Harvesting Large-Scale Weakly-Tagged Image Databases from the Web

Jianping Fan¹, Yi Shen¹, Ning Zhou¹, Yuli Gao²
¹Department of Computer Science, UNC-Charlotte, NC28223, USA
²Multimedia Interaction and Understanding, HP Labs, Palo Alto, CA94304, USA

Abstract

To leverage large-scale weakly-tagged images for computer vision tasks (such as object detection and scene recognition), a novel cross-modal tag cleansing and junk image filtering algorithm is developed for cleansing the weakly-tagged images and their social tags (i.e., removing irrelevant images and finding the most relevant tags for each image) by integrating both the visual similarity contexts between the images and the semantic similarity contexts between their tags. Our algorithm can address the issues of spams, polysemes and synonyms more effectively and determine the relevance between the images and their social tags more precisely, thus it can allow us to create large amounts of training images with more reliable labels by harvesting from large-scale weakly-tagged images, which can further be used to achieve more effective classifier training for many computer vision tasks.

1. Introduction

For many computer vision tasks, such as object detection and scene recognition, machine learning techniques are usually involved to learn the classifiers from a set of labeled training images [1]. The size of the labeled training images must be large-scale due to: (1) the number of object classes and scenes of interest could be very large; (2) the learning complexity for some object classes and scenes could be very high because of visual ambiguity; and (3) a small number of labeled training images are incomplete or insufficient to interpret the diverse visual properties of large amounts of unseen test images. However, hiring professionals to label large amounts of training images is cost-sensitive and poses a key limitation for the practical use of some advanced computer vision techniques. On the other hand, large-scale digital images and their associated text terms are available on the Internet, thus it is very attractive to leverage large-scale online images for computer vision tasks [2].

Some pioneering works have been done to leverage Internet images for computer vision tasks [2, 4-8]. Fergus et al. [4] and Li et al. [6] dealt with the precision problem by re-ranking the images which are downloaded from an image search engine. Recently, Schroff et al. [7] have developed a new algorithm for harvesting image databases from the web by combining text, meta-data and visual information. All these existing techniques have made a hidden assumption, e.g., image semantics have an explicit correspondence with the associated texts or nearby texts. Unfortunately, such an assumption may not always be true.

Collaborative image tagging system, such as Flickr [3], is now a popular way to obtain large set of labeled images easily by relying on the collaborative effort of a large population of Internet users. In a collaborative image tagging system, people can tag the images according to their social or cultural backgrounds, personal expertise and perception. We call the collaboratively-tagged images as weakly-tagged images because their social tags may not have exact correspondences with the underlying image semantics. With the exponential growth of the weakly-tagged images, it is very attractive to develop new algorithms that can leverage large-scale weakly-tagged images for computer vision tasks (such as learning the classifiers for object detection and scene recognition). Without controlling the word vocabulary, many text terms for image tagging may be synonyms or polysemes or even spams. The appearances of synonyms, polysemes and spams may either return incomplete sets of the relevant images or result in large amounts of ambiguous images or even junk images. Thus it is not a trivial task to leverage large-scale weakly-tagged images for computer vision tasks.

In this paper, we focus on collecting large-scale weakly-tagged images from collaborative image tagging systems such as Flickr by addressing the following crucial issues:

(a) **Synonymous Tags**: Different people may use different tags, which have the same or close meanings (synonyms), to tag their images. For example, *car, auto*, and *automobile* are a set of synonyms. The synonyms may result in incomplete returns of the relevant images in the image crawling process, and most tag clustering algorithms cannot incorporate the visual similarities between the relevant images to deal with the issue of synonyms more effectively.

(b) **Polysemous Tags**: Collaborative image tagging is an ambiguous process. Without controlling the vocabulary, different people may apply the same tag in different ways (i.e., the same tag may have different meanings under different contexts), which may result in large amounts of am-
biguous images. For example, the text term “bank” can be used to tag “bank office”, “river bank” and “cloud bank”. Word sense disambiguation is one potential solution for addressing this ambiguity issue, but it cannot incorporate the visual properties of the relevant images to deal with the issue of polysems more effectively [9-10].

(c) Spam Tags: Spam tags, which are used to drive traffic to certain images for fun or profit, are done by inserting the text terms that are more related to popular query terms rather than the text terms related to the actual image content. Spam tags are problematic because the junk images may mislead the underlying machine learning tools for classifier training. Junk image filtering is an interesting direction for dealing with the issue of spam tags, but it is worth noting that the scenario for junk image filtering in a collaborative image tagging space is significantly different.

