Extracting Multiple Visual Senses for Web Learning

Yazhou Yao, Fumin Shen, Jian Zhang, *Senior Member, IEEE*, Li Liu, Zhenmin Tang, and Ling Shao, *Senior Member, IEEE*

***Abstract*—Labeled image datasets have played a critical role in high-level image understanding. However, the process of manual labeling is both time consuming and labor intensive. To reduce the dependence on manually labeled data, there have been increasing research efforts on learning visual classifiers by directly exploiting web images. One issue that limits their performance is the problem of polysemy. Existing unsupervised approaches attempt to reduce the influence of visual polysemy by filtering out irrelevant images, but do not directly address polysemy. To this end, in this paper, we present a multimodal framework that solves the problem of polysemy by allowing sense-specific diversity in search results. Specifically, we first discover a list of possible semantic senses from untagged corpora to retrieve sense-specific images. Then, we merge visual similar semantic senses and prune noise by using the retrieved images. Finally, we train one visual classifier for each selected semantic sense and use the learned sense-specific classifiers to distinguish multiple visual senses. Extensive experiments on classifying images into sense-specific categories and reranking search results demonstrate the superiority of our proposed approach.**

***Index Terms*—Multiple visual senses, visual polysemy, polysemous words.**

I. Introduction

IN THE past few years, labeled image datasets have played a critical role in high-level image understanding [1]–[4]. For example, ImageNet [5] has acted as one of the most important factors in the recent advance of developing and deploying visual representation learning models (e.g., deep CNN [46]). However, the process of constructing ImageNet is both time-consuming and labor-intensive. To construct ImageNet, thousands of people have taken several years to complete [5].

To reduce the time and labor costs of manual annotation, some works also focused on active learning. For example, a method in [6] proposed to label some seed images to train the initial classifiers. Then these classifiers were used to do image categorization on other unlabeled images, to find low confidence images for manual labeling. The process was iterated until sufficient classification accuracy was achieved. In [7], a system for online learning of object detectors was proposed. This system refined its models by actively requesting annotations on images. However, active learning methods require pre-existing annotations, which often results in one of the most significant limitations to overcome the scalability.

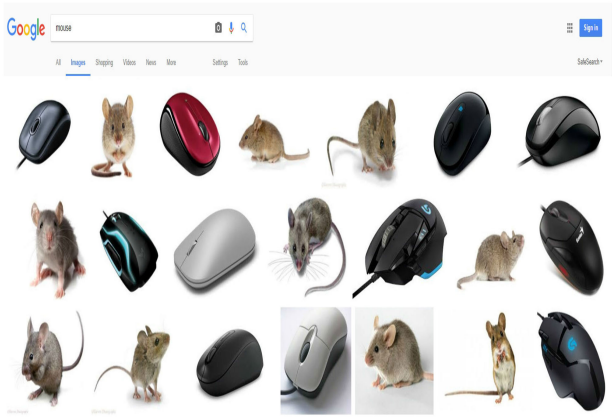


Fig. 1. Visual polysemy. For example, the query "mouse" returns multiple visual senses on the first page of results. The retrieved web images suffer from the low precision of any particular visual sense.

To further reduce the cost of manual annotation, learning directly from the web images has attracted more and more people's attention [8], [26], [38], [39]. Compared to manual-labeled image datasets, web images are a rich and free resource. For arbitrary categories, the potential training data can be easily obtained from the image search engines like Google or Bing. Unfortunately, due to the error index of image search engine, the precision of returned images from image search engine is still unsatisfactory. For example, Schroff *et al.* in [8] reported that the average precision of the top 1000 images for 18 categories from Google Image Search engine is only 32%. One of the most important reasons for the noisy results is the inherent ambiguity in the user query. As shown in Fig. 1, when we submit the query "mouse" into the Google Image Search engine, the returned results can refer to the animal "mouse", or the electronic product "mouse". The retrieved web images suffer from the low precision of any particular visual sense.

Visual polysemy means that a word has several semantic senses that are visually distinct. Some existing unsupervised

Manuscript received January 24, 2018; revised April 19, 2018 and May 27, 2018; accepted May 27, 2018. Date of publication June 15, 2018; date of current version December 20, 2018. This work was supported by the National Natural Science Foundation of China under Grant 61473154. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Raouf Hamzaoui. *(Corresponding author: Jian Zhang.)*

Y. Yao and J. Zhang are with the Global Big Data Technologies Center, University of Technology Sydney, Ultimo NSW 2007, Australia (e-mail: yazhou.yao@outlook.com; jian.zhang@uts.edu.au).

F. Shen is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: fumin.shen@gmail.com).

Z. Tang is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: Tzm.cs@njust.edu.cn).

L. Liu and L. Shao are with the Inception Institute of Artificial Intelligence, Abu Dhabi 999041, UAE (e-mail: liuli1213@gmail.com; ling.shao@ieee.org).

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.2018.2847248

approaches attempt to reduce the influence of visual polysemy by filtering out irrelevant images [8], [10], [18], [25], [33]. For example, one approach in [10] utilized the few top-ranked images returned from an image search engine to learn the initial classifier. The classifier refined its model through incremental learning strategy. With the increase in the number of positive images accepted by the classifier, the learned classifier would reach a robust level. The method in [18] leveraged the clustering based strategy to remove "group" noisy images and propagation based strategy to filter individual noisy images. These methods have the advantage of eliminating manual intervention. However, all of these methods do not directly address the problem of polysemy.

The traditional way to handle polysemy is text-based methods [16], [17]. However, all of these methods have no information about the visual senses and still need manual annotation to bridge the semantic and visual senses. Some works also leverage the human-developed knowledge such as Wikipedia [48] or dictionaries [13], [15] to handle polysemy. However, this human-developed knowledge still suffers from the problem of information missing [42]. For example, the machine-readable dictionary has a large coverage of NOUN category, but it contains very few entities (e.g., organizations, locations). Wikipedia can help to bridge this gap, but a great deal of information is still missing [28].

Since the semantic and visual senses of a given query are highly related, recent works also concentrated on jointly leveraging text and images [40], [44], [47]. Most of these methods assume that there exists a one-to-one mapping between semantic and visual sense towards to the given query. However, this assumption is not always true in practice. To deal with the multiple visual senses, Chen *et al.* in [28] adopt a one-to-many mapping between semantic and visual spaces. This approach can help us to find multiple visual senses from the web but overly depends on the collected web pages. If we can not collect web pages that contain multiple semantic and visual senses for the given query, the effect of this method will be significantly reduced.

Inspired by the situation described above, we seek to automate the process of discovering and distinguishing multiple visual senses for polysemous words. We propose an unsupervised method that resolves visual polysemy by allowing sense-specific diversity in search results. We take a three-step approach. Firstly, we discover a list of possible semantic senses through Google Books Ngram Corpus [36], to retrieve sense-specific images. Secondly, we merge visual similar semantic senses and prune noise by using the retrieved sense-specific images. Thirdly, we learn one visual classifier for each selected semantic sense and use the learned sense-specific classifiers to group and re-rank the polysemous images into its specific senses. To verify the effectiveness of our approach, we conduct experiments on the tasks of classifying images into sense-specific categories and re-ranking search results. The experimental results demonstrate the superiority of our proposed approach. The main contributions of this work can be summarized as follows:

1) We propose a novel approach for discovering and distinguishing multiple visual senses for polysemous words without explicit supervision.

2) Our work can be used as a pre-step before directly learning from the web, which helps to choose appropriate visual senses for sense-specific images collection, thereby improving the efficiency of learning from the web.

3) Our work shows substantial improvement over existing weakly supervised state-of-the-art methods.

This paper is an extended version of [50]. The extensions include: Comparing the sense-specific image classification ability of our approach with the human-developed knowledge-based methods; comparing the search results re-ranking ability of our approach with human-developed knowledge-based methods; analyzing the role of different steps contributing to the final results; and analyzing the parameter sensitivity and time complexity of our proposed approach.

The rest of the paper is organized as follows: In Section II, a brief discussion of related works is given. We propose our framework and associated algorithms in Section III. The experimental evaluations and discussions are presented in Section IV. Lastly, the conclusions are offered in Section V.

II. Related Works

Automatically discovering and distinguishing multiple visual senses for polysemous words is an extremely difficult problem. Several authors proposed to clean the retrieved images and learn visual classifiers, although none have specifically addressed the problem of polysemy [10], [25], [32], [33]. Fergus *et al.* [32] proposed the use of visual classifiers learned from Google Image Search engine to re-rank the images based on the visual consistency. Subsequent methods [10], [25] have employed similar removing mechanisms to automatically construct clean image datasets for training visual classifiers. Berg *et al.* [33] discovered topics using LDA in the text domain, and then use them to cluster the images. This approach requires manual intervention by the user to sort the topics into positive and negative for each category. However, these methods are category-independent and do not learn which words are predictive of a specific sense.

