C. Käding, A. Freytag, E. Rodner, P. Bodesheim, and J. Denzler: Active Learning and Discovery of Object Categories in the Presence of Unnameable Instances. Proc. CVPR 2015.


# Active Learning and Discovery of Object Categories in the Presence of Unnameable Instances

Christoph Käding, Alexander Freytag, Erik Rodner, Paul Bodesheim, Joachim Denzler
Computer Vision Group, Friedrich Schiller University Jena, Germany
{firstname.lastname}@uni-jena.de | www.inf-cv.uni-jena.de

## Abstract

*Current visual recognition algorithms are "hungry" for data but massive annotation is extremely costly. Therefore, active learning algorithms are required that reduce labeling efforts to a minimum by selecting examples that are most valuable for labeling. In active learning, all categories occurring in collected data are usually assumed to be known in advance and experts should be able to label every requested instance. But do these assumptions really hold in practice? Could you name all categories in every image?*

*Existing algorithms completely ignore the fact that there are certain examples where an oracle can not provide an answer or which even do not belong to the current problem domain. Ideally, active learning techniques should be able to* discover new classes *and at the same time cope with queries an expert is* not able or willing to label.

*To meet these observations, we present a variant of the expected model output change principle for active learning and discovery in the presence of unnameable instances. Our experiments show that in these realistic scenarios, our approach substantially outperforms previous active learning methods, which are often not even able to improve with respect to the baseline of random query selection.*

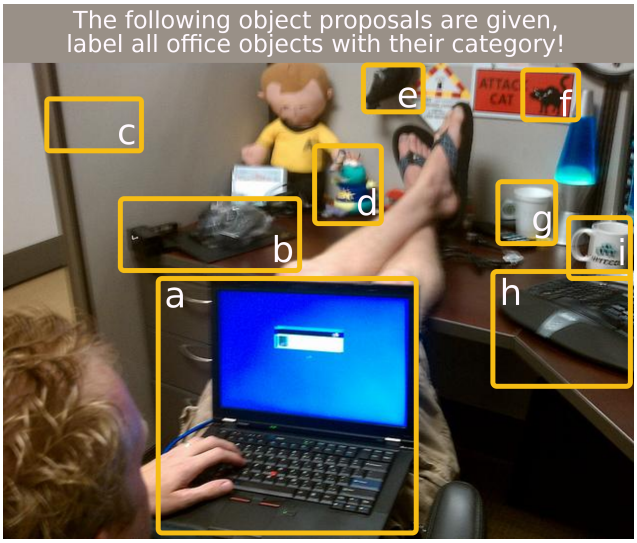



Figure 1: The answers of an annotator are likely to be: *(a) laptop (b) what the heck is that? (c) no object! (d) some toy figure (e) no idea (f) cat (g) cup (h) keyboard (i) cup.* In an active learning and discovery scenario, an annotator might reject examples that do not show valid objects, that cannot be identified, or that are not part of the problem domain, which should be considered when requesting annotations.

## 1. Introduction

Visual data is omnipresent and a flood of free digital data became available with the existence of smart phones, cheap cameras, and internet sharing platforms. While advantageous in several aspects, the resulting data is often unstructured and with none or wrong annotations. As an illustrative example, consider the amount of video data uploaded to YouTube, which currently is about 100 hours of video data per minute [36]. While computer vision systems would greatly benefit from using these massive amounts of *unlabeled* data resources, manual annotation is costly. Thus, only a fraction of the data is actually available for supervised training of object classification systems.

Not surprisingly, researchers have been interested in understanding what makes individual examples informative for a given task. Given that knowledge, we could automatically select a small, unlabeled subset containing as much information as possible which is then labeled by an expert and still stays within a specified labeling budget. This field of research is known as active learning and gained popularity within the computer vision community over the last years [20, 17, 21, 35, 22, 10, 25, 11].

Existing active learning algorithms assume that an annotator always knows an answer, *e.g.,* the requested label of a given image. From our experience, this assumption does *not* hold in practice and annotators are likely to reject labeling in two scenarios: either the sample does

not belong to a valid category (*e.g.,* lens artifacts, motion blur, or segmentations covering parts of multiple objects) or the sample is categorial, but unknown to the annotator or unrelated to the current labeling task. A visual example is given in Fig. 1. We therefore present an active learning technique that allows for rejecting an example and for providing answers like *"I can not name the object."*, *"This is not a real object."*, or *"This object is not of interest."*. The latter is especially crucial in the case of an increasing number of categories in the unlabeled data in order to avoid learning unrelated concepts. In general, we consider active learning in open set multi-class scenarios and we are therefore not restricted to binary classification [24, 4, 34, 21, 28, 5, 11] and also not restricted to multi-class classification scenarios with a pre-defined set of classes as assumed by [32, 26, 20, 17, 22, 35, 25]. Thus, we are studying a more realistic active and life-long learning scenario in which not all categories are given in advance and the expert is allowed to refuse the labeling.

The remainder of the paper is organized as follows. We start with a detailed introduction to active learning from the theoretical perspective of risk minimization in Section 2. Given these concepts, we put existing approaches into perspective and highlight resulting drawbacks. Our multi-class technique for active learning, which circumvents these burdens, is introduced in Section 3, and we extend it for handling unnameable instances in Section 4. After a brief review of related work in Section 5, results in several application areas, including face identification, digit recognition, and object classification, are presented in Section 6. The results obtained show that our approach substantially outperforms previous active learning methods.

## 2. Active learning: goals and problems

To better understand active learning tasks and occurring problems, we start in this section from the ultimate goal in active learning. We then figure out resulting drawbacks, explain how common methods tackle these issues, and why they are inappropriate in realistic scenarios.

