Multi-class Multi-annotator Active Learning with Robust Gaussian Process for Visual Recognition

Chengjiang Long*
*Stevens Institute of Technology
Hoboken, NJ, USA 7030
clong@stevens.edu

Gang Hua*,†
†Microsoft Research Asia
Haidian District Beijing, P.R. China 100080
ganghua@gmail.com

Abstract

Active learning is an effective way to relieve the tedious work of manual annotation in many applications of visual recognition. However, less research attention has been focused on multi-class active learning. In this paper, we propose a novel Gaussian process classifier model with multiple annotators for multi-class visual recognition. Expectation propagation (EP) is adopted for efficient approximate Bayesian inference of our probabilistic model for classification. Based on the EP approximation inference, a generalized Expectation Maximization (GEM) algorithm is derived to estimate both the parameters for instances and the quality of each individual annotator. Also, we incorporate the idea of reinforcement learning to actively select both the informative samples and the high-quality annotators, which better explores the trade-off between exploitation and exploration. The experiments clearly demonstrate the efficacy of the proposed model.

1. Introduction

Most of the current recognition systems are based on supervised learning with large quantity of labeled training data [15, 29]. In recent years, crowd-sourcing is has been explored to collect large-scale labeled image datasets, such as ImageNet [4] and LabelMe [28].

There are still several issues raised when using the current crowd-sourcing systems like Amazon Mechanical Turk. First of all, the collected labels could be very noisy from irresponsible or low-quality annotators. Secondly, there is no mechanism to control the label quality online. Last but not least, there is no mechanism to prioritize the data to be labeled. To make the most use of the scarce human resource and facilitate more efficient data labeling, active learning has been explored in some previous works [5, 13, 14, 23, 32, 33] to enhance the efficacy of the labeled data for a generalizable model.

However, most previous active learning approaches only investigate the case with a single human oracle on the assumption that the provided labels are noise-free. Theoretically, due to human perception variations, multiple annotators are likely to provide diverse labels for some inherent ambiguous examples even if they are responsible, not to mention random behaviors from the irresponsible annotators. Hence, the problem of active learning with multiple annotators under the condition that multiple annotators may provide noisy labels has not been fully explored, although previous works such as Hua *et al.* [9] and Long *et al.* [21, 22] have studied it under the context of binary classification.

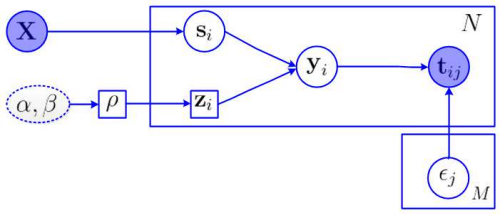


Figure 1: The graphical model of the proposed Gaussian process classifier, with multiple noisy labels from multiple annotators in crowdsourcing.

Indeed, most existing research works in active learning explore merely on binary classification [7, 9, 10, 21, 27, 40, 42]. Relatively fewer approaches have been investigated for multi-class active learning as discussed in [11] and many of them are direct extension of binary active learning approaches to the multi-class scenario [12]. However, many real visual recognition are multi-class application problems and it is possible that the performance of active learning will be degraded by decomposing a multi-class problem as several independent binary classification subproblems. Therefore, the problem multi-class active learning algorithms with collaborative multiple annotators deserves further exploration.

In this paper, we propose a Bayesian multi-class classification model which explicitly models the expertise level of

each individual annotator from crowds, as shown in Figure 1. Expectation propagation (EP) is adopted for efficient approximate Bayesian inference of our probabilistic model for classification. Based on the EP approximation inference, a generalized Expectation Maximization (GEM) algorithm is derived to estimate both the parameters for instances and the expertise of each individual annotator. Active learning is also explored to select the high-quality annotators and guide them to label the most informative visual examples. We incorporate the ideas of reinforcement learning [5] to determine the optimal strategy to actively select both the samples and higher quality annotators, which effectively strikes for a balance between exploitation and exploration.

Several aspects distinguish our work from previous multi-class active learning based labeling [5, 11, 19, 31, 37, 41]: first of all, our proposed Gaussian process classification model uses a back-up mechanism [8], which is robust when the label errors occur far away from the decision boundaries; Secondly, we enable the active learning with multiple annotators who may label an example incorrectly, a topic which has not been sufficiently explored before. Thirdly, we achieve better exploration-exploitation trade-off for collaborative active learning, which finally leads to a unified Gaussian porcess model that simultaneously model the noise from multiple annotators.

2. Related work

The related works fall into 4 categories: *Crowd-sourcing labels, Multi-class Active learning, Active learning with Gaussian process, and Reinforcement learning*.

Crowd-sourcing labels. To handle the label noises in crowd-sourcing, besides Zhao *et al.*'s incremental relabeling mechanism [43], there are existing research works that explore to model the annotator's quality [9, 30, 34, 38, 39]. For modeling multiple annotators' quality for image labeling from crowds, the most relevant works to our research is Welinder *et al.*'s Bayesian model for the annotation process [34, 35]. We shall emphasize that they rely on the noisy labels for each images from multiple annotators and never extract visual features. In contrast, our work in this paper is designed directly on visual features extracted from images and models the annotators' quality to enable the active selection of high-quality annotators to obtain the reliable labels.

Multi-class active learning. The existing multi-class active learning approaches can be divided to two categories. One type of methods decompose the multi-class problem to binary cases [11,12,19,25,26,37], and the other type of methods deal with the multi-class scenario directly [1, 31, 41]. Three most recent works are Yang *et al.*'s a multi-class active learning algorithm [41] that explores the uncertain evaluation with diversity maximization, Aodha *et al.*'s graph-based active learning framework [1] to effectively explore the potential of Expected Error Reduction (EER), and Vasisht *et al.*'s non-myopic and near-optimal active learning with sparse Bayesian multi-label graphical model [31]. However, all these existing algorithms haven't considered the multiple-annotator scenario, which is the focus of this paper.