Our work is related to the text-based word senses discovering methods [16], [17]. Pantel *et al.* in [16] presented a clustering algorithm called Clustering By Committee (CBC) that automatically discovers word senses from text. It firstly discovers a set of tight clusters called committees that are well scattered in the similarity space. Then proceed by assigning words to their most similar clusters. It allows CBC to discover the less frequent senses of a word and to avoid discovering duplicate senses. Each cluster that a word belongs to represents one of its senses. A subsequent method in [17] has also employed similar Clustering by Committee algorithm to congregate similar words.

Our work is also related to the human-developed knowledge-based works [13]–[15], [48]. Yarowsky in [14] proposed to disambiguate word senses in unrestricted corpora using statistical models of the major Roget's Thesaurus categories. Roget's categories serve as approximations of conceptual classes. The categories listed for a word in Roget's index tend to correspond to sense distinctions; thus selecting the most likely category provides a useful level of sense disambiguation. The selection of categories is accomplished by identifying and weighing words

that are indicative of each category when seen in context, using a Bayesian theoretical framework. Then Yarowsky in [15] proposed an unsupervised word senses disambiguation method but relied on the use of dictionary definition as an initial seed. Mihalcea *et al.* in [48] and Veronis *et al.* in [13] proposed to use Wikipedia and dictionary for disambiguating word senses.

Our work is more related to the combination of text and images based methods. To discover multiple semantic and visual senses for polysemous words, previous works have also concentrated on clustering both of the text and image sources on the web [28], [40], [44], [47]. Method in [40] involves two major steps: (1) extracting and weighting text features from the web pages, visual features from the retrieved images, (2) running spectral clustering on both of the text features and visual features to derive the multiple semantic senses. Wan *et al.* in [47] and Saenko *et al.* in [44] proposed a latent model to learn multiple visual senses from a large collection of unlabeled web data, but rely on Wikipedia and WordNet's sense inventory respectively. Chen *et al.* [28] proposed a one-to-many mapping between the text-based feature space and image-based visual space to discover multiple semantic and visual senses of a Noun Phrase. However, clustering presents a scalability issue for this problem. The reason is that our images are sourced directly from the web and have no bounding boxes, every image creates millions of data points, the majority of which are outliers. In addition, this approach overly depends on the quality of the collected web pages, and the effect will be greatly reduced when we can not collect web pages that contain enough useful semantic and visual senses.

Our work is related to the following work. A visual concept learning system was recently proposed in [26] and achieved impressive performance for object detection. It discovers an exhaustive vocabulary explaining all the appearance variations from Google Books Corpora, and trains full-fledged detection models for it. The differences between us lie in three aspects. First, we have different goals. The purpose of [26] is to train a robust detection model while our approach aims to address the problem of polysemy in the process of learning from the web. Second, we adopt different approaches to filter out the noisy semantic senses. Third, we leveraged different strategies to purify the collected web images.

III. Framework and Methods

The inspiration for our work stems from the fact that web images indexed by a polysemous word are often rich in diversity. Our main idea of solving the problem of polysemy is allowing sense specific diversity in search results. Specifically, our proposed framework consists of three major steps: 1) discovering a list of possible semantic senses, to retrieve sense-specific images, 2) merging and pruning semantic senses, 3) distinguishing multiple visual senses for polysemous words.

A. Discovering Possible Semantic Senses

Inventories of manually compiled dictionaries (e.g., WordNet [41], ConceptNet [22]) usually serve as a source for word senses. However, they often include many rare senses while missing corpus/domain-specific senses. In addition, the process of constructing manually compiled dictionaries is time-consuming and labor-intensive. To ease the limitations of missing information, as well as to reduce the dependence on manually labeled data, Pantel *et al.* in [16] and Chatterjee *et al.* in [17] proposed to discover semantic senses from text via clustering. The disadvantage is that these methods overly depend on the quality of the collected text. The performance of these methods will be greatly reduced when we failed to collect enough useful text.

Inspired by recent works [26], [36], we can use untagged Google Books Ngram Corpus to discover an exhaustive vocabulary explaining all the appearance variations for the given query. Compared to manually labeled WordNet [41] and ConceptNet [22], it is not only much richer but also more general and exhaustive. Following [37, Sec. 4.3], we specifically use the dependency gram data with parts-of-speech (POS) for possible semantic senses discovering. For example, given a word (e.g., "mouse") and its corresponding POS tag (e.g., 'mighty, ADJ'), we find all its occurrences annotated with POS tag within the dependency gram data. Of all the ngram dependencies retrieved for the given word, we choose those whose modifiers are tagged as NOUN, VERB, ADJECTIVE, and ADVERB as the possible semantic senses. Our motivation is to find all the possible semantic senses the human race has ever written down in books. We use these discovered semantic senses to retrieve sense-specific web images from the image search engine.

B. Merging and Pruning Semantic Senses

As shown in Fig. 2, among the list of possible semantic senses, some of them are sharing visually similar distributions (e.g., "jerry mouse", "Minnie mouse" and "cartoon mouse"). To avoid training separate models for visually similar semantic senses, and to pool valuable training data across them, we need to merge and sample these visually similar semantic senses. In addition, not all the discovered semantic senses are useful, some noise may also be included (e.g., "figure mouse" and "flying mouse"). To avoid training meaningless visual models and to better distinguish multiple visual senses, we need to prune these noisy semantic senses.

1) Merging Visual Similar Semantic Senses: The traditional way to merge senses is calculating the semantic similarity of texts [42], [45]. These methods usually calculate the semantic similarity by calculating the frequency of their simultaneous appearance. Semantically similar senses usually have a smaller semantic distance. However, this assumption is not always true from the perspective of computer vision. For example, the semantic distance (Normalized Google Distance [45]) between "hot dog" and "dog" is relatively smaller (0.213). But visually speaking, they are two completely different objects that should not be merged. Different from previous works which merge semantic senses from the viewpoint of textual similarity, we propose to merge them from the viewpoint of visual consistency.

For each possible semantic sense, we use the top N images from image search engine to represent its visual distribution. We denote the visual similarity space of all discovered semantic

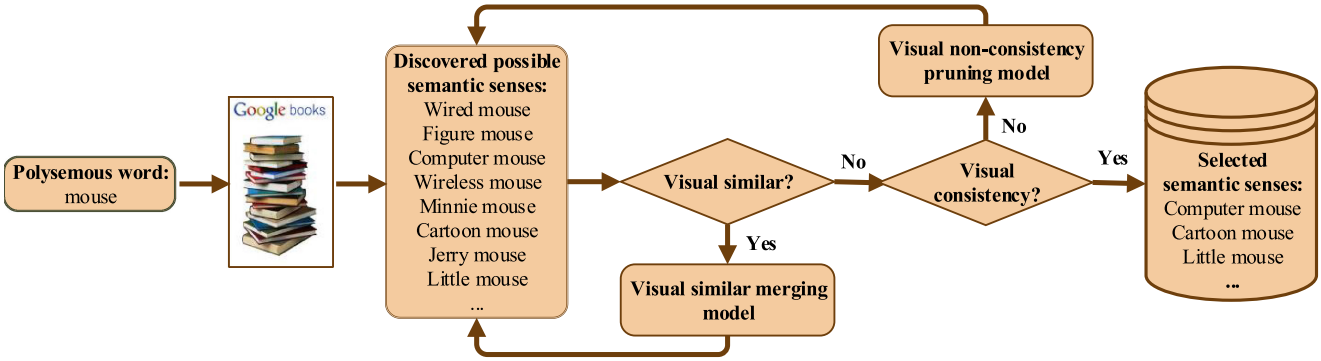


Fig. 2. Illustration of the process for obtaining selected semantic senses. The input is a textual word that we would like to find multiple visual senses for. The output is a set of selected semantic senses which will be used for distinguishing multiple visual senses.

senses by a graph $G = \{V, W\}$, where each node represents a semantic sense and each edge represents the visual similarity between two nodes. Each node has a score S_i which corresponds to the quality of its classifier. Specifically, we assume the top N images are positive instances, then these images were randomly split into a training set and validation set $I_i = \{I_i^t, I_i^v\}$. A random pool of negative images was collected and split into a training set and validation set $\overline{I} = \{\overline{I}^t, \overline{I}^v\}$. We learn the linear SVM classifier f_i with I_i^t and $\overline{I}^t$ using the 4096 dimensional deep features (based on AlexNet [46]). We then use $\{I_i^v, \overline{I}^v\}$ as validation images to calculate the classification results. We set the score S_i equal to the classification results on its own validation set $\{I_i^v, \overline{I}^v\}$. The edge weights $W_{i,j}$ correspond to the visual similarity between two nodes, and is measured by the score of the ith node classifier f_i on the jth node validation set $\{I_j^v, \overline{I}^v\}$.