### 2.1. Active learning: seeking for the smallest risk

From the most general point of view, our goal is to train models $f : \Omega \to \mathcal{Y}$ that result in highest classification accuracy or, similarly, in the smallest risk $\mathcal{R}(f)$:

$$\mathcal{R}(f) = \int_{\Omega} \int_{\mathcal{Y}} \mathcal{L}\left(y, f\left(\boldsymbol{x}\right)\right) p(\boldsymbol{x}, y) \, \mathrm{d}y \, \mathrm{d}\boldsymbol{x} \quad , \tag{1}$$

where the loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ penalizes mismatching outputs. Since the joint probability $p(\boldsymbol{x}, y)$ for evaluating Eq. (1) is not available in practice, collected labeled data $\mathfrak{L} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\} \subset \Omega \times \mathcal{Y}$ is used and assumed to be representative

$$\mathcal{R}_{\text{emp}}(f) = \frac{1}{|\mathfrak{L}|} \sum_{(\boldsymbol{x}_i, y_i) \in \mathfrak{L}} \mathcal{L}\left(y_i, f\left(\boldsymbol{x}_i\right)\right) \quad , \tag{2}$$

which is known as *empirical risk*. In active learning, our labeled set $\mathfrak{L}$ is usually small, since annotations are assumed to be costly. However, we can access an additional unlabeled dataset $\mathfrak{U} \subset \Omega$ and we are allowed to select a subset for labeling. Thus, for active learning, Eq. (2) translates to selecting a subset $\mathfrak{S}^* \subset \mathfrak{U}$ with highest decrease in risk:

$$\mathfrak{S}^* = \underset{\mathfrak{S}' \subset \mathfrak{U} \text{ with } \text{cost}(\mathfrak{S}') < \xi}{\operatorname{argmax}} \Delta\mathcal{R}_{\text{emp}}\left(\mathfrak{L}, (\mathfrak{S}', \boldsymbol{y}')\right) \quad , \tag{3}$$

$$\Delta\mathcal{R}_{\text{emp}}\left(\mathfrak{L}, (\mathfrak{S}', \boldsymbol{y}')\right) = \mathcal{R}_{\text{emp}}\left(f_{\mathfrak{L}}\right) - \mathcal{R}_{\text{emp}}\left(f_{\mathfrak{L}'}\right) \quad , \tag{4}$$

where $f_{\mathfrak{L}'}$ is the model $f$ trained on the labeled set $\mathfrak{L}' = \mathfrak{L} \cup (\mathfrak{S}', \boldsymbol{y}')$. The $\text{cost}(\mathfrak{S}')$ reflects label costs induced by $\mathfrak{S}'$ which should be less than a pre-defined maximum $\xi$ and which are mostly measured in terms of the number of labels requested, *i.e.,* number of examples in $\mathfrak{S}'$.

### 2.2. Substantial problems in active learning

Starting from Eq. (2)-(4), we observe three drawbacks preventing us from implementing an optimal active learner:

(P1) the labeled set $\mathfrak{L}$ is too small for being representative,[1] thus, replacing $\mathcal{R}$ by $\mathcal{R}_{\text{emp}}$ in Eq. (2) is cumbersome,

(P2) exponentially many subsets $\mathfrak{S}' \subset \mathfrak{U}$ in Eq. (3) make optimization computationally intractable,

(P3) the absence of labels $\boldsymbol{y}'$ for samples in $\mathfrak{U}$ makes evaluating Eq. (4) impossible.

The first drawback (P1) can be tackled by leveraging the unlabeled set $\mathfrak{U}$, and we will see possible solutions in Section 2.3. With respect to computational tractability (P2), a common approximation is to perform an iterative optimization by selecting one example at a time, *i.e.,* $|\mathfrak{S}'| = 1$, which is known as *myopic* learning. The third problem (P3) of missing labels, however, is among the most crucial aspects and prevents from reliably judging the *informativeness* of a selected subset with respect to Eq. (4).

### 2.3. Approximating the expected information gain

An ad-hoc solution for tackling the first problem (P1) mentioned in Section 2.2 is to include unlabeled examples in Eq. (2) to obtain a large and representative dataset $\mathfrak{D} = \mathfrak{L} \cup \mathfrak{U}$. Missing labels can be estimated using outputs of the current model (compare with Eq. (2))

$$\mathcal{R}_{\mathbb{E}}(f_{\mathfrak{L}}) = \frac{1}{|\mathfrak{D}|} \sum_{\boldsymbol{x} \in \mathfrak{D}} \sum_{y \in \mathcal{Y}} \mathcal{L}\left(y, f_{\mathfrak{L}}\left(\boldsymbol{x}\right)\right) p(y | f_{\mathfrak{L}}\left(\boldsymbol{x}\right)) \quad , \tag{5}$$

which is called *estimated empirical risk*. Intuitively, the same label estimates can also be applied to tackle (P3).

[1] If a representative labeled dataset was available, we would not need to query new examples.

Thus, we replace in Eq. (4) the empirical risk $\mathcal{R}_{\text{emp}}$ (see Eq. (2)) with its estimated pendant $\mathcal{R}_{\mathbb{E}}$ from Eq. (5):

$$\Delta\mathcal{R}_{\mathbb{E}}\left(\boldsymbol{x}'\right) = \sum_{y'\in\mathcal{Y}} p(y'|f\left(\boldsymbol{x}'\right)) \cdot \frac{1}{|\mathfrak{D}|} \cdot \sum_{\boldsymbol{x}\in\mathfrak{D}} \sum_{y\in\mathcal{Y}} \dots \quad (6)$$
$$\Big(\mathcal{L}\left(y, f\left(\boldsymbol{x}\right)\right) p(y|f\left(\boldsymbol{x}\right)) - \mathcal{L}\left(y, f'\left(\boldsymbol{x}\right)\right) p(y|f'\left(\boldsymbol{x}\right))\Big) ,$$

where we introduced short-hands $f = f_{\mathfrak{L}}$ and $f' = f_{(\mathfrak{L}\cup(\boldsymbol{x}',y'))}$ for referring to the model before and after including $(\boldsymbol{x}', y')$ as additional training example. For active selection, this results in expected loss reduction for the $0/1$-loss [32, 33] or expected entropy minimization for the log-loss [32, 22, 33, 35, 25].

