such a solution may be feasible, but unfortunately, it does not scale. Recent studies on ImageNet dataset using SVMs or CNN require days to train their system [30, 17], e.g. 5-6 CPU/GPU days in case of CNN for 1000 category image

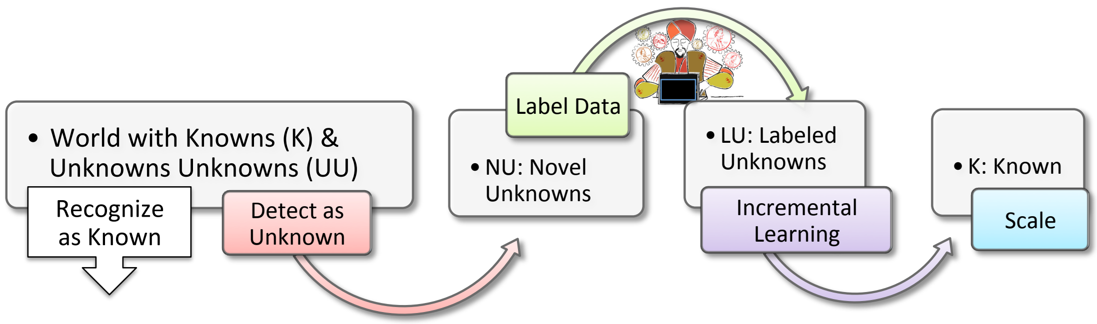


Figure 1: In open world recognition, the system must be able to recognize objects and associate them with known classes while also being able to label classes as unknown. These "novel unknowns" must then be collected and labeled (e.g. by humans). When there are sufficient labeled unknowns for new class learning, the system must incrementally learn and extend the multi-class classifier, thereby making each new class "known" to the system. Open World recognition moves beyond just being robust to unknown classes and toward a scalable system that is adapting itself and learning in an open world.

classification task. Distance based classifiers like Nearest Class Mean (NCM) [16, 28, 32] offer a natural choice for building scalable system that can learn new classes incrementally. In NCM-like classifiers, incorporating new images or classes in implies adjusting the existing means or updating the set of class means. However, NCM classifier in its current formulation is not suited for open set recognition because it uses close-set assumptions for probability normalization. Handling unknowns in open world recognition requires gradual decrease in the value of probability (of class membership) as the test point moves away from known data into open space. The softmax based probability assignment used in NCM does not account for open space.

The *first contribution* of this paper is a formal definition of the problem of open world recognition, which extends the existing definition of open set recognition which was defined for a static notion of set. In order to solve open world recognition, the system needs to be robust to unknown classes, but also be able to move through the stages and knowledge progression summarized in Fig. 1. *Second contribution* of the work is a recognition system that can continuously learn new object categories in an open world model. In particular, we show how to extend Nearest Class Mean type algorithms (NCM) [28], [32], to a Nearest Non-Outlier (NNO) algorithm that can balance open space risk and accuracy.

To support this extension, our *third contribution* is showing that thresholding sums of monotonically decreasing functions of distances of linearly transformed feature space can have arbitrarily small "open space risk". Finally, we present a protocol for evaluation for open world recognition, and use this protocol to show our NNO algorithm perform significantly better on open world recognition evaluation using Image-Net [1].

2 Related Work

Our work addresses an issue that is related to and has received attention from various communities such as incremental learning, scalable learning and open set learning.

Incremental Learning: As SVMs rose to prominence in for object recognition applications [46, 22], many incremental extensions to SVMs were proposed. Cauwenberghs *et al*. [3] proposed an incremental binary SVM by means of saving and updating KKT conditions. Yeh *et al*. [45] extended the approach to object recognition and demonstrated multi-class incremental learning. Pronobis [31] proposed a memory-controlled online incremental SVM for visual place recognition. Although incremental SVMs might seem natural for large scale incremental learning for object recognition, they suffer from multiple drawbacks. The update process is extremely expensive (quadratic in the number of training examples learned [19]) and depends heavily on the number of support vectors stored for performing updates [19]. To overcome the update expense, Crammer *et al*. [4] and Shalev-Shwartz *et al*. [36] proposed classifiers with fast and inexpensive update process along with their multi-class extensions. However, the multi-class incremental learning methods and other incremental classifiers [4, 36, 43, 21] are incremental in terms of additional training samples but not additional training categories..

Scalable Learning: Researchers like [27, 23, 9] have proposed label tree based classification methods to address scalability (# of object categories) in large scale visual recognition challenges [10, 1]. Recent advances in deep learning community [17, 38], has resulted in state of the art performance on these challenges. Such methods are extremely useful when the goal is to obtain maximum classification/recognition performance. These systems assume a priori availability of entire training data (images and categories). However, adapting such methods to a dynamic learning scenario becomes extremely challenging. Adding object categories requires retraining the entire system, which could be infeasible for many applications. Thus, these methods are scalable but not incremental (Fig 2)

Open Set Learning: Open set recognition assumes there

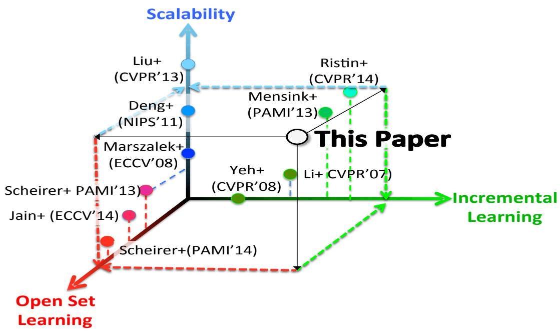


Figure 2: Putting the current work in context by depicting locations of prior work with respect to three axes of the major issues for open world recognition: open set learning, incremental learning and scalability. In this work, we present a system that is scalable, can handle open set recognition and can learn new categories incrementally without having to retrain the system every time a new category arrives. The works depicted include Ristin *et al.* [32], Mensink *et al.* [28], Scheirer *et al.* [34], [33], Jain *et al.* [15], Yeh *et al.*, [45], Marszalek *et al.* [27], Liu *et al.* [23], Deng *et al.* [9], and Li *et al.* [21]. This papers advances the state of the art in open set learning and incremental learning while providing reasonable scalability.

