Image Classification by Cross-Media Active Learning With Privileged Information

Yan Yan, Feiping Nie, Wen Li, Chenqiang Gao, Yi Yang, and Dong Xu, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a novel cross-media active learning algorithm to reduce the effort on labeling images for training. The Internet images are often associated with rich textual descriptions. Even though such textual information is not available in test images, it is still useful for learning robust classifiers. In light of this, we apply the recently proposed supervised learning paradigm, learning using privileged information, to the active learning task. Specifically, we train classifiers on both visual features and privileged information, and measure the uncertainty of unlabeled data by exploiting the learned classifiers and slacking function. Then, we propose to select unlabeled samples by jointly measuring the cross-media uncertainty and the visual diversity. Our method automatically learns the optimal tradeoff parameter between the two measurements, which in turn makes our algorithms particularly suitable for real-world applications. Extensive experiments demonstrate the effectiveness of our approach.**

***Index Terms*—Active learning, cross-media analysis, image classification, image-text joint modeling.**

I. Introduction

WITH the advance of network technology and web services, numerous photos are uploaded to the internet every day, which makes the internet becomes a huge repository of images. Therefore, collecting web images as the training data has become a popular way to learn models for image classification [1]–[4]. Labeling large scale images is time consuming and labour intensive. A more practical way is to actively sample and label a small subset of training images which are the most informative [5]–[8].

In this paper, we propose a novel active sample selection approach (a.k.a. active learning) for image classification by using web images. Previous research has shown that cross-media modeling of various media types is beneficial for multimedia content analysis [9]–[13]. The web images are often associated with rich textual descriptions (e.g., surrounding texts, captions, etc). While such text information is not available in testing images, we show that text features are useful for learning robust classifiers, enabling better active learning performance of image classification. Typical active sampling methods only deal with one media type [14]–[18], which cannot simultaneously utilize different media types. The new supervised learning paradigm, namely learning using privileged information (LUPI), can be used to solve this problem [4], [19]. In a LUPI scenario, in addition to main features, there is also privileged information available in the training procedure. Privileged information can only be used in training, and is not available in testing.

Uncertainty sampling is the most frequently used strategy in the active learning [17]. In this work, we propose to exploit both visual and text features for active sample selection by taking text as privileged information. By LUPI, we train SVMs on visual features and slacking function on text features. We present five strategies to combine the uncertainty measure of these two classifiers.

To ensure the selected samples to be representative, we exploit the diversity measurement, such that the selected samples are less similar to each other. We formulate a ratio objective function to maximize cross-media uncertainty and minimize the similarity of selected data. Then we propose to measure uncertainty and diversity for training sample selection [17]. A new optimization method is proposed to solve the proposed model, which automatically learns the optimal ratio of uncertainty to similarity. In this way, we avoid introducing the trade-off parameter between the two types of measurements. We summarize the main contributions of this paper as follows:

1) By exploiting privileged information, we propose a new notion of cross-media uncertainty measurement, which measures the uncertainty of unlabeled images by jointly considering visual features as the main information and text features as the privileged information.
2) We propose a new method to optimize the objective without using the trade-off parameter between diversity and uncertainty.

II. Related Work

Active learning aims to obtain better performance when learning with fewer labeled training samples by actively selecting a portion of the training data from a pool of unlabeled data [5]. Uncertainty sampling is the most frequently used approach to active sample selection [5], [7], [20], which selects queries the unlabeled data that the learner is most uncertain.

Manuscript received September 18, 2015; revised May 18, 2016 and July 6, 2016; accepted August 10, 2016. Date of publication August 25, 2016; date of current version November 15, 2016. This work was supported in part by the Data to Decisions Cooperative Research Centre (http://www.d2dcrc.com.au), in part by the ARC DECRA, and in part by the National Natural Science Foundation of China under Grant 61571071. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Winston Hsu.

Y. Yan and Y. Yang are with the Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney, Sydney, NSW 2007, Australia (e-mail: yan.yan-3@student.uts.edu.au; yi.yang@uts.edu.au).

F. Nie is with the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: feipingnie@gmail.com).

W. Li is with the Computer Vision Laboratory, ETH Zurich, Zurich, CH 8092, Switzerland (e-mail: liwenbnu@gmail.com).

C. Gao is with the School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: gaochenqiang@gmail.com).

D. Xu is with School of Electrical and Information Engineering, University of Sydney, Sydney, NSW 2006, Australia (e-mail: dong.xu@sydney.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2602938

There are some other criteria in addition to uncertainty, such as diversity [17], representativeness [8], [16] and density [15], [21]. In [21], pre-clustering method was proposed to avoid repeatedly labeling samples in the same cluster, by which diversity can be introduced. The authors in [8] propose an active sampling strategy based on a hierarchical clustering of unlabeled data. However, the performance of these methods likely depends on the performance of clustering. If the result of clustering is not consistent with the target model, their active learning performance may degrade accordingly [16]. Some works consider representativeness. Representative unlabeled data are those that best represent the underlying distribution of data [5], [16]. In [16], the authors proposed an algorithm that takes both informativeness and representativeness of unlabeled data into consideration. A probabilistic variant of K-Nearest-Neighbor is used to extend active learning when the number of classes is large [15].

In a multi-view scenario, each sample is represented by multiple features. It is assumed that a concept is possible to learn from a single feature type [18], [22]. A web page on Wikipedia, for example, may contain various types of features, including images and texts. Co-testing is studied in [18]. It queries the samples that cause disagreement of the learners from various views, which are named contention points [18]. The motivation of co-testing is that at least one learner can lead to improvement from the queried data. A combination of multi-view active sampling and semi-supervised learning is proposed in [23].

LUPI is proposed in [19]. Compared to conventional multi-view learning, in the LUPI scenario, privileged information is only available as auxiliary features in the training process rather than the testing process. LUPI has shown promising results in many works. Various types of privileged information can be exploited to assist learning. Image attributes can be used as middle-level semantic features bridging the gap between visual features and high-level object classes [24]. Textual descriptions, which are rather abundant particularly for Web data, are frequently leveraged in classification tasks [4] and retrieval tasks [25]. In contrast to the traditional computer vision tasks such as image classification, the authors in [26] proposed a new framework by infering knowledge in the multimedia domain from the semantic domain.

III. The Proposed Model

In our task, the training data consists of an active seed set containing a few labeled samples, and a pool set containing unlabeled samples. We aim to select the most useful unlabeled samples, and query the Oracle to label them. Thus, we have a new training set for training.

