Diverse Expected Gradient Active Learning for Relative Attributes

Xinge You, *Senior Member, IEEE,* Ruxin Wang, Dacheng Tao, *Senior Member, IEEE*

***Abstract*—The use of relative attributes for semantic understanding of images and videos is a promising way to improve communication between humans and machines. However, it is extremely labor- and time-consuming to define multiple attributes for each instance in large amount of data. One option is to incorporate active learning, so that the informative samples can be actively discovered and then labeled. However, most existing active-learning methods select samples one at a time (serial mode), and may therefore lose efficiency when learning multiple attributes. In this paper, we propose a batch-mode active-learning method, called *Diverse Expected Gradient Active Learning (DEGAL)*. This method integrates an informativeness analysis and a diversity analysis to form a diverse batch of queries. Specifically, the informativeness analysis employs the expected pairwise gradient length as a measure of informativeness, while the diversity analysis forces a constraint on the proposed *diverse gradient angle*. Since simultaneous optimization of these two parts is intractable, we utilize a two-step procedure to obtain the diverse batch of queries. A heuristic method is also introduced to suppress imbalanced multi-class distributions. Empirical evaluations of three different databases demonstrate the effectiveness and efficiency of the proposed approach.**

***Index Terms*—Batch Mode, Active Learning, Diverse Expected Gradient, Relative Attributes.**

I. Introduction

Semantic understanding of scenes aims to narrow the gap between what humans and computers understand by providing the meanings of elements in text, speech, images, or videos (e.g. ”the sky in the image is blue” and ”the boy’s hand in the video is waving”) in a format that is understandable to humans. From a practical perspective, semantic understanding is highly relevant in systems that organize personal and professional information, and for this reason the approach has received much attention in the computer vision community. However, several important research challenges still exist for various vision tasks, including image/video classification, annotation, and retrieval. Techniques to organize, annotate, and retrieve digital media data are lagging behind the exponential growth in the amount of that data, and some researchers believe that perfecting semantic understanding is an urgent need in order to gain access to the content of images and videos [1].

Previous research has mainly focused on building a semantic vocabulary, i.e., embedding the semantic information into a visual vocabulary using either unsupervised or supervised methods. Of the unsupervised methods, the ’topic models’ provide an intuitive approach for researchers; this approach typically includes *Probabilistic Latent Semantic Analysis (pLSA)* [2], [3], [4], *Latent Dirichlet Allocation (LDA)* [5], [6], and *Diffusion Maps (DM)* [7]. These methods attempt to discover the mixture distribution of hidden topics, each of which can then be related to a meaningful concept, and a recent study suggested the use of a randomized visual vocabulary for action search [8]. In the supervised case, there have been attempts to utilize local patch information or image/video annotation to explore a visual vocabulary. For example, Vogel et al. [9] constructed a vocabulary with explicit semantic meanings by labeling certain semantic concepts (e.g. *sky, rocks, sand*) to each local image patch. However, the applicability of this approach is limited due to the large labeling cost when allocating each patch. Ji et al. [10] proposed the use of the *Hidden Markov Random Field (HMRF)* model to integrate both local visual features and semantic labels to guide vocabulary construction. In this study, the similarities between different local visual features were extracted from 60,000 labeled Flickr images, while the semantic label correlations were provided by WordNet. Similar to previous methods, this approach still required a large number of manually provided labels to produce a general vocabulary.

Rather than constructing a vocabulary, the recent literature pays increasing attention to visual semantic attributes. Farhadi et al. [11], Lampert et al. [12], and Kumar et al. [13] have all proposed the use of a set of visual semantic attributes to describe various objects and human faces. Due to their robustness to visual variations, attributes have been applied to different vision tasks, including classification [14], [15], recognition [16], and retrieval [17]. These methods treat attributes as binary values to indicate their existence. On the other hand, *relative attributes*, as proposed by Parikh et al. [18], are designed to provide a richer mode of communication and detailed access with human supervision. Due to the intrinsic properties of binary and relative attributes, it is intuitive that they can either be user defined (from a professional human perspective) or discovered from the data itself, in order to complement human deficiencies [16], [19]. However, this also means that each object or scene has many attributes that need to be labeled manually. In addition, training a robust classifier or recognizer for a real-world application requires thousands of samples, and obtaining these samples and attributes is an extremely time- and labor-consuming task.

Semantic understanding would therefore benefit from high-volume semantic learning with restricted time and labor costs, and progress in this area has been seen over the past two

Xinge You is with the Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhuan 430074, China.

Ruxin Wang and Dacheng Tao are with the Centre for Quantum Computation and Intelligent Systems and the Faculty of Engineering and Information Technology, University of Technology, Sydney, NSW 2007, Australia.

years. For instance, Parkash et al.'s [20] and Biswas et al.'s [21] adopted an active learning framework to select the sample (known as the '*query*' in the active learning field) that was most uncertain to the attribute learner. Instead of simply demanding the label for the image, the learner conveyed its current belief about the image to the oracle and demanded a response and explanation in return, the image classifiers simultaneously benefited from this feedback process. However, in each iteration of active learning, the learner only selected the most uncertain query to be labeled, i.e., only one query was chosen, and therefore many iterations were required to reach stability. Rather than selecting the samples important to the classification task, Xu et al. [22] studied the issue of deciding which semantics (i.e., attributes) are pivotal. By defining a data-driven *Category-Attribute Matrix*, they automatically designed discriminative attributes in a principled way and in doing so avoided the use of large-scale, but redundant, attribute sets. Similarly, Choi et al. [23] proposed a novel joint optimization framework in which the attribute learner, category recognizer, and sample selector were simultaneously optimized. To ensure discrimination, they learned all attributes from the data in order to identify which unlabeled sample was critical to the category boundaries and, in this way, both the attribute learner and category recognizer were trained on a relatively small set. As well as the attribute-related work, other active learning methods have also been proposed to improve image/video semantic tasks [24], [25], [26].

Incorporating the active learning framework to solve the above problem is clearly effective. Active learning evaluates the informativeness of unlabeled instances so that more informative instances are more likely to be queried [27]. However, as in Parkash et al. [20], most active learning approaches serially select queries, i.e., they are selected one at a time [27]. The time required to induce a semantic model can be slow or expensive, especially when multiple annotators work on different labeling workstations in a network at the same time, which is the case in attributes learning. Under these conditions, *batch-mode* active learning, which allows the learner to select queries in groups, is more suitable for serial labelling environments. By picking up several queries during one iteration, batch-mode active learning results in less iterations and faster convergence.

Here we aim to improve the training efficiency of a type of semantics learning, namely the recently proposed relative attributes method. We present a novel batch-mode active learning approach called *Diverse Expected Gradient Active Learning* (DEGAL), which addresses the following two objectives: to collect batches of the most informative queries, and 2) to enforce the selected queries to be diverse with respect to each other in the training procedure. Our main contributions include:

1) Inspired by [28] and [29], we use the expected pairwise gradient length as the informativeness measure. The most informative query should provide a large number of confusing pairwise relationships and cause a large change on the model parameters. To show that this is reasonable, we demonstrate equivalence between this strategy and Tongs widely accepted result [30].

2) We extend serial-mode active learning based on gradient length to the batch-mode case. To measure the diversity of a query set, the *Diverse Gradient Angle* is defined, based on the expected gradient direction. By imposing a constraint on the angular differences between queries in the set, we prove that the satisfied queries can result in different model parameters if they are separately added to the training set.

3) The proposed active learning method suffers from a multi-class imbalance issue, which might result in poor performance. We therefore design a heuristic method by introducing a balance constraint to suppress the imbalanced multi-class distributions.

We perform empirical evaluations on three datasets equipped with relative attributes and demonstrate that our method performs favorably compared to other batch-mode active learning and random-sampling baseline methods. Our approach is similar to [31]; however, our study differs in that we handle the diversity analysis in the gradient space, rather than in a projected feature space characterized by a kernel.

The remainder of this paper is organized as follows. Section II provides the background to the relative attributes model, as well as a detailed analysis from an active learning perspective. In Section III, we introduce our approach, followed by an outline of the experimental results demonstrating the efficiency of our strategy in Section IV. Finally, we summarize our method and briefly discuss future research directions.

II. Preliminary

The content of the relative attributes model is briefly reviewed in this section, before providing a detailed analysis of the model that inspired our proposed algorithm.

A. Relative Attributes

Attribute-based vision tasks, such as image classification and object recognition, are an embedded mapping that can be decomposed as follows [32]:

$$\begin{array}{rcl} H & = & L(S(\cdot)) \\ S & : & \mathbb{R}^d \rightarrow \mathbb{A}^M \\ L & : & \mathbb{A}^M \rightarrow \mathbb{L} \end{array} \tag{1}$$

where S is composed of M individual attribute learners $\{b_m(\mathbf{x})\}_{m=1}^{M}$, each learner $b_m(\mathbf{x})$ maps a raw feature $\mathbf{x} \in \mathbb{R}^d$ to the corresponding m-th attribute a_m of $\mathbb{A}^M$, L maps a semantic attribute point $\mathbf{a} \in \mathbb{A}^M$ to a class label $l \in \mathbb{L}$. $\mathbb{R}^d$, $\mathbb{A}^M$, and $\mathbb{L}$ denote the d-dimensional real-value space, the M-dimensional attribute space, and the label space, respectively.

The relative attributes model, which differs from the binary attributes model, may provide a promising method to deeply exploit human cognizance and build a wider information bridge between humans and machines. This model encodes each image with the strength of different attributes with respect to other images, and can be modeled as follows [18].

Suppose a set of training images I are represented by raw feature vectors $\{\mathbf{x} \in \mathbb{R}^d\}^1$, and a set of attributes $\{a_1, ..., a_M\}$

[1] In the following, we denote the raw features as $\mathbf{x}_i^1$, $\mathbf{x}_i^2$, $\mathbf{x}_j^1$, $\mathbf{x}_j^2$, where the subscripts i and j are the indexes of pairs in O_m and S_m respectively, and the superscripts 1 and 2 reveal the relative order.