In this paper, a novel cross-modal tag cleansing and junk image filtering algorithm is developed by integrating both the visual properties of the weakly-tagged images and their social tags to deal with the issues of spams, polysems and synonyms more effectively, so that we can create large amounts of training images with more reliable labels for computer vision tasks by harvesting from large-scale weakly-tagged images. The paper is organized as follows. In section 2, an automatic algorithm is introduced for image topic extraction. In section 3, a mixture-of-kernels algorithm is introduced for image similarity characterization. In section 4, a spam tag detection technique is introduced for junk image filtering. In section 5, a cross-modal tag cleansing algorithm is introduced for addressing the issues of synonyms and polysems. The algorithm evaluation results are given in section 6. We conclude this paper at section 7.

2. Image Topic Extraction

Each image in a collaborative tagging system is associated with the image holder’s taggings of the underlying image content and other users’ taggings or comments. It is worth noting that entity extraction can be done more effectively in a collaborative image tagging space. In this paper, we first focus on extracting the social tags which are strongly related to the most popular real-world objects and scenes or events. The social tags, which are related to image capture time and place, are also very attractive, but they are beyond the scope of this paper. Thus the image tags are first partitioned into two categories: noun phrases versus verb phrases. The noun phrases are further partitioned into two categories automatically: content-relevant tags (i.e., tags that are relevant to image objects and scenes) and content-irrelevant tags. The verb phrases are further partitioned into two categories automatically: event-relevant tags (i.e., tags that are relevant to image events) and event-irrelevant tags.

The occurrence frequency for each content-relevant tag and each event-relevant tag is counted automatically by using the number of relevant images. The misspelling tags may have low frequencies (i.e., different people may make different typing mistakes), thus it is easy for us to correct such the misspelling tags and their images are added into the relevant tags automatically. Two tags, which are used for tagging the same image, are considered to co-occur once without considering their order. A co-occurrence matrix is obtained by counting the frequencies of such pairwise tag co-occurrences.

The content-relevant tags and the event-relevant tags are further partitioned into two categories according to their interestingness scores: interesting tags and uninteresting tags. In this paper, multiple information sources have been exploited for determining the interesting tags more accurately. For a given tag $C$, its interestingness score $\omega(C)$ depends on: (1) its occurrence frequency $t(C)$ (e.g., higher occurrence frequency corresponds to higher interestingness score); and (2) its co-occurrence frequency $\vartheta(C)$ with any other tag in the vocabulary (e.g., higher co-occurrence frequency corresponds to higher interestingness score). The occurrence frequency $t(C)$ for a given tag $C$ is equal to the number of images that are tagged by the given tag $C$. The co-occurrence frequency $\vartheta(C)$ for the given tag $C$ is equal to the number of images that are tagged jointly by the given tag $C$ and any other tag in the vocabulary.

The interestingness score $\omega(C)$ for a given tag $C$ is defined as:

$$\omega(C) = \xi \log(t(C)) + \sqrt{\vartheta^2(C) + 1} + \zeta \log(\vartheta(C)) + \sqrt{\vartheta^2(C) + 1}$$

where $\xi$ and $\zeta$ are the weighting factors, $\xi + \zeta = 1$.

All the interesting tags, which have large values of $\Omega(\cdot)$ (i.e., top 5000 tags in our current experiments), are treated as image topics. In this work, only the interesting tags, which are used to interpret the most popular real-world object classes and scenes or events, are treated as the image topics. It is worth noting that one single weakly-tagged image may be assigned into multiple image topics when the relevant tags are used for tagging the image jointly. Collecting large-scale training images for the most popular real-world object classes and scenes or events and learning their classifiers more accurately are crucial for many computer vision tasks.

3. Image Similarity Characterization

To achieve more sufficient characterization of various visual properties of the images, both global and local visual features are extracted for image content representation. In our current experiments, the following visual features are extracted: (1) 36-bin RGB color histogram to characterize the global color distributions of the images; (2) 48-dimensional
texture features from Gabor filter banks to characterize the global visual properties (i.e., global structures) of the images; and (3) a number of interest points and their SIFT (scale invariant feature transform) features to characterize the local visual properties of the underlying salient image components.