Let us denote $\{(\mathbf{x}_1, \tilde{\mathbf{x}}_1, \mathbf{y}_1), (\mathbf{x}_2, \tilde{\mathbf{x}}_2, \mathbf{y}_2), ..., (\mathbf{x}_{n_s}, \tilde{\mathbf{x}}_{n_s}, \mathbf{y}_{n_s})\}$ as the seed set and $\{(\mathbf{x}_{n_s+1}, \tilde{\mathbf{x}}_{n_s+1}), (\mathbf{x}_{n_s+2}, \tilde{\mathbf{x}}_{n_s+2}), ..., (\mathbf{x}_{n_s+n_p}, \tilde{\mathbf{x}}_{n_s+n_p})\}$ as the pool set, where n_s and n_p represent the numbers of data in the active seed set and pool set respectively. For the i-th instance, we denote $\mathbf{x}_i \in \mathbb{R}^d$ as the main feature and $\tilde{\mathbf{x}}_i \in \mathbb{R}^{\tilde{d}}$ as the privileged feature, and also denote $\mathbf{y}_i \in \{-1, +1\}^t$ as the label, where t is the number of tasks or classes.

A typical active learning procedure consists of two main phases. The first step is to train an initial model (e.g., a classifier or a regression predictor) based on the labeled data, which can be used to measure the importance of the unlabeled data. In the second step, we generate the ranking scores for all unlabeled samples based on the importance measure. Given the ranking scores, the unlabeled samples that would be incorporated to the training set can thus be simply determined. We follow this process, and propose a new active learning algorithm with privileged information.

Besides the uncertainty, the proposed model additionally consider the diversity of the queried unlabeled data. While uncertain samples would benefit the performance, the underlying distribution of unlabeled data could not always be presented correctly by a few labeled data. We would miss some important information if the attention is only paid to the most uncertain samples. Therefore, we propose to find the samples that best explain the distribution of the data at the same time with considering the most uncertain ones. This can be intuitively achieved by making the unlabeled query samples as dissimilar as possible. In the second phase, we propose to consider the diversity based on the similarity matrix of the unlabeled data when computing the ranking scores. The similarity measurement can be obtained from the kernel matrix. We propose an efficient optimization method to solve the objective function with uncertainty and similarity jointly.

A. Uncertainty Measurement With Privileged Information

1) Privileged Information and Uncertainty: In the real world data, particularly Web images, auxiliary information such as text information is often approachable. LUPI is proposed in [19] as a new learning paradigm. LUPI assumes that additional features, namely privileged information, are contained in the training phase, but not in the test data. It is similar to the teacher in a class who offers extra explanations to students. On the Internet, people also tend to write some additional texts to facilitate the management of their multimedia repository, which usually includes images and videos. The text could provide more detailed descriptions for understanding the visual content in their repository. Several computer vision tasks like image classification can be benefited from the surrounding texts of web images. In this work, we show we can benefit from the additional text information associated with web images during the learning procedure.

However, the auxiliary text data usually cannot be applied in the image training procedure directly since it is in another feature space. To involve such privileged information when learning, in [19], the authors introduce slack functions into the formulation of a non-separable support vector machine (SVM) as follows:

$$\begin{aligned}
&\min_{\mathbf{w}, \tilde{\mathbf{w}}, b, \tilde{b}} \frac{1}{2}(||\mathbf{w}|| + \gamma||\tilde{\mathbf{w}}||) + C\sum_{i=1}^{n_s}\left(\tilde{\mathbf{w}}^\top \tilde{\phi}(\tilde{\mathbf{x}}_i) + \tilde{b}\right) \\
&s.t.\ \ y_i\left(\mathbf{w}^\top \phi(\mathbf{x}_i) + b\right) \geq 1 - \left(\tilde{\mathbf{w}}^\top \tilde{\phi}(\tilde{\mathbf{x}}_i) + \tilde{b}\right), i = 1, ..., n_s, \\
&\quad\ \left(\tilde{\mathbf{w}}^\top \tilde{\phi}(\tilde{\mathbf{x}}_i) + \tilde{b}\right) \geq 0, \quad i = 1, ..., n_s,
\end{aligned} \tag{1}$$

where $\phi(\mathbf{x}_i)$ is the feature mapping function for main information, and $\tilde{\phi}(\tilde{\mathbf{x}}_i)$ is the feature mapping function for privileged information. C is the trade-off parameter between data loss and model regularization, and γ is the trade-off parameter between the influence of main information and privileged information. $\mathbf{w}$ and $\tilde{\mathbf{w}}$ are the weight vectors for main features and privileged features respectively. The above problem is rather similar to a non-separable SVM problem, which can be formulated as

$$\begin{aligned} &\min_{\mathbf{w},b} = \frac{1}{2}||\mathbf{w}|| + C\sum_{i=1}^{n_s}\xi_i \\ &s.t.\ \ y_i\big(\mathbf{w}^\top \mathbf{z}_i + b\big) \geq 1 - \xi_i, i = 1, ..., n_s, \\ &\qquad \xi_i \geq 0, \quad i = 1, ..., n_s. \end{aligned} \tag{2}$$

The difference between LUPI and the conventional SVM comes with the slack variable ξ_i. In SVM, the slack variables can be optimized by the quadratic solver. In LUPI, the slack variables are replaced by the slack function $\hat{\xi}_i = (\tilde{\mathbf{w}}^\top \tilde{\phi}(\tilde{\mathbf{x}}_i) + \tilde{b})$. This slack function is defined for the correcting (text) feature instead of the main (visual) feature.

LUPI aims to determine the value of the slack function by leveraging the privileged information, rather than directly learning slack variables for the main feature. Although privileged information is not in the same space with the principal feature, it can still assist to obtain slack variables. If the learned slack function yields a larger value for the sample $\mathbf{x}_i$, then $\mathbf{x}_i$ is allowed to maintain a larger distance to the decision boundary. In other words, this implies that it would be more difficult for the learner to classify this sample correctly. Hence it is natural to measure the uncertainty of instances by the learned slack function. For example, suppose there are two samples m_1 and m_2 and $\hat{\xi}_1$ and $\hat{\xi}_2$ are their corresponding values returned by the learned slack function. If $\hat{\xi}_1 > \hat{\xi}_2$ then m_1 is likely more uncertain to classify correctly.

2) Cross-Media Uncertainty Measurement: Our method measures cross-media uncertainty by simultaneously learning from images and their surrounding texts. Next, we detail the strategies of cross-media uncertainty measurement.

It is natural to obtain two simple ways to measure the uncertainty. The first one is based on the prediction confidence associated with the predictor according to the visual feature, which is a traditional measurement, while the other one exploits the correcting function of text features.

1) Prediction confidence uncertainty measurement is based on the predictions of the visual feature, which is commonly used in many active sample selection algorithms. In a multiclass scenario, as mentioned, several strategies can be applied to generate the uncertainty. We focus on the margin sampling, which is a simple and effective sampling approach, and can be written as

$$p_i = 1/(\hat{y}_{i,k_1} - \hat{y}_{i,k_2} + \epsilon) \tag{3}$$

where $\hat{y}_{i,k_j} = \mathbf{w}_{k_j}^\top \phi(\mathbf{x}_i) + b_{k_j}$ is the prediction of the i-th unlabeled sample, p_i presents the corresponding uncertainty measurement, $\hat{y}_{i,k_1} \geq \hat{y}_{i,k_2} \geq ... \geq \hat{y}_{i,k_K}$ are the predicted values of the pool data $\mathbf{x}_i$ for K classes which are sorted in descending order, k_i presents the sorted index of classes, and $1 \leq i \leq n_p$. ϵ is a small constant that avoids division by zero. There are also other uncertainty sampling approaches. For instance, *least confidence* [5] strategy simply samples the data with least prediction confidence. However, this approach only considers the least confident class and ignores other classes. To exploit the remaining label distribution, margin sampling method considers more information about other class labels.