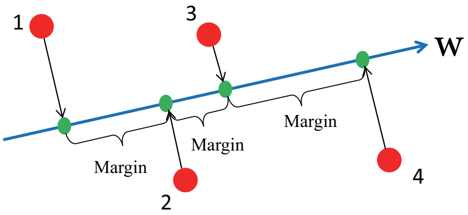


Fig. 1. Ranking Direction. For a certain attribute, samples 1, 2, 3, and 4 are sorted in increasing order according to the strength of that attribute. Under SVM conditions, $\mathbf{w}$ is the optimal ranking direction on which the cumulative margin between all adjacent samples are maximized.

are accordingly defined. Then, for each attribute a_m, two kinds of image pair modes, O_m and S_m, are constructed by comparing the corresponding attribute in two images. $O_m = \{(\mathbf{x}_i^1, \mathbf{x}_i^2) | \mathbf{x}_i^1 \succ \mathbf{x}_i^2\}$ is the ordered image pair mode, indicating that image $\mathbf{x}_i^1$ has a stronger strength on attribute a_m than image $\mathbf{x}_i^2$, while $S_m = \{(\mathbf{x}_j^1, \mathbf{x}_j^2) | \mathbf{x}_j^1 \sim \mathbf{x}_j^2\}$ is the un-ordered image pair mode, denoting that image $\mathbf{x}_j^1$ has a similar strength of attribute a_m as image $\mathbf{x}_j^2$.

The goal is to learn M attribute ranking functions, each of which is

$$r_m(\mathbf{x}) = \mathbf{w}_m^T \mathbf{x} \tag{2}$$

for $m = 1, ..., M$, such that most image pairs satisfy the corresponding image pair modes, i.e.

$$\begin{aligned} \forall (\mathbf{x}_i^1, \mathbf{x}_i^2) \in O_m &\quad : \quad \mathbf{w}_m^T \mathbf{x}_i^1 > \mathbf{w}_m^T \mathbf{x}_i^2 \\ \forall (\mathbf{x}_j^1, \mathbf{x}_j^2) \in S_m &\quad : \quad \mathbf{w}_m^T \mathbf{x}_j^1 = \mathbf{w}_m^T \mathbf{x}_j^2 \end{aligned}$$

where $\mathbf{w}_m \in \mathbb{R}^d$ is the direction parameter and $(\cdot)^T$ denote the transposition. In other words, we aim to find the optimal projected direction on which all training samples are ranked in an accurate order in the feature space, as shown in Fig. 1.

To deal with the above NP-hard problem, its solution can be approximated by introducing: (1) the non-negative slack variables δ_i and γ_j, and (2) a regularization term to maximize the margin between the closest pair's projection on $\mathbf{w}_m$. This leads to the following optimization problem, named *Ranking SVM with Similarity* (*RankSVM-with-Sim*):

$$\begin{aligned} \underset{\mathbf{w}_m}{\text{minimize}} : \quad & \frac{1}{2}\|\mathbf{w}_m\|_2^2 + C(\sum_i \delta_i^2 + \sum_j \gamma_j^2), \qquad (3) \\ \text{subject to} : \quad & \forall (\mathbf{x}_i^1, \mathbf{x}_i^2) \in O_m : \mathbf{w}_m^T \mathbf{x}_i^1 \geq \mathbf{w}_m^T \mathbf{x}_i^2 + 1 - \delta_i; \\ & \forall (\mathbf{x}_j^1, \mathbf{x}_j^2) \in S_m : |\mathbf{w}_m^T \mathbf{x}_j^1 - \mathbf{w}_m^T \mathbf{x}_j^2| \leq \gamma_j; \\ & \forall \ i, j : \delta_i \geq 0; \ \ \gamma_j \geq 0, \end{aligned}$$

where C is a free parameter that allows a trade-off between margin and training error. Rearranging the above constraints, we can rewrite them as:

$$\begin{aligned} &\forall (\mathbf{x}_i^1, \mathbf{x}_i^2) \in O_m : \mathbf{w}_m^T (\mathbf{x}_i^1 - \mathbf{x}_i^2) \geq 1 - \delta_i; \\ &\forall (\mathbf{x}_j^1, \mathbf{x}_j^2) \in S_m : |\mathbf{w}_m^T \mathbf{x}_j^1 - \mathbf{w}_m^T \mathbf{x}_j^2| \leq \gamma_j. \end{aligned}$$

By handling the above optimization problem, we can acquire M attribute ranking functions $\{r_m(\cdot)\}_{m=1}^M$. To further induce proper active learning for this model, certain properties should be considered, as detailed in the next sub-section.

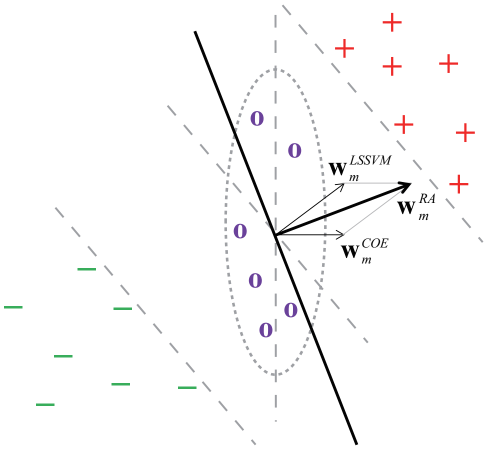


Fig. 2. Decomposition of RankSVM-with-Sim. Symbol ”o” denotes the transformed samples in $\tilde{S}_m$, while ”+” and ”-” indicate those in $\tilde{O}_m$. Optimization on problem (5) results in $\mathbf{w}_m^{LSSVM}$, and problem (6) leads to $\mathbf{w}_m^{COE}$. Thus, the overall solution $\mathbf{w}_m^{RA}$ can be expressed as a combination of $\mathbf{w}_m^{LSSVM}$ and $\mathbf{w}_m^{COE}$. Obviously, if the sample closest to the hyper-plane (solid line) belongs to $\tilde{S}_m$, it has no optimization benefit for the current model parameters. Samples that could result in a significant change in the direction of $\mathbf{w}_m^{RA}$ should be treated as the most informative and not necessarily close to any plane.

B. *Active Learning Analysis*

Even though RankSVM-with-Sim is designed to find the optimal direction in which each sample is assigned to a correct order, it can be equivalent to a classification SVM on pairwise difference vectors $(\mathbf{x}_i^1 - \mathbf{x}_i^2)$ and $(\mathbf{x}_j^1 - \mathbf{x}_j^2)$. For attribute a_m, let us denote $\mathbf{d}_i^m = (\mathbf{x}_i^1 - \mathbf{x}_i^2) \in \tilde{O}_m$ to be assigned with a label y_i^m, where $y_i^m = 1$ if $r_m(\mathbf{x}_i^1) > r_m(\mathbf{x}_i^2)$ and $y_i^m = -1$ if $r_m(\mathbf{x}_i^1) < r_m(\mathbf{x}_i^2)$, while $\mathbf{d}_j^m = (\mathbf{x}_j^1 - \mathbf{x}_j^2) \in \tilde{S}_m$, which is the difference between the un-ordered image pairs in S_m. For completeness, we define the label $y_j^m = 0$ for S_m.

Based on the above definition and transferring RankSVM-with-Sim into an unconstrained scenario, we can get

$$\underset{\mathbf{w}_m}{\text{minimize}} : \quad \frac{1}{2}\|\mathbf{w}_m\|_2^2 + C\left(\sum_{i=1}^{T_O} \max(0, 1 - y_i \mathbf{w}_m^T \mathbf{d}_i^m)^2 + \sum_{j=1}^{T_S} (\mathbf{w}_m^T \mathbf{d}_j^m)^2\right) \tag{4}$$

where T_O denotes the number of ordered pairs in $\tilde{O}_m$, and T_S is the number of un-ordered pairs in $\tilde{S}_m$. We can now make an approximate decomposition of problem (4) into two parts, namely using *least-squares SVM* (LS-SVM) and constrained optimization on the ellipse (COE):

$$\text{LS-SVM} \quad \min_{\mathbf{w}_m} \quad \frac{1}{2}\|\mathbf{w}_m\|_2^2 + C\sum_{i=1}^{T_O} \max(0, 1 - y_i \mathbf{w}_m^T \mathbf{d}_i^m)^2, \tag{5}$$

$$\text{COE} \quad \min_{\mathbf{w}_m} \quad \sum_{j=1}^{T_S} (\mathbf{w}_m^T \mathbf{d}_j^m)^2 \quad s.t. \ \|\mathbf{w}_m\| \geq \rho, \tag{6}$$

where ρ is a positive scalar that constrains the minimum norm of $\mathbf{w}_m$ in COE. The above decomposition is possible because the first two parts of (4) form the LS-SVM, while the third part of (4) becomes the COE. The constraint in (6) emerges from LS-SVM in (5), and it is used to avoid the solution of (6) to be all zeros. In other words, if we regard ρ as the norm of the solution of (5), the optimal solution $\mathbf{w}_m^*$ can be obtained by iteratively optimizing (5) and (6) until converged. For convenience, we provide an illustration of this in Fig. 2, where we assume the solutions of the two problems are $\mathbf{w}_m^{LSSVM}$ and $\mathbf{w}_m^{COE}$, and the solution to problem (4) is therefore a compromise.

Given this model, we need to find the most informative queries (or the queries that the learner finds most confusing) and add them into the training set. In the active learning field, several strategies have been proposed for SVM and Rank SVM. In Tong's result [30] for binary SVM, the most informative query was the one closest to the classification hyperplane, but this is not the case in a relative attributes scenario, since the queries closest to the classification hyperplane belong to $\tilde{S}_m$ and cannot significantly affect the optimization of the plane. In Donmez et al.'s [29] and Settles et al.'s [28] methods, the query with the largest gradient length once added into training set under current model parameters was selected as the most informative. Meanwhile, Parkash et al. [20] and Biswas et al. [21] proposed selecting samples with the most entropy or the largest entropy variation. However, the critical issue with these methods is, as stated in Section I, that they are serial-mode. In order to design a batch-mode strategy, our contribution accounts for the informativeness, while efficiently computing the diversity of different queries relative to the current model.

III. Proposed Approach

We next describe our *Diverse Expected Gradient Active Learning (DEGAL)* method, which improves on the serial-mode approach presented in the previous section. A flowchart of the proposed method is shown in Fig. 3. The theoretical foundation of our batch-mode active learning method is based on Settle et al. [28]. Given that the expected gradient length is a measure of sample uncertainty, our goal is to assist the attribute learner to actively find a batch of informative queries that possess relatively large expected gradient length (not only the largest one), and simultaneously maintain diversity from each other.