By using high-dimensional visual features (color histogram, wavelet textures, and SIFT features) for image content representation, it is able for us to characterize various visual properties of the images more sufficiently. On the other hand, the statistical properties of the images in the high-dimensional feature space may be heterogeneous because different feature subsets are used to characterize different visual properties of the images, thus the statistical properties of the images in the high-dimensional feature space may be heterogeneous and sparse. Therefore, it is hard to use only one single type of kernel to characterize the diverse visual similarity contexts between the images precisely.

Based on these observations, the high-dimensional visual features are first partitioned into multiple feature subsets and each feature subset is used to characterize one certain type of visual properties of the images, thus the underlying visual similarity contexts between the images are more homogeneous and can be approximated more precisely by using one particular type of kernel. For each feature subset, a suitable base kernel is designed for image similarity characterization. Because different base image kernels may play different roles on characterizing the diverse visual similarity contexts between the images, the optimal kernel for diverse image similarity context characterization can be approximated more accurately by using a linear combination of these base image kernels with different importance.

For a given image topic \( C \) in the vocabulary, different base image kernels may play different roles on characterizing the diverse visual similarity relationships between the images. Thus the diverse visual similarity contexts between the images are characterized more precisely by using a mixture-of-kernels [13-14]:

\[
\kappa(x, y) = \sum_{l=1}^{\tau} \beta_l \kappa_l(x, y), \quad \sum_{l=1}^{\tau} \beta_l = 1 \tag{2}
\]

where \( \tau \) is the number of feature subsets (i.e., the number of base image kernels), \( \beta_l \geq 0 \) is the importance factor for the \( l \)th base image kernel \( \kappa_l(x, y) \). Combining multiple base kernels can allow us to achieve more precise characterization of the diverse visual similarity contexts between the weakly-tagged images.

4. Spam Tag Detection

Some popular image topics in the vocabulary may consist of large amounts of junk images because of spam tagging, and incorporating the junk images for classifier training may seriously mislead the underlying machine learning tools. Obviously, the junk images, which are induced by spam tagging, may make a significant difference on their visual properties with the relevant images. Thus the junk images can be filtered out effectively by performing visual-based image clustering and relevance analysis.

4.1 Image Clustering

A K-way min-max cut algorithm is developed to achieve more effective image clustering, where the cumulative inter-cluster visual similarity contexts are minimized while the cumulative intra-cluster visual similarity contexts (summation of pairwise image similarity contexts within a cluster) are maximized. These two criteria can be satisfied simultaneously with a simple K-way min-max cut function [11].

For a given image topic \( C \), a graph is first constructed for organizing all its weakly-tagged images according to their visual similarity contexts [11-12], where each node on the graph is one weakly-tagged image for the given image topic \( C \) and an edge between two nodes is used to characterize the visual similarity contexts between two weakly-tagged images, \( \kappa(\cdot, \cdot) \).

All the weakly-tagged images for the given image topic \( C \) are partitioned into \( K \) clusters automatically by minimizing the following objective function:

\[
\min \left\{ \Psi(C, K, \beta) = \sum_{i=1}^{K} \frac{s(G_i, G/G_i)}{s(G_i, G_i)} \right\} \tag{3}
\]

where \( G = \{ G_i | i = 1, \cdots, K \} \) is used to represent \( K \) image clusters, \( G/G_i \) is used to represent other \( K - 1 \) image clusters in \( G \) except \( G_i \), \( K \) is the total number of image clusters, \( \beta \) is the set of the optimal kernel weights. The cumulative inter-cluster visual similarity context \( s(G_i, G/G_i) \) is defined as:

\[
s(G_i, G/G_i) = \sum_{u \in G_i} \sum_{v \in G/G_i} \kappa(u, v) \tag{4}
\]

The cumulative intra-cluster visual similarity context \( s(G_i, G_i) \) is defined as:

\[
s(G_i, G_i) = \sum_{u \in G_i} \sum_{v \in G_i} \kappa(u, v) \tag{5}
\]

We further define \( X = [X_1, \cdots, X_l, \cdots, X_k] \) as the cluster indicators, and its component \( X_l \) is a binary indi-
solving multiple eigenvalue equations:

For the given image topic \( C \) by: components are defined as:

\[
X_l(u) = \begin{cases} 
1, & u \in G_l \\
0, & \text{otherwise}
\end{cases}
\]  

(6)