Then the problem of merging visually similar semantic senses can be formulated as sampling a representative subset of space $v \subseteq V$ which maximizes the quality of the subset:

$$\max_{v} \sum_{i \in V} S_i \cdot \phi(i, v) \quad \text{s.t.} \quad |v| \leqslant k \tag{1}$$

where k is the number of semantic senses for the given word. ϕ is a soft coverage function that implicitly ensure the diversity of representative subset:

$$\phi(i, v) = \begin{cases} 1 & i \in v \\ 1 - \prod_{j \in v}(1 - W_{i,j}) & i \notin v \end{cases} \tag{2}$$

Similar to recent work [9], our formulation is to find a subset of representative space v which can cover the space of variance within the space V. Since our objective function is sub-modular, we can get a constant approximation of the optimal solution. We use an iterative mechanism for discovering the most representative subset. Particularly, we add one semantic sense i at each iteration by maximizing the current space:

$$\arg\max_{i} S(v \cup i) - S(v). \tag{3}$$

Fig. 3. A snapshot of the retrieved images for visual consistency and non-consistency semantic senses.

By setting the cost of adding semantic sense in v to a large value, each new semantic sense can be merged to its closest member in v.

2) Pruning Noisy Semantic Senses: After we merge the visual similar semantic senses, we set the rest as candidate semantic senses. Among these candidate semantic senses, some noise may also be included. To avoid training meaningless visual models and to better distinguish multiple visual senses, we prune these noisy semantic senses. As shown in Fig. 3, our basic idea is that noisy semantic senses have no specific visual patterns (e.g., "figure mouse", "flying mouse"). Thus, we can prune noise from the perspective of visual consistency.

We represent each discrete semantic sense as a "bag" and the retrieved images therein as "instances". In particular, we represent each semantic sense G_I with the compound feature $\delta_{f,k}$ of its top k positive images:

$$\delta_{f,k}(G_I) = \frac{1}{k} \sum_{x_i \in \Phi^*_{f,k}(G_I)} x_i \tag{4}$$

with

$$\Phi^*_{f,k}(G_I) = \underset{\Phi \subseteq G_I, |\Phi| = k}{\arg\max} \sum_{x_i \in \Phi} f(x_i). \tag{5}$$

The images in $\Phi^*_{f,k}(G_I)$ are referred to the top k positive instances of G_I according to the SVM classifier f_i (obtained in previous step). The closer of images to the center of the bag, the higher probability to be associated with the bag. The assignment of relatively heavier weights to these images would increase the accuracy of classifying semantic sense G_I to be positive or negative, then increase the efficiency of pruning noisy semantic senses. Following [27], the form of weighting function is assumed as

$$\rho_i = [1 + \exp(\alpha \log d(x_i) + \beta)]^{-1}. \tag{6}$$

$d(x_i)$ is the visual distance of image x_i to the bag center, $\alpha \in \mathbb{R}_{++}$ and β are scaling and offset parameters. Then the representation of (4) for semantic sense G_I can be represented as a weighted compound feature:

$$\delta_{f,k}(G_I) = \delta(X, h^*) = \frac{Xh^*}{\rho^\top h^*} \tag{7}$$

with

$$h^* = \underset{h \in \mathrm{H}}{\arg\max}\ f\left(\frac{Xh}{\rho^\top h}\right) \quad \text{s.t.} \sum_i h_i = k. \tag{8}$$

$X = [x_1, x_2, x_3.., x_i] \in \mathbb{R}^{D \times i}$ is a matrix whose columns are the instances of bag G_I, and $h^* \in \mathrm{H} = \{0,1\}^i \backslash \{0\}$ ($\sum_i h_i = k$) is an indicator function for the top k positive instances of bag G_I. $\rho = [\rho_1, \rho_2, \rho_3 \ldots \rho_i]^\top \in \mathbb{R}^i_{++}$ are the vectors of weights. Then the decision rule of semantic sense G_I to be selected or pruned is:

$$f_{\mathrm{w}}(X) = \max_{h \in \mathbf{H}} \mathrm{w}^\top \delta(X, h) \quad \sum_i h_i = k \tag{9}$$

where $\mathrm{w} \in \mathbb{R}^D$ is the vector of classifying coefficients, $\delta(X, h) \in \mathbb{R}^D$ is the feature vector of (7), h is a vector of latent variables and H is the hypothesis space $\{0,1\}^i \backslash \{0\}$. In order to solve the classifying rule of (9), we need to solve the below following problem:

$$\max_{h \in \mathbf{H}} \frac{\mathrm{w}^\top Xh}{\rho^T h} \quad \text{s.t.} \sum_i h_i = k. \tag{10}$$

This is an integer linear-fractional programming problem. Since $\rho \in \mathbb{R}^i_{++}$, (10) is identical to the relaxed problem:

$$\max_{h \in \lambda^i} \frac{\mathrm{w}^\top Xh}{\rho^\top h} \quad \text{s.t.} \sum_i h_i = k. \tag{11}$$

where $\lambda^i = [0,1]^i$ is a unit box in $\mathbb{R}^i$. (11) is a linear-fractional programming problem. We can reduce it to be a linear programming problem with $i+1$ variables and $i+2$ constraints [24]. Given a training set $\{G_I, Y_I\}_{I=1}^N$, the learning problem is to determine the parameter vector w in (9). This is a latent SVM problem:

$$\min_{\mathrm{w}} \frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{I=1}^{N} \max(0, 1 - Y_I f_{\mathrm{w}}(X_{G_I})). \tag{12}$$

The objective of (12) can be rewrited as two convex functions:

$$\min_{\mathrm{w}} \left[\frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{I \in D_N} \max(0, 1 + f_{\mathrm{w}}(X_{G_I})) + C\sum_{I \in D_P} \max(f_{\mathrm{w}}(X_{G_I}), 1)\right] - \left[C\sum_{I \in D_P} f_{\mathrm{w}}(X_{G_I})\right] \tag{13}$$

where D_P and D_N are positive and negative training sets respectively. Here we leverage the concave-convex procedure (CCCP) algorithm [49] to address (13). Finally, we obtain the pruning rule as (9) to remove noisy semantic senses which have no specific visual senses.

C. Distinguishing Visual Senses

After pruning the noisy semantic senses, we set the rest as the final selected semantic senses. As shown in Fig. 4, due to the error index of image search engine, even we retrieve the sense-specific images, some instance-level noise may also be included. The last step of our approach is to prune these instance-level noisy images and train visual classifiers for distinguishing multiple visual senses. Particularly, we train one optimal classifier for each semantic sense based on the selected images.

By treating each selected semantic sense as a "bag" and the retrieved images therein as "instances", we formulate noisy images pruning and classifiers learning as an instance-level multi-instance learning problem. Our objective is to select a subset of images from each bag to learn the optimal classifier for the selected semantic sense. As the accuracy of images retrieved from an image search engine is relatively high, we define each positive bag has a portion of δ positive instances.

Each instance was denoted as x_i with its label $y_i \in \{\pm 1\}$, where $i = 1, \ldots, n$. The label of each bag was denoted as $Y_I \in \{\pm 1\}$. The decision function is assumed in the form of $f(x) = \mathrm{w}^\top \varphi(x) + b$ and it will be used to prune instance-level noisy images. We apply the formulation of Lagrangian SVM. Then the decision function can be learned by minimizing the following structural risk functional:

$$\begin{aligned} \min_{\mathbf{y},\mathbf{W},b,\rho,\varepsilon_i} \ & \frac{1}{2}\left(\|\mathrm{w}\|^2 + b^2 + C\sum_{i=1}^{n} \varepsilon_i^2\right) - \rho \\ \text{s.t.} \ & y_i(\mathrm{w}^\top \varphi(x_i) + b) \geq \rho - \varepsilon_i, i = 1, \ldots n, \\ & y_i = -1 \quad \text{for} \quad Y_I = -1, \\ & \sum_{i:x_i \in G_I} \frac{y_i + 1}{2} \geq \delta |G_I| \quad \text{for} \quad Y_I = 1, \end{aligned} \tag{14}$$