Unfortunately, reliable estimates of class labels are crucial, which is an ill-posed problem when labeled data for model training is rare and thus prone to errors. Given poor label estimates, samples are selected that change the model's decisions towards the assumed-to-be correct direction or towards the direction with assumed-to-have small entropy, respectively, although estimations of correct directions are hardly reliable. In our recent paper [11], we meet this observation by proposing to instead select samples that are likely to change model outputs in *any* direction. In comparison with Eq. (6), the resulting expected model output change (EMOC) criterion can be expressed by

$$\Delta f(\boldsymbol{x}') = \sum_{y'\in\mathcal{Y}} p(y'|f\left(\boldsymbol{x}'\right)) \cdot \frac{1}{|\mathfrak{D}|} \sum_{\boldsymbol{x}_j\in\mathfrak{D}} \mathcal{L}\left(f(\boldsymbol{x}_j), f'(\boldsymbol{x}_j)\right) . \quad (7)$$

A sample that induces large changes should be preferred for being labeled. From a theoretical point of view, this maximizes an upper bound on error reduction [11] and thereby guarantees that the possible gain in accuracy is not limited in advance by a poor query selection.

### 2.4. Model output changes for binary scenarios

While we introduced the EMOC principle in [11] for arbitrary classification and regression settings (see Eq. (7)), we were only able to propose a concrete algorithm for binary classification scenarios, which uses Gaussian process (GP) regression as a model with continuous outputs together with label regression for performing classification. In [11], we suggested the absolute difference of continuous model outputs as a suitable loss function

$$\mathcal{L}_{|\cdot|}\left(f\left(\boldsymbol{x}\right), f'\left(\boldsymbol{x}\right)\right) = \left|f\left(\boldsymbol{x}\right) - f'\left(\boldsymbol{x}\right)\right| \quad (8)$$

rather than differences of resulting classification decisions which, thus, circumvents estimation of classification thresholds. Although working with continuous outputs, the *a-priori* knowledge of an underlying binary classification task $\mathcal{Y} = \{-1, 1\}$ allows for marginalization over unknown labels $y'$ by simply evaluating the expected model output change for $y' = 1$ and $y' = -1$. As an estimate for the label probability, the probit model of Gaussian process regression is used [11]:

$$p(y' = 1|f\left(\boldsymbol{x}'\right)) = \frac{1}{2} - \frac{1}{2} \cdot \text{erf}\left(-f(\boldsymbol{x}')/\sqrt{2\sigma_f^2}\right) , \quad (9)$$

where the error function $\text{erf}(z)$ represents the cumulative Gaussian noise model for binary classification with Gaussian process models as presented in [30]. The predictive variance $\sigma_f^2$ of $f$ evaluated on $\boldsymbol{x}'$ is given as part of the Gaussian process regression prediction and reflects uncertainty in model decisions. Since our application scenario requires a multi-class system, we show how to extend the EMOC principle in the following section.

## 3. Multi-class expected model output change

In the previous section, we reviewed the EMOC principle for approximating the information gain in binary classification as proposed by [11]. Extending this to multi-class scenarios with labels $y \in \mathcal{Y} = \{1, \dots, C\}$ requires: (1) the definition of a proper loss function $\mathcal{L}$ and (2) the estimation of multi-class probabilities $p(y' = c|\boldsymbol{x}')$. We introduce techniques for both aspects in the following.

**Model output loss function for one-vs-all classifiers** For multi-class classification, we use Gaussian process regression with the one-vs-all principle [21], since it allows for efficient computation of model updates and it is strongly related to one-vs-all SVM, which is the most prominent technique for multi-class classification. However, also other supervised classification techniques could be used in general. The difference between GP regression and SVM boils down to using the hinge loss in the SVM case and a quadratic loss in the GP regression case [21].

Since we use the one-vs-all technique, we learn $C$ binary classifiers $f_c$ with GP regression when a classification problem with $C$ classes is given. Each of the classifiers gives a continuous classification score $f_c(\boldsymbol{x}) \in \mathbb{R}$, which is used to perform classification decisions according to:

$$\bar{y}(\boldsymbol{x}) = \underset{c=1\dots C}{\text{argmax}}\ f_c(\boldsymbol{x}) \ . \quad (10)$$

To measure the model output change, we use the $L_1$-loss on the class-specific scores:

$$\mathcal{L}_{|\cdot|}\left(f\left(\boldsymbol{x}\right), f'\left(\boldsymbol{x}\right)\right)\right) = \sum_{c=1}^{C} \left|f_c(\boldsymbol{x}) - f'_c(\boldsymbol{x})\right| \ . \quad (11)$$

We also experimented with other multi-class losses, *e.g.,* number of label-flips $\mathcal{L}_{0/1}\left(\bar{y}(\boldsymbol{x}), \bar{y}'(\boldsymbol{x})\right) = 1 - \delta_{\bar{y}(\boldsymbol{x}),\bar{y}'(\boldsymbol{x})}$ where $\delta_{\cdot,\cdot}$ is the Kronecker delta, but we did not observed superior performance[2].

[2]Due to the lack of space, we add a small experimental comparison of loss functions and probability estimates in the supplementary material.

**Multi-class classification probabilities** Let $y_c \in \{-1, 1\}$ be the random variable for the binary label $y_c = 1 - 2 \cdot \delta_{y,c}$ for class $c \in \{1, \dots, C\}$. Naively computing multi-class classification probabilities $p(y = c|\boldsymbol{x})$ could be done by computing probabilities $p(y_c = 1|\boldsymbol{x})$ for each binary classification problem as done in Eq. (9). However, these probabilities are derived from the binary problems and not sufficiently normalized. Furthermore, normalizing them afterwards often leads to poor results due to imbalanced scores [29]. A very common alternative is a *multi-class logistic regression model*:

$$p(y = c|\boldsymbol{x}) \propto \exp(\alpha_c \cdot f_c(\boldsymbol{x}) + \beta_c) \tag{12}$$

with class-specific parameters $\alpha_c$ and $\beta_c$ estimated from training data [6]. In our case, we can even make use of leave-one-out estimates to learn the parameters [30].