is incomplete knowledge of the world at training time, and unknown classes can be submitted to an algorithm during testing [20, 34]. Scheirer *et al.* [34] formulated the problem of open set recognition for static one-vs-all learning scenario by balancing open space risk while minimizing empirical error. Scheirer *et al.* [33, 15] extended the work to multi-class settings by introducing compact abating probability model. Their work offers insights into building robust methods to handle unseen categories. However, class specific Weibull based calibration of SVM decision scores does not scale. Fragoso *et al.* [12] proposed a scalable Weibull based calibration for hypothesis generation for modeling matching scores, but they do not address it in the context of general recognition problem.

The final aspect of related work is nearest class mean (NCM) classifiers. NCM classification, in which samples undergo a Mahalanobis transform and then are are associated with a class/cluster mean, is a classic pattern recognition approach [13]. NCM classifiers have a long history of use in vision systems [6] and have multiple extensions, adaptations and applications [7, 39, 44, 18, 24]. Recently the technique has been adapted for use in larger scale vision problems [42, 41, 28, 32], with the most recent and most accurate approaches combining NCM with metric learning [28] and with random forests[32].

Since we extend NCM classification, we briefly review the formulation including a probabilistic interpretation. Consider an image represented by a d-dimensional feature vector $x \in \mathbb{R}^d$. Consider $\mathcal{K}$ object categories with their corresponding centroids μ_k, where $k \in \mathcal{K}$. Let $\mathcal{I}_k$ be images for each object category. The centroid is given by $\mu_k = \frac{1}{|\mathcal{I}_k|}\sum_{i \in \mathcal{I}_k} x_i$. NCM classification of a given image instance I with a feature vector x is formulated as searching for the closest centroid in feature space as $c^* = \underset{k \in \mathcal{K}}{\operatorname{argmin}} \ \mathbf{d}(x, \mu_k)$. Here $\mathbf{d}(.)$ represents a distance operator usually in Euclidean space. Mensink *et al.* [28] replace Euclidean distance with a low-rank Mahalanobis distance optimized on training data. The Mahalanobis distance is induced by a weight matrix $W \in \mathbb{R}^{d \times D}$, where D is the dimensionality of the lower dimensional space. Class conditional probabilities $p(c|x)$ using an NCM classifier are obtained using a probabilistic model based on multi-class logistic regression as follows:

$$p(c|x) = \frac{exp(-\frac{1}{2}\mathbf{d}_W(x, \mu_k))}{\sum_{k'=1}^{\mathcal{K}} exp(-\frac{1}{2}\mathbf{d}_W(x, \mu_{k'}))} \tag{1}$$

In the above formulation, class probabilities $p(c)$ are set to be uniform over all classes. During metric learning optimization, Mensink *et al.* [28] considered non-uniform probabilities given by:

$$p'(c|x) = \frac{1}{Z} exp(x^T W^T W \mu_c + s_c) \tag{2}$$

where Z denotes the normalizer and s_c is a per class bias.

3 Open World Recognition

We first establish preliminaries related to open world recognition, following which we formally define the problem. Let classes be labeled by positive integers $\mathbb{N}^+$ and let $\mathcal{K}_t \subset \mathbb{N}^+$ be the set of labels of known classes at time t. Let the zero label (0) be reserved for (temporarily) labeling data as unknown. Thus $\mathbb{N}$ includes unknown and known labels.

Let our features be $x \in \mathbb{R}^d$. Let $f \in \mathcal{H}$ be a measurable recognition function, i.e. $f_y(x) > 0$ implies recognition of the class y of interest and $f_y(x) \leq 0$ when y is not recognized, where $\mathcal{H} : \mathbb{R}^d \mapsto \mathbb{R}$ is a suitably smooth space of recognition functions.

The objective function of open set recognition, including multi-class formulations, must balance open space risk against empirical error. As a preliminary we adapt the definition of open space and open space risk used in [34]. Let open space, the space sufficiently far from any known positive training sample $x_i \in \mathcal{K}, i = 1 \ldots N$, be defined as:

$$\mathcal{O} = S_o - \bigcup_{i \in N} B_r(x_i) \tag{3}$$

where $B_r(x_i)$ is a closed ball of radius r centered around any training sample x_i. Let S_o be a ball of radius r_o that includes all known positive training examples $x \in \mathcal{K}$ as well as the open space $\mathcal{O}$. Then probabilistic *Open Space Risk* $R_\mathcal{O}(f)$ for a class y can be defined as

$$R_\mathcal{O}(f_y) = \frac{\int_\mathcal{O} f_y(x)dx}{\int_{S_o} f_y(x)dx} \tag{4}$$

That is, the open space risk is considered to be the relative measure of positively labeled open space compared to the overall measure of positively labeled space.

Given an empirical risk function $R_{\mathcal{E}}$, e.g. hinge loss, the objective of *open set recognition* is to find a measurable recognition function that manages (minimizes) the **Open Set Risk**:

$$\underset{f \in \mathcal{H}}{\operatorname{argmin}} \{R_{\mathcal{O}}(f) + \lambda_r R_{\mathcal{E}}(f)\} \tag{5}$$

where λ_r is a regularization constant.