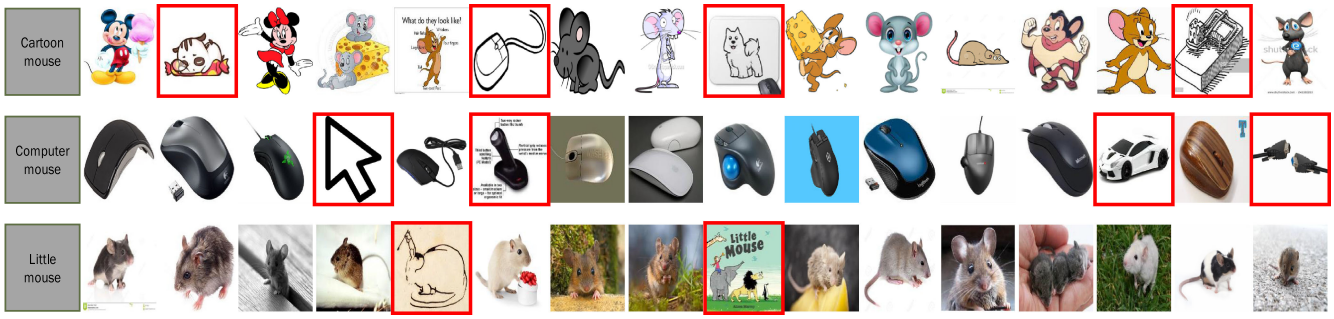


Fig. 4. A snapshot of the retrieved images for selected semantic senses. Due to the error index of image search engine, even we retrieve the sense-specific images, some instance-level noise may also be included. The noisy images are marked with red bounding boxes.

where φ is a mapping function that maps x from the original space into a high dimensional space $\varphi(x)$, $C > 0$ is a regularization parameter and ε_i values are slack variables. The margin separation is defined as $\rho/\|\mathrm{w}\|$. $\mathbf{y} = [y_1 \dots y_n]^\top$ means the vector of instance labels, $\lambda = \{\mathbf{y}|y_i \in \{\pm 1\}\}$ and $\mathbf{y}$ satisfies constraint in (14).

We employ the cutting-plane algorithm [43] to solve the optimization problem (14). Finally, we can derive the decision function for the selected semantic sense as:

$$f(x) = \sum_{i:\alpha_i \neq 0} \alpha_i \widetilde{y}_i \widetilde{k}(x, x_i) \tag{15}$$

where $\widetilde{y}_i = \sum_{t:\mathbf{y}^t \in \lambda} u_t y_i^t$ and $\widetilde{k}(x, x_i) = k(x, x_i) + 1$. The decision function will be used to prune instance-level noisy images in each selected semantic sense. In addition, it will also be leveraged to distinguish different visual senses.

IV. Experiments

To verify the effectiveness of our proposed approach, in this section, we first conduct experiments on the task of classifying images into sense-specific categories. Then we compare the search results re-ranking ability of our approach with baseline methods. In addition, we quantitatively analyze the role of different steps contributing to the final results. We also analyze the parameter sensitivity and time complexity of our proposed approach in this section.

A. Classifying Sense-Specific Images

The goal of this experiment is to compare the image sense-specific categorization ability of our proposed approach with two sets of baseline works.

1) Experimental Setting: We follow the setting of baseline methods [40], [47] and exploit web images as the training set, human-labeled images as the testing set. Instead of using co-clustering on web text and images, we use general corpus information and web images to discover and distinguish multiple visual senses for polysemous words. Particularly, we evaluate the performance on following datasets:

CMU-Poly-30 [28]: The CMU-Poly-30 dataset consists of 30 polysemy categories. Each category contains a varying number of images.

MIT-ISD [44]: The MIT-ISD dataset contains 5 categories. Each of which has three sizes. We are concerned with the "keyword" based size as it has the ground truth.

For each category, we first discover the possible semantic senses by searching in the Google Books Ngram Corpus. Then we retrieve the top $N = 100$ images from the Google Image Search engine for each discovered semantic sense. We assume the retrieved images as the positive instances (in spite of the fact that noisy images might be included). We randomly split the retrieved 100 images for each semantic sense into a training set and validation set $I_i = \{I_i^t = 50, I_i^v = 50\}$. We gather a random pool of negative images and split them into a training set and validation set $\overline{I} = \{\overline{I}^t = 50, \overline{I}^v = 50\}$. We train the SVM classifier f_i and calculate the score S_i using the validation set. The edge weights $W_{i,j}$ are obtained by calculating the score of the ith node classifier f_i on the jth node validation set $\{I_j^v, \overline{I}^v\}$. We merge the visually similar semantic senses and sample the representative subset of space by setting the cost to be 0.3.

To prune noisy semantic senses, we retrieve the top 500 images for each semantic sense. We then use the previously trained classifier f_i to select the most positive $k = 200$ images from the rest 450 images (the training data and testing data have no duplicates). We represent the selected semantic sense G_I with the compound feature $\delta_{f,k}$ of the most positive 200 images. There are multiple methods for learning the weighting function (e.g., cross-validation or logistic regression), we follow [27] and take cross-validation to learn the weighting function. To this end, we label $D_P = 500$ positive bags and $D_N = 500$ negative bags. Labeling work only needs to be done once to learn the weighting function and the bag classification rule (9). The learned classification rule (9) will also be used to prune noisy bags (corresponding to noisy semantic senses) which have no specific visual senses.

After pruning the noisy semantic senses, we set the rest as the final selected semantic senses. For each selected semantic sense, we collect the training data (500 images) from the image search engine. We take the MIL based method to handle instance-level noisy images and select the positive training data, to train the visual classifier. The negative training data is drawn from a "background" category, which in our case is the union of all other categories that we are asked to classify. The visual feature in our experiment is 4096 dimensional deep features (based on AlexNet [46]).

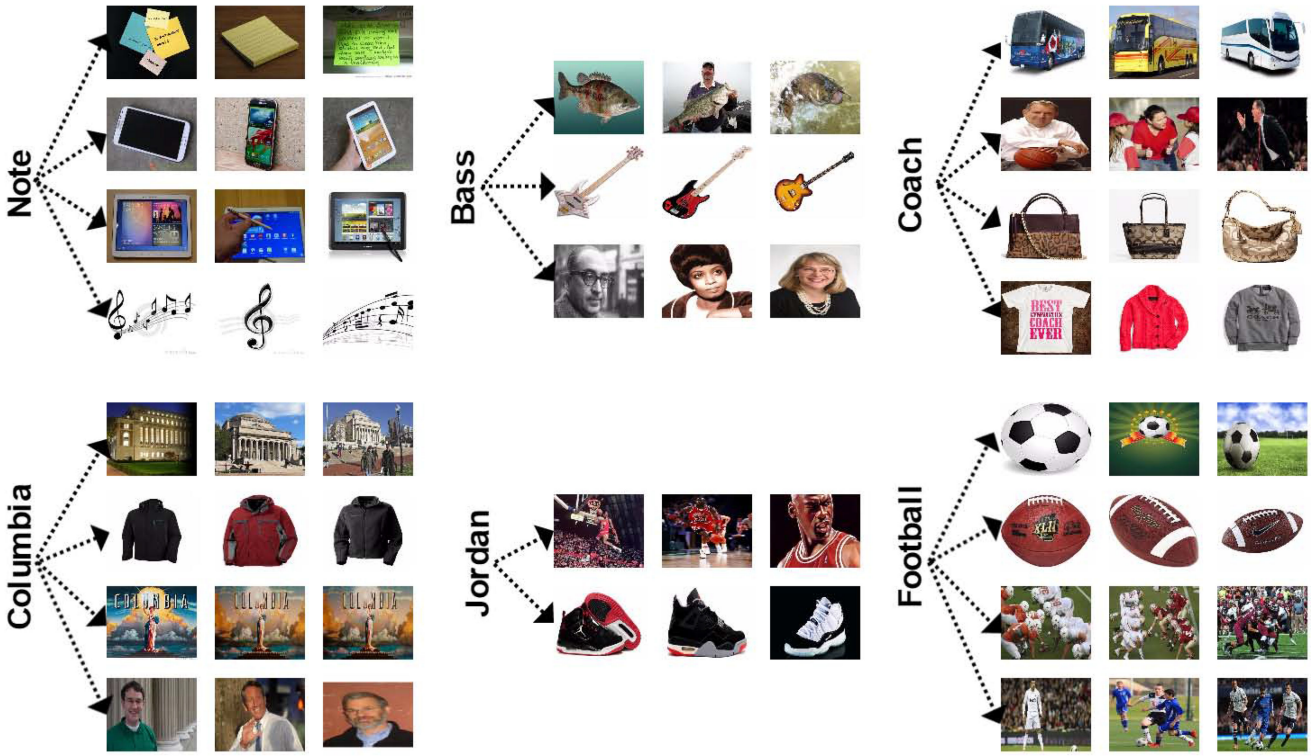


Fig. 5. Examples of multiple visual senses discovered by our proposed approach. For example, our approach automatically discovers and distinguishes four senses for "$Note$": notes, galaxy note, note tablet and music note. For "$Bass$", it discovers multiple visual senses of: bass fish, bass guitar and Mr./Mrs. Bass etc.