Active learning with Gaussian process. Regarding active learning with Gaussian processes, Kapoor *et al.* [13] extended Lawrence *et al.*'s work [16] by introducing a heuristic confidence criterion for active selection of the informative instance based on the variance of the posterior prediction for active learning. Recently, Long *et al.* [21] and Rodrigues *et al.* [27] proposed general Gaussian process classifiers in multiple-annotator settings. However, both of these two active learning algorithms focus only on the binary classification case. In contrast, our approach aims to directly deal with multi-class cases, which integrates the potential diverse opinions from multiple annotators and introduces a new heuristic for actively selecting the high-quality annotators to label the informative instances.

Reinforcement learning. The Markov decision process (MDP) provides the general framework to make a decision with respect to the discrimination dynamics. Ebert *et al.* [5] formulated the active selection criteria based on the MDP-based reinforcement learning to adapt the trade-off between exploration and exploitation and obtained promising experimental results. In this paper, we extend the MDP-based framework to determine the criteria for active selection of both the informative samples and the high-quality annotators.

3. Formulation, inference, and learning

Given a set of N data points $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, where $\mathbf{x}_i \in \Re^D$. We let M annotators label each $\mathbf{x}_i$. With s_i denoting the latent random variable with a Gaussian process prior, s_i can be interpreted intuitively as the soft score for the corresponding data point $\mathbf{x}_i$. We denote the hidden true label of $\mathbf{x}_i$ as y_i, and the observed label of $\mathbf{x}_i$ from annotator j as t_{ij}. Note that t_{ij} could be noisy, *i.e.*, t_{ij} may not be consistent with the hidden true label y_i. We denote $\mathbf{t}_i = \{t_{ij}\}_{j=1}^M$ as the set of labels from the M annotators for $\mathbf{x}_i$.

Assuming there are C categories and $\mathbf{S} = \{\mathbf{S}^k | k = 1, \ldots, C\}$, then the Gaussian process prior of the overall function value $\mathbf{S}^k$ is defined as a normal distribution. For notation convenience, we denote $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_N\}$, $\mathbf{Y} = \{y_1, y_2, ..., y_N\}$ and $\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_N\}$.

3.1. Probabilistic Model

As illustrated in the graphical model in Figure 1, the conditional joint probability of our proposed probabilistic mod-

el is defined as

$$p(\mathbf{S},\mathbf{Y},\mathbf{T},\rho,\mathbf{z}|\mathbf{X},\Theta) \propto p(\mathbf{S}|\mathbf{X})p(\rho)p(\mathbf{z}|\rho)\prod_{i=1}^{N}\left\{p(y_i|\mathbf{s}_i,z_i)\prod_{j=1}^{M}p(t_{ij}|y_i,\epsilon_j)\right\}, \tag{1}$$

where Θ is the hyperparameter and $\mathbf{z}=\{z_1,\ldots,z_N\}$ is a set of binary latent variables for each visual instance to indicate whether $s_i^{y_i}\geq s_i^k$ for any $k\neq y_i$ ($z_i=0$) or not ($z_i=1$).

In this paper, $p(\mathbf{S}|\mathbf{X})$ is defined as a Gaussian process prior [36] with kernel tricks, *i.e.*,

$$p(\mathbf{S}|\mathbf{X})=\prod_{k=1}^{C}\mathcal{N}(\mathbf{S}^k|\mathbf{0},\mathbf{K}^k), \tag{2}$$

where $\mathbf{K}^k=[\mathbf{k}(\mathbf{x}_i,\mathbf{x}_j)]_{i,j=1}^N$ is a kernel matrix defined over the set of all N data samples. This treatment ensures that similar data samples to have similar prediction scores. Usually, any valid kernel can be used in our formulation.

$p(y_i|\mathbf{s}_i,z_i)$ is a *back-up* mechanism to deal with label noises and is defined as

$$p(y_i|\mathbf{s}_i,z_i)\left[\prod_{k\neq y_i}H(s_i^{y_i}-s_i^k)\right]^{1-z_i}\left[\frac{1}{C}\right]^{z_i}, \tag{3}$$

where $H(x)=1$ if $x>0$ and $H(x)=0$, otherwise. Note that the first term directly depends on the accuracy of $\mathbf{s}^{y_i}$. In particular, it takes value 1 when the corresponding instance is correctly classified and 0 otherwise. Our model is robust when the observed data contain labeling errors far from the decision boundaries, because the likelihood function described in Equation 3 considers only the total number of prediction errors made by $\mathbf{s}^{y_i}$, rather than the distance of these errors to the decision boundary.

The conditional probability $p(t_{ij}|y_i,\epsilon_j)$ is assumed as a flipping noise model [24], *i.e.*,

$$p(t_{ij}|y_i,\epsilon_j)=\epsilon_j H(y_i=t_{ij})+(1-\epsilon_j)H(y_i\neq t_{ij}). \tag{4}$$

The intuition here is that we assume that t_{ij} will be a flipped version of y_i with probability $(1-\epsilon_j)$. Obviously, the larger ϵ_j leads to the higher the probability that t_{ij} will agree with the true label y_i, and vice versa. Therefore, we can use ϵ_j to naturally represents the quality of the labels given by annotator j. Note that different from [24], we parameterize the flip model based on label quality, which equals one minus the label noise.

The prior $p(\mathbf{z}|\rho)$ is defined as a factorizing multivariate Bernoulli distribution

$$p(\mathbf{z}|\rho)=Bern(\mathbf{z}|\rho)=\prod_{i=1}^{N}\rho^{z_i}(1-\rho)^{1-z_i}, \tag{5}$$

where ρ is the prior fraction of training instances expected to be outliers. And the prior for ρ is set to be a conjugate beta distribution, *i.e.*,

$$p(\rho)=Beta(\rho|\alpha,\beta)=\frac{\rho^{\alpha-1}(1-\rho)^{\beta-1}}{B(\alpha,\beta)}, \tag{6}$$

where $B(\cdot,\cdot)$ is the the beta function and α and β are hyperparameters.