With the background in place, we formalize the problem of open world recognition.

Definition 1 (Open World Recognition): *A solution to open world recognition is a tuple $[F, \varphi, \nu, L, I]$ with:*

1. *A* **multi-class open set recognition function** $F(x) : \mathbb{R}^d \mapsto \mathbb{N}$ *using a vector function $\varphi(x)$ of i per-class measurable recognition functions $f_i(x)$, also using a* **novelty detector** $\nu(\varphi) : \mathbb{R}^i \mapsto [0, 1]$. *We require the per class recognition functions $f_i(x) \in \mathcal{H} : \mathbb{R}^d \mapsto \mathbb{R}$ for $i \in \mathcal{K}_t$ to be open set recognition functions that manage open space risk as Eq.4. The novelty detector $\nu(\varphi) : \mathbb{R}^i \mapsto [0, 1]$ determines if results from vector of recognition functions is from an unknown (0) class.*

2. *A labeling process $L(x) : \mathbb{R}^d \mapsto \mathbb{N}^+$ applied to novel unknown data U_t from time t, yielding labeled data $D_t = \{(y_j, x_j)\}$ where $y_j = L(x_j) \forall x_j \in U_t$. Assume the labeling finds m new classes, then the set of known classes becomes $\mathcal{K}_{t+1} = \mathcal{K}_t \cup \{i + 1, \ldots i + m\}$.*

3. *An incremental learning function $I_t(\varphi; D_t) : \mathcal{H}^i \mapsto \mathcal{H}^{i+m}$ to scalably learn and add new measurable functions $f_{i+1}(x) \ldots f_{i+m}(x)$, each of which manages open space risk, to the vector φ of measurable recognition functions.*

Ideally, all of these steps should be automated, but herein we presume supervised learning with labels obtained by human labelling.

If we presume that each $f_k(x)$ reports a likelihood of being in class k and $f_k(x)$ is normalized across the respective classes. Let $\varphi = [f_1(x), \ldots, f_k(x)]$. For this paper we let the multi-class open set recognition function be given as

$$y^* = \underset{y \in \mathcal{K}, f_y(x) \in \varphi(x)}{\operatorname{argmax}} f_y(x), \tag{6}$$

$$F(x) = \begin{cases} 0 & \text{if } \nu(\varphi(x)) = 0 \\ y^* & otherwise \end{cases} \tag{7}$$

With these definitions, a simple approach for the novelty detection is to set a minimum threshold τ for acceptance, e.g. letting $\nu(\varphi(x)) = f_{y^*}(x) > \tau$. In the following section we will prove this simple approach can manage open space risk and hence provide for item 1 in the open world recognition definition.

4 Opening existing algorithms

The series of papers [34, 33, 15] formalized the open set recognition problem and proposed 3 different algorithms for managing open set risk. It is natural to consider these algorithms for open world recognition. Unfortunately, these algorithms use EVT-based calibration of 1-vs-rest RBF SVMs and hence are not well suited for incremental updates or scalability required for open world recognition. In this paper we pursue an alternative approach better suited to open world using non-negative combinations of abating distance. Using this we develop the Nearest Non-Outlier (NNO) algorithm to inexpensively extend NCM for open world recognition.

The authors of [33] show if a recognition function is decreasing away from the training data, a property they call abating, then thresholding the abating function limits the labeled region and hence can manage/limit open space risk. The Compact Abating Probability (CAP) model presented in that paper is a sufficient model, but it is not necessary. In particular we build on the concept of a CAP model but generalize the model showing that any non-negative combination of abating functions, e.g., a convex combination of decreasing functions of distance, can be thresholded to have zero open space risk. We further show we can work in linearly transformed spaces, including projection onto subspaces, and still manage open space risk and NCM type algorithms manage open space risk.

Theorem 1 (Open space risk for model combinations): *Let $M_{\tau,y}(x)$ be a recognition function that thresholds a non-negative weighted sum of η CAP models ($M_{\tau,y}(x) = \sum_{j=1}^{\eta} c_j M_{j,\tau_j,y}(x)$) over a known training set for class y, where $1 \geq c_j \geq 0$ and $M_{j,\tau,y}(x)$ is a CAP model. Then for $\delta \geq 0 \ \exists \tau^*$ s.t. $R_{\mathcal{O}}(M_{\tau^*,y}) \leq \delta$, i.e. one can threshold the probabilities $M_{\tau,y}(x)$ to limit open space risk to any level.*

Proof: It is sufficient to show it holds for $\delta = 0$, since similar to Corr. 1 of [33], larger values of δ allow larger labeled regions with larger open space risk. Considering each model $M_{j,\tau_j,y}(x) j = 1..\eta$ separately, we can apply Theorem 1 of [33] to each $M_{j,\tau_j,y}(x)$ yielding a τ_j such that the function $M_{j,\tau_j,y}(x) > 0$ defines a labeled region $l_j(\tau_j) \subset X$ with zero open space risk. Letting $\tau^* = \min_j \tau_j$ it follows that $M_{\tau^*,y}(x) > 0$ is contained within $\cup_j l_j(\tau^*)$, which as a finite union of compact regions with zero risk, is itself a compact labeled region with zero open space risk. *Q.E.D*

The theorem/proof trivially holds for a max over classes but can be generalized to combinations via product or to combinations of monotonic transformed recognition functions, with appropriate choice of thresholds. For this paper we need max over models using data from metric learned tranformed features, i.e. lower-dimensional projected spaces.