A disadvantage of the method of [6] is that the predictive variance of the test example is not taken into account. Therefore, we compute *multi-class probabilities directly derived from uncertainty estimates* [12], which leads to higher performance in our experiments[2]. The underlying idea of the uncertainty technique is that for label regression with Gaussian processes [21], we do not only have the model prediction $f_c(\boldsymbol{x})$ but rather the whole posterior distribution $\mathcal{N}(f_c(\boldsymbol{x}), \sigma^2(\boldsymbol{x}))$ independently for each classification score. The probability of class $c$ achieving the maximum score in Eq. (10) can therefore be expressed by

$$p(\bar{y}(\boldsymbol{x}) = c|\boldsymbol{x}) = p\left(c = \underset{c'=1\dots C}{\operatorname{argmax}} f_{c'}(\boldsymbol{x})\right) \quad . \tag{13}$$

To estimate the probabilities, we apply a Monte-Carlo technique and sample $Z$ times from all $C$ Gaussian distributions $\mathcal{N}(f_c(\boldsymbol{x}), \sigma^2(\boldsymbol{x}))$ and estimate the probability of each class

$$p(y = c|\boldsymbol{x}) = p(\bar{y}(\boldsymbol{x}) = c|\boldsymbol{x}) \approx \frac{Z_c}{Z} \quad , \tag{14}$$

with $Z_c$ denoting the number of times where the draw from the distribution of class $c$ was the maximum value. A large variance $\sigma^2$, *i.e.,* a high uncertainty of the estimate, leads to a nearly uniform distribution $p(y = c)$, whereas a zero variance results in a distribution which is equal to one for the class which corresponds to the highest posterior mean.

We now have everything properly defined to apply the EMOC principle in multi-class classification problems.

## 4. Active learning with unnameable instances

**Problem setting** A very common assumption in active learning is that the oracle (*e.g.,* a human annotator) can provide a label for every instance of the set of unlabeled examples. Especially for tasks that involve a large set of categories, this assumption is not reasonable. Therefore, we have to deal with cases where the oracle rejects to label the example that the active learning algorithm just selected. From our experience, there are basically two main scenarios in which a rejection can possibly happen:

1. **Rejection of non-categorial examples:** The unlabeled example does not show a valid object. Possible reasons are *noise* during image acquisition (*e.g.,* sensor noise, motion blur, or JPEG artifacts), segments covering *multiple objects*, or *background* regions. Examples are bounding boxes (b) and (c) in Fig. 1.

2. **Rejection of categorial examples:** The unlabeled example is a valid object, but the annotator is *not able* to name it or he decides that it is *not part* of the problem domain. Examples are bounding boxes (d), (e), and (f) in Fig. 1.

Both cases need to be considered during active learning and we present solutions and adaptations of the EMOC principle for each of them in the following.

**Dealing with non-categorial rejections (GP-EMOC$_{\text{PDE}}$)** The number of images showing no valid objects is vast. However, it is unlikely that during dataset acquisition and proposal generation, the same non-object sample is obtained several times. Thus, samples that do not show valid objects are characterized by a low data density.[3] In contrast, samples from object categories should cluster since different samples from the same category are likely to be recorded over time. Therefore, the examples we query should be in a high density region to ensure a high impact on examples nearby. However, our previous work [11] (see Section 2.3) followed the idea behind empirical risk estimation by exploiting the empirical density using the Dirac function $\delta(\cdot)$:

$$p_\delta(\boldsymbol{x}) = \frac{1}{|\mathfrak{D}|} \sum_{\boldsymbol{x}_j \in \mathfrak{D}} \delta(\boldsymbol{x} - \boldsymbol{x}_j) \quad , \tag{15}$$

induced by all available samples (second term in Eq. (7)). Thus, EMOC values of each example are not taking the local data density into account. In contrast, we propose to use the local data density $p(\boldsymbol{x}')$ obtained with a Parzen estimate

$$p_{\text{PDE}}(\boldsymbol{x}') \propto \frac{1}{|\mathfrak{D}|} \sum_{\boldsymbol{x}_j \in \mathfrak{D}} \mathcal{K}(\boldsymbol{x}_j, \boldsymbol{x}') \quad , \tag{16}$$

where $\mathcal{K}$ is a kernel function measuring sample similarity. The resulting GP-EMOC$_{\text{PDE}}$ can then be expressed by

$$\begin{aligned} \Delta f(\boldsymbol{x}') = & \sum_{y' \in \mathcal{Y}} p(y'|f(\boldsymbol{x}')) \cdot p_{\text{PDE}}(\boldsymbol{x}') \\ & \cdot \Big( \sum_{\boldsymbol{x}_j \in \mathfrak{D}} \mathcal{L}(f(\boldsymbol{x}_j), f'(\boldsymbol{x}_j)) \Big) \quad . \end{aligned} \tag{17}$$

[3] A low data density for non-objects is reasonable, *e.g.,* sensor noise should happen rarely, or segment proposals should by algorithmic design favor objects over non-objects.

This is essential in order to focus on examples in high-density regions rather than on less frequent non-categorial samples. Interestingly, it turns out that integrating the data density is sufficient for allowing to discover new categories. This is reasonable since examples of new categories are located in high density areas and lead to a high expected model output change for similar samples. We also experimented with explicitly incorporating the possibility of new classes into Eq. (17), but found no superior behavior.