2) Correcting function uncertainty measurement is based on the optimized slack function of text features. As mentioned, the value of the slack function implies how difficult a sample may be correctly classified by the model. Hence, after training the LUPI model, we simply use the correcting function $p_i = \tilde{\xi}_i = (\tilde{\mathbf{w}}^\top \tilde{\phi}\tilde{\mathbf{x}}_i + \tilde{b})$ to measure the uncertainty of unlabeled data $\tilde{\mathbf{x}}_i$ in the privileged feature.

Suppose $\mathbf{p}_{\text{pred}} \in \mathbb{R}^{n_p}$ and $\mathbf{p}_{\text{corr}} \in \mathbb{R}^{n_p}$ denote the prediction confidence uncertainty measurement and correcting function uncertainty measurement on pool data. Based on these two simple uncertainty measurements we present five strategies to combine them together, which are listed as follows.

1) LUPI-sum: element-wise sum of two measurements. We use the sum of them as the uncertainty measurement: $\mathbf{p}_{\text{sum}} = \mathbf{p}_{\text{corr}} + \mathbf{p}_{\text{pred}}$.
2) LUPI-max: element-wise maximum of two measurements. We use the max of them as the uncertainty measurement: $\mathbf{p}_{\max} = \max(\mathbf{p}_{\text{corr}}, \mathbf{p}_{\text{pred}})$.
3) LUPI-min: element-wise minimum of two measurements. We use the min of them as the uncertainty measurement: $\mathbf{p}_{\min} = \min(\mathbf{p}_{\text{corr}}, \mathbf{p}_{\text{pred}})$.
4) LUPI-pro: Hadamard product of two measurements. We use the Hadamard product of them as the uncertainty measurement: $\mathbf{p}_{\text{had}} = \mathbf{p}_{\text{corr}} \odot \mathbf{p}_{\text{pred}}$, where $\odot$ denotes the Hadamard product.
5) LUPI-dis: element-wise distance of two measurements. We use the distance of them as the uncertainty measurement, namely $\mathbf{p}_{\text{sub}} = |\mathbf{p}_{\text{corr}} - \mathbf{p}_{\text{pred}}|$, where $|\cdot|$ computes the element-wise absolute values.

In this paper, we study all of the above five strategies.

B. Active Sampling With Uncertainty and Similarity Measurement

Typically, active sample selection aims to sample the most uncertain instances from the active pool set which most confuse the trained classifiers. In this way, the decision boundary can be refined and hopefully closer to optimal. However, as the number of initial labeled data only accounts for a small proportion of the entire data, there could be sample bias when selecting unlabeled samples from active pool set for labeling [16]. Thus, to achieve promising performance, researchers propose to combine more strategies together. For example, in [16], the authors consider both uncertainty and representativeness. In [27], multiple criteria are taken into consideration.

Similar to previous works, to obtain better performance, we consider to combine the uncertainty component and diversity component together in this paper. A very simple and straightforward way is to optimize an objective function that combines various components by trade-off parameters, which is similar to other works [16], [17], [27]. This objective function may be written as

$$\min Uncert + \lambda_1 Similar$$

where $Uncert$ and $Similar$ denote some measurement of uncertainty and similarity respectively. However, in the real world, it is not practical to tune the trade-off parameter λ_1 very well for all datasets. Therefore, in this paper, we propose a ratio objective function which computes ranking scores for all unlabeled data that maximizes the uncertainty, and meanwhile minimizes the similarity. This hence avoids tuning the trade-off parameter. Our proposed objective function is as follows:

$$\max_{\mathbf{r}} \frac{\mathbf{r}^\top \mathbf{p}}{\mathbf{r}^\top \mathbf{A}\mathbf{r}} \quad s.t. \sum_{i=1}^{n_p} r_i = 1,\ r_i \geq 0 \tag{4}$$

where $\mathbf{r} \in \mathbb{R}^{n_p}$ is the ranking score vector for samples in active pool set, $\mathbf{p} \in \mathbb{R}^{n_p}$ is the uncertainty measurement computed by one of the strategies in III-A2 and $\mathbf{A} \in \mathbb{R}^{n_p \times n_p}$ is the kernel matrix of the samples in active pool set, which represents the similarity among the unlabeled samples. Let $\lambda = \frac{\mathbf{r}^\top \mathbf{p}}{\mathbf{r}^\top \mathbf{A}\mathbf{r}}$ denotes the ratio of the uncertainty component to the similarity component. Here we use the radial basis function kernel. By this objective function, a higher ranking score tends to be assigned to a sample with higher uncertainty and less similarity with other data. In the next section, we propose an optimization scheme to solve the objective function efficiently.

When solving this objective function, it is not necessary to consider the entire pool set. In other words, after obtaining the uncertainty vector $\mathbf{p}$, we can find a specific portion of unlabeled data and input them into Problem 4. For instance, there are 1,000 unlabeled data in the pool set, while the target number of queries is 100. We may only use the most uncertain 200 instances from the pool set, instead of using the full uncertainty vector $\mathbf{p}$ and the entire kernel matrix $\mathbf{A}$. Since the size of the pool set is often huge in real world, this method is natural to speed up the optimization.

C. A Brief Overview of Augmented Lagrangian Method

We present an efficient optimization method based on Augmented Lagrangian method (ALM) [28] in the next subsection. ALM is to solve the following problem:

$$\min_T g(T) \quad s.t.\ \ h(T) = 0 \tag{5}$$

where $g : \mathbb{R}^t \rightarrow \mathbb{R}$, $h : \mathbb{R}^t \rightarrow \mathbb{R}^s$ and $T \in \mathbb{R}^t$ is the optimization variable. To solve the above constrained problem, one can construct an augmented Lagrangian function as

$$L(T, Z, \mu) = g(T) + \langle Z, h(T)\rangle + \frac{\mu}{2}||h(T)||_F^2 \tag{6}$$

where Z is the Lagrangian coefficient and μ is a scalar. The general approach to update T, Z and μ is briefed in Algorithm 1.

Algorithm 1: An overview of ALM algorithm [28].

Initialize $\rho > 1$, $t = 0$, $Z_t = 0$, and $\mu_t > 0$.
1: $T_t = \arg\min_{T, Z_t, \mu_t}$.
2: $Z_{t+1} = Z_t + \mu_t h(T_{t+1})$.
3: $\mu_{t+1} = \rho\mu_t$.
4: t = t + 1.
5: Go to step 1 until convergence.