3.2. Inference

By integrating y_i out, we can reach the probability on Equation 3 and 4 as,

$$p(\mathbf{t}_i|\mathbf{s}_i,z_i,\epsilon)=\sum_{y_i=1}^{C}p(y_i|\mathbf{s}_i,z_i)\prod_j p(t_{ij}|y_i,\epsilon_j) \tag{7}$$

and then the joint probability in Equation 1 can be rewriten as

$$p(\mathbf{T},\mathbf{S},\mathbf{z},\rho|\mathbf{X},\epsilon)\propto p(\mathbf{S}|\mathbf{X})p(\rho)p(\mathbf{z}|\rho)\prod_i p(\mathbf{t}_i|\mathbf{s}_i,z_i,\epsilon). \tag{8}$$

Such a collapsed joint probability enable us to conveniently derive the EP inference algorithm.

Given a set of labeled data samples $\mathbf{X}_L=\{\mathbf{x}_1,...,\mathbf{x}_L\}$, the set of labels are denoted as $\mathbf{T}_L=\{t_{ij}|1\leq i\leq N,1\leq j\leq M\}$ and an unlabeled data sample $\mathbf{x}_u$, we need to solve the following Bayesian inference problem to predict the label y_u of a $\mathbf{x}_u$. We denote $\mathbf{D}_L=\{\mathbf{X}_L,\mathbf{T}_L\}$, $\mathbf{S}=\{\mathbf{S}_L,\mathbf{s}_u\}$, and $\mathbf{X}=\{\mathbf{X}_L,\mathbf{x}_u\}$, and then we arrive at,

$$\begin{aligned}p(y_u|\mathbf{x}_u,\mathbf{D}_L)&=\sum_{z_u}\int_{\mathbf{S}}p(y_u|\mathbf{s}_u,z_u)p(z_u|\rho)p(\mathbf{S}|\mathbf{D}_L,\mathbf{x}_u)d\rho d\mathbf{S}\\&=\sum_{z_u}\int_{\mathbf{s}_u}p(y_u|\mathbf{s}_u,z_u)p(z_u|\rho)\int_{\mathbf{S}_L}p(\mathbf{S}|\mathbf{D}_L,\mathbf{x}_u)d\rho d\mathbf{S}_L d\mathbf{s}_u\end{aligned} \tag{9}$$

where

$$p(\mathbf{S}|\mathbf{D}_L,\mathbf{x}_u)\propto p(\mathbf{S}|\mathbf{X})p(\rho)\prod_{\mathbf{s}_i\in\mathbf{S}_L}p(z_i|\rho)p(\mathbf{t}_i|\mathbf{s}_i,z_i,\epsilon). \tag{10}$$

Let $\mathbf{\Psi}$ be the set that contains all these exact factors, and then we can rewrite $p(\mathbf{S}|\mathbf{D}_L,\mathbf{x}_u)$ in Equation 10 as $\left[\prod_{k=1}^{C}\psi_k\right]\psi_\rho\left[\prod_{s_i\in\mathbf{S}_L}\psi_i\right]\left[\prod_{s_i\in\mathbf{S}_L}\psi_{it}\right]$, where $\psi_k=\mathcal{N}(\mathbf{S}^k|\mathbf{0},\mathbf{K}^k)$, $\psi_i=p(\mathbf{z}|\rho)$, $\psi_\rho=p(\rho)$ and $\psi_{it}=p(\mathbf{t}_i|\mathbf{s}_i,z_i,\epsilon)$. The integral in Equation 9 is intractable as neither $p(\mathbf{S}|\mathbf{D}_L,\mathbf{x}_u)$ nor $p(y_u|\mathbf{s}_u,z_u)$ can be integrated in closed form.

We resort to Expectation Propagation [24] to obtain an approximate integral by approximating each $\psi \in \Psi$ using a corresponding simple factor $\tilde{\psi}$ such that $\psi_\rho \left[\prod_{s_i \in \mathbf{S}_L} \psi_i\right]\left[\prod_{s_i \in \mathbf{S}_L} \psi_{it}\right] \approx \tilde{\psi}_\rho \left[\prod_{s_i \in \mathbf{S}_L} \tilde{\psi}_i\right]\left[\prod_{s_i \in \mathbf{S}_L} \tilde{\psi}_{it}\right]$ with the constraint that all the approximate factors $\tilde{\psi}$ belong to the same family of exponential distributions. And then we can approximate the posterior distribution of as the normalized product of the approximate factors, *i.e.*,

$$Q(\mathbf{S}, \mathbf{z}, \rho) = \frac{1}{Z}\left[\prod_{k=1}^{C} \tilde{\psi}_k\right] \tilde{\psi}_\rho \left[\prod_{s_i \in \mathbf{S}_L} \tilde{\psi}_i\right]\left[\prod_{s_i \in \mathbf{S}_L} \tilde{\psi}_{it}\right] \tag{11}$$

where Z is a normalization constant that approximates $p(\mathbf{S}|\mathbf{D}_L, \mathbf{x}_u)$. Note that exponential distributions are preserved with product and division operations. Since all the approximate factors belong to the exponential family, Q has the same form as the approximate factors and Z can be readily computed.

In implementation, we select the form of Q first and then constrain the approximate factors to reach the same form as Q. The relation between each approximate factor $\tilde{\psi}$ and the corresponding exact factor ψ is defined $Q^{\setminus\tilde{\psi}} \propto Q/\tilde{\psi}$. By minimizing the Kullback-Leibler (KL) divergence between $\psi Q^{\setminus\tilde{\psi}}$ and $\tilde{\psi} Q^{\setminus\tilde{\psi}}$, all the $\tilde{\psi}$ can be iteratively updated one by one. The steps of the EP algorithm are summarized in Algorithm 1.

Algorithm 1 The Expectation Propagation Algorithm

1: Initiate each approximate factor $\tilde{\psi}$ and the posterior approximation Q .
2: **repeat**
3: Choose one $\tilde{\psi}$ to refine and compute $Q^{\setminus\tilde{\psi}} \propto Q/\tilde{\psi}$.
4: Update $\tilde{\psi}$ by minimizing $KL(\psi Q^{\setminus\tilde{\psi}} || \tilde{\psi} Q^{\setminus\tilde{\psi}})$.
5: Update the posterior approximation Q to the normalized version of $\tilde{\psi} Q^{\setminus\tilde{\psi}}$.
6: **until** convergence
7: Evaluate $Z \approx p(\mathbf{Y}|\mathbf{X})$ as the integral of the product of all the approximate factors.