\( W \) is defined as an \( n \times n \) symmetrical matrix (i.e., \( n \) is the total number of web images), and its component is defined as:

\[ W_{u,v} = \kappa(u,v) \]  

(7)

\( D \) is defined as an \( n \times n \) diagonal matrix, and its diagonal components are defined as:

\[ D_{u,u} = \sum_{v=1}^{n} W_{u,v} \]  

(8)

For the given image topic \( C \), an optimal partition of its weakly-tagged images (i.e., image clustering) is achieved by:

\[
\min \left\{ \Psi(C, K, \beta) = \sum_{l=1}^{K} \left( \frac{1}{X_l^T(D - W)X_l} \right) \right\}
\]  

(9)

\[ X_l^T(D - W)X_l \]

Let \( \overline{W} = D^{-\frac{1}{2}}WD^{-\frac{1}{2}} \) and \( \overline{X}_l = \frac{D^{\frac{1}{2}}X_l}{\|D^{\frac{1}{2}}X_l\|} \), the objective function for our K-way min-max cut algorithm can further be refined as:

\[
\min \left\{ \Psi(C, K, \beta) = \sum_{l=1}^{K} \frac{1}{X_l^T\overline{W} \cdot \overline{X}_l} - K \right\}
\]  

(10)

subject to:

\[ \overline{X}_l^T \cdot \overline{X}_l = I, \quad \overline{X}_l^T \cdot \overline{W} \cdot \overline{X}_l > 0, \quad l \in [1, \cdots, K] \]

The optimal solution for Eq. (10) is finally achieved by solving multiple eigenvalue equations:

\[
\overline{W} \cdot \overline{X}_l = \lambda_l \overline{X}_l, \quad l \in [1, \cdots, K]
\]  

(11)

The objective function for kernel weight determination is to maximize the inter-cluster separability and the intra-cluster compactness. For one specific cluster \( G_l \), its inter-cluster separability \( \mu(G_l) \) and its intra-cluster compactness \( \sigma(G_l) \) are defined as:

\[ \mu(G_l) = X^T_l (D - W) X_l, \quad \sigma(G_l) = X^T_l W X_l \]  

(12)

For one specific cluster \( G_l \), we can refine its cumulative intra-cluster pairwise image similarity contexts \( s(G_l, G_l) \) as \( W(G_l) \):

\[ W(G_l) = \sum_{u \in G_l \cap \nu \in G_l} \kappa(u, v) = \sum_{i=1}^{\tau} \beta_i \omega_i(G_l) \]  

(13)

\[ D(G_l) - W(G_l) = \sum_{i=1}^{\tau} \beta_i [\epsilon_i(G_l) - \omega_i(G_l)] \]  

(14)

where \( \omega_i(G_l) \) and \( \epsilon_i(G_l) \) are defined as:

\[ \omega_i(G_l) = \sum_{u \in G_l \cap \nu \in G_l} \kappa_i(u, v), \quad \epsilon_i(G_l) = \sum_{v=1}^{n_l} \omega_v(G_l) \]  

(15)

The optimal weights \( \beta = [\beta_1, \cdots, \beta_{\tau}] \) for kernel combination are determined automatically by maximizing the inter-cluster separability and the intra-cluster compactness:

\[
\max_{\beta} \left\{ \frac{1}{K} \sum_{l=1}^{K} \frac{\sigma(G_l)}{\mu(G_l)} \right\}
\]  

(16)

subject to: \( \sum_{i=1}^{\tau} \beta_i = 1, \forall i : \beta_i \geq 0 \)

The optimal kernel weights \( \beta = [\beta_1, \cdots, \beta_{\tau}] \) are determined automatically by solving the following quadratic programming problem:

\[
\min_{\beta} \left\{ \frac{1}{2} \beta^T \left( \sum_{l=1}^{K} \Omega(G_l) \Omega(G_l)^T \right) \beta \right\}
\]  

(17)

subject to: \( \sum_{i=1}^{\tau} \beta_i = 1, \forall i : \beta_i \geq 0 \)
\[ \Omega(G_i) = \frac{\omega(G_i)}{\epsilon(G_i) - \omega(G_i)} \]  (18)

In summary, our K-way min-max cut algorithm takes the following steps iteratively for image clustering and kernel weight determination: (1) \( \beta \) is set equally for all these feature subsets at the first run of iterations. (2) Given the initial values of kernel weights, our K-way min-max cut algorithm is performed to partition the weakly-tagged images into \( K \) clusters according to their pairwise visual similarity contexts. (3) Given an initial partition of the weakly-tagged images, our kernel weight determination algorithm is performed to estimate more suitable kernel weights, so that more precise characterization of the diverse visual similarity contexts between the images can be achieved. (4) Go to step 2 and continue the loop iteratively until \( \beta \) is convergent. As shown in Fig. 1(a) and Fig. 2(a), our image clustering algorithm can achieve a good partition of large amounts of weakly-tagged images and determine their global distributions and inter-cluster correlations effectively. Unfortunately, such image clustering process cannot directly identify the clusters for the junk images.