A. Informativeness Analysis

In this section we incorporate the expected gradient length, which measures the informativeness of each unlabeled instance, into the model. This is feasible because the relative attributes model can be optimized in the primal using the gradient descent method [33]. Once a query is added into the training set, it will create the greatest change in the pairwise gradient length under the objective function. Here is a mathematical explanation:

First, we need to impose a general constraint on the pairwise differences, which is

$$\begin{aligned}&\forall(\mathbf{x}_i^1,\mathbf{x}_i^2)\in O_m : \|\mathbf{x}_i^1-\mathbf{x}_i^2\|=1;\\ &\forall(\mathbf{x}_j^1,\mathbf{x}_j^2)\in S_m : \|\mathbf{x}_j^1-\mathbf{x}_j^2\|=1.\end{aligned} \tag{7}$$

Then following equation 4, we define the objective function as

$$\begin{aligned}\underset{\mathbf{w}_m}{\text{minimize}} :\ \hbar_{\mathbf{w}_m} &= \lambda\|\mathbf{w}_m\|_2^2\\ &+\sum_{(\mathbf{x}_i^1,\mathbf{x}_i^2)\in O_m}[1-y_i\mathbf{w}_m^T(\mathbf{x}_i^1-\mathbf{x}_i^2)]_+^2\\ &+\sum_{(\mathbf{x}_j^1,\mathbf{x}_j^2)\in S_m}(\mathbf{w}_m^T(\mathbf{x}_j^1-\mathbf{x}_j^2))^2,\end{aligned} \tag{8}$$

where $\lambda = 1/2C$ and $[\cdot]_+$ is the hinge loss function. If an instance $\mathbf{x}$ is added into the training set with a certain rank label $\mathbf{y}$, the above function becomes

$$\begin{aligned}\underset{\mathbf{w}_m}{\text{minimize}} :\ \hbar^{new}_{\mathbf{w}_m} &= \lambda\|\mathbf{w}_m\|_2^2\\ &+\sum_{(\mathbf{x}_i^1,\mathbf{x}_i^2)\in O_m}[1-y_i\mathbf{w}_m^T(\mathbf{x}_i^1-\mathbf{x}_i^2)]_+^2\\ &+\sum_{(\mathbf{x}_j^1,\mathbf{x}_j^2)\in S_m}(\mathbf{w}_m^T(\mathbf{x}_j^1-\mathbf{x}_j^2))^2+Q(\mathbf{w}_m,\mathbf{x}),\end{aligned} \tag{9}$$

where

$$\begin{aligned}Q(\mathbf{w}_m,\mathbf{x}) &= \sum_{(\mathbf{x}_u,\mathbf{x})\in O_m}[1-y_u\mathbf{w}_m^T(\mathbf{x}_u-\mathbf{x})]_+^2\\ &+\sum_{(\mathbf{x}_u,\mathbf{x})\in S_m}(\mathbf{w}_m^T(\mathbf{x}_u-\mathbf{x}))^2\end{aligned}$$

encodes the total loss caused by the instance $\mathbf{x}$ and other relevant instances. $(\mathbf{x}_u,\mathbf{x})\in O_m$ ($(\mathbf{x}_u,\mathbf{x})\in S_m$) means that the instance $\mathbf{x}_u$ has a different order (same order) to $\mathbf{x}$.

Let $\nabla\hbar_{\mathbf{w}_m}$ and $\nabla\hbar^{new}_{\mathbf{w}_m}$ be the gradient of the original objective function (8) and the new objective function (9) with respect to model parameter $\mathbf{w}_m$, respectively. In most cases, the uniqueness of the SVM solution can be guaranteed [34]. So let us assume that the unique optimal solution of function (8) is $\mathbf{w}^*$. Equation $\nabla\hbar_{\mathbf{w}_m^*}=0$ holds due to the optimality. Then, the change of gradient induced by instance $\mathbf{x}$ is

$$\begin{aligned}\phi(\mathbf{x},\mathbf{y}) &= \nabla\hbar^{new}_{\mathbf{w}_m^*}-\nabla\hbar_{\mathbf{w}_m^*}\\ &= \nabla Q(\mathbf{w}_m^*,\mathbf{x})\\ &= 2\sum_{\mathbf{x}_u} g(\mathbf{x}_u,\mathbf{x},y_u),\end{aligned} \tag{10}$$

where

$$\forall(\mathbf{x}_u,\mathbf{x})\in O_m :$$

$$g(\mathbf{x}_u,\mathbf{x},y_u)=\begin{cases}-y_u(\mathbf{x}_u-\mathbf{x})[1-y_u\mathbf{w}_m^{*T}(\mathbf{x}_u-\mathbf{x})], & \\ \qquad\qquad\qquad \text{if } y_u\mathbf{w}_m^{*T}(\mathbf{x}_u-\mathbf{x})<1\\ 0, \qquad\qquad\quad \text{if } y_u\mathbf{w}_m^{*T}(\mathbf{x}_u-\mathbf{x})\geq 1\end{cases}$$

$$\forall(\mathbf{x}_u,\mathbf{x})\in S_m : g(\mathbf{x}_u,\mathbf{x},y_u)=(\mathbf{x}_u-\mathbf{x})[\mathbf{w}_m^{*T}(\mathbf{x}_u-\mathbf{x})].$$

and $\mathbf{y}$ is a label vector, each dimension of which indicates the relative order between $\mathbf{x}$ and the corresponding training sample. Now, according to [28], we can define the informativeness of the instance $\mathbf{x}$ as the accumulated pairwise gradient lengths, i.e.,

$$\psi(\mathbf{x},\mathbf{y})=2\sum_{\mathbf{x}_u}\|g(\mathbf{x}_u,\mathbf{x},y_u)\|. \tag{11}$$

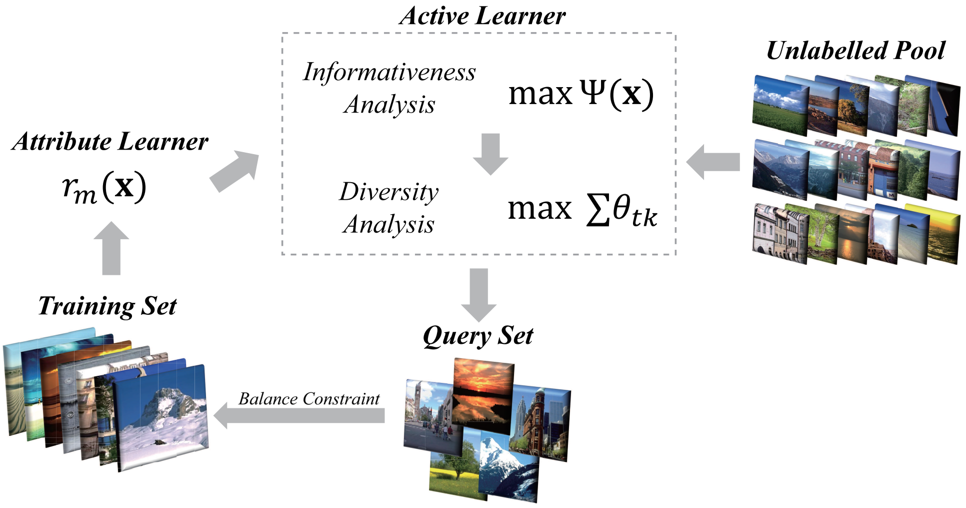


Fig. 3. A flowchart of the proposed method.

From definitions (10) and (11) above, we know that $\phi(\mathbf{x},\mathbf{y})$ is the direction of gradient descent after $\mathbf{x}$ is added into the training set. This will be used in the diversity analysis in Section III-B. Furthermore, the informativeness $\psi(\mathbf{x},\mathbf{y})$ reveals the uncertainty of the current model with respect to $\mathbf{x}$. The larger $\psi(\mathbf{x},\mathbf{y})$, the more confusing $\mathbf{x}$ will be to the model.

However, one issue to consider when evaluating the informativeness of unlabeled instances is that an active learner cannot know the true label $\mathbf{y}$ in advance. The expectations of both $\phi(\mathbf{x},\mathbf{y})$ and $\psi(\mathbf{x},\mathbf{y})$ therefore need to be calculated over the attribute learner's current belief $P(\mathbf{y}|\mathbf{x})$, that is

$$\Phi(\mathbf{x}) = \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x})\phi(\mathbf{x},\mathbf{y}), \tag{12}$$

$$\Psi(\mathbf{x}) = \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x})\psi(\mathbf{x},\mathbf{y}). \tag{13}$$

The belief $P(\mathbf{y}|\mathbf{x})$ can be estimated using Platt's method [35], in which the posterior probability $P(\mathbf{y}|\mathbf{x})$ is regressed using a sigmoid function. Then, the instance with the largest expected informativeness is the optimal one to be added into the training set. In fact, $\max_{\mathbf{x}} \Psi(\mathbf{x})$ is equivalent to Tong's result [30], which is proven in Appendix A. In our setting, we select a batch of K queries, $X^* = \{\mathbf{x}\}^K$, each of which has a relatively large value of $\Psi(\mathbf{x})$, i.e.,

$$X^* = \underset{X\subseteq U \cap |X|=K}{\operatorname{argmax}} \sum_{\mathbf{x}\in X} \Psi(\mathbf{x}) \tag{14}$$

where U is the unlabeled set.

B. Diversity Analysis

The query set selected in the above section captures a large amount of information to be discovered by the current model, which we refer to as query candidates. However, as described in Section II-B, each query candidate may have similar information to other candidates, and the effect of adding one similar query into the training set will therefore be the same as adding other similar queries. In this case, the expanded training set will result in slow convergence to the overall optimal solution; we therefore need to diversify the query set. Different diversity measurements possess specific properties and can result in different sets of query candidates. Despite this, one particular case needs to be emphasized that cannot be guaranteed: for any two queries in the query set, one will still be selected as an informative sample if the other was added into the training set and the model was updated. This is because the data used in real applications always changes, and the updated model cannot be predicted unless the label of the added query is known. However, according to the definition of informativeness in (13), the queries selected in each iteration are always useful for updating the model.

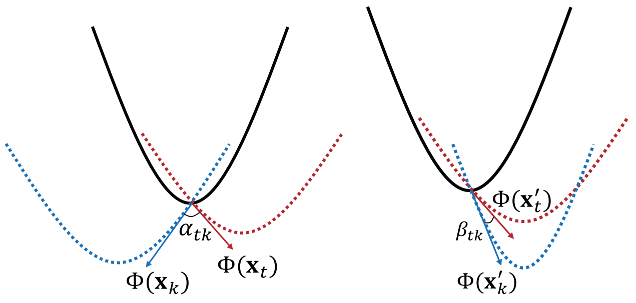
The theory behind our approach is that the selected queries can greatly change the model parameters if they are used for training. When considering diversity, it might be expected that different queries could lead the current model parameters in different directions, as shown in Fig. 4. To realize this, we propose the *diverse gradient angle (DGA)* to measure the diversity of candidates in a query set, a detailed description of which is given below.