2) Baselines: To quantify the performance of our proposed approach, we compare the sense-specific image classification ability of our approach with two sets of baseline methods. For all the baseline methods, we adopt the same parameter configuration as described in their original works. Our baseline methods include:

Knowledge-based methods: The knowledge-based methods consist of Wikipedia method Wiki-MD [48], dictionary method Dict-MD [13] and corpora method Copr-MD [14]. For all of these three methods, we obtain the multiple semantic senses from human-developed knowledge. We directly retrieve the images from image search engine to learn the visual classifier for each semantic sense (without noisy images removing).

Combination of text and images based methods: This set of baselines include ISD [40], VSD [47], ULVSM [44], SDCIT [28] and LEAN [26]. The ISD [40] approach and SDCIT [28] approach involve two major steps: (1) extracting and weighting text features from the web pages, visual features from the retrieved images, (2) running spectral clustering or co-clustering mechanism on both of the text features and visual features to derive the multiple semantic senses. The VSD [47] approach and ULVSM [44] approach consist of three steps: (1) discovering multiple semantic senses and using the discovered semantic senses to retrieve images, (2) learning probabilistic models for discovered semantic senses, (3) using the probabilistic models to construct visual classifiers. The LEAN [26] approach contains three steps: (1) using Google Books Ngram Corpus to discover multiple semantic senses, (2) using the iterative mechanism to filter noisy semantic senses and images, (3) learning visual classifiers.

3) Experimental Results: Fig. 5 presents the examples of multiple visual senses discovered by our proposed approach on the CMU-Poly-30 dataset. Figs. 6 and 7 demonstrate the detailed performance comparison of classification accuracy on the CMU-Poly-30 and MIT-ISD dataset respectively. Table I shows the average performance comparison of classification accuracy on the CMU-Poly-30 and MIT-ISD dataset. It should be noted that the annotations used in the experiment are derived from the Google Corpus instead of the contents of the original web page.

From Figs. 6 and 7, we achieved the best results in 26 categories on the CMU-Poly-30 dataset. In the 5 categories of dataset MIT-ISD, we obtained the best results in all 5 categories. By observing Table I, the best average performance is achieved by our approach, which produces significant improvements over two sets of baseline methods. One possible explanation is that the automatically generated sense-specific terms by our approach could return relatively high-precision web images. Meanwhile, our proposed MIL model can effectively filter out the retrieved noisy images.

It is interesting to note in Fig. 5, our proposed approach not only discovers and distinguishes the sense of "notes" for "$Note$", but also "galaxy note", "note tablet" and "music note". For "$Bass$", in addition to "bass fish" and "bass guitar", our approach also discovers and distinguishes the sense of "Mr./Mrs. Bass". Compared to knowledge-based methods which discover possible semantic senses through Wikipedia or WordNet, our proposed approach that adopts untagged Google Books Ngram Corpus to discover possible semantic senses is much more exhaustive and general. Method ISD [40] and SDCIT [28] which

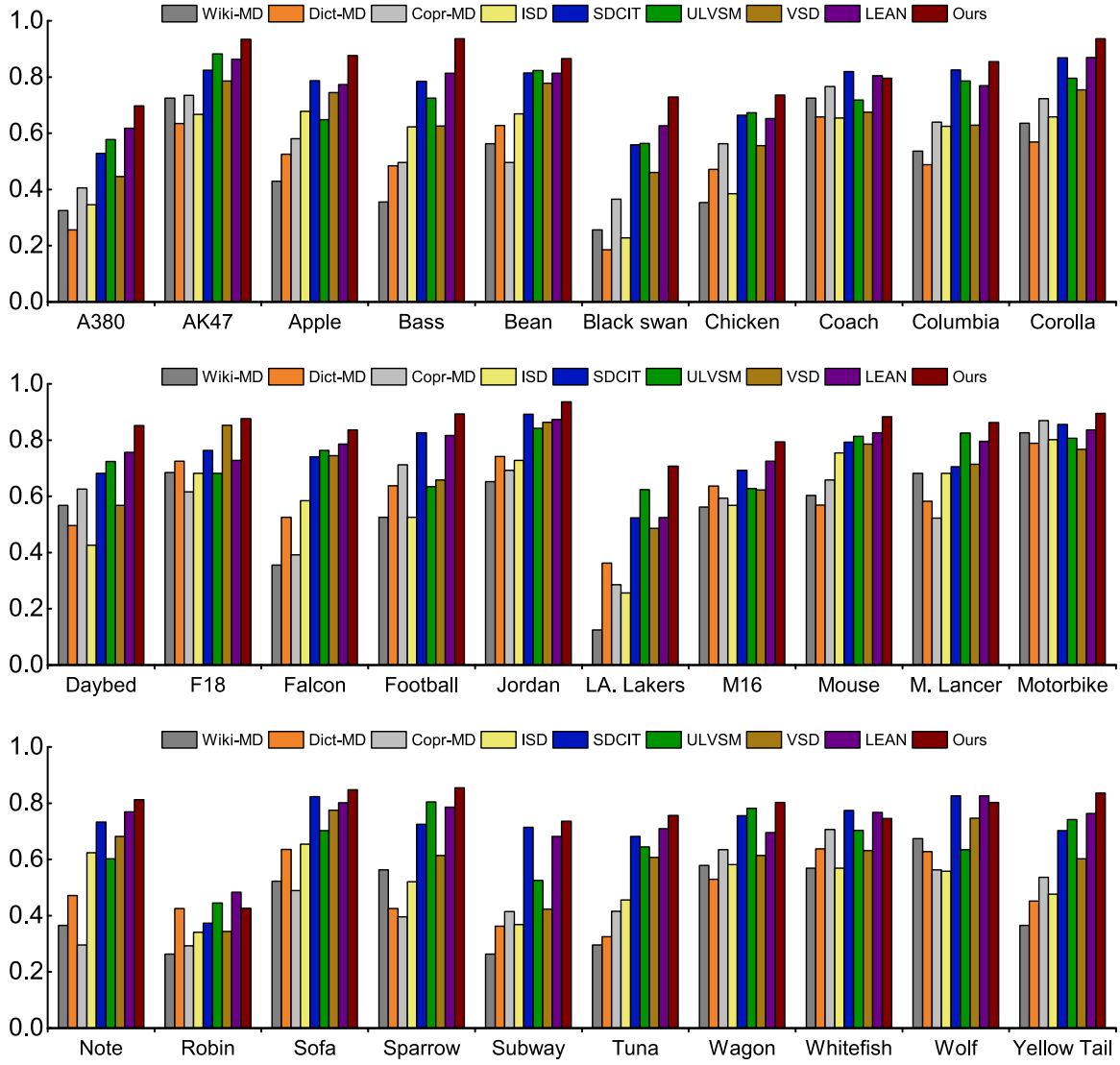


Fig. 6. The detailed performance comparison of classification accuracy over 30 categories on the CMU-Poly-30 dataset.

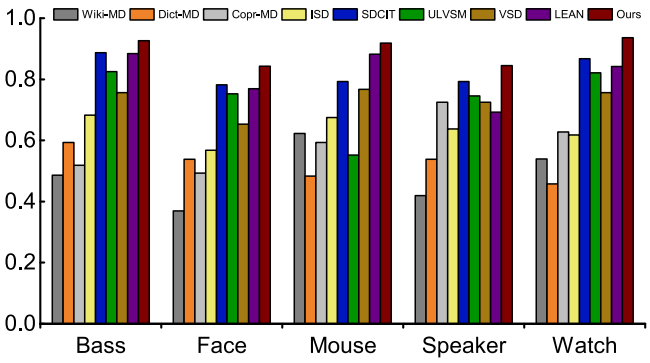


Fig. 7. The detailed performance comparison of classification accuracy over 5 categories on the MIT-ISD dataset.

TABLE I
THE AVERAGE PERFORMANCE COMPARISON OF CLASSIFICATION ACCURACY ON THE CMU-POLY-30 AND MIT-ISD DATASET

Method	Dataset	
	CMU-Poly-30	MIT-ISD
Wiki-MD [49]	0.498	0.487
Dict-MD [15]	0.529	0.522
Copr-MD [16]	0.549	0.593
ISD [41]	0.555	0.634
VSD [48]	0.728	0.786
ULVSM [45]	0.772	0.803
SDCIT [31]	0.839	0.853
LEAN [29]	0.827	0.814
Ours	**0.884**	**0.897**

uses webpages can discover multiple semantic senses but overly depends on the collected data. For example, method ISD [40] fails to collect webpages that contain enough semantic senses and visual senses for the given query, it can be seen that in Table I, the performance of this method is greatly reduced.