Theorem 2 (Open Space Risk for Transformed Spaces): *Given a linear transform $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ let $x' =$*

$T(x), \forall x \in X$, yields X' a linearly transformed space of features derived from feature space $X \subset \mathbb{R}^n$. Let $\mathcal{O}' = \cup_{x \in \mathcal{O}} T(x)$ be the transformation of points in open space $\mathcal{O}$. Let $M'_{\tau,y}(x')$ be a probabilistic CAP recognition function over $x' \in X'$ and let $M_{\tau,y}(x) = M'_{\tau,y}(Tx)$ be a recognition function over $x \in X$. Then $\exists \epsilon : R_{\mathcal{O}'}(M'_{\tau',y}) \leq \delta \implies R_{\mathcal{O}}(M_{\tau,y}) < \epsilon\delta$, i.e. managing open set risk in X' will also manage it in the original feature space X.

Proof: If T is dimensionality preserving, then the theorem follows from the linearity of integrals in the definition of risk. Thus we presume T is projecting away $n-m$ dimensions. Since the open space risk in the projected space is δ we have $\lambda_m(M'_{\tau',y} \cap \mathcal{O}') = c\delta$ where λ_m is the Lebesgue measure in $\mathbb{R}^m$ and $c < \infty$. Since $\mathcal{O} \subset S_o$, i.e. $\mathcal{O}$ is contained within a ball of radius r_o, it follows from the properties of Lebesgue measure that $\lambda_n(M_{\tau,y} \cap \mathcal{O}) \leq \lambda_m\left(M'_{\tau',y} \cap (\mathcal{O}' \times [-r_o, r_o]^{n-m})\right) = c{*}\delta{*}(2r_o)^{n-m} = 0$ and hence the open space risk in $\mathbb{R}^m$ is bounded. *Q.E.D.*

It is desirable for open world problems that we consider the error in the original space. We note that ϵ varies with dimension and the above bounds are generally not tight. While the theorem gives a clean bound for zero open space risk, for a solution with non-zero δ risk in the lower dimensional space, when considered in the original space, the solution may have open space risk that increases exponentially with the number of missing dimensions.

We note that these theorems are not a license to claim that any algorithms with rejection manage open space risk. While many algorithms can be adapted to compute a probability estimate of per class inclusion and can threshold those probabilities to reject, not all such algorithms/rejections manage open space risk. Thresholding Eq 2, which [28] minimizes in place of 1, will not manage risk because the function does not always decay away from known data. Similarly, rejecting a decision close to the plane in a linear SVM does not manage open space risk, nor does the thresholding layers in a convolution neural network [35].

On the positive side, these theorems show that one can adapt algorithms that linearly transforms feature space and use a *probability/score mapping that combines positive scores that decrease with distance from a finite set of known samples*. In the following section, we demonstrate how to generalize an existing algorithm while managing open space risk. Open world performance, however, greatly depends on the underlying algorithm and the rejection threshold. While theorems 1 and 2 say there exists a threshold with zero open space risk, at that threshold there may be minimal or no generalization ability.

4.1 Nearest Non-Outlier (NNO)

As discussed previously (sec 1) one of the significant contributions of this paper is combining theorems 1 and 2 to provide an example of open space risk management and move toward a solution to open world recognition. Before moving on to defining open world NCM, we want to add a word of caution about "probability normalization" that presumes all classes are known. e.g. softmax type normalization used in eqn 1. Such normalization is problematic for open world recognition where there are unknown classes. In particular, **in open world recognition the Law of Total Probability and Bayes' Law cannot be directly applied** and hence cannot be used to normalize scores. Furthermore, as one adds new classes, the normalization factors and hence probabilities, keep changing and thereby limiting interpretation of the probability. For an NCM type algorithm, normalization with the softmax makes thresholding very difficult since for points far from the class means the nearest mean will have a probability near 1. Since it does not decay, it does not follow Theorem 1.

To adapt NCM for open world recognition, we introduce Nearest Non-Outlier (NNO) which uses a measurable recognition function consistent with Theorems 1 and 2. Let NNO represent its internal model as a vector of means $\mathcal{M} = [\mu_1, \ldots \mu_k]$. Let $W \in \mathbb{R}^{d \times m}$ be the linear transformation dimensional reduction weight matrix learned by the process described in [28]. Then given τ, let

$$\hat{f}_i(x) = \frac{\Gamma(\frac{m}{2}+1)}{\pi^{\frac{m}{2}}\tau^m}\left(1 - \frac{1}{\tau}\|W^\top x - W^\top \mu_i\|\right) \quad (8)$$

be our measurable recognition function with $\hat{f}_i(x) > 0$ giving the probability of being in class i, where Γ is the standard gamma function which occurs in the volume of a m-dimensional ball. Intuitively, the probability is a tent-like function in the sphere and the first fraction in eqn 8 comes from volume of m-sphere and ensures that the probability integrates to 1.

Let $\hat{\varphi} = [\hat{f}_1(x), \ldots, \hat{f}_k(x)]$ with y^* and $F(x)$ given by Eq. 7. Let with $\hat{\nu}(\hat{\varphi}(x)) = \hat{f}_{y^*}(x) > 0$. That is, NNO rejects x as an outlier for class i when $\hat{f}_i(x) = 0$, and NNO labels input x as unknown/novel when all classes reject the input.

Finally, after collecting novel inputs, let D_t be the human labeled data for a new class $k+1$ and let our incremental class learning $I_t(\hat{\varphi}; D_t)$ compute $\mu_{k+1} = mean(D_t)$ and append μ_{k+1} to $\mathcal{M}$.