Definition 1: Suppose that instances $\mathbf{x}_t$ and $\mathbf{x}_k$ are two candidates selected in the query set, and their associated expected gradient changes are $\Phi(\mathbf{x}_t)$ and $\Phi(\mathbf{x}_k)$, respectively. Then, the *Diverse Gradient Angle* between $\mathbf{x}_t$ and $\mathbf{x}_k$ is

$$\theta_{tk} = \arccos \frac{\langle \Phi(\mathbf{x}_t), \Phi(\mathbf{x}_k)\rangle}{\|\Phi(\mathbf{x}_t)\|\|\Phi(\mathbf{x}_k)\|} \tag{15}$$

where $\langle\cdot\rangle$ is the inner product and $\|\cdot\|$ is L^2-norm.

Given the definition of DGA, we know that if θ_{tk} is larger than a specific value, say α, $\mathbf{x}_t$ and $\mathbf{x}_k$ lead to different directions of the gradient descent. In other words, when $\mathbf{x}_t$ is added into the training set, the solution of (9) is quite different to that when $\mathbf{x}_k$ is added, which is expected. On the other hand, if θ_{tk} is smaller than α or leans more towards 0, the two query candidates would result in a similar direction of gradient descent. Obviously, in this case, $\mathbf{x}_k$ may be useless after $\mathbf{x}_t$ is labeled as a training sample.



(a) (b)

Fig. 4. Take a quadratic form as an example. In both (a) and (b), the black solid line indicates the objective function under the original training set, while the blue and red dotted lines denote the objective function under the enlarged training set with $\mathbf{x}_k$ and $\mathbf{x}_t$, respectively. Obviously, in (a), the DGA α_{tk} is large, so $\mathbf{x}_k$ and $\mathbf{x}_t$ would lead to quite different updates of the optimal solution. In (b), DGA β_{tk} is small, which means labeling $\mathbf{x}_k$ or $\mathbf{x}_t$ would possibly result in the same solution.

To appropriately choose the value of α, the following proposition and its proof are proposed.

Proposition 1: Suppose $\mathbf{x}_t$ and $\mathbf{x}_k$ are two instances selected by (14) as informative queries, and their associated expected gradient changes are $\Phi(\mathbf{x}_t)$ and $\Phi(\mathbf{x}_k)$. If $\forall\ \alpha \geq \frac{\pi}{3}$, such that the DGA θ_{tk} between $\mathbf{x}_t$ and $\mathbf{x}_k$ is larger than α, i.e., $\theta_{tk} > \alpha$, the difference between the updated model parameters by $\mathbf{x}_t$ and $\mathbf{x}_k$ is expected to be larger than $s \cdot \min(\|\Phi(\mathbf{x}_t)\|, \|\Phi(\mathbf{x}_k)\|)$, where s is a positive scalar.

Proof: To prove this proposition, we first assume $\mathbf{w}^*$ is the optimal solution of (8) and $\mathbf{w}_t^*$ (or $\mathbf{w}_k^*$) is the optimal solution of (9) when $\mathbf{x}_t$ (or $\mathbf{x}_k$) is added into the training set. Then,

$$\mathbf{w}_t^* = \mathbf{w}^* - \Delta\mathbf{w}_t = \mathbf{w}^* - s \cdot \phi(\mathbf{x}_t, \mathbf{y}_t) \tag{16}$$

$$\mathbf{w}_k^* = \mathbf{w}^* - \Delta\mathbf{w}_k = \mathbf{w}^* - s \cdot \phi(\mathbf{x}_k, \mathbf{y}_k) \tag{17}$$

where $\mathbf{y}_t$ and $\mathbf{y}_k$ are true labels, and s is a scalar controlling the step size in the gradient descent method. The above equations hold because the step size in each iteration is proportional to the magnitude of the current gradient. The difference between $\mathbf{w}_t^*$ and $\mathbf{w}_k^*$ is calculated as

$$\begin{aligned} \mathbf{w}_k^* - \mathbf{w}_t^* &= s \cdot (\phi(\mathbf{x}_t, \mathbf{y}_t) - \phi(\mathbf{x}_k, \mathbf{y}_k)) \\ &= s \cdot \Delta\phi_{tk}, \end{aligned} \tag{18}$$

where we denote $\Delta\phi_{tk} = \phi(\mathbf{x}_t, \mathbf{y}_t) - \phi(\mathbf{x}_k, \mathbf{y}_k)$. Since the true labels are all unknown, the expected difference should be calculated over the distribution of $P(\mathbf{y}_t|\mathbf{x}_t)$ and $P(\mathbf{y}_k|\mathbf{x}_k)$, i.e.,

$$\begin{aligned} &\|\mathbb{E}(\mathbf{w}_k^* - \mathbf{w}_t^*)\|^2 \\ &= s^2 \|\mathbb{E}(\Delta\phi_{tk})\|^2 \\ &= s^2 \| \sum_{\mathbf{y}_t, \mathbf{y}_k} P(\mathbf{y}_t|\mathbf{x}_t) P(\mathbf{y}_k|\mathbf{x}_k) (\phi(\mathbf{x}_t, \mathbf{y}_t) - \phi(\mathbf{x}_k, \mathbf{y}_k)) \|^2 \\ &= s^2 (\| \sum_{\mathbf{y}_t} P(\mathbf{y}_t|\mathbf{x}_t) \phi(\mathbf{x}_t, \mathbf{y}_t) \|^2 + \| \sum_{\mathbf{y}_k} P(\mathbf{y}_k|\mathbf{x}_k) \phi(\mathbf{x}_k, \mathbf{y}_k) \|^2 \\ &\quad - 2 \langle \sum_{\mathbf{y}_t} P(\mathbf{y}_t|\mathbf{x}_t) \phi(\mathbf{x}_t, \mathbf{y}_t), \sum_{\mathbf{y}_k} P(\mathbf{y}_k|\mathbf{x}_k) \phi(\mathbf{x}_k, \mathbf{y}_k) \rangle) \\ &= s^2 \left(\|\Phi(\mathbf{x}_t)\|^2 + \|\Phi(\mathbf{x}_k)\|^2 - 2\|\Phi(\mathbf{x}_t)\| \|\Phi(\mathbf{x}_k)\| \cos\theta_{tk} \right). \end{aligned} \tag{19}$$

Therefore, $\forall\ \alpha \geq \frac{\pi}{3}$ and $\theta_{tk} > \alpha$,

$$\begin{aligned} &\|\mathbb{E}(\mathbf{w}_k^* - \mathbf{w}_t^*)\|^2 \\ &> s^2 \left(\|\Phi(\mathbf{x}_t)\|^2 + \|\Phi(\mathbf{x}_k)\|^2 - \|\Phi(\mathbf{x}_t)\| \|\Phi(\mathbf{x}_k)\| \right) \\ &\geq s^2 \cdot \min(\|\Phi(\mathbf{x}_t)\|^2, \|\Phi(\mathbf{x}_k)\|^2), \end{aligned} \tag{20}$$

and

$$\|\mathbb{E}(\mathbf{w}_k^* - \mathbf{w}_t^*)\| > s \cdot \min(\|\Phi(\mathbf{x}_t)\|, \|\Phi(\mathbf{x}_k)\|). \tag{21}$$

The proposition is proved. ■

Proposition 1 states that when DGA θ_{tk} is large enough, the difference between the updated model parameters by $\mathbf{x}_t$ and $\mathbf{x}_k$ is proportional to $\min(\|\Phi(\mathbf{x}_t)\|, \|\Phi(\mathbf{x}_k)\|)$. It is therefore reasonable to believe that instances $\mathbf{x}_t$ and $\mathbf{x}_k$ can result in quite different model parameters under this condition. Furthermore, $\mathbf{x}_t$ and $\mathbf{x}_k$ become more diverse in relation to each other when θ_{tk} continues to increase. To select a batch of queries with an expected diversity, we incorporate the following objective function to modify the query set:

$$\begin{aligned} X^D &= \underset{X \subseteq U \cap |X| = K}{\text{argmax}} \sum_{\mathbf{x}_t, \mathbf{x}_k \in X} \theta_{tk} \\ &s.\ t. \quad \forall\ \mathbf{x}_t, \mathbf{x}_k \in X: \quad \cos\theta_{tk} > \alpha. \end{aligned} \tag{22}$$

The free parameter α controls the diversity of the query set, i.e., when the value of α increases, the constraint $\theta_{tk} > \alpha$ could result in a more diverse set. However, the large value of α will lead to only a limited number of available queries, since only a small number of queries can satisfy the constraint. This might result in no queries being selected in subsequent iterations. To overcome this issue, we use an intuitive method to set α in each iteration, which is detailed in the next subsection.

C. Diverse Expected Gradient Active Learning

In this section, we develop our *Diverse Expected Gradient Active Learning (DEGAL)* method as a combination of the methods proposed in Sections III-A and III-B. By integrating both the informativeness and diversity analyses, the overall objective function becomes

$$\begin{aligned} X^{DEGAL} &= \underset{X \subseteq U \cap |X| = K}{\text{argmax}} \sum_{\mathbf{x} \in X} \Psi(\mathbf{x}) + \sum_{\mathbf{x}_t, \mathbf{x}_k \in X} \theta_{tk} \\ &s.\ t. \quad \forall\ \mathbf{x}_t, \mathbf{x}_k \in X: \quad \cos\theta_{tk} > \alpha \end{aligned} \tag{23}$$

where the first item provides the measurement of informativeness of the query set and the second item accumulates the set's diversity.

Optimization of the above problem would undoubtedly produce a diverse query set. However, to the best of our knowledge, it is intractable due to the need to enumerate all possible combinations of queries in the unlabeled set to achieve the optimal solution. To tackle this, we propose a two-step heuristic method to discover an approximate optimal query set.