EP obtains a Gaussian approximation $Q(\mathbf{S})$ to the posterior distribution $p(\mathbf{S}|\mathbf{D}_L, \mathbf{x}_u)$. Hence the integral over $\mathbf{S}_L$ in Equation 9 can also be approximated by a Gaussian distribution over $\mathbf{s}_u$, i.e., $\mathcal{N}(s_u^k | m_u^k, v_u^k)$, where m_u^k and v_u^k are mean and variance, respectively. Then the predictive distribution of $\mathbf{x}_u$ can be approximated as:

$$p(y_u|\mathbf{x}_u, \mathbf{D}_L) \approx \frac{\bar{\rho}}{C} + (1-\bar{\rho}) \int \mathcal{N}(s_u | m_u^{y_u}, v_u^{y_u}) \prod_{k \neq y_u} \Phi\left(\frac{s_u - m_u^k}{\sqrt{v_u^k}}\right) ds_u \tag{12}$$

where $\bar{\rho} = \frac{\alpha}{\alpha+\beta}$, $m_u^{y_u}$ and $v_u^{y_u}$ indicate the corresponding predictive mean and variance, respectively, and $\Phi(.)$ is the step function.

3.3. Learning Θ with Expectation Maximization

In order to online evaluate the quality of multiple annotators, we need to estimate the parameters $\Theta = \{\alpha, \beta, \{\varepsilon_j\}_{j=1}^M\}$. We further introduce a generalized Expectation-Maximization algorithm for estimating it. With the lower bound F of the log likelihood defined as

$$\begin{aligned} &\log p(\mathbf{T}_L, \mathbf{S}_L | \mathbf{X}_L, \Theta) \\ &\geq \sum_{z_i} \int_{\mathbf{S}_L} Q(\mathbf{S}_L) \log \frac{p(\mathbf{z}|\rho) p(\rho) p(\mathbf{T}_L, \mathbf{S}_L | \mathbf{X}_L, \Theta)}{Q(\mathbf{S}_L)} \\ &= \mathrm{C} + \sum_{z_i} \sum_{i=1}^{L} \int_{\mathbf{s}_i} q(\mathbf{s}_i) \log \{p(z_i|\rho) p(\rho) p(\mathbf{t_i}|\mathbf{s}_i, z_i, \epsilon)\} \, d\mathbf{s}_i, \end{aligned} \tag{13}$$

where C is a constant, the EM algorithm is formed with the following two iterative steps:

1. **E-Step:** Given the current parameter Θ_p, conduct the EP inference to obtain an approximate inference of $Q(\mathbf{S}_L) \sim p(\mathbf{S}_L | \mathbf{X}_L, \mathbf{T}_L, \Theta_p)$.
2. **M-Step:** Maximize the lower bound of $\log p(\mathbf{T}_L, \mathbf{S}_L | \mathbf{X}_L, \Theta)$ over Θ to obtain a new parameter Θ. $\Theta_p \leftarrow \Theta$, goto the **E-Step** and iterate until convergence.

For the **M-Step**, since closed-form solution of Θ is not tractable, we resort to the L-BFGS-B algorithm [44] to find a numerical estimation of them to maximize the lower bound F by gradient ascent, which is guaranteed to obtain a local optimal solution.

4. Reinforcement Learning for Active selection

To fully explore the trade-off between exploitation and exploration, we adopt the *entropy (Ent)* and *margin (Mar)* as two exploitation criteria, *Graph density (Gra)* as the exploration criterion. For the exact definitions of these criteria, we refer the readers to look into [5] due to the limit of space. Regarding the annotator selection, we provide two criteria as follows.

Label rate (LR). ϵ_j directly models the annotator j's quality, which can be interpreted as the probability that annotator j would label the data correctly. In other words, the higher ϵ_j is, the better quality the annotator has. In our active learning process, we can naturally select the top $K (K < M)$ annotators with the top K ϵ_j to label a selected data sample. The joint active selection of both annotators and data samples greatly facilitates to obtain higher quality labels.

Label correct likelihood (LCL). We identify the annotators who are more likely to label correctly given our current

state of knowledge, *i.e.*, given the predictive probability of the selected unlabeled sample and the levels of expertise of the different annotators. Therefore, we pick the top K annotators based on the measurement $\epsilon_j p(y_u^*|\mathbf{x}_u^*, \mathbf{D}_L) + (1 - \epsilon_j)(1 - p(y_u^*|\mathbf{x}_u^*, \mathbf{D}_L))$, where $p(y_u^*|\mathbf{x}_u^*, \mathbf{D}_L)$ is the predictive probability of the selected unlabeled sample.

Inspired by [5], we formulate the active learning as a Markov decision precess to incorporate the accumulated feedback to deal with multiple selection criteria. The 4-tuple (S, A, Q, R) is defined as follows: (a) $S = \{U + D + L\}$ with $U \in \{Mar, Ent\}$, $D \in \{Gra\}$ and $L \in \{LR, LCL\}$ is a mixture of two sampling criteria and one annotator selection criterion; (b) $A = \{\beta_1(t) = a_1, \ldots, \beta_n(t) = a_n\} \times S$ with $a_i \in [0, 1]$, represents n different fixed trade-offs among U and D to fully explore the trade-off between exploitation and exploration; (c) R is the reward for executing action a_i in state s_j; and (d) Q are the transition weights that action a_i is selected in state s_j .