Algorithm 2: ALM based optimization to solve the problem in (4).

Initialize r with random. Set $\lambda = \frac{\mathbf{r}^\top \mathbf{p}}{\mathbf{r}^\top \mathbf{A}\mathbf{r}}$. It is obvious that $0 \leq \lambda \leq \lambda^*$ where λ^* is the optima of λ.
1: Update $\mathbf{r}$ by solving problem 7
1.1 Set $\hat{\mathbf{A}} = 2\lambda\mathbf{A}$ and $\hat{\mathbf{p}} = -\mathbf{p}$.
1.2 Set $\tilde{\mathbf{A}} = \hat{\mathbf{A}} + \mu\mathbf{1}_{n_p}\mathbf{1}_{n_p}^\top + \mu\mathbf{I}_{n_p}$.
1.3 Set $\tilde{\mathbf{p}} = \mu\mathbf{v} + \mu\mathbf{1}_{n_p} - \hat{\mathbf{p}} - \delta\mathbf{1}_{n_p} - \gamma$.
1.4 Set $\mathbf{r}^* = \tilde{\mathbf{A}}^{-1}\tilde{\mathbf{p}}$.
2: Update λ by $\lambda = \frac{\mathbf{r}^\top \mathbf{p}}{\mathbf{r}^\top \mathbf{A}\mathbf{r}^\top}$.
3: Go to step 1 until convergence.

D. Optimization of Problem 4

The objective function is the ratio of the uncertainty to the similarity and it is not feasible to directly maximize the ratio. In this section, we propose an optimization approach to update this ratio, λ, iteratively, which is summarized in Algorithm 2. By this optimization approach, we can obtain ranking scores of unlabeled data that maximizes the uncertainty and minimizes the similarity meantime.

As illustrated in Algorithm 2, to start with, we exploit the ratio variable λ to rewrite Problem 4 as follow, which is a subproblem of our objective function:

$$\min_{\mathbf{r}} \lambda\mathbf{r}^\top \mathbf{A}\mathbf{r} - \mathbf{r}^\top \mathbf{p} \quad s.t. \sum_{i=1}^{n_p} r_i = 1,\ r_i \geq 0. \tag{7}$$

The constraints on $\mathbf{r}$ aims to limit the scale of the ranking scores. This objective function is now a general form that combines the uncertainty component and the diversity component by the variable λ. Problem 7 is a quadratic programming problem, but for the efficiency, we propose to use a faster optimization method based on augmented Lagrange multiplier (ALM) [28], [29], which is analogous to [17]. Introducing a new variable v, Problem 7 can be rewritten as

$$\min_{\mathbf{r}} \frac{1}{2}\mathbf{r}^\top \hat{\mathbf{A}}\mathbf{r} + \mathbf{r}^\top \hat{\mathbf{p}} \quad s.t.\ \ \mathbf{r}^\top \mathbf{1}_{n_p} = 1, \mathbf{r} = \mathbf{v}, \mathbf{v} \succeq 0 \tag{8}$$

where $\mathbf{1}_{n_p} \in \mathbb{R}^{n_p}$ is a vector with all elements of 1, $\hat{\mathbf{A}} = 2\lambda\mathbf{A}$ and $\hat{\mathbf{p}} = -\mathbf{p}$.

Then we can obtain the augmented Lagrangian function of Problem 8 as below

$$
\begin{aligned}
&L(\mathbf{r}, \mathbf{v}, \mu, \delta, \gamma) \\
&= \langle \delta, \mathbf{r}^\top \mathbf{1}_{n_p} - 1 \rangle + \langle \gamma, \mathbf{r} - \mathbf{v} \rangle \\
&\quad + \frac{\mu}{2} ||\mathbf{r}^\top \mathbf{1}_{n_p} - 1||_F^2 + \frac{\mu}{2} ||\mathbf{r} - \mathbf{v}||_F^2 + \frac{1}{2} \mathbf{r}^\top \hat{\mathbf{A}} \mathbf{r} + \mathbf{r}^\top \hat{\mathbf{p}} \\
&= \frac{\mu}{2} \left[||\mathbf{r} - \mathbf{v}||_F^2 + \frac{2}{\mu} \langle \gamma, \mathbf{r} - \mathbf{v} \rangle + \frac{\gamma^2}{\mu^2} \right] \frac{\mu}{2} \left[||\mathbf{r}^\top \mathbf{1}_{n_p}||_F^2 \right. \\
&\quad \left. + \frac{2}{\mu} \langle \delta, \mathbf{r}^\top \mathbf{1}_{n_p} - 1 \rangle + \frac{\delta^2}{\mu^2} \right] + \frac{1}{2} \mathbf{r}^\top \hat{\mathbf{A}} \mathbf{r} + \mathbf{r}^\top \hat{\mathbf{p}} + \frac{\gamma^2}{2\mu} \\
&\quad + \frac{\delta^2}{2\mu} = \frac{\mu}{2} (\mathbf{r}^\top \mathbf{1}_{n_p} - 1 + \frac{1}{\mu} \delta)^2 + \frac{\mu}{2} ||\mathbf{r} - \mathbf{v} + \frac{1}{\mu} \gamma||_F^2 \\
&\quad + \frac{1}{2} \mathbf{r}^\top \hat{\mathbf{A}} \mathbf{r} + \mathbf{r}^\top \hat{\mathbf{p}} + \frac{\gamma^2}{2\mu} + \frac{\delta^2}{2\mu}
\end{aligned} \tag{9}
$$

where $\mathbf{v} \succeq 0$, δ and γ are the Lagrangian coefficients, and μ is a scalar. According to [28], the Lagrangian function and the original problem have the same local minimization solution. Let

$$\tilde{\mathbf{A}} = \hat{\mathbf{A}} + \mu \mathbf{1}_{n_p} \mathbf{1}_{n_p}^\top + \mu \mathbf{I}_{n_p}$$

and

$$\tilde{\mathbf{p}} = \mu \mathbf{v} + \mu \mathbf{1}_{n_p} - \hat{\mathbf{p}} - \delta \mathbf{1}_{n_p} - \gamma$$

and then the Lagrangian function can be

$$
\begin{aligned}
L(\mathbf{r}, \mathbf{v}, \mu, \delta, \gamma) = \; & \frac{1}{2} \mathbf{r}^\top \tilde{\mathbf{A}} \mathbf{r} - \mathbf{r}^\top \tilde{\mathbf{p}} + \frac{\mu}{2} (-1 + \frac{\delta}{\mu})^2 \\
& + \frac{\mu}{2} || - \mathbf{v} + \frac{\gamma}{\mu} ||_F^2.
\end{aligned} \tag{10}
$$