In each selection iteration, we first calculate the informativeness of all unlabeled samples in U, and select K' $(> K)$ most informative instances, which form the candidate set Q. The second step eliminates all candidates in Q that do not

Algorithm 1 *DEGAL*

```
Input:
    Optimized parameter $\mathbf{w}$, unlabeled set $U$.
Output:
    Diverse query set $X$ comprising $K$ queries, and the updated model parameter $\mathbf{w}^{new}$.
Calculate, according to (13), the informativeness $\Psi(\mathbf{x}_t)$ of each sample in $U$ with $\mathbf{w}$;
Collect $K'$ $(> K)$ most informative samples with a decreasing order, to form candidate set $Q$;
Calculate, according to (15), the Diverse Gradient Angle between each pair of samples in $Q$, resulting the matrix $\Theta$ where $[\Theta]_{tk} = \theta_{tk}$;
Initialize $\alpha$;
repeat
    Process $\mathbf{x}_t \in Q$ from the most informative, to least informative;
    Find the samples whose $\theta_{t\cdot}$(w.r.t. $\mathbf{x}_t$) violate the constraint of $\theta_{t\cdot} > \alpha$;
    Eliminate these samples from $Q$, and the corresponding rows and columns of $\Theta$;
    if $(\#$ of $Q) > K$ then
        Increase $\alpha = \alpha + \epsilon$;
    end if
until There are $K$ samples remaining in $Q$;
$X = Q$;
Add X into the training set;
Optimize the model parameter $\mathbf{w}^{new}$ on the enlarged training set, by using $\mathbf{w}$ as initialization;
return $X = Q$ and $\mathbf{w}^{new}$.
```

satisfy the constraints in (22). The overall *DEGAL* is shown in Algorithm 1.

In the above procedure, the issue remains of how to initialize and increase the value of α. Due to the unknown structure of the feature space, we cannot evaluate what α could properly diversify different samples. Also, when the iterations continue for some time, the informativeness of the remainder of the unlabeled samples may not be as strong as those previously selected, because the direction parameter $\mathbf{w}$ tends to be globally optimal. In this case, any fixed value of α would cause an empty set of Q. Here, we incorporate an exploratory trick that starts with a relatively small α at initialization and then increase α by ϵ if all values of Θ satisfy the constraints; meanwhile, more K samples remain in Q.

D. The Multi-class Imbalance Issue

In the relative attributes model, the task requires learning a ranking direction, along which the pairs of images in the training set are ordered as correctly as possible. The image pairs in O have different attribute strengths, while those in S are similar. From this viewpoint, this task can be cast as a multi-class classification scenario, where each class label corresponds to a specific attribute strength, and different images with different attribute strengths belong to different classes.

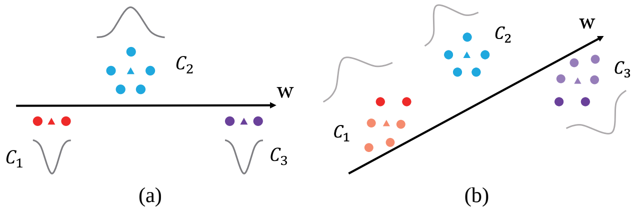


Fig. 5. Multi-class imbalance. Circles indicate training samples, while triangles denote the class center. Figure (a) illustrates the poor solution produced by an imbalanced training set formed according to an active learning scheme. Figure (b) shows the optimal solution if all data and their labels are observed. Indeed, once the solution is affected by imbalance, it becomes even more severe in the subsequent iteration and cannot be self-tuned into a regular case.

However, the multi-class imbalance problem, which is a severe limitation of multi-class classification [36], [37], also affects relative attributes performance and, as a consequence, the proposed active learning scheme. In fact, in an active learning procedure, all query selection is blind. If the diversity of one class is large, then it is very likely to choose queries belonging to this class in every iteration, leading to a serious imbalance in the distribution of the training data. Even if an acceptable balance between all classes of all data exists, a random initialization of the training data would result in a poor selection of queries in the subsequent iteration and eventually form an imbalanced training set. For an example, see Fig. 5.

We therefore propose to control the imbalance of the training set in each iteration by first defining the balance constraint, as follows.

Definition 2: The balance constraint denotes that the number of training samples in each class differ from each other no more than κ.

Given this definition, we utilize a Gaussian model to simulate the class center of each attribute strength, which is estimated according to the ranking score under the current model parameters. In a new selection iteration, once the query set X is obtained, the active learner predicts the label of each query in X and discards those that violate the balance constraint. Note that in this step the oracle has no responsibility to inform the active learner whether the predicted label is correct or incorrect, which is reasonable because, if the label is correct, either keeping or eliminating this sample is exactly what we intend to do. If the label is incorrect, retaining this sample would, as expected, adjust the model parameters, while removing it does not harm the overall procedure, except for slightly affecting the rate at which the optimal solution is reached.

One further issue deserves consideration: if some queries are discarded by the balance constraint in the current selection iteration, they could still be selected as queries in the next iteration, and continuing execution would likely result in an infinite loop. To ensure that the active learner does not fall into this trap, we construct a backup set to store the discarded queries, instead of returning them into the unlabeled pool. After a certain number of iterations (say 20 or 30), this backup set is returned to the unlabeled pool.

The whole procedure for controlling the multi-class imbalance issue is shown in Algorithm 2.

Algorithm 2 *Multi-class Balancing*

```
Input:
    The labelled training set $L$, unlabeled set $U$, $\kappa$ and MAX_BACKUP_ITER.
Optimize the model parameter $\mathbf{w}$ on $L$;
Initialize the backup set $B = \emptyset$;
$i = 1$;
repeat
    Predict the attribute scores $r(\mathbf{x})$ in $L$ using $\mathbf{w}$;
    Calculate the Gaussian parameters for each class by using the predicted scores;
    $(X, \mathbf{w}^{new})$=DEGAL$(\mathbf{w}, U)$;
    $U = U \backslash X$;
    Predict the attribute scores $r(\mathbf{x})$ in $X$ using $\mathbf{w}^{new}$;
    Estimate each query's label as the one with the highest Gaussian probability;
    Eliminate the queries from $X$ which would violate the balance constraint if added to $L$;
    Store the discarded queries in $B$;
    $L = L \bigcup X$, $\mathbf{w} = \mathbf{w}^{new}$;
    $i = i + 1$;
    if $i >$MAX_BACKUP_ITER then
        Return $B$ to $U$;
        $i = 1$, $B = \emptyset$;
    end if
until Get an expected model parameter.
```

IV. Experiments

A. Dataset

To empirically investigate performance, we evaluate our approach on three datasets: the Outdoor Scene Recognition (OSR) dataset [38], the Public Figures Face (PubFig) dataset [13], and the Shoes from the Attribute Discovery dataset [39]. The OSR data contains 2688 images from eight categories and has six attributes *('natural', 'open', 'perspective', 'large-objects', 'diagonal-plane', 'close-depth')* described by a 512-dimensional GIST descriptor. For the PubFig dataset, a subset of images was selected from the original dataset in [13], consisting of 772 images from eight people with eleven attributes *('masculine-looking', 'white', 'young', 'smiling', 'chubby', 'visible-forehead', 'bushy-eyebrows', 'narrow-eyes', 'pointy-nose', 'big-lips', 'round-face')*. The feature for describing face instances is a concatenation of 512-dimensional GIST features and 30-dimensional color histogram features. The third dataset, Shoes, is a relatively large-scale dataset that contains 14658 shoe images structured by ten classes and ten attributes *('pointy-at-the-front', 'open', 'bright-in-color', 'covered-with-ornaments', 'shiny', 'high-at-the-heel', 'long-on-the-leg', 'formal', 'sporty', 'feminine')*, and 960-dimensional GIST features and 30-dimensional color histogram features are also utilized to describe a shoe instance. These datasets cover diverse domains of interest, including natural scenes, human faces, and products (Fig. 6) and provide an ideal test-bed for our approach.

Scenes

Faces

Shoes

Fig. 6. Example images. First row: OSR examples. Second row: PubFig examples. Third row: Shoes examples.

B. Experimental Setup

This section provides a detailed description of our baselines and experimental settings.

Dataset Splitting: Every dataset is equally divided into two sets, i.e., 50% of samples in the training set and 50% of samples in the testing set. For each trial, the samples used for training are selected from the training set, either actively or randomly, while the testing set is used to evaluate performance.

Baselines: We include four active learning baselines and two randomized baselines for comparison. The first active learning baseline is the batch mode proposed in [31]. Since this method is realized by minimizing the version space of the model, while maximizing the diversity measurement (which we call the *Kernel-based Angle* (KBA)) of the query batch, we denote it as "MVS+KBA". The second active learning baseline ("MS+KBA") is a combination of multi-class uncertainty sampling (called *margin sampling* (MS)) and KBA [40], [41]. The third ("Entropy-QBC") is an extension of query-by-committee algorithms from the entropy viewpoint [42]. The final active learning baseline ("EGL+KBA") is an integration of EGL and KBA. The two randomized baselines are as follows: the first ("Random") is where training samples are randomly selected from the training set, while the quantity in each class is the same. The second ("RandomC") is designed to illustrate the effect of the imbalance issue, and differs from the first in that all classes are randomly and equally split into two cliques. We randomly choose $90\% \times N$ samples for classes in the first clique and $10\% \times N$ for classes in the second clique, where N denotes the total number of samples used to train the model parameter. Note that in this situation, the number of samples in all the classes is not necessarily equal.

Active Setting: For active learning, an initial labeled set L

is randomly selected from the training set, and the remainder form the unlabeled set U. Since a good initialization of a model parameter (i.e., the model is trained on the initial labeled set) is favorable for active learning, we set the number of initial labeled samples to: 1) 32 for OSR and PubFig with four samples per class; and 2) 100 for Shoes, with ten samples per class. For each iteration, we utilize a fixed batch size of 5, meaning that 5 queries are chosen each time. The angular parameter α is set to $\pi/4$, and ϵ to $\pi/36$. Furthermore, the balance constraint parameter κ is set at 5. The influence of different parameter settings can be found in Section IV-D.

Model training: In all experiments, problem (8) is optimized using Joachims' method [43], and the parameter C is determined by cross-validation. All experiments are conducted 10 times, and the performance is then averaged.

Performance evaluation: Performance is measured by the accuracy of predicted relative strength. For two test samples $\mathbf{x}_i$ and $\mathbf{x}_j$, the comparison $r_m(\mathbf{x}_i) > r_m(\mathbf{x}_j)$ (or $r_m(\mathbf{x}_i) < r_m(\mathbf{x}_j)$) is correct if it is consistent with the ground truth. Then, the total accuracy for attribute m is calculated by the rate of correct comparison to total comparison.