### 4.2 Relevance Re-Ranking

For different users, their motivations for spam tagging are significantly different and their images for spam tagging should contain different content and have different visual properties. Thus the clusters for the junk images (which come from different users with different motivations) could be in small sizes. Based on this observation, it is reasonable for us to define the relevance score \( \rho(C, G_i) \) for a given image cluster \( G_i \) with the image topic \( C \) as:

\[ \rho(C, G_i) = \frac{\sum_{x \in G_i} P(x, C)}{\sum_{y \in C} P(y, C)} \]  (19)

where \( x \) and \( y \) are used to represent particular weakly-tagged images for the image topic \( C \), \( P(x, C) \) and \( P(y, C) \) are used to indicate the co-occurrence probabilities for the images \( x \) and \( y \) with the image topic \( C \).

In order to leverage the inter-cluster correlations for achieving more effective relevance re-ranking, a random walk process is performed for automatic relevance score refinement [15]. For a given image topic \( C \), our image clustering algorithm can automatically determine a cluster correlation network (i.e., \( K \) image clusters and their inter-cluster correlations) as shown in Fig. 1(a) and Fig. 2(a). We use \( \rho_l(G_i) \) to denote the relevance score for the \( l \)th image cluster \( G_i \) at the \( l \)th iteration. The relevance scores for all these \( K \) image clusters at the \( l \)th iteration will form a column vector \( \rho_l(G_i) \equiv [\rho_l(G_i)]_{K \times 1} \). We further define \( \Phi \) as an \( K \times K \) transition matrix, its element \( \phi_{G_i, G_j} \) is used to define the probability of the transition from the image cluster \( G_i \) to its inter-related image cluster \( G_j \). \( \phi_{G_i, G_j} \) is defined as:

\[ \phi_{G_i, G_j} = \frac{s(G_i, G_j)}{\sum_{G_h \in C} s(G_i, G_h)} \]  (20)

where \( s(G_i, G_j) \) is the inter-cluster visual similarity context between two image clusters \( G_i \) and \( G_j \) as defined in Eq. (4).

The random walk process is then formulated as:

\[ \rho_l(G_i) = \theta \sum_{j \in \Omega_j} \rho_{l-1}(G_j) \phi_{G_i, G_j} + (1 - \theta) \rho(C, G_i) \]  (21)

where \( \Omega_j \) is the first-order nearest neighbors of the image cluster \( G_j \) on the cluster correlation network, \( \rho(C, G_i) \) is the initial relevance score for the image cluster \( G_i \) and \( \theta \) is a weight parameter. This random walk process will promote the image clusters which have many connections on the cluster correlation network, e.g., the image clusters which have close visual properties (i.e., stronger visual similarity contexts) with other image clusters. On the other hand, this random walk process will also weaken the isolated image clusters on the cluster correlation network, e.g., the image clusters which have weak visual correlations with other image clusters. This random walk process is terminated when the relevance scores converge.
By performing random walk over the cluster correlation network, our relevance score refinement algorithm can re-rank the relevance between the image clusters and the image topic $C$ more precisely. Thus the top-k image clusters, which have higher relevance scores with the image topic, are selected as the most relevant image clusters for the given image topic $C$. Through integrating the cluster correlation network and random walk for relevance re-ranking, our spam tag detection algorithm can filter out the junk images effectively as shown in Fig. 1(b) and Fig. 2(b). By filtering out the junk images, we can automatically create large-scale training images with more reliable labels to learn more accurate classifiers for object detection and scene recognition.