We resort to a fast and adaptive reinforcement learning algorithm, *i.e.*, Q-Learning, to learn our transition table $Q \in R^{|S|\times|A|}$ online during the active learning process. After each transition $s^{(t-1)} \rightarrow a \rightarrow s^{(t)}$, we update Q given the current reward, *i.e.*,

$$Q(s^{(t-1)}, a) \leftarrow Q(s^{(t-1)}, a) + \lambda(r^{(t)} + \gamma(\max_{a_i} Q(s^{(t)}, a_i) - Q(s^{(t-1)}, a)), \tag{14}$$

where λ is the learning rate that controls the influence of the current reward $r^{(t)}$, and γ is the discount factor that weights the future reward. $r^{(t)}$ is defined as the difference of the overall entropy (alternatively, either the mean predicted output or the KL-divergence) of the class posteriors between two consecutive steps during the learning process.

During the active learning process, the optimal action $a = \max_{a_i} Q(s^{(t-1)}, a_i)$ leads to the adaptive trade-off. Then we can combine the current state $U + D + L$ and the obtained trade-off between U and D to determine which unlabeled samples to select and which K annotators to query. In this paper, we set $K = 3$.

5. Experiments

Our experiments are carried out on three image collections, *i.e.*, the E-Album [3] and the G-Album [6], and the ImageNet dataset [4]. We measure our proposed method and the competing methods with the recognition accuracy in both the active learning pool and the hold-out testing dataset with the progress of the learning process, and the results reported in this paper are average over multiple rounds.

5.1. Datasets

The E-Album consists of 108 photos taken with 15 different people in 145 detected faces. The G-Album has 312 photos taken with 13 different people in 441 detected faces. The detected faces in both albums are labeled by 7 nonprofessional annotators. For each annotator, we measure the annotator quality with the label accuracy which is defined as the percentage of his/her labels which are correct. In the E-Album, the label accuracies of the 7 annotators are 95.17%, 75.17%, 84.83%, 94.48%, 96.55%, 92.41% and 91.72%, respectively. And in the G-Album, the corresponding label accuracies are 98.41%, 79.37%, 94.33%, 75.06%, 97.96%, 87.07% and 94.10%, respectively.

The third dataset is composed of 3 classes of images from the ImageNet grand challenge [4], which includes 2 categories of dogs, *i.e.*, "Yorkshire terrier", "English setter" plus the "Meerkat, meerkat" category. These three classes are among the top 10 in the ImageNet grand challenge in terms of number of labeled images, with 3047, 2426 and 2341 images, respectively. We put these images back to Amazon Mechanical Turk and obtained 7 copies of labels for each image. The label accuracies of the 7 annotators are 92.03%, 92.62%, 91.89%, 92.41%, 92.68%, 92.08% and 92.50%, respectively.

For the readers' convenience, we summarize the abovementioned information in Table 1.

Table 1: The summarization of the 3 visual datasets used in the paper.

	♯classes	♯instances	annotator quality
E-Album	15	145	84.83% - 95.17%
G-Album	13	441	75.06% - 98.41%
ImageNet	3	7814	91.89% - 92.68%

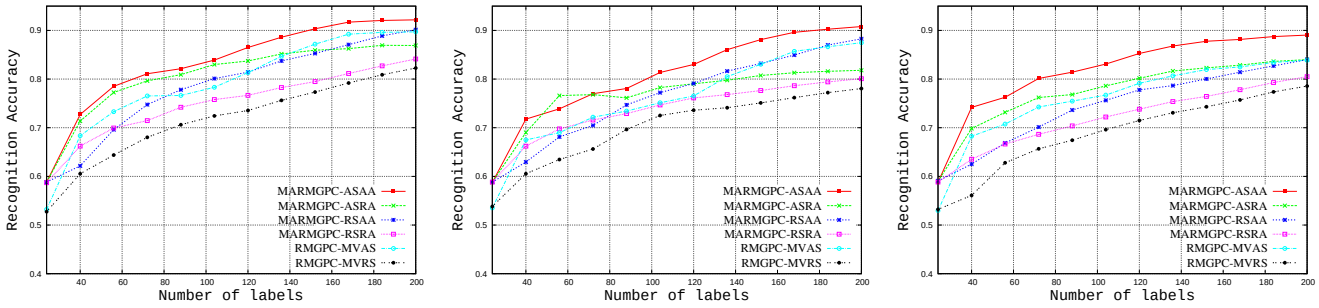
5.2. Visual features and kernels

On both the E-Album and the G-Album, we use the 100-dimensional feature with the Eigen-PEP representation [17] extracted from detected faces after resizing the images to 150 pixels by 150 pixels. To make it simple, we adopt the RBF kernel. As for the similarity or distance measurements, we adopt the similarity score from the Joint-Bayesian classifier [18].

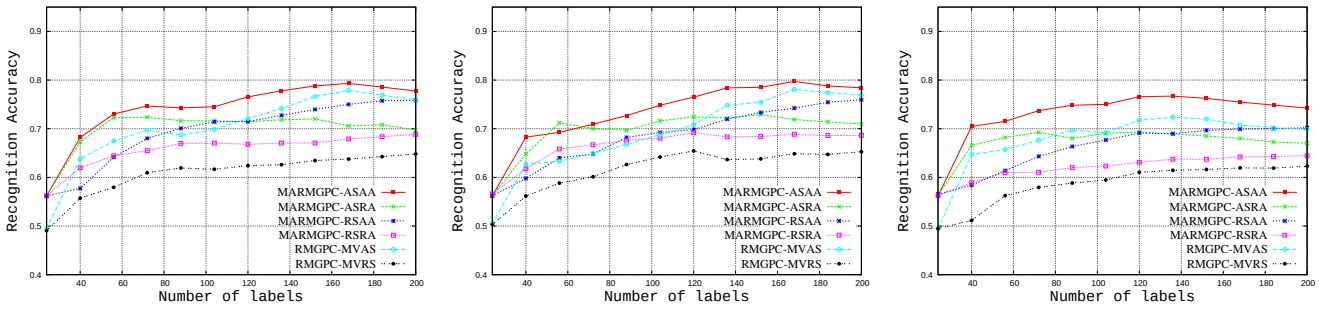
The features we use in the ImageNet dataset is the local coordinate coding (LCC) [20] on densely extracted HoG features with 4096 codewords. The LCC features are pooled in 10 spatial cells, resulting a 40960 dimensional feature. We use the dot-product kernel.