By setting the derivative of $L(\mathbf{r}, \mathbf{v}, \mu, \delta, \gamma)$ w.r.t. $\mathbf{r}$ as zero, the objective function can be solved as follows:

$$\mathbf{r}^* = \tilde{\mathbf{A}}^{-1} \tilde{\mathbf{p}}. \tag{11}$$

After updating the ranking score vector $\mathbf{r}$, we need to update the auxiliary variable $\mathbf{v}$ by

$$\min_{\mathbf{v} \geq 0} L(\mathbf{r}, \mathbf{v}, \mu, \delta, \gamma) \Rightarrow \min_{\mathbf{v} \geq 0} ||\mathbf{v} - (\mathbf{r} + \frac{1}{\mu} \gamma)||^2. \tag{12}$$

Specifically, we can solve this problem by

$$v_i = \max(0, r_i + \frac{1}{\mu} \gamma_i), \quad 1 \leq i \leq n_p. \tag{13}$$

Once completing the ALM subproblem and obtaining the resultant ranking scores $\mathbf{r}^*$, we consider to update the ratio variable λ as follows:

$$\lambda = \frac{\mathbf{r}^{*\top} \mathbf{p}}{\mathbf{r}^{*\top} \mathbf{A} \mathbf{r}^*}. \tag{14}$$

We repeat the process for optimizing $\mathbf{r}$ and updating λ until convergence.

Theorem 1: Algorithm 2 increases monotonously in each iteration until the algorithm converge.

Proof: Let $\mathbf{r}_{k+1}$ be the updated $\mathbf{r}$ in the iteration $k+1$, $\mathbf{r}_k$ be the $\mathbf{r}$ computed in the iteration k and λ_k be the λ computed in the iteration k. Firstly, according to the step 2 in Algorithm 2, we have

$$\lambda_k \mathbf{r}_k^\top \mathbf{A} \mathbf{r}_k - \mathbf{r}_k^\top \mathbf{p} = 0.$$

Since $\mathbf{r}^*$ is the solution of Problem (7), it clear that $\lambda_k \mathbf{r}_{k+1}^\top \mathbf{A} \mathbf{r}_{k+1}^\top - \mathbf{r}_{k+1}^\top \mathbf{p} \leq \lambda_k \mathbf{r}_k^\top \mathbf{A} \mathbf{r}_k^\top - \mathbf{r}_k^\top \mathbf{p} = 0$. Therefore, we can easily obtain

$$\frac{\mathbf{r}_{k+1}^\top \mathbf{p}}{\mathbf{r}_{k+1}^\top \mathbf{A} \mathbf{r}_{k+1}} \geq \lambda_k = \frac{\mathbf{r}^\top \mathbf{p}}{\mathbf{r}^\top \mathbf{A} \mathbf{r}}.$$

■

Theorem 2: Let

$$g(\lambda) = \min_{\mathbf{r}^\top \mathbf{1}_{n_p}, \mathbf{r} \geq 0} \lambda \mathbf{r}^\top \mathbf{A} \mathbf{r} - \mathbf{r}^\top \mathbf{p}.$$

If $g(\lambda^*) = 0$, then λ^* is the global solution of problem 4.

Proof: From $g(\lambda^*) = 0$, we can easily obtain

$$\min_{\mathbf{r}^\top \mathbf{1}_{n_p}, \mathbf{r} \geq 0} \lambda^* \mathbf{r}^\top \mathbf{A} \mathbf{r} - \mathbf{r}^\top \mathbf{p} = 0.$$

Hence, for all $\mathbf{r}$

$$
\begin{aligned}
\lambda^* \mathbf{r}^\top \mathbf{A} \mathbf{r} - \mathbf{r}^\top \mathbf{p} \geq 0 \\
\Rightarrow \frac{\mathbf{r}^\top \mathbf{p}}{\mathbf{r}^\top \mathbf{A} \mathbf{r}} \leq \lambda^*.
\end{aligned} \tag{15}
$$

Thus λ^* is the global solution. ■

Theorem 3: Algorithm 2 can converge to the global solution.

Proof: Let λ^* and $\mathbf{r}^*$ are the values for λ and $\mathbf{r}$ when Algorithm 2 converges. According to the step 2 of Algorithm 2, we have

$$\lambda^* = \frac{\mathbf{r}^\top \mathbf{p}^*}{\mathbf{r}^* \mathbf{A} \mathbf{r}^*}.$$

Based on the step 1 of Algorithm 2, we have

$$\mathbf{r}^* = \arg \min_{\mathbf{r}^\top \mathbf{1}_{n_p}, \mathbf{r} \geq 0} \lambda^* \mathbf{r}^\top \mathbf{A} \mathbf{r} - \mathbf{r}^\top \mathbf{p}.$$

Thereby $g(\lambda^*) = 0$. According to Theorem 2, λ^* is the global solution. ■

E. Application to Large Datasets

We present a simple strategy to apply our proposed method to large scale datasets. Note that our goal is to query a small number of data from the large pool set (when n_p is large). Those unlabeled instances with high certainty would not be useful in active selection. As a result, we propose to only select a small portion of the unlabeled data as candidates for query. For example, if we aim to query n unlabeled data, we can only consider the top $6n$ uncertain instances in Problem (4) ($6n \ll n_p$). Therefore, our proposed method can be efficient enough to handle large scale datasets.

IV. Experiments

A. Datasets and Setups

In this section, we study the performance of the proposed active sample selection algorithm on the following four datasets which contains both image data and text data. We randomly sample 70% instances as the training set and use remaining 30%

instances as the test set. In the training set, we again randomly select 10% data as the seed set and 90% data as the pool set. Four public datasets are used in experiments.

WebQueries [1] contains 71,478 images with metadata of 353 queries collected from the Internet. The text data in the metadata files are captured up to 10 words before and after the corresponding image on the HTML web page. We remove the instances that do not contain any text data, select the 19 queries that contains more than 150 positive instances, and extract all the positive instances from these 19 queries as the new dataset, namely WebQueries19. Eventually we obtain a dataset with 19 classes and 3,323 instances and name it WebQueries19. The seed set of WebQueries19 consists of 12 instances for each categories. There are 2,099 and 996 instances in the pool set and test set, respectively.

Wikipedia articles [2] contains 2,866 Wikipedia articles including images and texts. We randomly sample 20 instances from each class into the seed set. Then we random select 1,807 instances as the pool set. The rest 860 instances are in the test set.

Pascal Sentences [30] contains 1,000 images with captions for 20 classes. There are 50 images for each class and 5 caption sentences for each image. We randomly select 8 images from every class into the seed set. The pool set consists of 620 images by random sampling. The rest 300 images are used as the test set.

MSCOCO[1] is a multi-label dataset and includes 82,783 samples in the training set, 40,504 samples in the validation set and 40,775 samples in the test set. Each instance contains an image and several caption sentences. Since the labels on the test set is not available, we use the training set and the validation set. We randomly select 15 classes that contains 3,000 to 5,000 instances from the entire dataset, and randomly sample 3,000 data from the selected subset, which is called as MSCOCO15. On MSCOCO15, there are around 240 positive labels for each class on average and the average number of labels for each instance is about 1.2. In addition, to investigate the efficiency of the proposed method and the competing baselines, we construct a larger binary dataset from MSCOCO by randomly selecting two disjoint classes, which contains 15,056 instances in total.