5. Cross-Modal Tag Cleansing

The appearance of synonyms may result in insufficient image collections, which may prevent the underlying machine learning techniques from learning reliable classifiers for the synonymous image topics. On the other hand, the appearance of polysemes may result in the image sets with huge visual diversity, which may also prevent the underlying machine learning tools from learning precise classifiers for the polysemous image topics. To leverage large-scale weakly-tagged images for computer vision tasks, it is very attractive to develop cross-modal tag cleansing techniques for addressing the issues of synonyms and polysemes more effectively.

5.1 Combining Synonymous Topics

When people tag their images, they may use multiple text terms with similar meanings to tag their images alternatively. Thus the image tags are inter-related and such inter-related tags and their relevant images should be considered jointly. Based on this observation, a topic network is constructed automatically for characterizing such inter-tag (inter-topic) similarity contexts more precisely. Our topic network consists of two key components: (a) a large number of image topics; and (b) their cross-modal inter-topic correlations. The cross-modal inter-topic correlations consist of two components: (1) inter-topic co-occurrence correlations; and (2) inter-topic visual similarity contexts.

For two given image topics $C_i$ and $C_j$, their visual similarity context $\gamma(C_i, C_j)$ is defined as:

$$\gamma(C_i, C_j) = \frac{1}{2|C_i||C_j|} \sum_{u \in C_i} \sum_{v \in C_j} [\hat{\kappa}(u, v) + \bar{\kappa}(u, v)]$$

(22)

where $|C_i|$ and $|C_j|$ are the numbers of the weakly-tagged images for the image topics $C_i$ and $C_j$, $\hat{\kappa}(u, v)$ is the kernel-based visual similarity context between two weakly-tagged images $u$ and $v$ by using the kernel weights for the image topic $C_i$, and $\bar{\kappa}(u, v)$ is the kernel-based visual similarity context between two weakly-tagged images $u$ and $v$ by using the kernel weights for the image topic $C_j$.

The co-occurrence correlation $\beta(C_i, C_j)$ between two image topics $C_i$ and $C_j$ is defined as:

$$\beta(C_i, C_j) = -P(C_i, C_j) \log \frac{P(C_i, C_j)}{P(C_i) + P(C_j)}$$

(23)

where $P(C_i, C_j)$ is the co-occurrence probability for two image topics $C_i$ and $C_j$, $P(C_i)$ and $P(C_j)$ are the occurrence probability for the image topics $C_i$ and $C_j$.

The cross-modal inter-topic correlation between two image topics $C_i$ and $C_j$ is finally defined as:

$$\varphi(C_i, C_j) = \alpha \cdot \gamma(C_i, C_j) + (1 - \alpha) \cdot \beta(C_i, C_j)$$

(24)

where $\alpha$ is the weighting factor and it is determined through cross-validation. The topic network for our image collections is shown in Fig. 3, where each image topic is linked with multiple most relevant image topics with larger values of $\varphi(\cdot, \cdot)$.

Our K-way min-max cut algorithm is further performed on the topic network for topic clustering, thus the synonymous topics are grouped into the same cluster and can be combined as one super-topic. The images for these synonymous topics may share similar visual properties and semantics, thus they are combined and assigned to the super-topic automatically and a more comprehensive set of the relevant images can be obtained. Multiple tags for interpreting these synonymous topics are combined as one unified phrase for tagging the super-topic. Through combining the synonymous topics and their similar images, we can obtain more sufficient images to achieve more reliable learning of the classifier for the corresponding super-topic.

5.2 Splitting Polysemous Topics

Some image topics may be polysemous, which may result in large amounts of ambiguous images with diverse visual properties. Using the ambiguous images for classifier training may result in the classifiers with high variance and low generalization ability. To address the issue of polysemes, automatic image clustering is performed to split the polysemous topics by partitioning their ambiguous images into
multiple K-way min-max cut algorithm is used to partition
the ambiguous images under the same polysemous topic
into multiple clusters automatically and each group may cor-
respond to one certain sub-topic with more homogeneous
visual properties and smaller semantic gap.

To address the issue of the polysemous topics more ef-
effectively, WordNet is first incorporated to identify the can-
didates of the polysemous topics. For a given candidate of
the polysemous topics $P$, all its weakly-tagged images are
first partitioned into multiple clusters according to their vi-
sual similarity contexts by using our K-way min-max cut
algorithm. The visual diversity $\Omega(P)$ for the given can-
didate $P$ is defined as:

$$\Omega(P) = \sum_{G_i, G_j \in P} \left( \frac{\mu(G_i) - \mu(G_j)}{\sigma(G_i) + \sigma(G_j)} \right)^2$$

(25)

where $\mu(G_i)$ and $\mu(G_j)$ are the means of the image clusters
$G_i$ and $G_j$, $\sigma(G_i)$ and $\sigma(G_j)$ are the variances of the image clusters $G_i$ and $G_j$.