5.3. Experiments with synthetic label noise

The simulated experiments we conducted is on the G-Album. We randomly select 60% of all the detected faces to form the active learning pool, and the rest of faces are put together as the hold-out testing dataset. To demonstrate the effectiveness of our model in the situation when there are irresponsible annotators, we simulated the case that there are



(a) Active-2 irresponsible annotators (c) Active-3 irresponsible annotators (e) Active-4 irresponsible annotators



(b) Test-2 irresponsible annotators (d) Test-3 irresponsible annotators (f) Test-4 irresponsible annotators

Figure 2: Recognition performance with 2, 3 and 4 irresponsible annotators on the G-Album. "Active" and "Test" refer to active learning pool and hold-out testing dataset, respectively.

2, 3 and 4 irresponsible annotators, who would randomly assign a label to the sample, so there is 50% chance that the label from them will be erroneous. For responsible annotators, we use noisy labels obtained by crowd-scourcing. We run our proposed active learning algorithm with the active selection of both informative data samples and high-quality annotators. The top 3 annotators are selected to provide the labels for the actively selected samples.

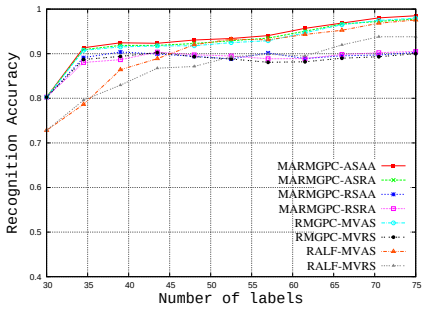
We name our algorithm as MARMGPC-ASAA, which stands for multi-annotator robust multi-class Gaussian process classifier (MARMGPC) with active selection of both samples and annotators. We compare MARMGPC-ASAA with a combination of other learning strategies, *i.e.*, active selection of samples but random selection of annotators, random selection of samples but active selection of annotators, and random selection of both samples and annotators. We call these algorithms MARMGPC-ASRA, MARMGPC-RSAA and MARMGPC-RSRA, respectively. All theses online learning algorithms are based on the same classification model and we select 3 annotators to provide the labels using the corresponding criterion for the annotator selection.

One algorithm we compare against is an active learning algorithm with the RMGPC classification model [8]. At each round, the method is performed on a single copy of labels which is obtained via majority voting among all 7 copies labels. In brief, we name such an active learning RMGPC with majority voting labels as RMGPC-MVAS. And the random learning counterpart as RMGPC-MVRS.

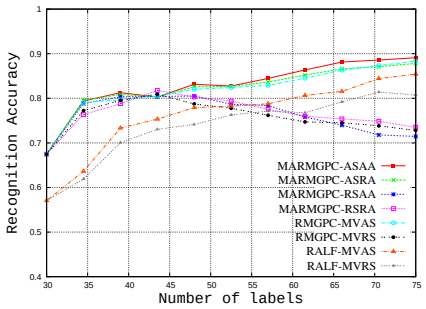
We report the results in Figure 2. As we can observe, (1) in all the cases, our proposed MARMGPC-ASAA outperforms all the competing algorithms in both the active learning pool and the hold-out testing dataset; (2) MARMGPC-ASRL performs better than MARMGPC-RSAA at the early stage and then it works worse than MARMGPC-RSAA, which suggests the active selection of both samples and high quality annotators benefits the improvement of recognition accuracy; and (3) with the increasing number of irresponsible annotators, the performance of all the competing algorithms are affected, but our MARMGPC-ASAA still performs the best. All these demonstrate that our proposed MARMGPC-ASAA is robust to deal with label noises in the learning progress.

5.4. Experiments with real crowd-sourced labels

We further run experiments with all real crowd-sourced labels on the E-Album, G-Album and ImageNet dataset. On both two albums, we use 60% of all the detected faces to form the active learning pool, and the rest of faces as the hold-out testing dataset. While on the ImageNet dataset, the active learning pool and the hold-out testing pool consist of the same number of images. Besides the above-mentioned competing methods, we also compare our method with a reinforced active learning formulation proposed by Ebert *et*

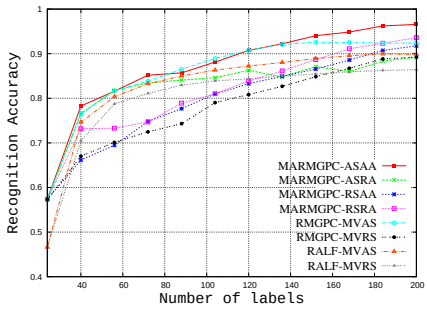


(a) Active learning pool

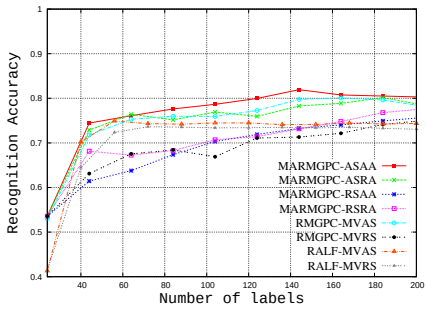


(b) Hold-out testing dataset

Figure 3: Recognition performance with real crowd-sourced labels on the E-Album.

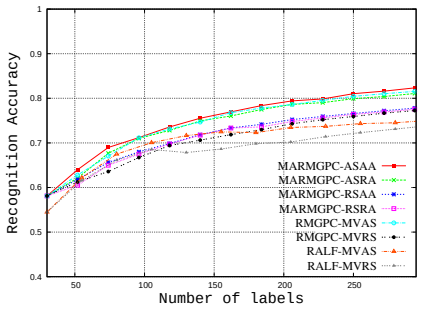


(a) Active learning pool

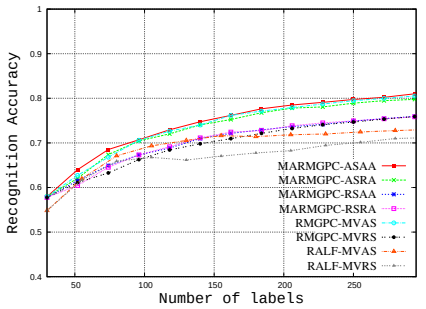


(b) Hold-out testing dataset

Figure 4: Recognition performance with real crowd-sourced labels on the G-Album.