**Dealing with categorial rejections (GP-EMOC$_{\text{PDE+R}}$)** It can also be the case that some of the unlabeled examples belong to *unknown or unrelated categories*. These examples are referred to as "blind spots" by [8] and we model them as one big class $r$. In particular, $y' = r$ denotes the event when an annotator would reject the example $\boldsymbol{x}'$ and we need to take this into account when computing the EMOC scores. We make use of the fact that we would not get an additional training example in this case. Thus, the classification model would simply not change, *i.e.*, $\forall \boldsymbol{x} : f'(\boldsymbol{x}) = f(\boldsymbol{x})$, which results in zero expected model output change for the case of $y' = r$. The EMOC value for an example $\boldsymbol{x}'$ under the assumption that there exists a rejection class $r$ is therefore given by:

$$\begin{aligned}\Delta f^r(\boldsymbol{x}') &= \mathbb{E}_{y' \in \mathcal{Y} \cup \{r\}} \, \mathbb{E}_{\boldsymbol{x} \in \Omega} \left( \mathcal{L}\left(f(\boldsymbol{x}), f'(\boldsymbol{x})\right)\right) \\ &= p(y' \neq r | \boldsymbol{x}') \cdot \Delta f(\boldsymbol{x}') + p(y' = r | \boldsymbol{x}') \cdot 0 \\ &= (1 - p(y' = r | \boldsymbol{x}')) \cdot \Delta f(\boldsymbol{x}') \ . \end{aligned} \tag{18}$$

In practice, we estimate the probability $p(y' = r|\boldsymbol{x}')$ of an example $\boldsymbol{x}'$ being an unnameable instance by using a GP regression classifier learned with previously rejected instances as positive examples and all samples of known classes as negatives. The classification score predicted by the classifier is transformed into a valid probability value using the probit model of Eq. (9) with the binary classifier corresponding to class $r$. As a byproduct, this allows to also model rejections for non-categorial samples. In addition, we also add all rejected examples as negatives to each of the one-vs-all binary classifiers, a strategy that has shown to be valuable also for task adaptation with large-scale datasets [18].

Before we experimentally validate our introduced techniques on several applications in Section 6, we give a short overview on related work in the following section.

## 5. Related work on active learning

Active learning is a well-known field of research for several years. Thus, a large variety of techniques has been developed to estimate the information of unlabeled data to then select most informative samples. We briefly review established approaches and put our work into perspective.

Existing criteria are often based on intuitive assumptions about what makes a sample informative. Focusing on a rapid exploration of feature space is one prominent example [2, 21], or selecting samples which likely result in changes of model parameters [33, 10, 3]. A completely different field approaches active learning by selecting samples the current model is most uncertain about, known as uncertainty or entropy sampling [24, 4, 34, 17, 20, 21, 9, 5]. Implications on changes in classification accuracy are taken into account in expected error reduction or expected entropy minimization [32, 22, 33, 35, 25]. Finally, combinations of multiple strategies are investigated in [2, 15, 7].

While estimated error- and entropy-reduction techniques require reliable label estimates, we proposed a less stringent alternative based on expected model output changes (EMOC) in [11]. The approach prefers samples that likely result in *any* change of classification decisions, even if expected errors given current model estimates would be increased. The resulting technique in [11], however, is restricted to binary classification scenarios. In this paper, we extend the EMOC principle to *multi-class* active learning settings and *discovery* of new classes. For dealing with noisy oracles in crowd-sourcing, [28] combine the maximum entropy framework with estimates about expected labeling quality. In contrast, we assume experts to be correct, but with the additional *option of rejecting examples* during labeling. Similar in spirit is the work in [8], which models the possibility of oracle rejections in binary classification scenarios. Their assumption of zero classification entropy for rejected samples however is likely to fail in practice. Therefore, we differ both in tackling the more general multi-class setting and a sound theoretical model for selective oracles. Closest in terms of targeted application is the inspiring work of [14] which introduces a framework for joint active learning and class discovery. Unfortunately, their approach does not scale to applications with several thousands of dimensions and unlabeled samples (see Section 6.3). In addition, we empirically show that no explicit modeling of unknown classes is required when incorporating data density instead.

## 6. Experiments

In this section, we evaluate our approach on three datasets, including face identification, digit recognition, and object classification, and compare against several baselines. Our proposed methods are the following:

1. GP-EMOC$_{\text{MC}}$ (see Eq. (7)) is our multi-class extension of the EMOC principle,
2. GP-EMOC$_{\text{PDE}}$ (see Eq. (17)) incorporates the data density for resistance to far-off non-categorial samples,
3. GP-EMOC$_{\text{PDE+R}}$ (see Eq. (18)) additionally models possible rejections of unnameable instances.

### 6.1. Baselines and general setup

For evaluating our active learning criterion, we used the source code of [11], and we extended it to multi-class scenarios with unnameable instances[4]. We compare our approach with the predictive variance (GP-var) as well as uncertainty (GP-unc) of Gaussian processes [21], the best-vs-second-best strategy (1–vs–2) proposed in [20], the multi-class query strategy based on probabilistic KNN classifiers (PKNN)[5] [17], the empirical risk minimization approach of [32] applied to GP (ERM), and the Dirichlet process expected accuracy (DPEA)[6] [14]. Furthermore, we also include the baseline of random querying.

In all our experiments, we start with an initial set of 2 known classes and 5 training samples per class, both randomly selected but identical for each method. We randomly select 10 tasks by splitting classes in known and unknown, and each task is randomly initialized 10 times, resulting in 100 individual test scenarios to average over. After querying and labeling a sample, the classification model is updated and evaluated on a held out test set of 30 samples per class. Note that in the beginning, the test set also contains samples of classes that are not known to the system since the total number of classes is larger than the number of classes in the initial training set.

All samples that are neither in the test set nor in the initial training set are treated as the unlabeled pool. Furthermore, unnameable samples are added individually to each dataset as described in the following sections and are additionally cross-checked by human annotators. In all settings, we are interested in fast discovery of all classes as well as high recognition accuracy. Further experiments with a focus on labeling times and qualitative evaluations can be found in the supplementary material.

### 6.2. Proof-of-Concept on USPS

For a proof-of-concept, we use the well known USPS dataset [1] which results in recognizing handwritten digits.