C. Results

The comparative results for different relative attributes on the OSR dataset, PubFig dataset and Shoes dataset are shown in Fig. 7, Fig. 8 and Fig. 9, respectively. *DEGAL* performs better than both the active learning baselines and the random baselines. Surprisingly, *Random* works well in some of our experiments, and may even be comparable to *DEGAL* in some cases. This phenomenon helps us to understand the meaning of each attribute in each of the datasets. Prior to analysis, we defined two types of attributes, namely the global attributes and the local attributes. Global attributes are related to the whole image, while local attributes are determined only by a local region in the image. In the OSR dataset, all six attributes are global because they need to be assigned by observing the whole image. In this case, *DEGAL* always performs better than *Random* since the features used are globally related to the attribute value. However, in the Pubfig dataset, some attributes are global (*male*, *white*, *young*, *chubby*, and *round-face*), while the others are local (*smiling*, *visible-forehead*, *bushy-eyebrows*, *narrow-eyes*, *pointy-nose* and *big-lips*). The results indicate that *DEGAL* still works well on the global attributes but exhibits some limitations on the local attributes. One reason for this is that our method cannot localize the attributes to a specific region, and without this information, i.e., the localization, the active learner might be confused and easily affected by the distinct features in other regions; the algorithm subsequently picks up queries that lead the attribute learner in the wrong direction. In subsequent iterations, the attribute learner performs normally until there are enough training samples, such that the localization can be recovered. For example, the attribute *smiling* can be regarded as a global case, since a smile changes the profile of the whole face even though it is only related to the mouth shape; *DEGAL* therefore works well on *smiling*. Taking the attribute *narrow-eyes* into consideration, even if a person has a pair of large eyes, laughing can narrow the eyes. If this type of image existed, the active learner's belief could be misled towards *narrow-eyes* (see Fig. 8(e)). The Shoes dataset also has the same issue, such as with the local attribute of *high-at-the-heel*. However, for most attributes, *DEGAL* performs well. These results suggest that *DEGAL* can discover key information in learning the relative attributes models, but lacks the ability of self-tuning when it gets into a trap.

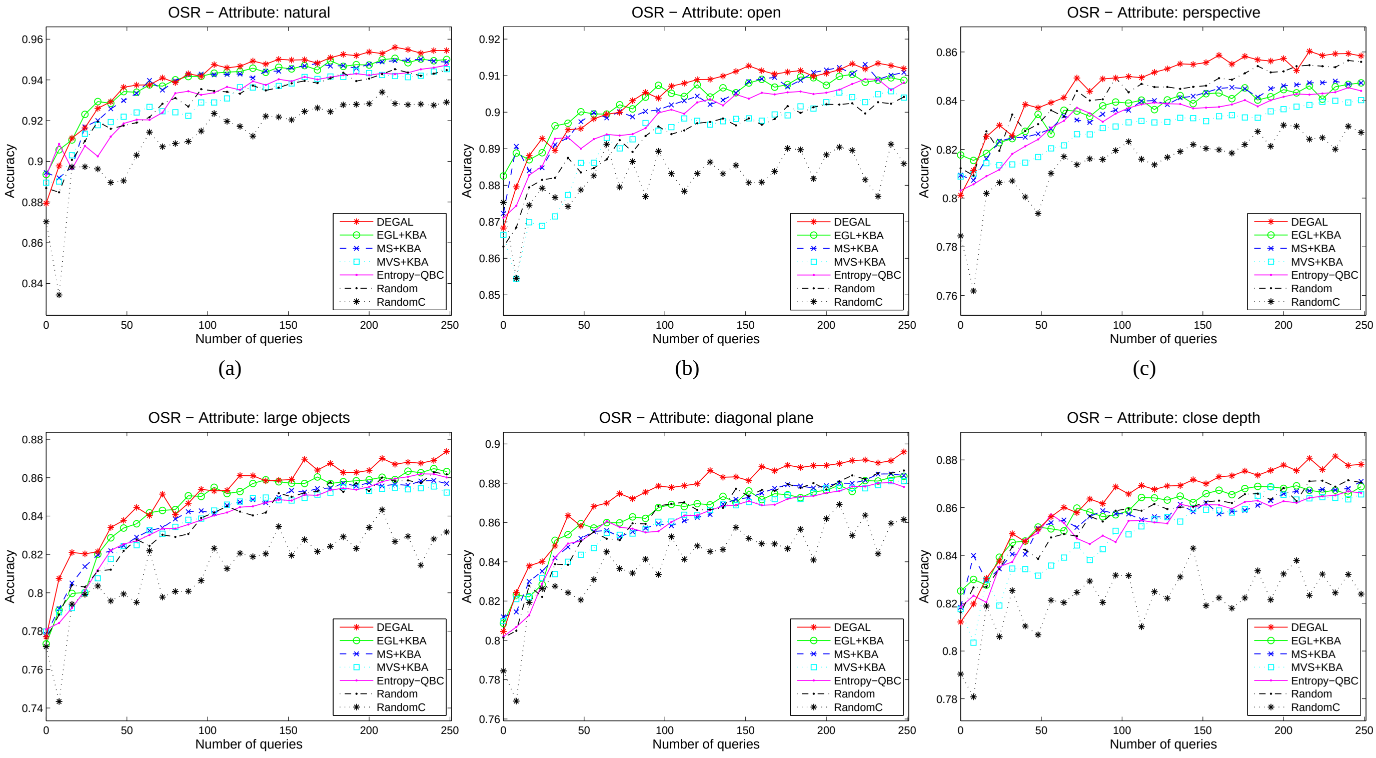
DEGAL is superior to the four active learning baselines in most cases. It can be seen that MVS+KBA has very limited performance, consistent with the analysis in Section II-B, i.e., the samples closest to the classification hyperplane are not necessarily the most informative. *DEGAL* outperforms both MS+KBA and Entropy-QBC, because our method intends to find a query that can produce a large number of uncertain pairwise relationships, whereas the other two methods treat the relative attributes as a multi-class problem and only select a query uncertain to all classes. At this point, the information induced by one query in *DEGAL* is larger than that in MS+KBA and Entropy-QBC. Finally, *DEGAL* offers an improvement over EGL+KBA by modifying the original EGL as the expected pairwise gradient length, which is more suitable for the relative attributes model. Furthermore, the proposed diversity measurement in the gradient space is an alternative to that in the primal feature space or the kernelized space.

The comparison of *DEGAL* with *RandomC* illustrates the importance of the *balance constraint*. *RandomC* becomes extremely unstable and has limited performance due to the imbalanced distribution between different classes. Loss of information is very harmful to the training model, particularly for *DEGAL*. As stated earlier, the initialization of *DEGAL* involves only minimal information and produces significant uncertainty that could blind the subsequent selection procedure and result in an imbalanced scenario that performs similarly to *RandomC*. This also accounts for the slight turbulence on *DEGAL* curves and the severe turbulence seen for *RandomC*. The same conclusion can also be reached from the results of the experiments that test different settings of κ, presented in Section IV-D. Therefore, by using the balance constraint, *DEGAL* can outperform *RandomC*.

D. Effects of the Parameters

In this section we evaluate the effects of changing the K, α, and κ parameters. Each is tested by fixing the others. All experiments are conducted on the OSR dataset and the results are averaged over all six attribute learners' performances.

Fig. 10 shows the influence of K on the performance of *DEGAL*. Overall, the different settings have similar performances and produce consistent results for the final selections. However, for the middle selections, there is slight divergence, as shown in the enlarged box. Smaller values of K result in a better performance, while increasing K decreases the accuracy in this range, suggesting that smaller K helps the active learner to precisely explore the feature space, while larger K results in redundant information in the selected query set. Although smaller K works better, this comes at computational



(a) (b) (c)

(d) (e) (f)

Fig. 7. Results on OSR Dataset

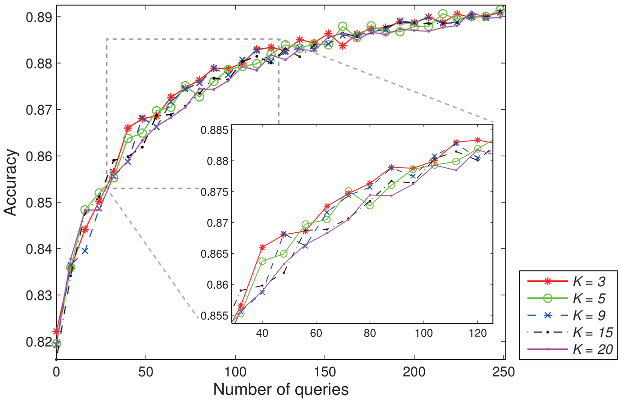


Fig. 10. The effects of the parameter K when setting $\alpha = \pi/4$ and $\kappa = 5$.

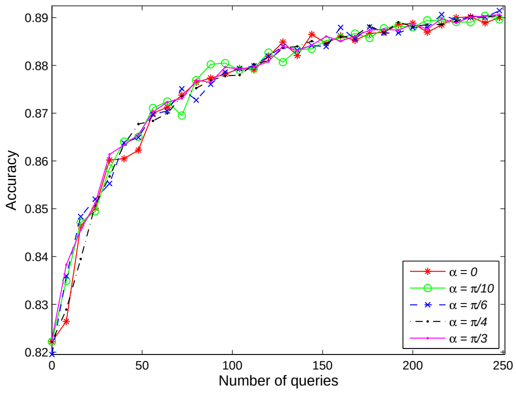


Fig. 11. The effects of the parameter α when setting $K = 5$ and $\kappa = 5$.

cost due to the need for more iterations for selections and retraining. There is therefore a trade-off between K and the computational cost, which needs to be considered in the implementation.

Fig. 11 shows the effects of angular parameter α. Different values of α have almost the same performance, suggesting that the proposed algorithm is insensitive to this parameter. Note that this is partially due to the incremental strategy used in Algorithm 1, i.e., increasing α by ϵ.