For the three multiclass datasets, namely WebQueries19, Wikipedia Articles and Pascal Sentences, accuracy is naturally used as the evaluation measurement. On the multi-label dataset, MSCOCO15, accuracy is not an appropriate evaluation. Hence we use average Macro F1 score and Micro F1 score over 10 runs as the evaluation.

As for the visual features, we extract CNNs features using the VGG16 model [31] and the Caffe package [32] to obtain the fc7 layer feature. As for the texts, we convert all the words into word vector using the public dictionary,[2] and then obtain the document-level features by the Bag-Of-Words (BoW). The dictionary size of the BoW model is set to 300.

We compare the proposed algorithm with other five baseline methods below.

[1] [Online]. Available: http://mscoco.org/dataset/
[2] [Online]. Available: https://code.google.com/p/word2vec/

1) pKNN [15] is a probabilistic variant of the K-Nearest-Neighbor method. We use radial basis function kernel (RBF kernel) as its input distance measurement.
2) LOD [14] is an unsupervised active sample selection method. We only compare it in Pascal Sentences dataset due to its high computational complexity.
3) Quire [16] queries the unlabeled data with combining uncertainty and representativeness. We use RBF kernel as the input for Quire. We turn it to a batch mode method by querying according to ranking scores for a fair comparison.
4) HSE [33] is a hierarchical subquery evaluation algorithm. The number of nearest neighbor is 10.
5) Aggressive Co-testing [18] is a Co-testing active learning algorithm which adopts an aggressive strategy.
6) Conservative Co-testing [18] is a Co-testing active learning algorithm which adopts a conservative strategy.
7) Random strategy selects unlabeled data randomly.
8) Initial results are calculated by the model trained on the initial labeled data.

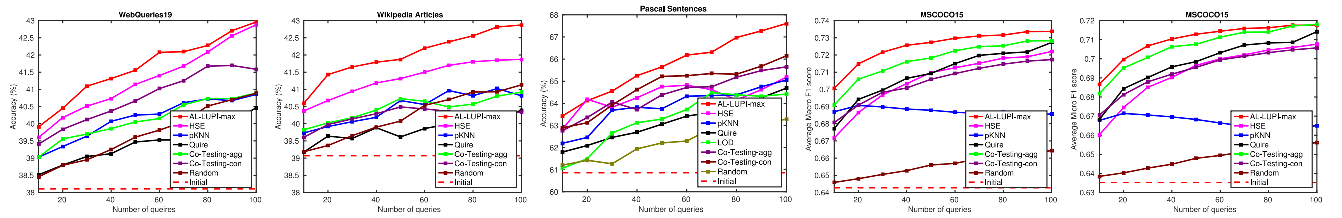
For all methods, we randomly sample the seed set, pool set and test set for 10 times and report the average results. For each run, all methods query 10,20,...,100 instances into the training set and linear SVMs are trained on the selected training set. The parameters in all methods are tuned from 10^{-5} to 10^{+5}. We select LUPI-max as the uncertainty measurement (denoted as AL-LUPI-max).

B. Comparison With Other Active Sampling Methods

Resulting accuracy of AL-LUPI-max and competing methods on WebQueries19, Pascal Sentences and Wikipedia Articles are illustrated in Fig. 1(a), Fig. 1(c) and Fig. 1(b). Resulting Macro F1 score and Micro F1 score of of AL-LUPI-max and competing methods on MSCOCO15 are reported in Fig. 1(d) and Fig. 1(e) respectively.

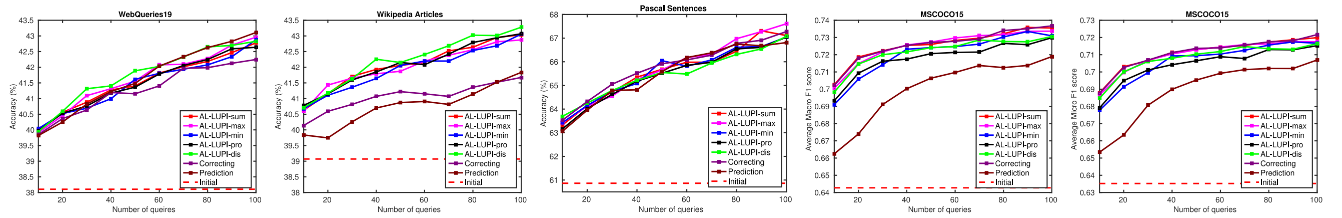
On the three multiclass datasets, namely WebQueries19, Pascal Sentences and Wikipedia Articles, we can observe that our proposed algorithm outperforms the competing methods from the results in Fig. 1. Note that pKNN, LOD, Quire and HSE are the state-of-the-art active sampling algorithms in the single-view cases. Unlike these single-view methods, our proposed method, aggressive Co-testing and conservative Co-testing are based on multiple features, and often obtain better performance. This performance improvement is likely from the benefit of additional textual features.

On the MSCOCO15, which is a multi-label dataset, our method also obtain the best performance compared to all other methods in terms of both macro-F1 score and micro F1 score. Aggressive Co-testing performs second best in this experiment. These results further demonstrate that it would be beneficial to exploit multiple categories of information for multimedia tasks. Note that conservative Co-testing performs much worse than aggressive Co-testing in MSCOCO15. This would imply that it is difficult to find a strategy that always outperforms other competitors on all datasets, which is similar to the summary on different multi-view active sample selection strategies in [18] However, as can be observed, AL-LUPI-max is more stable



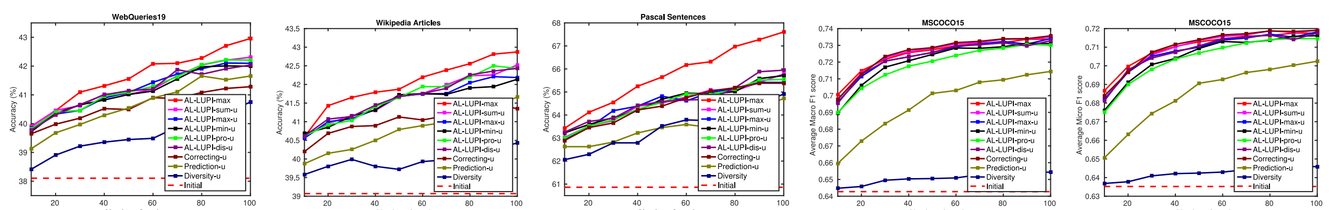
(a) WebQueries19 (b) Wikipedia Articles (c) Pascal Sentences (d) MSCOCO15 (e) MSCOCO15

Fig. 1. Average results on various datasets.



(a) WebQueries19 (b) Wikipedia Articles (c) Pascal Sentences (d) MSCOCO15 (e) MSCOCO15

Fig. 2. Average results on various datasets.