**Dataset and unnameable instances** The USPS dataset contains $16\times16$ grayscale images of handwritten digits $(0-9)$. Features are extracted by concatenating gray values which are quantized into levels between 0 and 16 and an RBF-kernel serves as ad-hoc measure of similarity. In addition to the unlabeled samples that come from the setting described in Section 6.1, unnameable samples are artificially created by randomly rearranging gray values of every second image and including them in the unlabeled pool.

[4]Source code available at https://github.com/cvjena/.

[5]Source code obtained from http://research.microsoft.com/en-us/um/people/akapoor/cvpr2009/

[6]Source code obtained from http://www.eecs.qmul.ac.uk/~tmh/

Table 1: Experimental results for USPS [1], LFW [16], and COCO [27] showing accuracy (in %) after 100 queries averaged over 100 experiments. $^{(*)}$ not possible due to excessive memory demand. See text for further details.

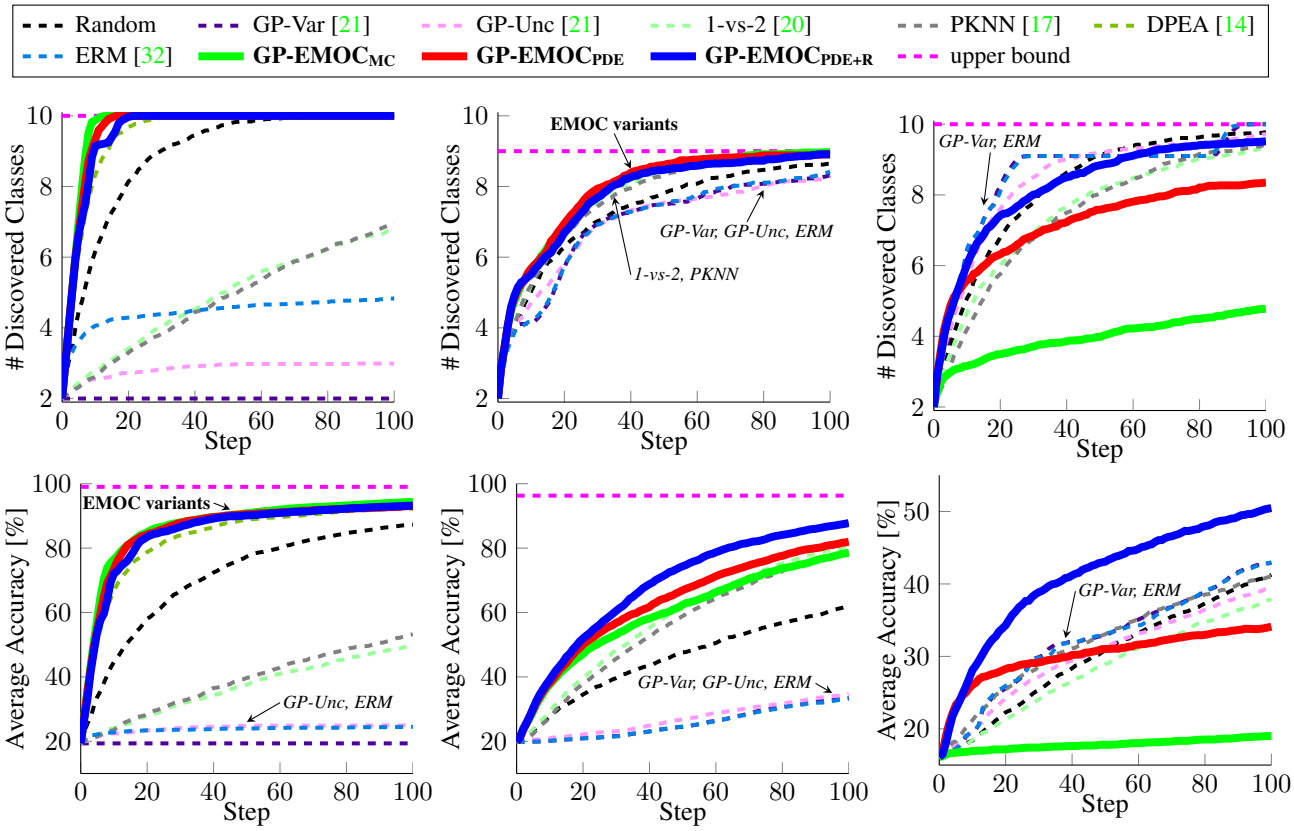
| Strategy | USPS [1] | LFW [16] | COCO [27] |
|---|---|---|---|
| Random | 87.27 | 62.01 | 41.27 |
| GP-Var [21] | 19.35 | 33.45 | 42.94 |
| GP-Unc [21] | 24.94 | 34.74 | 39.49 |
| 1-vs-2 [20] | 49.54 | 80.58 | 37.87 |
| PKNN [17] | 53.26 | 78.88 | 41.03 |
| DPEA [14] | 92.20 | $-^{(*)}$ | $-^{(*)}$ |
| ERM [32] | 24.49 | 33.21 | 42.98 |
| **GP-EMOC$_{\text{PDE+R}}$** | **93.21** | **87.79** | **50.48** |

**Evaluation** The results in terms of discovered classes and change in classification accuracy are shown in Fig. 2a and Table 1. We clearly observe that all EMOC strategies rapidly discover unknown classes and substantially improve recognition rates. The DPEA method [14] is slightly inferior and needs roughly 40 queries to discover all classes. In direct comparison, our EMOC variants reliably find all classes after only ten to twenty queries – a reduction by roughly one third – without the necessity of explicitly modeling potentially occurring classes. Notably, the elegant method PKNN [17], the established 1–vs–2 scheme [20], and expected risk minimization [32] perform worse than random query selection. This is due to the fact that they are not designed for discovering new classes, which can be clearly observed from Fig. 2a. Interestingly, GP-Var and GP-unc mainly queries unnameable samples, although proven to be valuable in other scenarios [21].