Fig. 12 shows that parameter κ has a relatively significant influence on performance. $\kappa = 5$ and $\kappa = 10$ have similar performance, while when $\kappa = 20$ the accuracy slightly decreases for the middle selections. When κ is set to 40, the performance significantly drops as the number of queries

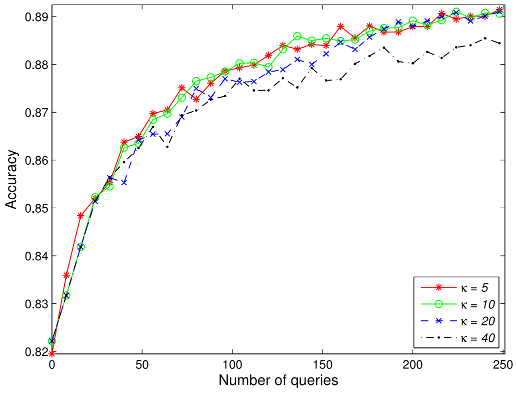
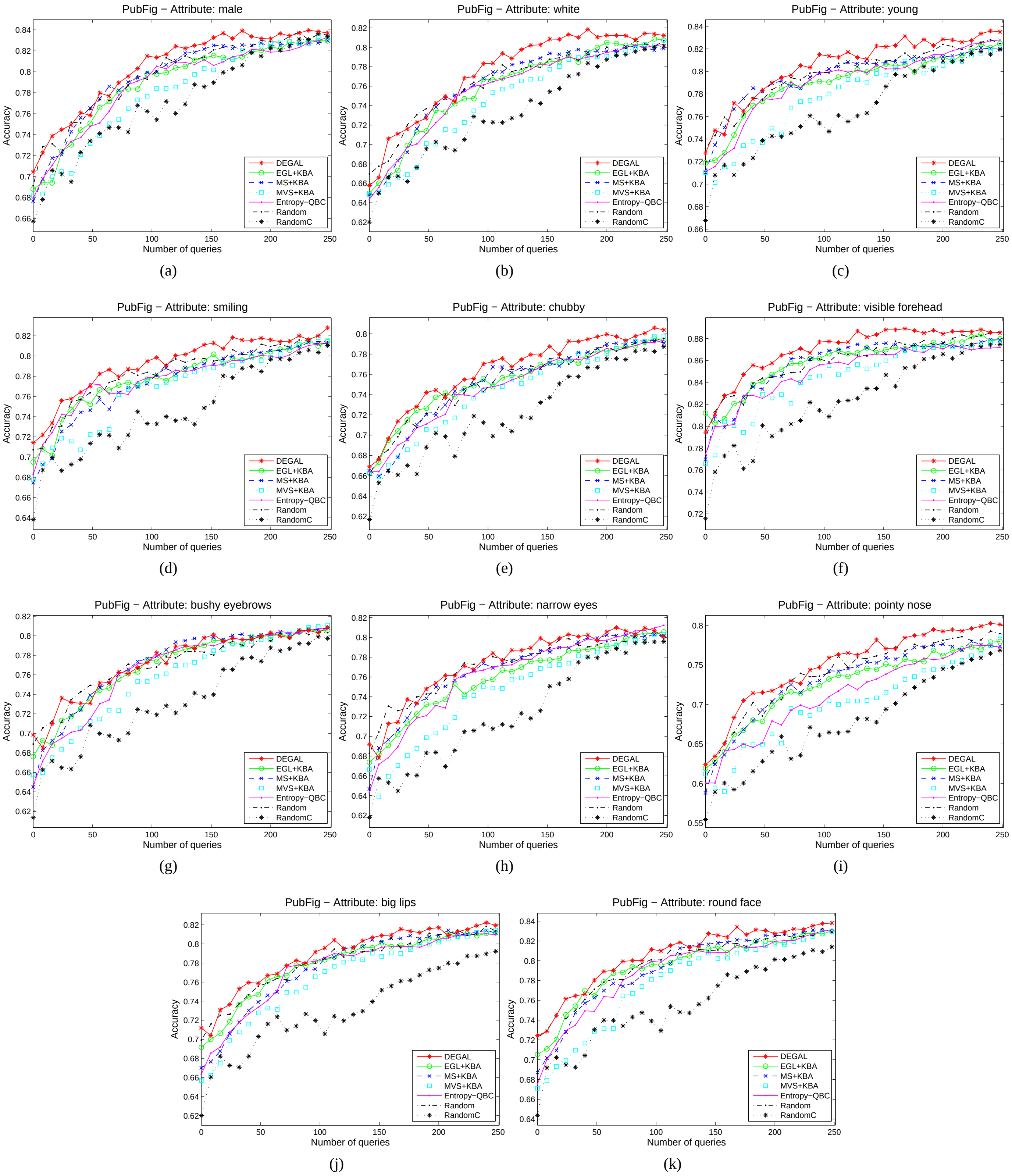


Fig. 12. The effects of the parameter κ when setting $K = 5$ and $\alpha = \pi/4$.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

(j)

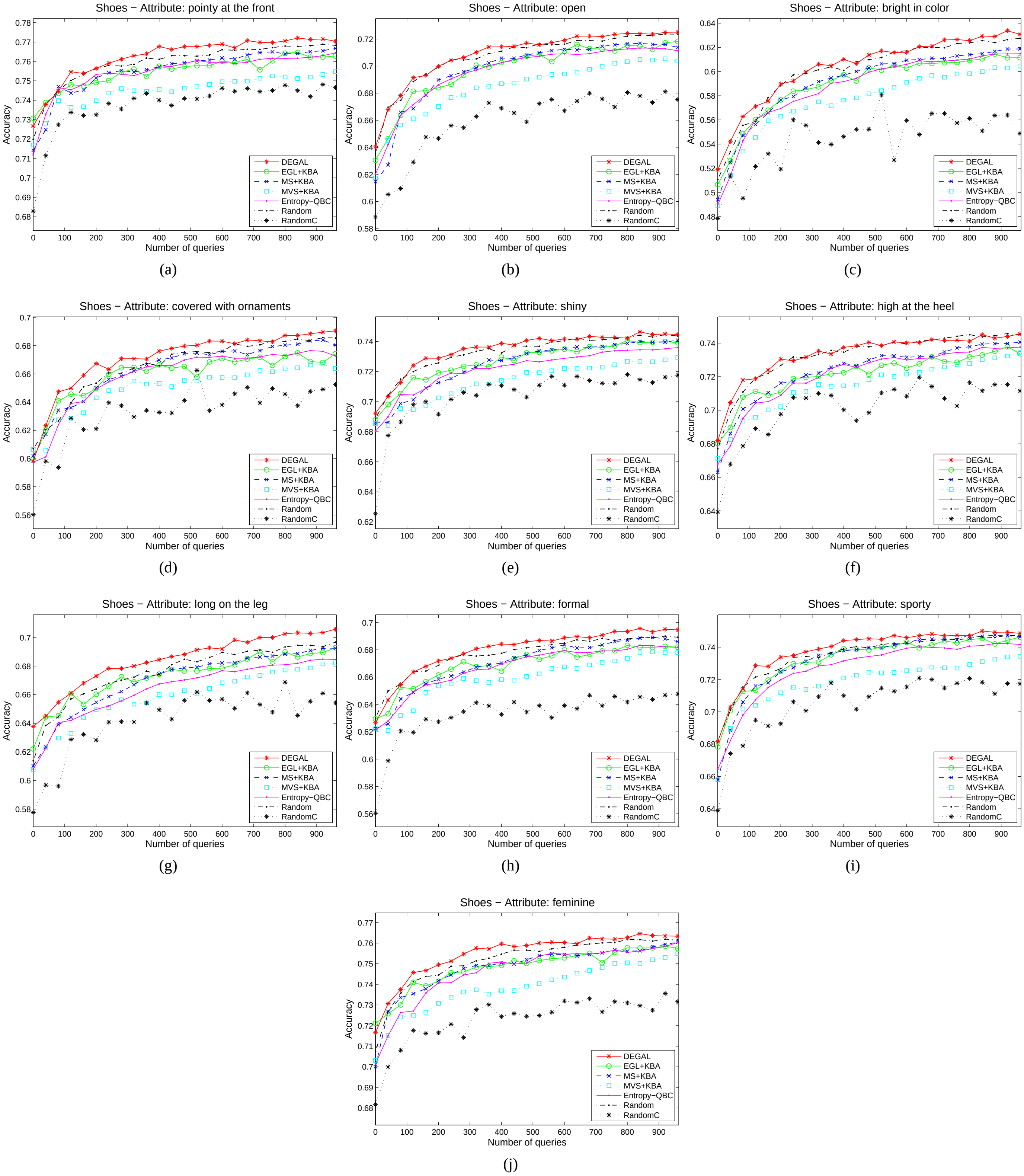
(k)

Fig. 8. Results on PubFig Dataset

increases, suggesting that the imbalance issue between the classes is severe enough to negatively affect the active learner. Keeping κ small is therefore an effective way to control the imbalance of multi-class distributions.

E. Computational Complexity

The computational complexity of the proposed method is analyzed based on the cost in one iteration, which consists of the training model parameters and active selection. Assume



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j)

Fig. 9. Results on Shoes Dataset

that the training set has m samples, the unlabeled pool has n samples, the batch size is K, and the attribute has l levels of strength. Since C_m^2 pairs need to be considered in the training phase, the complexity for training is of order $O((C_m^2)^3)$ in the worst case [33]. In the active selection phase, the time taken to compute the informativeness and the angle values is $O(nm^2)$ and $O(C_K^2)$, respectively. The calculation involved in multi-class balancing is of order $O(m+K)$. Thus, the total complexity in one iteration is $O((C_m^2)^3 + nm^2 + C_K^2 + m + K)$. The corresponding computational complexities of

TABLE I
COMPARISON OF COMPUTATIONAL COMPLEXITY

Method Name	**Computational Complexity**
DEGAL	$O(nm^2 + C_K^2 + m + K)$
EGL+KBA	$O(nm^2 + nm)$
MS+KBA	$O(lnC_n^2 + nm)$
MVS+KBA	$O(nC_n^2 + nm)$
Entropy-QBC	$O(l^2 n)$

the different methods are shown in Table I. For clarity, we omit the complexity for training since it is the same for all approaches. Our method takes less time than other batch-mode active learning strategies, since m and K are generally smaller than n. Entropy-QBC has the lowest computational complexity because it is a serial-mode method, but this method may take more iterations to reach an expected solution.

V. CONCLUSION

Incorporating an active learning scheme into semantic learning is a promising method to efficiently improve various semantic learners, especially when faced with a large amount of internet data. In order to focus on the improvement of relative attributes learning with limited label information, here we present a novel batch-based active learning method, called *Diverse Expected Gradient Active Learning (DEGAL)*. We use the expected gradient length as the informativeness of each unlabeled sample, and illustrate its equivalence to Tong's result [30]. To collect a batch of queries of reasonable diversity, we constrain the *diverse gradient angles* between the queries to preserve different guidance on parameter optimization. Finally, a two-step optimization is formulated that ranges from informativeness analysis to diversity analysis. To address the problem of imbalanced class distribution, we exploit a simple method to minimize the issue using the *balance constraint*. The experimental results on three different kinds of datasets demonstrate that the proposed *DEGAL* is superior to other baselines.