(a) WebQueries19 (b) Wikipedia Articles (c) Pascal Sentences (d) MSCOCO15 (e) MSCOCO15

Fig. 3. Comparison of the contribution of uncertainty and diversity.

than other algorithms on the four datasets, and achieves the best performer on all the four datasets.

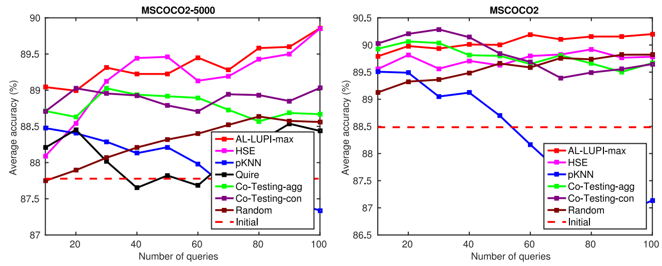
C. Comparison of Various Uncertainty Measurements

To investigate the performance of the proposed five strategies of the uncertainty measurement, we compare them on the four datasets. These results can be found in Fig. 2. We denote other four strategies, namely LUPI-sum, LUPI-min, LUPI-pro and LUPI-dis by AL-LUPI-sum, AL-LUPI-min, AL-LUPI-pro and AL-LUPI-dis. In general, we found that LUPI-max is stable on various datasets and achieves relatively better performance.

In addition, to investigate the contribution of privileged information in the active sampling task, we also compare two simple strategies of uncertainty measurement. The first strategy measures the uncertainty only according to the correcting function, while the other one measures the uncertainty only by the prediction confidence. We denote these two methods by "correcting" and "prediction" in Fig. 2. As observed, correcting function method achieves satisfactory performance on Pascal Sentences and MSCOCO15 and sometimes is competitive compared to LUPI-max. This demonstrates the effectiveness of privileged information in uncertainty measurement. However, on WebQueries and Wikipedia Articles, we observe that correcting function method performs much worse than LUPI-max. Therefore, it would be unstable if uncertainty measurement is only dependent on the privileged information. As for the prediction confidence strategy, it performs much worse than correcting function, particularly on Wikipedia Articles and MSCOCO15. Thus, we demonstrate that privileged information in the training procedure is very useful for the active sample selection task.

D. Contribution of Uncertainty and Diversity

In this section, we perform an experiment to investigate the contribution of uncertainty and diversity measurements. The proposed active sample selection algorithm samples unlabeled data based on two major measurements, namely uncertainty and diversity. To show the contribution of these two components respectively, we perform two baselines. For the first one, we drop the diversity measurement and sample unlabeled data only replying on the uncertainty measurement. Since we propose five various strategies, we denote those variants as "AL-LUPI-sum-u", "AL-LUPI-max-u", "AL-LUPI-min-u", "AL-LUPI-pro-u" and "AL-LUPI-dis-u". In addition, we examine two simple strategies which solely exploit the visual feature or text features respectively to measure uncertainty of unlabeled data. We denote them as "Prediction-u" and "Correction-u" respectively. As for the second baseline, we ignore the uncertainty measurement and only preserve diversity measurement. Then we sample unlabeled data directly according to the diversity ranking. We show these results in Fig. 3.



(a) Comparison on MSCOCO2-5000 (b) Comparison on MSCOCO2

Fig. 4. Comparison on MSCOCO2-5000 and MSCOCO2.

TABLE I
AVERAGE TRAINING TIME (IN SECONDS) OVER 10 RUNS

Dataset	AL-LUPI-max	HSE	pKNN	Quire
MSCOCO2-5000	17.5666	54.4269	0.2605	159.7137
MSCOCO2	18.3564	278.6959	0.2800	*

*Quire requires expensive computational cost, and does not finish a single run in 8 hours. Therefore, we did not include the training time in this table.

From the results, we observe that uncertainty measurement is more effective and beneficial for improving the performance of active sample selection, compared to diversity measurement. Even compared to AL-LUPI-max, the full version of our proposed method, uncertainty based baseline achieves relatively competitive performance. For example, on MSCOCO15, these baselines outperform AL-LUPI-max slightly in some cases. However, we note that it would be unreliable to depend on only one type of measurement in the active sampling procedure. On other three datasets, AL-LUPI-max outperforms all baselines significantly. Although diversity sampling is less effective than uncertainty sampling, our results demonstrate that by combining these two strategies, we achieve significant improvement. Therefore, we could conclude that the improvement of our proposed method comes from the combination of the uncertainty information and diversity measurements, which makes our algorithm robust and effective.

E. Comparison of Efficiency

We compare our method with other active learning baselines on MSCOCO2-5000 and MSCOCO2. We report the average accuracies of different number of actively selected samples over 10 runs in Fig. 4. We did not compare our method with LOD since it cannot finish a single run in 12 hours on both datasets. The corresponding average training time of active learning methods is listed in Table I. Note that active learning algorithm often relies on some auxiliary information, such as kernel matrices, graph or clustering. For instance, our method and pKNN requires kernel matrices on training data. Graph and cluster data are required in HSE. For the ease of comparison, Table I only contains the time spent on generating queries, instead of the time spent on the auxiliary information acquisition.

From Fig. 4, it is easy to observe that our method outperform other baselines in terms of average accuracy, even compared to the two multi-view baselines, namely Co-Testing-agg and Co-Testing-con. Table I shows that our algorithm is more efficient than HSE and Quire. Although pKNN is faster than our method, it cannot achieve satisfactory accuracy. We demonstrate our algorithm is efficient and can be applied to large dataset.