Corollary 1 (NNO solves open world recognition): *The NNO algorithm with human labeling $L(x)$ of unknown inputs is a tuple $[F(x), \hat{\varphi}, \hat{\nu}(\hat{\varphi}(x)), L, I_t(\hat{\varphi}; D_t)]$, consistent with Definition 1, hence NNO is a open world recognition algorithm.*

By construction theorems 1 and 2 apply to the measurable recognition functions $F(x)$ from Eq. 7 when using a vector of per classes functions given eq. 8. By inspection the NNO definitions of $\hat{\nu}(\hat{\varphi}(x))$ and $I_t(\hat{\varphi}; D_t)$ are consistent with Definition 1 and are scalable. *Q.E.D.*

5 Experiments

In this section we present our protocol for open world experimental evaluation of NNO, and a comparisonwithmultiple baseline classifiers including NCM, a liblinear SVM [11] and our liblinear version of the 1vSet algorithm of [34][1].

Dataset and Features: Our evaluation is based on the ImageNet Large Scale Visual Recognition Competition 2010 dataset. ImageNet 2010 dataset is a large scale dataset with images from 1K visual categories. The dataset contains 1.2M images for training (with around 660 to 3047 images per class), 50K images for validation and 150K images for testing. The large number of visual categories allow us to effectively gauge the performance of incremental and open world learning scenarios. In order to effectively conduct experiments using open set protocol, we need access to ground truth. ILSVRC'10 is the only ImageNet dataset with full ground truth, which is why we selected that dataset over later releases of ILSVRC (e.g. 2011-2014).

We used densely sampled SIFT features clustered into 1K visual words as given by Berg *et al*. [1]. Though more advanced features are available [30, 17, 37], extensive evaluation across features is beyond the scope of this work [2]. Each feature is whitened by its mean and standard deviation to avoid numerical instabilities. We report performance in terms of average classification accuracy obtained using top-1 accuracy as per the protocol provided for the ILSVRC'10 challenge. As our work involves initially training a system with small set of visual categories and incrementally adding additional categories, we shun top-5 accuracy.

Algorithms: The proposed Nearest Non-Outlier (NNO) extension of NCM classifier is compared with the baseline NCM algorithm in both incremental and open world settings. We use the code provided by Mensink *et al*. [28] as the NCM baseline. This algorithm has near state of the art results and while recent extension with random forests[32] improved accuracy slightly, [32] does not provide code. While not incremental, we also include a comparison with the state of the art open set algorithm by extending liblinear to provide a 1vSet SVM [34]. Details about our extension can be found in the supplemental material.

5.1 Open World Evaluation Protocol

Closed set evaluation is when a system is tested with allclassesknown during training and testingi.e. training, and testing use the same classes but different instances. In open set evaluation,training uses known classes and testing uses both known and unknownclasses.The open set recognition evaluation protocol proposed by Scheirer *et al*. [34] does not handle the open world scenario in which object categories are beingincrementallyadded to the system. Ristin *et al*. [32] presented an incremental closed set learning scenario where novel object categories are added continuously. We combined ideas from both of these approaches and propose a protocol that is suited for open world recognition in which categories are being added to the system continuously while the system is also tested with unknown categories.

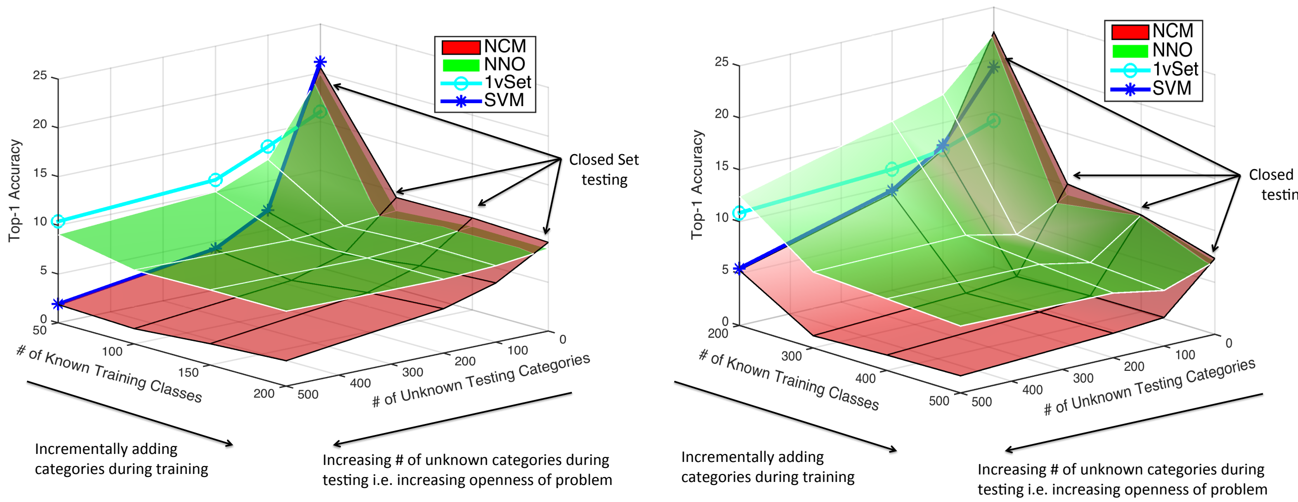
Training Phase: The training of the NCM classifier is divided into two phases: an initial metric learning/training phase and a growth/incremental learning phase. In the metric learning phase, a set of object categories are provided to the system uses iterative metric learning on these categories. Once the metric learning phase is completed, the incremental learning phase uses the fixed metrics and parameters. During the incremental learning phase, object categories are added to the system one-by-one. While for scalability one might measure time, both NCM and NNO add new categories in the same way, and it is extremely fast since it only consists of computing the means. Thus, so we do not measure/report timing here.