However, *DEGAL* still has some limitations. For example, how to actively discover the specific regions related to the local attributes remains open. Furthermore, the proposed method is fixed at the attribute level. How to define the joint informativeness of a single sample for different attributes still needs to be considered, and this will be investigated in future work.

APPENDIX A
EQUIVALENCE BETWEEN $\max_{\mathbf{x}} \Psi(\mathbf{x})$ AND TONG'S RESULT [30]

Remember that Tong's result indicates the most informative query is the one located on the classification hyperplane, which says that the label of this query has 50% probability to be 1 and 50% to be -1. To see equivalence between this and $\max_{\mathbf{x}} \Psi(\mathbf{x})$, we can write $\Psi(\mathbf{x})$ by substituting (11) into (13):

$$\begin{aligned}\Psi(\mathbf{x}) =& \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x})\psi(\mathbf{x},\mathbf{y})\\ =&2\sum_{\mathbf{x}_u}\sum_{y_u} P(y_u|\mathbf{x})\|g(\mathbf{x}_u,\mathbf{x},y_u)\|\\ =&2\sum_{\mathbf{x}_u}\left(P(1|\mathbf{x})\|g(\mathbf{x}_u,\mathbf{x},1)\| + P(-1|\mathbf{x})\|g(\mathbf{x}_u,\mathbf{x},-1)\ \|\right.\\ &\left.+P(0|\mathbf{x})\|g(\mathbf{x}_u,\mathbf{x},0)\|\right). \qquad (24)\end{aligned}$$

Due to the unit constraint in equation (7),

$$\begin{aligned}\Psi(\mathbf{x}) =&2\sum_{\mathbf{x}_u}\big(P(1|\mathbf{x})[1-\mathbf{w}_m^{*T}(\mathbf{x}_u-\mathbf{x})]_+\\ &+P(-1|\mathbf{x})[1+\mathbf{w}_m^{*T}(\mathbf{x}_u-\mathbf{x})]_+\\ &+P(0|\mathbf{x})|\mathbf{w}_m^{*T}(\mathbf{x}_u-\mathbf{x})|\big). \qquad (25)\end{aligned}$$

Define the above accumulation items as $f(\mathbf{x}_u,\mathbf{x})$:

$$\begin{aligned}&f(\mathbf{x}_u,\mathbf{x})\\ =&P(1|\mathbf{x})[1-\mathbf{w}_m^{*T}(\mathbf{x}_u-\mathbf{x})]_+ + P(-1|\mathbf{x})[1+\mathbf{w}_m^{*T}(\mathbf{x}_u-\mathbf{x})]_+\\ &+P(0|\mathbf{x})|\mathbf{w}_m^{*T}(\mathbf{x}_u-\mathbf{x})|\\ \le&(P(1|\mathbf{x})+P(-1|\mathbf{x}))+(P(-1|\mathbf{x})-P(1|\mathbf{x}))\mathbf{w}_m^{*T}(\mathbf{x}_u-\mathbf{x})\\ &+P(0|\mathbf{x})|\mathbf{w}_m^{*T}(\mathbf{x}_u-\mathbf{x})|. \qquad (26)\end{aligned}$$

The above equality is obtained when $|\mathbf{w}_m^{*T}(\mathbf{x}_u-\mathbf{x})| \le 1$. Note that $P(1|\mathbf{x})+P(-1|\mathbf{x})+P(0|\mathbf{x}) = 1$, and in an ideal case, $P(1|\mathbf{x})+P(-1|\mathbf{x}) = 1$ for $\mathbf{w}_m^{*T}(\mathbf{x}_u-\mathbf{x}) \neq 0$, while $P(0|\mathbf{x}) = 1$ for $\mathbf{w}_m^{*T}(\mathbf{x}_u-\mathbf{x}) = 0$. This is because of the definitions of O and S.

Without loss of generality, we assume $\exists\mathbf{x}_u$, $|\mathbf{w}_m^{*T}(\mathbf{x}_u-\mathbf{x})| \le 1$. Therefore, if $\mathbf{w}_m^{*T}(\mathbf{x}_u-\mathbf{x}) \neq 0$,

$$f(\mathbf{x}_u,\mathbf{x}) = 1 + (P(-1|\mathbf{x})-P(1|\mathbf{x}))\mathbf{w}_m^{*T}(\mathbf{x}_u-\mathbf{x}). \qquad (27)$$

In this case, $f(\mathbf{x}_u,\mathbf{x})$ reaches a maximum when $P(-1|\mathbf{x}) = P(1|\mathbf{x}) = 0.5$, which is exactly the same as Tong's result. On the other hand, if $\mathbf{w}_m^{*T}(\mathbf{x}_u-\mathbf{x}) = 0$,

$$f(\mathbf{x}_u,\mathbf{x}) = |\mathbf{w}_m^{*T}(\mathbf{x}_u-\mathbf{x})|. \qquad (28)$$

Under this condition, any large value of $f(\mathbf{x}_u,\mathbf{x})$ means that the pairwise difference $\mathbf{x}_u-\mathbf{x}$ is informative to the current model parameter $\mathbf{w}_m^*$, since it is supposed to be located on the classification hyperplane.

In conclusion, since $f(\mathbf{x}_u,\mathbf{x}) \ge 0$, the query $\mathbf{x}$ maximizing $\Psi(\mathbf{x})$ is the one which produces the greatest quantity of informative pairwise differences. Adding such a query to the training set can significantly improve leading the optimization procedure to the optimum.

REFERENCES

[1] N. Sebe, M. S. Lew, X. Zhou, T. S. Huang, and E. M. Bakker, "The state of the art in image and video retrieval," in *Image and Video Retrieval*, 2003, pp. 1–8.

[2] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via plsa," in *Proc. 9th European Conference on Computer Vision*, 2006, pp. 517–530.

[3] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *Proc. 10th IEEE International Conference on Computer Vision*, vol. 1, 2005, pp. 370–377.

[4] S.-F. Wong, T.-K. Kim, and R. Cipolla, "Learning motion categories using both semantic and structural information," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–6.

[5] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2005, pp. 524–531.

[6] J. Sivic, B. C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros, "Unsupervised discovery of visual object class hierarchies," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[7] J. Liu, Y. Yang, and M. Shah, "Learning semantic visual vocabularies using diffusion distance," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 461–468.

[8] G. Yu, J. Yuan, and Z. Liu, "Action search by example using randomized visual vocabularies," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 377–390, 2013.

[9] J. Vogel and B. Schiele, "Semantic modeling of natural scenes for content-based image retrieval," *International Journal of Computer Vision*, vol. 72, no. 2, pp. 133–157, 2007.

[10] R. Ji, H. Yao, X. Sun, B. Zhong, and W. Gao, "Towards semantic embedding in visual vocabulary," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 918–925.

[11] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1778–1785.

[12] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 951–958.

[13] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *Proc. 12th IEEE International Conference on Computer Vision*, 2009, pp. 365–372.

[14] Y. Su and F. Jurie, "Improving image classification using semantic attributes," *International Journal of Computer Vision*, vol. 100, no. 1, pp. 59–77, 2012.

[15] C. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PP, no. 99, pp. 1–14, 2013.

[16] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 3337–3344.

[17] M. Douze, A. Ramisa, and C. Schmid, "Combining attributes and fisher vectors for efficient image retrieval," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 745–752.

[18] D. Parikh and K. Grauman, "Relative attributes," in *Proc. 13th IEEE International Conference on Computer Vision*, 2011, pp. 503–510.

[19] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Attribute learning for understanding unstructured social activity," in *Proc. 12th European Conference on Computer Vision*, 2012, pp. 530–543.

[20] A. Parkash and D. Parikh, "Attributes for classifier feedback," in *Proc. 12th European Conference on Computer Vision*, 2012, pp. 354–368.

[21] A. Biswas and D. Parikh, "Simultaneous active learning of classifiers & attributes via relative feedback," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[22] X. Y. Felix, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang, "Designing category-level attributes for discriminative visual recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[23] J. Choi, M. Rastegari, A. Farhadi, and L. S. Davis, "Adding unlabeled samples to categories by learned attributes," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[24] J. Tang, Z.-J. Zha, D. Tao, and T.-S. Chua, "Semantic-gap-oriented active learning for multilabel image annotation," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2354–2360, 2012.

[25] L. Zhang, C. Chen, J. Bu, and X. He, "A unified feature and instance selection framework using optimum experimental design," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2379–2388, 2012.

[26] X.-Y. Wei and Z.-Q. Yang, "Coaching the exploration and exploitation in active learning for interactive video retrieval," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 955–968, 2013.

[27] B. Settles, "Active learning literature survey," *University of Wisconsin, Madison*, 2010.

[28] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," in *Advances in Neural Information Processing Systems*, 2007, pp. 1289–1296.

[29] P. Donmez and J. G. Carbonell, "Optimizing estimated loss reduction for active sampling in rank learning," in *Proc. 25th International Conference on Machine Learning*, 2008, pp. 248–255.

[30] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of Machine Learning Research*, vol. 2, pp. 45–66, 2002.

[31] K. Brinker, "Incorporating diversity in active learning with support vector machines," in *Proc. 20th International Conference on Machine Learning*, vol. 3, 2003, pp. 59–66.

[32] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *Advances in Neural Information Processing Systems*, 2009, pp. 1410–1418.

[33] O. Chapelle, "Training a support vector machine in the primal," *Neural Computation*, vol. 19, no. 5, pp. 1155–1178, 2007.

[34] C. J. Burges, D. J. Crisp *et al.*, "Uniqueness of the svm solution," in *Advances in Neural Information Processing Systems*, 1999, pp. 223–229.

[35] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[36] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[37] S. Wang and X. Yao, "Multiclass imbalance problems: Analysis and potential solutions," *IEEE Trans. Syst., Man, Cybern., Part B*, vol. 42, no. 4, pp. 1119–1130, 2012.

[38] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[39] T. L. Berg, A. C. Berg, and J. Shih, "Automatic attribute discovery and characterization from noisy web data," in *Proc. 11th European Conference on Computer Vision*, 2010, pp. 663–676.

[40] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden markov models for information extraction," in *Advances in Intelligent Data Analysis*, 2001, pp. 309–318.

[41] B. Demir, C. Persello, and L. Bruzzone, "Batch-mode active-learning methods for the interactive classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 3, pp. 1014–1031, 2011.

[42] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2218–2232, 2009.

[43] T. Joachims, "Optimizing search engines using clickthrough data," in *Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 133–142.