REFERENCES

[1] J. Krapac, M. Allan, J. Verbeek, and F. Jurie, “Improving web image search results using query-relative classifiers,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 1094–1101.
[2] N. Rasiwasia *et al.*, “A new approach to cross-modal multimedia retrieval,” in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 251–260.
[3] L. Duan, W. Li, I. W.-H. Tsang, and D. Xu, “Improving web image search by bag-based reranking,” *IEEE Trans. Image Process*, vol. 20, no. 11, pp. 3280–3290, Nov. 2011.
[4] W. Li, L. Niu, and D. Xu, “Exploiting privileged information from web data for image categorization,” in *Proc. Comput. Vis.*, 2014, pp. 437–452.
[5] B. Settles, “Active learning literature survey,” Univ. Wisconsin-Madison, Madison, WI, USA, Comput. Sci. Tech. Rep. 1648, Jan. 2010.
[6] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, “Selective sampling using the query by committee algorithm,” *Mach. Learn.*, vol. 28, nos. 2/3, pp. 133–168, 1997.
[7] S. Tong and D. Koller, “Support vector machine active learning with applications to text classification,” *J. Mach. Learn Res*, vol. 2, pp. 45–66, 2002.
[8] S. Dasgupta and D. Hsu, “Hierarchical sampling for active learning,” in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 208–215.
[9] Y. Yang, Y. -T. Zhuang, F. Wu, and Y. -H. Pan, “Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval,” *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 437–446, Apr. 2008.
[10] F. Wu *et al.*, “Sparse multi-modal hashing,” *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 427–439, Feb. 2014.
[11] S. Wang *et al.*, “Semi-supervised multiple feature analysis for action recognition,” *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 289–298, Feb. 2014.
[12] Y. Yang, Z. Ma, A. Hauptmann, and N. Sebe, “Feature selection for multimedia analysis by sharing information among multiple tasks,” *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 661–669, Apr. 2013.
[13] X. Chang, Z. Ma, Y. Yang, Z. Zeng, and A. G. Hauptmann, “Bi-level semantic representation analysis for multimedia event detection,” *IEEE Trans. Cybern.*, vol. PP, no. 99, pp. 1–18, 2016.
[14] X. He, W. Min, D. Cai, and K. Zhou, “Laplacian optimal design for image retrieval,” in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, 2007, pp. 119–126.
[15] P. Jain and A. Kapoor, “Active learning for large multi-class problems,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 762–769.
[16] S. Huang, R. Jin, and Z.-H. Zhou, “Active learning by querying informative and representative examples,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 1936–1949, Oct. 2014.
[17] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, “Multi-class active learning by uncertainty sampling with diversity maximization,” *Int. J. Comput. Vis.*, vol. 113, no. 2, pp. 113–127, 2015.
[18] I. Muslea, S. Minton, and C. A. Knoblock, “Active learning with multiple views,” *J. Artif. Intell. Res.*, vol. 27, pp. 203–233, 2006.
[19] V. Vapnik and A. Vashist, “A new learning paradigm: Learning using privileged information,” *Neural Netw.*, vol. 22, no. 5, pp. 544–557, 2009.
[20] M.-F. Balcan, A. Broder, and T. Zhang, “Margin based active learning,” in *Proc. Learn. Theory*, 2007, pp. 35–50.
[21] H. T. Nguyen and A. Smeulders, “Active learning using pre-clustering,” in *Proc. 21th Int. Conf. Mach. Learn.*, 2004.
[22] C. C. Aggarwal, *Data Classification: Algorithms and Applications*. Boca Raton, FL, USA: CRC Press, 2014.
[23] W. Wang and Z.-H. Zhou, “On multi-view active learning and the combination with semi-supervised learning,” in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 1152–1159.
[24] Y. Han *et al.*, “Image attribute adaptation,” *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1115–1126, Jun. 2014.
[25] L. Chen, D. Xu, I. W. Tsang, and J. Luo, “Tag-based image retrieval improved by augmented features and group-based refinement,” *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1057–1067, Aug. 2012.

[26] A. Lindner and S. Susstrunk, "Semantic-improved color imaging applications: It is all about context," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 700–710, May 2015.

[27] Y. Wu, I. Kozintsev, J.-Y. Bouguet, and C. Dulong, "Sampling strategies for active learning in personal photo retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2006, pp. 529–532.

[28] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA, USA: Athena Scientific, 1999.

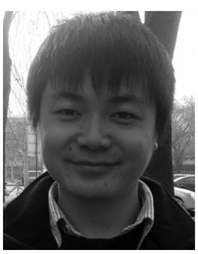
[29] F. Delbos and J. C. Gilbert, "Global linear convergence of an augmented lagrangian algorithm for solving convex quadratic optimization problems," *J. Convex Anal.*, vol. 12, pp. 45–69, 2005.

[30] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *Proc. NAACL HLT Workshop Creating Speech Lang. Data Amazon's Mech. Turk*, 2010, pp. 139–147.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

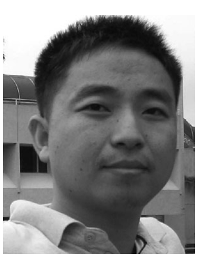
[32] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," *Proc. 22nd ACM Int. Conf. Multimedia*, pp. 675–678, 2014.

[33] O. Mac Aodha, N. D. Campbell, J. Kautz, and G. J. Brostow, "Hierarchical subquery evaluation for active learning on a graph," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 564–571.



Yan Yan received the B.E. degree in computer science from Tianjin University, Tianjin, China, in 2013, and is currently working toward the Ph.D. degree at the Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney, Ultimo, NSW, Australia.

His current research interest includes machine learning and computer vision.



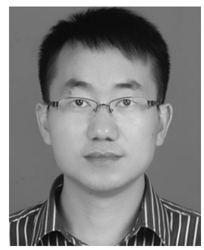
Feiping Nie received the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2009.

He is currently a Professor with the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Shaanxi, China. He has authored or coauthored more than 100 papers in the prestigious journals and conferences like the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the International Conference on Machine Learning, the Advance Neural Information Processing System, and the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. His research interests include machine learning and its applications fields, such as pattern recognition, data mining, computer vision, image processing, and information retrieval.

Prof. Nie is currently serving as an Associate Editor or PC Member for several journals and conferences.

Wen Li received the B.S. and M.Eng. degrees from the Beijing Normal University, Beijing, China, in 2010 and 2007, respectively, and the Ph.D. degree from the Nanyang Technological University, Singapore, in 2015.

He is a Postdoctoral Researcher with the Computer Visional Laboratory, ETH Zürich, Zürich, Switzerland. His main research interests include learning using privileged information, transfer learning, multiview learning, multiple kernel learning, and their applications in computer vision.



Chenqiang Gao received the B.S. degree in computer science from the China University of Geosciences, Wuhan, China, in 2004, and the Ph.D. degree in pattern recognition and intelligence systems from the Huazhong University of Science and Technology, Wuhan, China, in 2009.

He is currently a Professor with the Chongqing University of Posts and Telecommunications, Chongqing, China. His current research interests include image processing, infrared target detection, action recognition, and event detection.

Yi Yang received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2010.

He is currently an Associate Professor with the University of Technology Sydney, Ultimo, NSW, Australia. He was previously a Postdoctoral Researcher with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. His current research interest includes machine learning and its applications to multimedia content analysis and computer vision, such as multimedia indexing and retrieval, surveillance video analysis, and video content understanding.

Dong Xu (M'07–SM'13) received the B.Eng. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2001 and 2005, respectively.

He was with Microsoft Research Asia, Beijing, China, and the Chinese University of Hong Kong, Hong Kong, China, for more than two years. He was also a Postdoctoral Research Scientist with Columbia University, New York, NY, USA, from 2006 to 2007, and a Faculty Member with Nanyang Technological University, Singapore, from 2007 to 2015. He is currently a Professor and the Chair of Computer Engineering with the School of Electrical and Information Engineering, University of Sydney, Sydney, NSW, Australia. He has authored more than 100 papers in IEEE transactions and top-tier conferences.

Prof. Xu, as a coauthor, was a recipient of the Best Student Paper Award in the IEEE International Conference on Computer Vision and Pattern Recognition in 2010, as well as the IEEE TRANSACTIONS ON MULTIMEDIA Prize Paper Award in 2014.