Nearest Non-Outlier (NNO) is based on the CAP model and requires estimation of τ for eq. 8. To estimate τ, during theparameter estimation phase using the metric learned in that phase,we use a 3-fold cross-class validation [15] wherein ech fold dividesthetrainingdatainto two sets: training categories and validation categories. The τ for NNO is estimated with 3-fold cross-class validation optimizing for F1-measure over values for which there is at least 90% recall in a fold, yielding a value of 5000 – see the supplemental material for more details. An important point to note about estimating τ is that one has to balance the classification errors between known set of categories along with the errors between known and unknown set of categories. One could obtain high accuracy when testing with large number of samples from unknown categories by rejecting everything, but this compromises accuracy on the known set of categories. Hence our requirement of high recall rate and optimization over F1-measure rather than accuracy.

Testing Phase: To ensure proper open world evaluation, we do cross-class folds that split the ImageNet test data into two sets of 500 categories each: the known set and the unknown set. At every stage, the system is evaluated with a subset of the known set and the unknown set to obtain closed set and open set performance. This process is repeated as we continue to add categories to the system. The whole process is repeated across multiple dataset splits to ensure fair comparisons and estimate error. While [34] suggest a particular openness measure, it does not address the incremental learning paradigm. We fixed the number of unknown categories and report performance as series of known categories are incrementally added. Thus, open world evaluation involves varying of two variables: number of known categories in

[1]Code and data partitions for experiments can be found at http://vast.uccs.edu/OpenWorld

[2]The supplemental material presents experiments with additional ILSVRC'13 features, showing the gains of NNO are not feature dependent



(a) 50 categories in metric learning phase.

(b) 200 categories in metric learning phase.

Figure 3: Open World learning on data from ILSVRC'10 challenge. Top-1 accuracy is plotted as a function of known classes in the system and unknown classes used during testing. NNO performs at par with NCM in closed set testing (marked with arrows in above figure) as categories are added incrementally to the system. As number of unknown categories are increased during testing phase, NNO continues to remain robust while performance of NCM suffers significantly. The proposed Nearest Non-Outlier (NNO) approach of handling unknown categories based on extending NCM with Compact Abating Probabilities remains robust in both circumstances: as more number of categories are added to the system and as the system is tested with more unknown categories. The current state-of-the-art on open set recognition 1vSet algorithm [34] and standard SVM [11] is shown above as a line, as neither of them possess incremental learning capabilities. Fig 3a and Fig 3b shows results when 50 and 200 categories were used for initial metric learning and parameter estimation.

training (incremental learning) and number of unknown categories during testing (open set learning) leading to surface plots as shown in Fig 3.

Multi-class classification error [5] for a system $F_{\mathcal{K}}(.)$ with test samples $\{(x_i, y_i)\}_{i=1}^{N}, y_i \in \mathcal{K}$ is given as $\epsilon_{\mathcal{K}} = \frac{1}{N}\sum_{i=1}^{N}[\![F_{\mathcal{K}}(x_i) \neq y_i]\!]$. For open world testing the evaluation must keep track of the errors which occur due to standard multi-class classification over known categories as well as errors between known and unknown categories. Consider evaluation of N samples from $\mathcal{K}$ known categories and N' samples from $\mathcal{U}$ unknown categories leading to $(N+N')$ test samples and $\mathcal{K} \cup \mathcal{U} \in X$. Thus, open world error ϵ_{OW} for a system $F_{\mathcal{K}}(.)$ trained over $\mathcal{K}$ categories is given as:

$$\epsilon_{OW} = \epsilon_{\mathcal{K}} + \frac{1}{N'}\sum_{j=N+1}^{N'}[\![F_{\mathcal{K}}(x_j) \neq unknown]\!] \qquad (9)$$

5.2 Experimental Results

In the first experiment, we do incremental learning from a base of relatively few (50) categories and we add 50 categories incrementally. For NCM and NNO systems, we update the means of the categories. We repeat that expansion 3 times growing from 50 to 100 to 150 and finally 200 known classes. For close-set testing we therefore have training and testing with 50, 100, 150 and 200 categories. To test open world performance, we considering an additional set of 100, 200 and 500 unknown categories showing up during testing. For example, open world testing with 100 unknown categories for a system that is trained with 50 categories would have a total of 50 + 100 i.e. 150 categories in testing. The experiment is thus a function of two variables : total number of known set of categories learned during training (both metric learning and incremental learning phase) and unknown set of categories seen by the system during testing phase. Varying both of these leads to performance being shown as the surface plots shown in 3. The plot shows showing top-1 accuracy where we treat unknown as its own class.

We note that each testing phase is independent and for these experiments we do not provide feedback for the results of testing as labels for the incremental training – i.e. we still presume perfect labels for each incremental training round. In this model, misclassifications in testing only impact the workload level of human labelers. If the open-world results of testing were used in a semi-supervised manner, the compounding of errors would significantly amplify the difference in the algorithm accuracy.

The back edges of the plot provide 2 easy to separate views. To see pure incremental learning, we incrementally add categories to the system in closed set testing paradigm. This is shown in the back right portion of Fig. 3a, where the performance of NCM (red) and NNO (green) drop similarly and rather gracefully, which is expected. However, as we increase openness for a fixed number of training classes, the back left of edge of Fig. 3a, the impact on NCM is a dramatic loss of performance even for the non-incremental growth case. This is caused by errors for the unknown categories, something NCM was not designed to handle and the NCM error is dominated by the second term in Eqn 9. As we can see the standard SVM also has a dramatic drop in accuracy as openness increases. Both the NNO and 1vSet algorithm, designed for open set recognition, degrade gracefully. The 1vSet and SVM don't have a way to incrementally add classes and are curves not surfaces in the plots. So the 1vSet, while slight better than NNO for pure open set on 50 categories, does not support open world recognition.