(a) Active learning pool



(b) Hold-out testing dataset

Figure 5: Recognition performance with real crowd-sourced labels on the ImageNet dataset.

al. [5] with the majority voting labels, which we call RALF-MVAS, and its random learning counterpart is named as RALF-MVRS.

The results on the three datasets are summarized in Figure 3, 4 and 5, respectively. As we can observe, (1) on the G-Album, the results are consistent with the observations in section 5.3. Furthermore, our proposed MARMGPC-ASAA also performs better than RALF-MVAS and RALF-MVRS; (2) in both the E-Album and the ImageNet dataset, our proposed MARMGPC-ASAA outperforms all the competing methods both in the active learning pool and in the hold-out testing dataset; (3) all the Gaussian process algorithms obtain a higher start point than RALF-MVAS and RALF-MVRS, which demonstrates the advantage of Gaussian process in the classification problem with a small number of labeled examples. Again, all the observations on these three datasets further confirm the efficacy of our proposed method with the active selection of both the informative examples and the high-quality annotators.

It is worth mentioning that RALF-MVAS also uses the idea of reinforcement learning to obtain the adaptive trade-off between exploration and exploitation active samples selection strategies. The reasons why our proposed MARMGPC-ASAA achieves the better performance are: (1) our proposed MARMGPC has the ability to jointly treat multiple copies of labels from multiple annotators and make full use the diverse opinion among them, while the majority voting strategy ignore such informative and useful diverse opinions; and (2) the active selection of annotators with the active selection of samples in the joint reinforcement learning framework can prevent the low-quality annotators from participating in the labeling process for a long time so that we can obtain the relatively more reliable labels in the progress.

5.5. Discussion

To obtain a quantitative evaluation of the effectiveness of our proposed MARMGPC, we run experiments with all the labeled examples on both the E-Album and the G-Album. We compare the proposed model with the multi-class support vector machine (SVM) [2] and the standard multi-classes Guassian process classifier (SMGPC) with majority voting labels. Both SVM and SMGPC conduct the multi-classes classifications by reducing it into the binary cases. We also compare the proposed model with the robust multi-class Gaussian process classifier (RMGPC) with majority voting labels and the ground truth labels. In brief, we call them SVM-MV, SMGPC-MV, RMGPC-MV and RMGPC-GRD, respectively.

It is worth paying attention that the performance of RMGPC-GRD is the upper bound for both MARMGPC and RMGPC-MV. As apparent in Table 2, our proposed MAR-

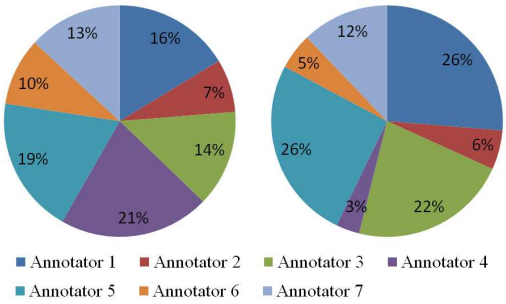


Figure 6: The fractions of the used labels from 7 annotators in the active learning progress of our proposed MARMGPC-ASAA on both the E-Album (left) and the G-Album (right).

MGPC obtains the comparable recognition performance to RMGPC-GRD and outperforms SVM-MV, SMGPC-MV and RMGPC-MV. On one hand, this demonstrates that MARMGPC is robust to multi-classes classification directly instead of decomposing into binary subproblems. On the other hand, MARMGPC jointly treats the labels from multiple annotators and can make full use of the diverse opinions to achieve higher recognition accuracy than the majority voting. All these further validate the efficacy of our proposed model.

Table 2: Comparison of recognition performance with different classification models

	E-Album	G-Album
SVM-MV	70.05%	71.93%
SMGPC-MV	71.64%	73.42%
RMGPC-MV	71.99%	74.30%
RMGPC-GRD	72.29%	75.69%
MARMGPC	72.08%	75.34%

To better understand the active selection of high quality annotators, we draw two pie charts in Figure 6 to display the fractions of the used labels from 7 annotators in the active learning progress on the two albums. As we can observe, on the E-Album, the top 3 annotators are Annotator 1, 4 and 5. On the G-Album, the corresponding top 3 annotators are Annotator 1, 3 and 5. Not surprisingly, the top 3 annotators in our experiments agree with the the label accuracies of the 7 annotators as described in Section 5.1. The most likely explanation to Figure 6 is that 3 high-quality annotators are selected to provide labels in each iteration so that the top 3 annotators are consistent as expected. This validates the ability of our proposed method to actively select high quality annotators to obtain more reliable labels.

We also visualize the selected samples in the active learning process on the G-Album. Figure 7 presents some examples that are selected actively in the early stages. As

Figure 7: Some examples selected by our proposed MARMGPC-ASAA in the early stages of the active learning on the G-Album.

we can see, the results are sensible as a lot of examples picked up in the early stage have cluttered background, heavy blurring, and several of them are baby faces. It is well known that it is not easy to recognize the identities of the baby from their facial images.

6. Conclusion

In this paper, we propose a novel multi-annotator Gaussian process model to deal with multi-class visual recognition in the collaborative active learning framework with multiple annotators. A generalized EM-EP algorithm is derived to estimate the parameters and approximate Bayesian inference. We also fully employ the idea of reinforcement learning and use Markov decision process to determine the optimal joint selection strategy of both the samples and annotators, and fully explore the trade-off between exploitation and exploration. The advantage of the proposed method over the state-of-the-art methods has been sufficiently validated through the experiments. Our future work includes extending the MARMGPC model to deal with the large-scale labeled data and further better exploring exploration-exploitation in the multi-annotator scenarios.