### 6.3. Face identification

The identification of faces in images is an important task in real-life applications, where unnameable samples (*e.g.,* false-positive face detections) or unrelated examples might occur. For example, a user is likely interested in only labeling and recognizing faces of friends in a huge collection of photos rather than annotating all faces in images that may belong to unknown persons in the background.

**Dataset and experimental setup** The Labeled-Faces-in-the-Wild (LFW) dataset [16] contains face images of 5,749 individuals. In our experiments, we use the face features provided by [13] consisting of SIFT descriptors extracted at nine detected landmark positions and three different scales. The 27 individual descriptors are concatenated and the whole vector is $L_1$-normalized. Consequently, a histogram intersection kernel [31] serves as similarity measure. We use the 9 classes that contain at least 55 images to represent the current problem domain. Additionally, we use 400 ran-

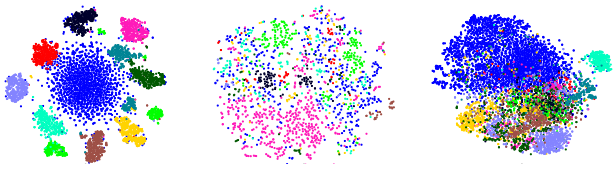



(a) USPS [1] (b) LFW [16] (c) COCO [27]

Figure 2: Evaluating active class discovery (*top*) and improving recognition accuracy with active learning (*bottom*). Results are obtained on the *USPS dataset* [1], the *Labeled-Faces-in-the-Wild dataset* [16], and the *COCO dataset* [27]. Baselines are indicated with dotted lines, whereas our techniques are plotted solidly. Overlapping learning curves of multiple strategies are additionally labeled in each plot. See text for details on the experimental setup. Best viewed in color.

domly selected classes that contain only a single image to provide unnameable samples for the unlabeled pool.

**Evaluation** For the face identification dataset, the results are shown in Fig. 2b and Table 1. We consistently observe that GP-var and GP-unc [21] as well as ERM [32] lead to discovery results significantly worse than random selection. Remaining techniques obtain roughly identical results slightly superior to passive learning. Note that no results for the close competitor DPEA are given, since the source code provided by [14] is not able to handle scenarios with that many samples and dimensions. Interestingly, no method is able to discover all classes after 50 queries and only some come close after 100 queries. We attitude this behavior to the cluttered, unordered distribution of data in space (see Fig. 3b) which hinders identification of new clusters. With respect to classification accuracy, GP-var, GP-unc, and ERM only obtain minor improvements, since requested samples are mostly rejected. In contrast, PKNN and 1-vs-2 can improve on random selection over time. However, they still require twice as much queries compared to our EMOC strategies to obtain equal performance. In addition, we notice that our simplest EMOC technique is significantly inferior to GP-EMOC$_{PDE+R}$, resulting in 10% performance gain.



(a) USPS [1] (b) LFW [16] (c) COCO [27]

Figure 3: T-SNE visualizations of datasets used in evaluations, colored according to ground truth class membership. Note how strongly dataset characteristics differ, *e.g.,* with respect to class distances or cluttered classes.

Figure 4: Object proposals extracted from the COCO dataset [27]. Our active learning approach can cope with these heterogeneous samples and with rejections of an annotator. *Upper*: nameable segments of the current problem domain. *Middle*: unknown or unrelated objects. *Lower*: segments not even corresponding to valid objects.

### 6.4. Active discovery with object proposals

Our active learning approach can handle heterogeneous unlabeled data containing unnameable examples and unknown categories, which would be rejected by an annotator. These properties are especially valuable when an annotator is provided with automatic object proposals which likely capture not only the classes of interest but also background artifacts. Thus, it provides another realistic scenario for our active discovery scheme.

**Dataset and experimental setup** For our experiments, we use a subset of the COCO training dataset [27] and extract object proposals with the geodesic object proposal method of [23]. The dataset for our experiment is created as follows: As a problem domain, we select all animal categories[7]. Segments that overlap with more than an intersection-over-union (IoU) value of 0.9 with a ground-truth object of one of these categories are considered as valid objects and labeled accordingly. Randomly chosen segments with no overlap with a ground-truth object are considered as unnameable segments, which would be rejected by an annotator. These segments can be categorial examples (objects of non-animal categories) and non-categorial instances (wrongly detected object proposals). In total, we use 10,000 random images of the dataset, which contain at least one of the objects of our problem domain. We start with 2 categories and 5 examples for each of them as in all other experiments. Features are extracted using outputs of `pool5`, a layer of a convolutional neural net (CNN) provided by the Caffe framework [19] and trained on ImageNet images. Given the high feature dimensionality, a simple linear kernel is applied. These features have shown to be powerful for scene understanding tasks, although they have been learned from internet images not related to scenes as contained in the COCO dataset.

[7] *bird*, *cat*, *dog*, *horse*, *sheep*, *cow*, *elephant*, *bear*, *zebra* and *giraffe*.


Figure 5: First three queried segments in our active discovery scenario with the COCO dataset. Successfully labeled queries are marked *green*, whereas rejected samples are indicated with an *orange* frame.

**Evaluation** Fig. 5 shows object detection results that we obtain after different numbers of additional active learning and discovery queries. A quantitative evaluation is given in Fig. 2c and Table 1. Note that for a better visualization, the upper bound (roughly 76%) is left out. In summary, we observe a similar pattern as in the previous experiments. The inferior results of GP-EMOC$_{\text{PDE}}$ can be explained by the structure of unnameable instances in the COCO dataset (categorial as well as homogenous segments) as can be seen in Fig. 3. However, we observe our GP-EMOC$_{\text{PDE+R}}$ strategy resulting in highest performance compared to all other methods. Thus, respecting data density and modeling possible rejections within evaluating information gain nicely pays off and boosts performance by roughly 30%.

