Folksonomy-Based Visual Ontology Construction and Its Applications

Quan Fang, Changsheng Xu, *Fellow, IEEE*, Jitao Sang, M. Shamim Hossain, *Senior Member, IEEE*, and Ahmed Ghoneim, *Member, IEEE*

***Abstract*—An ontology hierarchically encodes concepts and concept relationships, and has a variety of applications such as semantic understanding and information retrieval. Previous work for building ontologies has primarily relied on labor-intensive human contributions or focused on text-based extraction. In this paper, we consider the problem of automatically constructing a folksonomy-based visual ontology (FBVO) from the user-generated annotated images. A systematic framework is proposed consisting of three stages as concept discovery, concept relationship extraction, and concept hierarchy construction. The noisy issues of the user-generated tags are carefully addressed to guarantee the quality of derived FBVO. The constructed FBVO finally consists of 139 825 concept nodes and millions of concept relationships by mining more than 2.4 million Flickr images. Experimental evaluations show that the derived FBVO is of high quality and consistent with human perception. We further demonstrate the utility of the derived FBVO in applications of complex visual recognition and exploratory image search.**

***Index Terms*—Knowledge discovery, ontology, visual recognition.**

I. Introduction

AN ONTOLOGY typically contains a set of concepts and the relationships between concepts. The concept relationships are encoded to organize the concepts in a coarse-to-fine semantic hierarchy. Examples include WordNet [1] and LSCOM [2]. The ontology is recognized to compactly capture how people perceive and understand the world, and has been successfully used as a kind of high-level supervision to facilitate many difficult tasks, such as natural language processing [3], visual object recognition [4], video concept detection [5], [6], video search [7], [8], multimedia information retrieval [9], and health information system [10].

Regarding the significance of ontology in solving practical problems, extensive efforts have been undertaken in ontology or semantic hierarchy construction. One line of efforts is human-based, which relies on heavy human intervention from well-trained experts (e.g., WordNet [1] and LSCOM [2]) or crowdsourcing annotators (e.g., ImageNet [11]). While the obtained ontology has a guaranteed high quality, these approaches need significant human workload and have limited scalability to generalize to specific domains or update from importing new information. Another line is automatic or semi-automatic based, which collects and analyzes the data available on the Web to discover concepts and extract concept relationships [12]–[14]. Currently, most of the efforts in this line focus on utilizing the textual information and exploit the co-occurrence of the discovered concepts in textual documents. This can harvest rich concepts and relationships. However, the concept relationship only depends on the co-occurrence of these concepts in the textual documents on the Web, which cannot accurately describe the metonymy or concurrence relationship of two concepts. For example, "car" and "wheel" have close relationship according to human perception, but may deserve a low relevance score as they are not usually used simultaneously in the same textual document. The ontology constructed only with texts may not be consistent with human cognition. According to [15], 80% of the human perception comes from visual information, it is more reasonable to incorporate visual information for ontology construction rather than by concept co-occurrence in textual documents. With the constructed visual ontology, users can better understand concepts and harvest more knowledge by exploring the images along with concept hierarchy. In this work, we propose to automatically construct a folksonomy-based visual ontology (FBVO) by exploiting the large-scale user-generated images.

Typical photo sharing folksonomies, e.g., Flickr and Instagram, allow users to share personal photos and annotate them with textual descriptions like tags. This has resulted in huge number of weakly tagged images available online. Exploiting these folksonomy-based images for visual ontology construction generally enjoys two advantages: 1) the user-generated tags provide a natural correspondence between the visual content and the textual semantics. This visual-textual correspondence is more compact than that extracted from the surrounding text [16]–[19]; and 2) users are likely to embed their perception of the real-world objects and semantics in the associated tags. Fig. 1 shows an example image from Flickr and its associated tags. We can see that both synonym concepts (e.g., "portrait"

Manuscript received July 17, 2015; revised November 12, 2015 and January 2, 2016; accepted January 28, 2016. Date of publication February 11, 2016; date of current version March 15, 2016. This work was supported in part by the National Basic Research Program of China under Grant 2012CB316304, in part by the National Natural Science Foundation of China under Grant 61225009, Grant 61432019, Grant 61332016, Grant 61303176, and Grant U1435211, in part by the Beijing Natural Science Foundation under Grant 4131004, and in part by the Deanship of Scientific Research, King Saud University, Riyadh, Saudi Arabia, under the research group project RGP-229. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chengcui Zhang.

Q. Fang, C. Xu, and J. Sang are with the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: qfang@nlpr.ia.ac.cn; csxu@nlpr.ia.ac.cn; jtsang@nlpr.ia.ac.cn).