From Fig. 6, we found that all methods showed higher accuracy in both of the "AK47" and "Motorbike" categories. The explanation is perhaps that the visual patterns of polysemous words "AK47" and "Motorbike" are relatively simpler than other

TABLE II
WEB IMAGES FOR POLYSEMY TERMS WERE ANNOTATED MANUALLY

Query (#Annot. images)	Semantic senses	Visual senses	Numbers of images	Coverage
Bass (349)	1. **bass fish**	fish	159	45.6%
	2. **bass guitar**	musical instrument	154	44.1%
	3. Mr./ Mrs. Bass	people	20	5.7%
	Noise	unrelated	16	4.6%
Mouse (251)	1. **computer mouse**	electronic product	125	49.8%
	2. **little mouse**	animal	81	32.3%
	3. carton mouse	cartoon role	26	10.4%
	Noise	unrelated	19	7.5%

For each term, the number of annotated images, the semantic senses, the visual senses and their distributions are provided, with core semantic senses marked in boldface.

polysemous words. That is to say, the samples are densely distributed in the feature space, and the distribution of the training data and testing data overlaps much more easily.

B. Re-Ranking Search Results

The goal of this experiment is to compare the image search results re-ranking ability of our approach with two sets of baseline works.

1) Experimental Setting: We collect the top 500 images from Google Image Search engine for semantically ambiguous words: "bass" and "mouse". We perform a cleanup step for broken links, webpages, end up with 349 and 251 images for "bass" and "mouse" respectively. These images were annotated with one of the several semantic senses by one of the authors. The annotator tried to resist name influence, and make judgments based just on the image. For each query, 2 core semantic senses were distinguished from inspecting the data. The detailed information for these retrieved images is summarized in Table II.

We now evaluate how well the two sets of baseline methods and our method can re-rank the retrieved images. For each query, the sense-specific classifiers are trained on the sense-specific web images. Particularly, we use the previously trained sense-specific classifiers in the previous experiment. Retrieved images are then re-ranked by moving the negatively-classified images down to the last rank. For an image d, we compute the probability $P(S_i|d)$ of image d belonging to the ith sense S_i and rank the corresponding images according to the probability of each sense S. $P(S_i|d)$ provides a way to re-rank the images in the original polysemous order. Images belonging to some sibling sense are given lower probabilities and pushed to the back of the rank list.

2) Baselines: We compare the search results re-ranking ability of our approach with two sets of baseline methods which include knowledge-based methods and the combination of text and images based methods. The knowledge-based methods consist of Wiki-MD [48], Dict-MD [13] and Copr-MD [14]. The combination of text and images based methods contain ISD [40], VSD [47], ULVSM [44], LEAN [26], and SDCIT [28].

3) Experimental Results: Following [47], we evaluate the re-ranking performance by computing the Area Under Curve (AUC) of all senses for "bass" and "mouse". The results are shown in Table III.

From Table II, we observe that there are only 4.6% and 7.5% true noise in the retrieved images for "bass" and "mouse" respectively. Most of the retrieved images are different forms of visual senses for the given query. This indicates that we should first discover the multiple visual senses for the given query. So that we can choose appropriate visual senses as needed to carry out sense-specific images collection. By doing this, we can greatly improve the efficiency of collecting web images, thereby improving the efficiency of learning from the web images.

We observe that the combination of text and images based methods ISD [40], VSD [47], ULVSM [44], SDCIT [28], LEAN [26] and our method are generally better than knowledge-based methods Wiki-MD [48], Dict-MD [13] and Copr-MD [14] in Table III. In specific, methods SDCIT [28], LEAN [26] and our method achieve better results than other methods. The explanation is that it is necessary to remove noisy images from the training set during the process of classifier learning. Learning directly from the web images without noise removing may affect the performance of the classifier due to the presence of noisy images.

By observing Table III, we achieve the best average performance which is consistent with the results of sense-specific image classification. The reason can be explained by the generated sense-specific terms and filtered images of our approach. Compared to knowledge-based methods Wiki-MD [48], Dict-MD [13] and Copr-MD [14], our approach does not directly use web images for classifier learning. Instead, we filter the retrieved images to select useful data and then use the selected images to learn classifiers. By doing this, our approach can effectively overcome the impact of noise on the classifiers due to the error index of image search engine. Compared to the combination of text and images based methods ISD [40], VSD [47], ULVSM [44], LEAN [26] and SDCIT [28], the sense-specific terms generated by our approach are more accurate and exhaustive, using our sense-specific terms to retrieve images can return high precision web images, thereby can help us to train sense-specific classifiers to re-rank the search results.

TABLE III
AREA UNDER CURVE (AUC) OF ALL SENSES FOR "BASS" AND "MOUSE"

Method	Semantic senses						Average
	bass fish	bass guitar	M. Bass	Computer mouse	little mouse	carton mouse	
Wiki-MD [49]	0.364	0.429	0.132	0.536	0.623	0.114	0.366
Dict-MD [15]	0.443	0.635	0.205	0.464	0.573	0.186	0.418
Copr-MD [16]	0.504	0.486	0.305	0.624	0.675	0.263	0.476
ISD [41]	0.453	0.526	0.243	0.614	0.536	0.218	0.432
VSD [48]	0.547	0.538	0.239	0.684	0.652	0.226	0.481
ULVSM [45]	0.526	0.615	0.326	0.732	0.735	0.314	0.541
LEAN [29]	0.623	0.658	0.413	0.753	0.785	0.336	0.595
SDCIT [31]	0.658	**0.773**	0.386	0.815	0.845	0.337	0.636
Ours	**0.713**	0.736	**0.572**	**0.834**	**0.873**	**0.434**	**0.694**

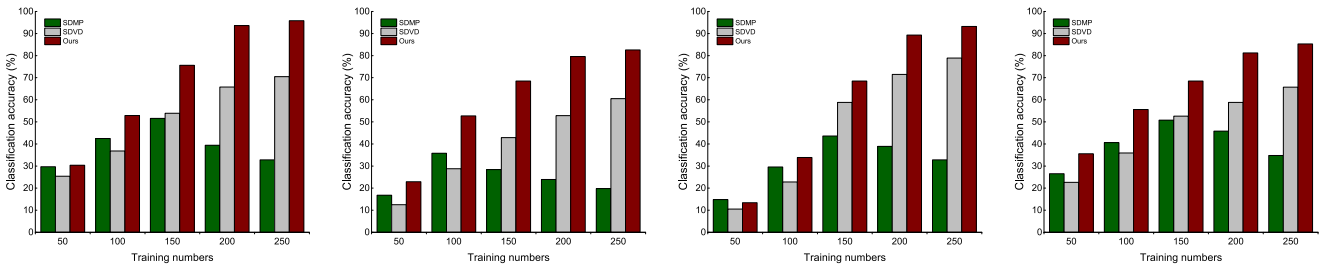


Fig. 8. Sense-specific image categorization ability of SDMP, SDVD and ours on CMU-Poly-30 dataset: (a) "bass", (b) "coach", (c) "football" and (d) "note".

C. Quantitative Analysis of Different Steps

Our proposed approach consists of three major steps: possible semantic senses discovering, semantic senses merging and pruning, and visual senses distinguishing. To quantify the role of different steps contributing to the final classifiers, we construct two new frameworks.

One is based on possible semantic senses discovering and semantic senses merging and pruning (which we refer to SDMP). Another one is based on semantic senses discovering and visual senses distinguishing (which we refer to SDVD). For framework SDMP, we first obtain the possible semantic senses through searching in the Google Books Ngram Corpus. Then we apply the semantic senses merging and pruning procedure to obtain the selected semantic senses. We directly retrieve the top images from the image search engine for selected semantic senses to train image classifiers (without noisy images removing). For framework SDVD, we also obtain the possible semantic senses by searching in the Google Books Ngram Corpus. Then we retrieve the top images from the image search engine for all the candidate semantic senses (without semantic senses merging and pruning procedure). We apply the MIL model to select images and train image classifiers.