Open world recognition needs to support increasing classes while handling unknowns, so is can be viewed as the performance as known training classes increase for non-zero number of unknowns. At all such points in the plot, NNO significantly dominates the NCM algorithm.

In second experiment, we consider 200 categories for metric learning and parameter estimation, and successively add 100 categories in each of three incremental learning phases. By the end of the learning process, the system needs to learn a total of 500 categories. Open world evaluation of the system is carried out as before by considering with 100, 200 and 500 additional unknown categories with results show in Fig 3b. In final stage of the learning process i.e 500 categories for training and 500 (known) + 500 (unknown) categories for open set testing, we use all 1000 categories from ImageNet for our evaluation process. We observe that NNO again dominates the baselines for open world recognition; this time even outperforming 1vSet for open set testing on 200 classes. On the largest scale task involving 500 categories in training and 1000 categories in testing, we observe that NNO provides almost 74% improvement over NCM. Also note performance starting with 200 classes (3b) is better than starting with 50 classes (3a), i.e. increased classes for the metric learning improves both NNO and NCM performance. We repeated the above experiments over three cross-class folds and found the standard deviation to be on the order of $\pm$ 1% which is not visible in the figure.

The training time required for the initial metric learning process depends on the SGD speed and convergence rate. We used close to 1M iterations which resulted in metric-learning time of 15 hours in case of 50 categories and 22 hours in case of metric learning for 200 categories. Given the metric, the learning of new classes via the update process is extremely fast as it is simply computation of means from labeled data. For this work we fix the metric, though future work might explore incrementally updating the metric as well. The majority of time in update process is dominated by feature extraction and file I/O. However, these operations could be easily optimized for real-time operations. The NNO Multi-class recognition and detecting novel classes is also easily done in real time.

6 Discussion

In this work, we formalize the problem of open world recognition and provide an open world evaluation protocol. We extend existing work on NCM classifiers and show how to adapt it for open world recognition. The proposed NNO algorithm consistently outperforms NCM on open world recognition tasks and is comparable to NCM on closed set – we gain robustness to the open world without much sacrifice.

There are multiple implications of our experiments. First, we demonstrate suitability of NNO for large scale recognition tasks in dynamic environments. NNO allows construction of scalable systems that can be updated incrementally with additional classes and that are robust to unseen categories. Such systems are suitable where minimum downtime is desired.

Second, as can be seen in Figs 3a and Fig 3b NNO offers significant improvement for open world recognition while for close set recognition NNO remains relatively close to NCM in performance.

We also noted that as the number of classes incrementally grew, the closed and open set performance NNO seems to converge, i.e. the front right edge of the plots in Figs. 3a an 3b are very flat. This observation suggests that adding classes in a system may also be limited by open space risk. We conjecture that as the number of classes grows, the close world performance converges to an open world performance and thus open world recognition is a more natural setting for building scalable systems.

While we provide one viable open world extension, the theory herein allows a broad range of approaches; more expressive models, improved CAP models and better open set probability calibration should be explored.

Open world evaluation across multiple features for a variety of applications is an important future work. Recent advances in deep learning and other areas of visual recognition have demonstrated significant improvements in absolute performance. The best performing systems on such tasks use a parallel system and train for days. Extending these systems to support incremental open world performance may allow one to provide a hybrid solution where one reuses the deeply learned features with a top layer of an open world multi-class algorithm. While scalable learning in the open world is critical for deploying computer vision applications in the real world, high performing systems enable adoption by masses. Pushing absolute performance on large scale visual recognition challenges [1], and development of scalable systems for the open world are essentially two sides of the same coin.

7 Acknowledgement

We would like to thank Thomas Mensink (ISLA, Informatics Institute, University of Amsterdam) for sharing code and Marko Ristin (ETH Zurich) for sharing features. The work support in part by NSF Research Grant IIS-1320956 (Open Vision - Tools for Open Set Computer Vision and Learning) and a UCCS Graduate School Fellowship.

References

[1] A. Berg, J. Deng, and L. Fei-Fei. Large scale visual recognition challenge, 2010. http://image-net.org/challenges/LSVRC/2010/index[Online; accessed 1-Nov-2013]. 2, 6, 8

[2] P. Bodesheim, A. Freytag, E. Rodner, M. Kemmler, and J. Denzler. Kernel null space methods for novelty detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3374–3381. IEEE, 2013. 1

[3] G. Cauwenberghs and . Poggio. Incremental and decremental support vector machine learning. In *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*, volume 13, page 409. MIT Press, 2001. 1, 2

[4] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. Online passive-aggressive algorithms. *The Journal of Machine Learning Research*, 7:551–585, 2006. 1, 2

[5] K. Crammer and Y. Singer. On the algorithmic implementations of multiclass kernel-based vector machines. *JMLR*, 2001. 7

[6] J. D. Crisman and C. E. Thorpe. Color vision for road following. In *1988 Robotics Conferences*, pages 175–185. International Society for Optics and Photonics, 1989. 3

[7] P. Datta and D. Kibler. Symbolic nearest mean classifiers. In *AAAI/IAAI*, pages 82–87, 1997. 3

[8] T. Dean, M. A. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, and J. Yagnik. Fast, accurate detection of 100,000 object classes on a single machine. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1814–1821. IEEE, 2013. 1