## 7. Conclusions and future work

We proposed a new method for multi-class active learning and class discovery which offers several advantages in real-world scenarios. Driven by the observation that human annotators are not always able or willing to label requested instances, we extended and generalized the expected model output change principle to handle resulting challenges. In extensive empirical evaluations, including a simple digit recognition experiment, a challenging face identification application, and a realistic object classification task, we compared our approach with established active learning methods. Results clearly indicate that our approach is able to outperform previous methods with respect to the number of new categories discovered, the average accuracy obtained by requesting influential unlabeled examples, and by avoiding label rejections of the annotator. Applying the same key ideas to support vector machines or neural network architectures is possible and will be future work.

## References


[1] K. Bache and M. Lichman. UCI machine learning repository, 2013. 6, 7

[2] Y. Baram, R. El-Yaniv, and K. Luz. Online choice of active learning algorithms. *Journal of Machine Learning Research (JMLR)*, 5:255–291, 2004. 5

[3] W. Cai, Y. Zhang, S. Zhou, W. Wang, C. H. Q. Ding, and X. Gu. Active learning for support vector machines with maximum model change. In *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML, PKDD)*, pages 211–226, 2014. 5

[4] C. Campbell, N. Cristianini, and A. Smola. Query learning with large margin classifiers. In *International Conference on Machine Learing (ICML)*, pages 111–118, 2000. 2, 5

[5] B. Demir and L. Bruzzone. A novel active learning method in relevance feedback for content-based remote sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2014. 2, 5

[6] K. Duan, S. S. Keerthi, W. Chu, S. K. Shevade, and A. N. Poo. Multi-category classification by soft-max combination of binary classifiers. In *Multiple Classifier Systems (MCS)*, pages 125–134, 2003. 4

[7] S. Ebert, M. Fritz, and B. Schiele. Ralf: A reinforced active learning formulation for object class recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3626–3633, 2012. 5

[8] M. Fang and X. Zhu. I dont know the label: Active learning with blind knowledge. In *International Conference on Pattern Recognition (ICPR)*, pages 2238–2241, 2012. 5

[9] A. Freytag, E. Rodner, P. Bodesheim, and J. Denzler. Rapid uncertainty computation with gaussian processes and histogram intersection kernels. In *Asian Conference on Computer Vision (ACCV)*, pages 511–524, 2012. 5

[10] A. Freytag, E. Rodner, P. Bodesheim, and J. Denzler. Labeling examples that matter: Relevance-based active learning with gaussian processes. In *German Conference on Pattern Recognition (GCPR)*, pages 282–291, 2013. 1, 5

[11] A. Freytag, E. Rodner, and J. Denzler. Selecting influential examples: Active learning with expected model output changes. In *European Conference on Computer Vision (ECCV)*, pages 562–577, 2014. 1, 2, 3, 4, 5, 6

[12] B. Fröhlich, E. Rodner, M. Kemmler, and J. Denzler. Large-scale gaussian process multi-class classification for semantic segmentation and facade recognition. *Machine Vision and Applications (MVA)*, 24(5):1043–1053, 2013. 4

[13] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *International Conference on Computer Vision (ICCV)*, pages 498–505, 2009. 6

[14] T. M. Hospedales, S. Gong, and T. Xiang. A unifying theory of active discovery and learning. In *European Conference on Computer Vision (ECCV)*, pages 453–466, 2012. 5, 6, 7

[15] T. M. Hospedales, S. Gong, and T. Xiang. Finding rare classes: Active learning with generative and discriminative models. *IEEE Transactions on Knowledge and Data Engineering*, 25(2):374–386, 2013. 5

[16] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 6, 7

[17] P. Jain and A. Kapoor. Active learning for large multi-class problems. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 762 –769, 2009. 1, 2, 5, 6, 7, 8

[18] Y. Jia and T. Darrell. Latent task adaptation with large-scale hierarchies. In *International Conference on Computer Vision (ICCV)*, pages 2080–2087, 2013. 5

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 8

[20] A. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2372 –2379, 2009. 1, 2, 5, 6, 7

[21] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Gaussian processes for object categorization. *International Journal of Computer Vision (IJCV)*, 88:169–188, 2010. 1, 2, 3, 4, 5, 6, 7

[22] A. Kovashka, S. Vijayanarasimhan, and K. Grauman. Actively selecting annotations among objects and attributes. In *International Conference on Computer Vision (ICCV)*, pages 1403–1410, 2011. 1, 2, 3, 5

[23] P. Krähenbühl and V. Koltun. Geodesic object proposals. In *European Conference on Computer Vision (ECCV)*, pages 725–739, 2014. 8

[24] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 3–12, 1994. 2, 5

[25] X. Li and Y. Guo. Multi-level adaptive active learning for scene classification. In *European Conference on Computer Vision (ECCV)*, pages 234–249, 2014. 1, 2, 3, 5

[26] X. Li, L. Wang, and E. Sung. Multilabel svm active learning for image classification. In *International Conference on Image Processing (ICIP)*, volume 4, pages 2207–2210, 2004. 2

[27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 6, 7, 8

[28] C. Long, G. Hua, and A. Kapoor. Active visual recognition with expertise estimation in crowdsourcing. In *International Conference on Computer Vision (ICCV)*, pages 3000–3007, 2013. 2, 5

[29] J. Milgram, M. Cheriet, and R. Sabourin. Estimating accurate multi-class probabilities with support vector machines. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1906–1911, 2005. 4

[30] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. The MIT Press, Cambridge, MA, USA, 2006. 3, 4

[31] E. Rodner, A. Freytag, P. Bodesheim, and J. Denzler. Large-scale gaussian process classification with flexible adaptive histogram kernels. In *European Conference on Computer Vision (ECCV)*, pages 85–98, 2012. 6

[32] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *International Conference on Machine Learning (ICML)*, pages 441–448, 2001. 2, 3, 5, 6, 7

[33] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009. 3, 5

[34] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research (JMLR)*, 2:45–66, 2002. 2, 5

[35] S. Vijayanarasimhan and K. Grauman. Cost-sensitive active visual category learning. *International Journal of Computer Vision (IJCV)*, 91:24–44, 2011. 1, 2, 3, 5

[36] YouTube. Youtube statistics, Nov. 2014. 1