We compare the sense-specific image categorization ability of these two new frameworks with our proposed framework. Specifically, "note", "bass", "coach" and "football" are selected as four target categories to compare the sense-specific image categorization ability. We sequentially collect [50, 100, 150, 200, 250] images for each selected semantic sense as the positive training samples and use 500 fixed irrelevant negative samples to learn image classifiers. We test the sense-specific image categorization ability of these three frameworks on the CMU-Poly-30 dataset. The results are shown in Fig. 8. From Fig. 8, we can observe:

Framework SDMP usually performs better than SDVD when the training number for each semantic sense is below 150. The explanation is that the first few retrieved images tend to have a relatively high accuracy. When the number of training images is below 150, the noisy images caused by noisy semantic senses are more serious than those induced by the image search engine. With the increase of image numbers for each semantic sense, the images retrieved from the image search engine contain more and more noise. In this condition, the noisy images caused by the image search engine have a worse effect than those induced by noisy semantic senses.

Our proposed framework outperforms both SDMP and SDVD. The reason is our approach, which takes a combination of noisy semantic senses and noisy images removing, can effectively remove the noise caused by both of the noisy semantic senses and the error index of image search engine.

D. Parameter Sensitivity Analysis

For the parameter sensitivity analysis, in the revision, we present the interaction between pairs of parameters. Specifically, we mainly concern the interaction between labeled positive and negative bags (D_P and D_N) in the process of noisy semantic senses purifying, the interaction between two parameters C and δ in our MIL model. In particular, we vary one parameter by fixing other parameters as a different value. MIT-ISD is selected

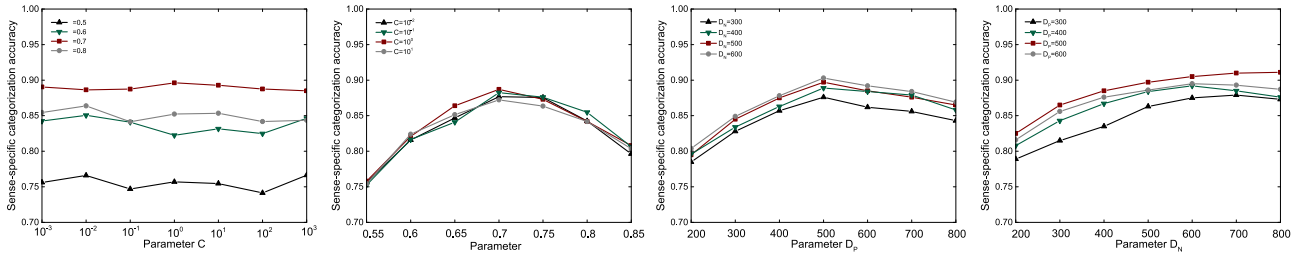


Fig. 9. The parameter sensitiveness of C, δ, D_P and D_N in terms of sense-specific image categorization accuracy.

as the benchmark testing dataset to evaluate the performance variation of our proposed approach. Fig. 9 presents the parameter sensitiveness of C, δ, D_P and D_N in terms of sense-specific image categorization accuracy on testing dataset.

By observing Fig. 9(a), we found that when the δ is fixed, our method is robust to the parameter C when it is varied in a certain range $[10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}]$. From Fig. 9(b), we noticed that when the parameter C is varied, the performance of our method is growing when δ increases but less than 0.7. The reason is perhaps that our training data was derived from image search engine. Due to the error index of image search engine, there may be too much noise in each bag which will result in decreasing the classification accuracy when $\delta \leqslant 0.7$. When δ increases over 0.7, the performance of our method decreases. One possible explanation is that with the increase of δ, the number of semantic senses is decreasing, which may lead to the degradation of our approach.

By observing Fig. 9(c), we found that when the D_N is fixed, the performance of our method is growing when D_P increases but less than 500. The explanation is that when $D_P \leqslant 500$, the performance of the noisy semantic senses purifying classifier increases, and when D_P increases over 500, the noisy semantic senses purifying classifier may have been over-fitted. In this condition, some positive semantic senses may be removed by mistake, resulting in a decrease in the performance of our approach. From Fig. 9(d), we observed that when the D_P is varied, the performance of our method shows a relatively rapid increase when the number of $D_N \leqslant 500$. When the number of negative semantic senses is larger than 500, the performance of our method increases at a relatively slower rate.

E. Time Complexity Analysis

For the time complexity analysis, we analyse the process of "merging visual similar semantic senses" and "distinguishing visual senses". The process of "pruning noisy semantic senses" was formulated as a traditional latent SVM problem and lots of works have analysed its time complexity [19], [20]. So the time complexity analysis of solving the latent SVM will not be a focus of our work.

In the process of merging visual similar semantic senses, our formulation is to find a subset of representative space v (suppose representative space v contains m semantic senses) which can cover the space of variance within the space V. We use an iterative algorithm for discovering the best and second-best solutions to the problem. After discovering the two best solutions, a new problem is formulated, so that the second-best solution to the original problem is the best solution to the new one. The second-best solution for the new problem is found, to become the overall third best solution and the procedure is repeated until all m solutions are found. The time complexity of the algorithm is $O(m \cdot T_2(n))$, where $T_2(n)$ is the complexity of finding the second-best solution to the problem. The worst case time complexity of the algorithm is $O(m \cdot T(n))$, where $T(n)$ is the complexity of finding a single solution.

In the process of distinguishing visual senses, we propose to address the problem in (14) by leveraging the cutting-plane algorithm [43]. We identify the most violating candidate and solve the MKL sub-problem at each iteration. The time complexity of (14) can be computed as $T \cdot O(\text{MKL})$, where T is the number of iterations and $O(\text{MKL})$ is the time complexity of the MKL sub-problem. According to [31], the time complexity of MKL is between $t \cdot O(LCM)$ and $t \cdot O((LCM)^{2.3})$, where M, L, C are the numbers of latent domains, bags and categories respectively. t is the number of iterations in MKL.

F. Failure Cases

Our semantic senses are derived from Google Books N-gram Corpus. Its latest version was built in July 2012. New semantic senses which emerge after July 2012 will not be found in the corpus. For these semantic senses, our method will fail.

V. Conclusion

In this work, we focused on one important yet often ignored problem: we argue that the current poor performance of some classification models learned from the web is due to the visual polysemy. We solved the problem of polysemy by allowing sense-specific diversity in search results. Specifically, we presented a new framework for discovering and distinguishing multiple visual senses for polysemous words. Our work could be used as a pre-step before directly learning from the web, which helped to choose appropriate visual senses for sense-specific images collection and thereby improve the efficiency of learning from the web. Compared to existing methods, our proposed method can not only figure out the right sense but also generates the right mapping between semantic and visual senses. We verified the effectiveness of our approach on the tasks of sense-specific image classification and search results re-ranking. The experimental results demonstrated the superiority of our proposed approach over existing weakly supervised state-of-the-art approaches.

REFERENCES

[1] F. Shen, *et al.*, "Asymmetric binary coding for image search," *IEEE Trans. Multimedia*, 19, no. 9, pp. 2022–2032, Sep. 2017.

[2] V. Liong, J. Lu, Y. Tan, and J. Zhou, "Deep video hashing," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1209–1219, Jun. 2017.

[3] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 598–608, Mar. 2017.

[4] Y. Yao, *et al.*, "Extracting privileged information from untagged corpora for classifier learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018.

[5] J. Deng, *et al.*, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[6] B. Collins, J. Deng, K. Li, and L. Fei-Fei, "Towards scalable dataset construction: An active learning approach," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 86–98.

[7] S. Vijayanarasimhan and K. Grauman, "Large-scale live active learning: Training object detectors with crawled data and crowds," *Int. J. Comput. Vis.*, vol. 108, no. 2, pp. 97–114, 2014.

[8] F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting image databases from the web," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 754–766, Apr. 2011.

[9] D. Batra, P. Yadollahpour, A. Guzman, and G. Shakhnarovich, "Diverse m-best solutions in markov random fields," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 1–16.

[10] L.-J. Li and L. Fei-Fei, "Optimol: Automatic online picture collection via incremental model learning," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 147–168, 2010.

[11] K. Barnard and M. Johnson, "Word sense disambiguation with pictures," *Artif. Intell.*, vol. 167, no. 2, pp. 13–30, 2005.

[12] Y. Yao, *et al.*, "Exploiting web images for dataset construction: A domain robust approach," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1771–1784, Aug. 2017.

[13] J. Veronis and N. Ide, "Word sense disambiguation with very large neural networks extracted from machine readable dictionaries," in *Proc. Assoc. Comput. Linguistics*, 1990, pp. 389–394.

[14] D. Yarowsky, "Word-sense disambiguation using statistical models of roget's categories trained on large corpora," in *Proc. Assoc. Comput. Linguistics*, 1992, pp. 454–460.

[15] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proc. Assoc. Comput. Linguistics*, 1995, pp. 189–196.