The candidates with large visual diversity between their
images are treated as the polysemous topics and are fur-
ther partitioned into multiple sub-topics. For a given poly-
semous topic, all its ambiguous images are partitioned into
multiple clusters automatically, and each cluster may cor-
respond to one certain sub-topic. By assigning the ambiguous
images for the polysemous topic into multiple sub-topics,
we can obtain multiple image sets with more homogeneous
visual properties, which may have better correspondences
between the tags (i.e., sub-topics) and the image semantics
(i.e., smaller semantic gaps). Through splitting the poly-
semous topics and their ambiguous images, we can obtain:
(a) multiple sub-topics with smaller semantic gaps and vi-
sual diversity; and (b) more precise image collections (with
smaller visual diversity) which can be used to achieve more
accurate learning of the classifiers for multiple sub-topics
with smaller semantic gaps.

## 6. Algorithm Evaluation

We have carried out our experimental studies by using
large-scale weakly-tagged Flickr images. We have down-
loaded more than 10 million Flickr images. Our algorithm
evaluation work focuses on evaluating how well our tech-
niques can address the issues of spams, polysemes and syn-
onyms. To evaluate the performance of our algorithms on
spam tag detection and cross-modal tag cleansing, we have
designed an interactive system for searching and exploring
large-scale collections of Flickr images. The benchmark
metric for algorithm evaluation includes precision $\rho$ and re-
call $\varrho$ for image retrieval. They are defined as:

$$\rho = \frac{\vartheta}{\vartheta + \xi}, \quad \varrho = \frac{\vartheta}{\vartheta + \nu}$$

(26)

where $\vartheta$ is the set of images that are relevant to the given
image topic and are returned correctly, $\xi$ is the set of images
that are irrelevant to the given image topic and are returned
incorrectly, and $\nu$ is the set of images that are relevant to the
given image but are not returned. In our experiments, only
top 200 images are used for calculating the precision and
recall rates.

The precision rate is used to characterize the accuracy of
our system for finding the particular images of interest, thus
it can be used to assess the effectiveness of our spam tag
detection algorithm. As shown in Fig. 4, one can observe
that our spam tag detection algorithm can filter out the junk
images effectively and result in higher precision rates for
image retrieval. On the other hand, the recall rate is used to
characterize the efficiency of our system for finding the par-
ticular images of interest, thus it can be used to assess the
effectiveness of our cross-modal tag cleansing algorithm on
addressing the issue of synonymous tags. As shown in Fig.
5, one can observe that our cross-modal tag cleansing algo-
rithm can combine the synonymous topics and their similar
images effectively and result in higher recall rates for image
retrieval.

To evaluate the effectiveness of our cross-modal tag
cleansing algorithm on dealing with the polysemous tags,
we have compared the performance differences on the pre-
cision rates before and after separating the polysmous tags
and their ambiguous images. Some results are shown in Fig.
6, one can obtain that our cross-modal tag cleansing algo-
rithm can tackle the issue of polysemous tags effectively.
By splitting the polysemous topics and their ambiguous im-
ages into multiple sub-topics, our system can achieve higher
precision rates for image retrieval.

We have also compared the precision and recall rates
between our system (i.e., which have provided techniques
to deal with the critical issues of spam tags, synonymous tags, and polysemous tags) and Flickr search system (which have not provided techniques to deal with the critical issues of spam tags, synonymous tags and polysemous tags). As shown in Fig. 7 and Fig. 8, one can observe that our system can achieve higher precision and recall rates for all these 5000 queries (i.e., 5000 tags of interest in our experiments) by addressing the critical issues of spams, synonyms and polysemes effectively.

7. Conclusions

The objective of this work is to create large amounts of training images with more reliable labels for computer vision tasks by harvesting from large-scale weakly-tagged images. A novel cross-modal tag cleansing and junk image filtering algorithm is developed by integrating both the visual similarity contexts between the images and the semantic similarity contexts between their tags for cleansing the weakly-tagged images and their social tags. Our experiments on large-scale collections of weakly-tagged Flickr images have provided very positive results. We will also release our image sets with more reliable labels on our web site.

References