Acknowledgement

Research reported in this publication was partly supported by the National Institute Of Nursing Research of the National Institutes of Health under Award Number R01NR015371. This work is also partly supported by US National Science Foundation Grant IIS 1350763, China National Natural Science Foundation Grant 61228303, GH's start-up funds form Stevens Institute of Technology, a Google Research Faculty Award, a gift grant from Microsoft Research, and a gift grant from NEC Labs America.

References

[1] O. M. Aodha, N. D. Campbell, J. Kautz, and G. J. Brostow. Hierarchical Subquery Evaluation for Active Learning on a Graph. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2

[2] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001. 7

[3] J. Cui, F. Wen, R. Xiao, Y. Tian, and X. Tang. Easyalbum: an interactive photo annotation system based on face clustering and re-ranking. In *Special Interest Group on Computer-Human Interaction*, 2007. 5

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 1, 5

[5] S. Ebert, M. Fritz, and B. Schiele. Ralf: A reinforced active learning formulation for object class recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1, 2, 4, 5, 7

[6] A. Gallagher. Clothing cosegmentation for recognizing people. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 5

[7] S. Hanneke and L. Yang. Minimax analysis of active learning. *ACM Computing Research Repository*, 2014. 1

[8] D. Hernández-Lobato, J. M. Hernández-Lobato, and P. Dupont. Robust multi-class gaussian process classification. In *Advances in Neural Information Processing Systems*, 2011. 2, 6

[9] G. Hua, C. Long, M. Yang, and Y. Gao. Collaborative active learning of a kernel machine ensemble for recognition. In *IEEE International Conference on Computer Vision*, 2013. 1, 2

[10] S. Huang, R. Jin, and Z. Zhou. Active learning by querying informative and representative examples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(10):1936–1949, 2014. 1

[11] P. Jain and A. Kapoor. Active learning for large multi-class problems. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 1, 2

[12] A. J. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 1, 2

[13] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Active learning with gaussian processes for object categorization. *IEEE International Conference on Computer Vision*, 2007. 1, 2

[14] A. Kovashka, S. Vijayanarasimhan, and K. Grauman. Actively selecting annotations among objects and attributes. In *IEEE International Conference on Computer Vision*, 2011. 1

[15] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012. 1

[16] N. D. Lawrence, M. Seeger, and R. Herbrich. Fast sparse gaussian process methods: The informative vector machine. In *Advances in Neural Information Processing Systems*, 2003. 2

[17] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic part model for unsupervised face detector adaptation. In *IEEE International Conference on Computer Vision*, 2013. 5

[18] H. Li, G. Hua, X. Shen, Z. Lin, and J. Brandt. Eigen-pep for video face recognition. In *Asian Conference on Computer Vision*, 2014. 5

[19] X. Li, L. Wang, and E. Sung. Multi-label svm active learning for image classification. In *IEEE International Conference on Image Processing*, 2004. 2

[20] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 5

[21] C. Long, G. Hua, and A. Kapoor. Active visual recognition with expertise estimation in crowdsourcing. In *IEEE International Conference on Computer Vision*, 2013. 1, 2

[22] C. Long, G. Hua, and A. Kapoor. A joint gaussian process model for active visual recognition with expertise estimation in crowdsourcing. *International Journal of Computer Vision*, 2015. 1

[23] C. Loy, T. Hospedales, T. Xiang, and S. Gong. Stream-based joint exploration-exploitation active learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1

[24] T. Minka. *A family of algorithms for approximate Bayesian inference*. Ph.d. thesis, MIT, 2001. 3, 4

[25] G. Qi, X. Hua, Y. Rui, J. Tang, and H. Zhang. Two-dimensional active learning for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 2

[26] G. Qi, X. Hua, Y. Rui, J. Tang, and H. Zhang. Two-dimensional multilabel active learning with an efficient online adaptation model for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1880–1897, 2009. 2

[27] F. Rodrigues, F. Pereira, and B. Ribeiro. Gaussian process classification and active learning with multiple annotators. In *Proceedings of International Conference on Machine Learning*, 2014. 1, 2

[28] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image. *International Journal of Computer Vision*, 77(1-3):157–173, 2008. 1

[29] J. Sanchez and F. Perronnin. High-dimensional signature compression for large-scale image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 1

[30] E. Simpson, S. J. Roberts, I. Psorakis, and A. Smith. Dynamic bayesian combination of multiple imperfect classifiers. In *Decision Making and Imperfection*, 2013. 2

[31] D. Vasisht, A. Damianou, M. Varma, and A. Kapoor. Active learning for sparse bayesian multilabel classification. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2014. 2

[32] A. Vezhnevets, J. Buhmann, and V. Ferrari. Active learning for semantic segmentation with expected change. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1

[33] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 1

[34] P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems*, 2010. 2

[35] P. Welinder and P. Perona. Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2010. 2

[36] C. Williams and D. Barber. Bayesian classification with gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998. 3

[37] R. Yan, J. Yang, and A. G. Hauptmann. Automatically labeling video data using multi-class active learning. In *IEEE International Conference on Computer Vision*, 2003. 2

[38] Y. Yan, R. Rosales, G. Fung, and J. Dy. Active learning from multiple knowledge sources. In *the Fifteenth International Conference on Artificial Intelligence and Statistics*, 2012. 2

[39] Y. Yan, R. Rosales, G. Fung, and J. G. Dy. Active learning from crowds. *Proceedings of International Conference on Machine Learning*, 2011. 2

[40] L. Yang and J. G. Carbonell. Buy-in-bulk active learning. In *Advances in Neural Information Processing Systems*, 2013. 1

[41] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, pages 1–15, 2014. 2

[42] L. Zhang, M. Mahdavi, and R. Jin. Improving the minimax rate of active learning. *ACM Computing Research Repository*, 2013. 1

[43] L. Zhao, G. Sukthankar, and R. Sukthankar. Incremental relabeling for active learning with noisy crowdsourced annotations. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust, and Social Computing*, 2011. 2

[44] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 1997. 4