[16] P. Pantel and D. Lin, "Discovering word senses from text," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2002, pp. 613–619.

[17] N. Chatterjee and S. Mohan, "Discovering word senses from text using random indexing," in *Proc. Comput. Linguistics Intell. Text Process.*, 2008, pp. 299–310.

[18] X. Hua and J. Li, "Prajna: Towards recognizing whatever you want from images without image labeling," in *Proc. AAAI Int. Conf. Artif. Intell.*, 2015, pp. 137–144.

[19] H. Azizpour, "Visual representations and models: From latent SVM to deep learning," Ph.D. dissertation, School Comput. Sci. Commun., KTH Roy. Inst. Technol., Stockholm, Sweden, 2016.

[20] C. N. Yu, "Improved learning of structural support vector machines: Training with latent variables and nonlinear kernels," Ph.D. dissertation, School Comput. Sci., Cornell Univ., Ithaca, NY, USA, 2011.

[21] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[22] R. Speer and C. Havasi, "Conceptnet 5: A large semantic network for relational knowledge," in *Proc. Peoples Web Meets NLP*, 2013, pp. 161–176.

[23] L. Niu, W. Li, D. Xu, and J. Cai, "Visual recognition by learning from web data via weakly supervised domain generalization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 22, no. 9, pp. 1985–1999, Sep. 2017.

[24] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[25] R. Fergus, Li Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from google's image search," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2005, pp. 1816–1823.

[26] S. Divvala, A. Farhadi, and C. Guestrin, "Learning everything about anything: Webly-supervised visual concept learning," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3270–3277.

[27] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 1462–1475, Mar. 2007.

[28] X. Chen, A. Ritter, A. Gupta, and T. Mitchell, "Sense discovery via co-clustering on images and text," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5298–5306.

[29] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[30] K. C. Kiwiel, "Proximity control in bundle methods for convex non differentiable minimization," *Math. Program.*, vol. 46, no. 1, pp. 105–122, 1990.

[31] J. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods*. Cambridge, MA, USA: MIT Press, 1999, pp. 185–208.

[32] R. Fergus, P. Perona, and A. Zisserman, "A visual category filter for google images," in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 242–256.

[33] T. Berg and D. Forsyth, "Animals on the web," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1463–1470.

[34] Y. Li, I. Tsang, J. Kwok, and Z. H. Zhou, "Tighter and convex maximum margin clustering," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2009, pp. 344–351.

[35] A. Mansur and Y. Kuno, "Improving recognition through object subcategorization," in *Proc. Int. Symp. Vis. Comput.*, 2008, pp. 851–859.

[36] J. Michel, *et al.*, "Quantitative analysis of culture using millions of digitized books," *Science*, vol. 331, no. 6014, pp. 176–182, 2011.

[37] Y. Lin, *et al.*, "Syntactic annotations for the Google books Ngram corpus," in *Proc. Assoc. Comput. Linguistics*, 2012, pp. 169–174.

[38] Y. Yao, X. Hua, F. Shen, J. Zhang, and Z. Tang, "A domain robust approach for image dataset construction," in *Proc. ACM Conf. Multimedia*, 2016, pp. 212–216.

[39] Y. Yao, *et al.*, "Automatic image dataset construction with multiple textual metadata," in *Proc. IEEE Conf. Multimedia Expo.*, 2016, pp. 1–6.

[40] N. Loef, C. O. Alm, and D. A. Forsyth, "Discriminating image senses by clustering with multimodal features," in *Proc. Assoc. Comput. Linguistics*, 2006, pp. 547–554.

[41] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[42] R. S. S. Prakash and A. Y. Ng, "Learning to merge word senses," in *Proc. Conf. Empirical Method. Natural Lang. Process. Comput. Natural Lang. Learn.*, 2007, pp. 1005–1015.

[43] J. E. Kelley, "The cutting-plane method for solving convex programs," *J. Soc. Ind. Appl. Math.*, vol. 8, no. 4, pp. 703–712, 1960.

[44] K. Saenko and T. Darrell, "Unsupervised learning of visual sense models for polysemous words," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1393–1400.

[45] R. Cilibrasi and P. Vitanyi, "The google similarity distance," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 370–383, Mar. 2007.

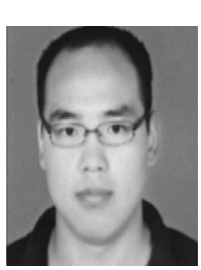
[46] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst*., 2012, pp. 1097–1105.

[47] K.-W. Wan, A.-H. Tan, J.-H. Lim, L.-T. Chia, and S. Roy, "A latent model for visual disambiguation of keyword-based image search," in *Proc. Brit. Mach. Vis. Conf*., 2009, pp. 2–7.

[48] R. Mihalcea, "Using wikipedia for automatic word sense disambiguation," in *Proc. Assoc. Comput. Linguistics*, 2007, pp. 196–203.

[49] A. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Comput*., vol. 15, no. 4, pp. 915–936, 2003.

[50] Y. Yao, *et al.*, "Discovering and distinguishing multiple visual senses for polysemous words," in *Proc. AAAI Conf. Artif. Intell*., 2018.



Yazhou Yao received the B.Sc. and M.Sc. degrees from Nanjing Normal University, Nanjing, China, in 2010 and 2013, respectively. He is currently working toward the Ph.D. degree in computer science with the Global Big Data Technologies Center, University of Technology Sydney, Ultimo, NSW, Australia. From 2013 to 2014, he studied at the School of Computer Science and Engineering, Nanjing University of Science and Technology, supervised by Prof. Z. Tang. With the support of the China Scholarship Council, he studied at the UTS from 2014, supervised by Prof. J. Zhang. His research interests include social multimedia processing and machine learning.

Fumin Shen received the Bachelor's degree from Shandong University, Jinan, China, in 2007, and the Ph.D. degree from the Nanjing University of Science and Technology, Nanjing, China, in 2014. He is currently a Professor with the University of Electronic Science and Technology of China, Chengdu, China. His major research interests include computer vision and machine learning, including face recognition, image analysis, and hashing methods.

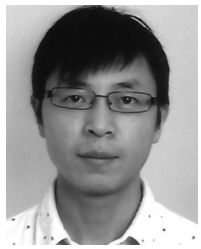
Jian Zhang (SM'04) received the B.Sc. degree from the East China Normal University, Shanghai, China, in 1982, the M.Sc. degree in computer science from Flinders University, Adelaide, SA, Australia, in 1994, and the Ph.D. degree in electrical engineering from the University of New South Wales (UNSW), Sydney, NSW, Australia, in 1999.

From 1997 to 2003, he was with the Visual Information Processing Laboratory, Motorola Labs, Sydney, as a Senior Research Engineer, and later became a Principal Research Engineer and a Foundation Manager with the Visual Communications Research Team. From 2004 to July 2011, he was a Principal Researcher and a Project Leader with National ICT Australia, Sydney, and a Conjoint Associate Professor with the School of Computer Science and Engineering, UNSW. He is currently an Associate Professor with the Advanced Analytics Institute and School of Software, Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW, Australia. He is the author or coauthor of more than100 paper publications, book chapters, and six issued patents filed in the U.S. and China. His current research interests include multimedia processing and communications, image and video processing, machine learning, pattern recognition, media and social media visual information retrieval and mining, human-computer interaction and intelligent video surveillance systems.

He was the General Chair and chaired the International Conference on Multimedia and Expo in 2012. He is an Associated Editor for the IEEE TRANSACTIONS ON MULTIMEDIA and the *EURASIP Journal on Image and Video Processing*.

Li Liu received the B.Eng. degree in electronic information engineering from Xian Jiaotong University, Xian, China, in 2011, and the Ph.D. degree from the Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield, U.K., in 2014. He is currently with the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. His current research interests include computer vision, machine learning, and data mining.

Zhenmin Tang received the Ph.D. degree from the Nanjing University of Science and Technology, Nanjing, China. He is currently a Professor with the Nanjing University of Science and Technology. His major research areas include intelligent system, pattern recognition, and image processing, embedded system. He has authored/coauthored more than 80 papers. He is also the Leader of several key programs of the National Nature Science Foundation of China.



Ling Shao (M'09–SM'10) is currently a Professor with the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. Previously, he was a Professor (2014–2016) with Northumbria University, a Senior Lecturer (2009–2014) with the University of Sheffield and a Senior Scientist (2005–2009) with Philips Research, Eindhoven, The Netherlands. His research interests include computer vision, image/video processing, and machine learning. He is an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and several other journals. He is a Fellow of the British Computer Society and the Institution of Engineering and Technology.