[9] J. Deng, S. Satheesh, A. C. Berg, and F. Li. Fast and balanced: Efficient label tree learning for large scale object recognition. In *Advances in Neural Information Processing Systems*, pages 567–575, 2011. 2, 3

[10] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010. 2

[11] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008. 6, 7

[12] V. Fragoso, P. Sen, S. Rodriguez, and M. Turk. EVSAC: accelerating hypotheses generation by modeling matching scores with extreme value theory. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2472–2479. IEEE, 2013. 3

[13] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic press, 1990. 3

[14] V. J. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004. 1

[15] L. P. Jain, W. J. Scheirer, and T. E. Boult. Multi-class open set recognition using probability of inclusion. In *Computer Vision–ECCV 2014*, pages 393–409. Springer, 2014. 1, 3, 4, 6

[16] A. Kapoor, S. Baker, S. Basu, and E. Horvitz. Memory constrained face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2539–2546. IEEE, 2012. 2

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*, pages 1097–1105, 2012. 1, 2, 6

[18] L. I. Kuncheva, C. J. Whitaker, and A. Narasimhamurthy. A case-study on naïve labelling for the nearest mean and the linear discriminant classifiers. *Pattern Recognition*, 41(10):3010–3020, 2008. 3

[19] P. Laskov, C. Gehl, S. Krüger, and K.-R. Müller. Incremental support vector learning: Analysis, implementation and applications. *The Journal of Machine Learning Research*, 7:1909–1936, 2006. 2

[20] F. Li and H. Wechsler. Open set face recognition using transduction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on (T-PAMI)*, 27(11):1686–1697, 2005. 3

[21] L.-J. Li and L. Fei-Fei. Optimol: automatic online picture collection via incremental model learning. *International Journal of Computer Vision (IJCV)*, 88(2):147–168, 2010. 2, 3

[22] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang. Large-scale image classification: fast feature extraction and svm training. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1689–1696. IEEE, 2011. 2

[23] B. Liu, F. Sadeghi, M. Tappen, O. Shamir, and C. Liu. Probabilistic label trees for efficient large scale image classification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 843–850. IEEE, 2013. 2, 3

[24] M. Loog. Constrained parameter estimation for semi-supervised learning: the case of the nearest mean classifier. In *Machine Learning and Knowledge Discovery in Databases*, pages 291–304. Springer, 2010. 3

[25] J. Lu, J. Zhou, J. Wang, T. Mei, X.-S. Hua, and S. Li. Image search results refinement via outlier detection using deep contexts. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3029–3036. IEEE, 2012. 1

[26] M. Markou and S. Singh. Novelty detection: a reviewpart 1: statistical approaches. *Signal processing*, 83(12):2481–2497, 2003. 1

[27] M. Marszałek and C. Schmid. Constructing category hierarchies for visual recognition. In *Computer Vision–ECCV 2008*, pages 479–491. Springer, 2008. 2, 3

[28] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *Pattern Analysis and Machine Intelligence, IEEE Transactions on (T-PAMI)*, 35(11):2624–2637, 2013. 2, 3, 5, 6

[29] V. Ordonez, V. Jagadeesh, W. Di, A. Bhardwaj, and R. Piramuthu. Furniture-geek: Understanding fine-grained furniture attributes from freely associated text and tags. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 317–324. IEEE, 2014. 1

[30] F. Perronnin, Z. Akata, Z. Harchaoui, and C. Schmid. Towards good practice in large-scale learning for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3482–3489. IEEE, 2012. 1, 6

[31] A. Pronobis, L. Jie, and B. Caputo. The more you learn, the less you store: Memory-controlled incremental svm for visual place recognition. *Image and Vision Computing*, 28(7):1080–1097, 2010. 2

[32] M. Ristin, M. Guillaumin, J. Gall, and L. V. Gool. Incremental learning of ncm forests for large-scale image classification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3654–3661. IEEE, 2014. 2, 3, 6

[33] W. Scheirer, L. Jain, and T. Boult. Probability models for open set recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on (T-PAMI)*, 36(11):2317–2324, Nov 2014. 1, 3, 4

[34] W. J. Scheirer, A. Rocha, A. Sapkota, and T. E. Boult. Towards open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 36, July 2013. 1, 3, 4, 6, 7

[35] P. Sermanet, D. Eigen, Z. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *ICLR*, 2014. 5

[36] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. *ICML*, 2007. 2

[37] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep fisher networks for large-scale image classification. In *Advances in neural information processing systems*, pages 163–171, 2013. 6

[38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arxiv*, 2014. 2

[39] M. Skurichina and R. P. Duin. Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis & Applications*, 5(2):121–135, 2002. 3

[40] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1701–1708. IEEE, 2014. 1

[41] C. J. Veenman and M. J. Reinders. The nearest subclass classifier: A compromise between the nearest mean and nearest neighbor classifier. *Pattern Analysis and Machine Intelligence, IEEE Transactions on (T-PAMI)*, 27(9):1417–1429, 2005. 3

[42] C. J. Veenman and D. M. Tax. A weighted nearest mean classifier for sparse subspaces. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 1171–1176. IEEE, 2005. 3

[43] Z. Wang, K. Crammer, and S. Vucetic. Multi-class pegasos on a budget. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 1143–1150, 2010. 2

[44] Y.-H. Yang, C.-C. Liu, and H. H. Chen. Music emotion classification: a fuzzy approach. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 81–84. ACM, 2006. 3

[45] T. Yeh and T. Darrell. Dynamic visual category learning. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 1, 2, 3

[46] H. Zhang, A. C. Berg, M. Maire, and J. Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2126–2136. IEEE, 2006. 2