M. Shamim Hossain is with the Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia (e-mail: mshossain@ksu.edu.sa).

A. Ghoneim is with the Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia, and also with the Department of Computer Science, College of Science, Menoufia University, Menoufia 32721, Egypt (e-mail: ghoneim@ccis.ksu.edu.sa).

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

Digital Object Identifier 10.1109/TMM.2016.2527602

1520-9210 © 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Fig. 1. Example image from Flickr.

with "portraiture") and subsumption concepts (e.g., "people" with "man" and "woman") are annotated. This collective knowledge is of significance and can be exploited to facilitate the ontology construction.

Despite the advantages, building FBVO from weakly labeled user-generated images also presents a number of challenges. In this paper, we identify and address three of them detailed as follows. 1) The user-generated raw tags suffer problems of imprecise, subjective, and incomplete, which prevents from direct utilization. For example, in Fig. 1, tags like "photographer" record the context of the image and do not well describe the visual content included. Also, some tags are missing, like "trees," "lake," etc. 2) Both the textual and visual information need to be considered to extract the concept relationships, where a concept hierarchy is expected as the final output. This requires to distinguish between the synonym and subsumption relationships from textual-visual co-occurrences. 3) A fully automatic FBVO is capable of updating over time by incorporating new images and concepts. Therefore, the FBVO construction solution needs to entail self-learning in a never-end way.

Our proposed framework for FBVO construction is illustrated in Fig. 2, which contains three key stages: (A) concept discovery, (B) concept relationship extraction, and (C) concept hierarchy construction. We call concept any tag that has a Wikipedia page. At the first stage, we first conduct Wikipedia-based identification to select the concept sets from Flickr tags. For the identified concept, we exploit the associated Flickr images and adopt max-margin hard instance learning [20] to learn concept models. This enables the automatic update of the constructed ontology by importing and recognizing new images. With the learned concept models, concept instance enrichment is conducted to refine the raw tags of the associated image instances. At the second stage, the undirected co-occurrent relationships between concepts are first extracted by exploiting both the visual exemplar similarity and tag co-occurrences. The directed subsumption relationships are then obtained by examining the frequency discrepancy, which measures the differences in the occurrence frequency distribution of concept tags and is calculated by the conditional probability between concepts. At the third stage, concept entropy is calculated to estimate the concept semantic broadness. An algorithm is introduced to traverse all the concepts from the highest entropy and transfer the concept-concept subsumption relationships to construct a directed acyclic graph (DAG). Associating each concept node with textual descriptions and exemplary image instances, the concept DAG well represents the semantic hierarchy and constitutes the final FBVO.

The proposed framework is implemented to generate a large-scale FBVO from more than 2.4 million Flickr images. The resultant FBVO contains 139,825 concepts, 12,433,209 concept co-occurrent relationships, and 1,545,854 subsumption relationships. We evaluate the quality of the derived FBVO by quantitatively examining the performance of both the concept models in concept recognition and the concept relationships versus human-based ontology. Moreover, two applications are designed to investigate the potentials of the derived FBVO, i.e., concept feature-based visual recognition and ontology-based exploratory image search. In concept feature-based visual recognition, the derived concept models are utilized to identify the involved concepts in the examined images, which then serve as the mid-level features for supervised visual recognition. In ontology-based exploratory image search, the derived concept hierarchy is incorporated to expand the user query and enable image search in an exploratory and interactive fashion. Experimental results have validated the quality of the derived FBVO as well as the effectiveness of the proposed FBVO construction framework.

We summarize the main contributions of this paper as follows.

1) We propose a simple framework to automatically construct a visual ontology from folksonomy-based images. The framework is capable of effectively leveraging user-generated noisy tags, exploiting both textual and visual information, and updating in a never-end way.
2) A large-scale visual ontology is constructed and evaluated with extensive experiments. The potential of the derived ontology is further examined with two novel applications.

II. Related Work

A. Visual Concept Modeling

Concept modeling has been extensively studied in multimedia [2] and computer vision (usually referred to as "attribute") [21] communities. The concepts being modeled are mostly objects [22], scenes [23], sentiments [24], locations [25], and events [26]. In benchmarks like TRVECVID [27] and PASCAL [28], researchers have investigated a variety of features and statistical learning models towards the task of concept modeling. Mylonas *et al.* [29] proposed to use visual context and region semantics for high-level concept detection on TRECVID and Corel data sets. Uijlings *et al.* [30] presented an evaluation of fast Bag-of-Words components for real-time visual concept classification. A key problem in visual concept modeling is to collect the training samples for the large number of concepts. One way is to issue the concepts as queries to the image search engines. For example, Ewerth *et al.* [31] proposed an incremental and scalable web-supervised learning system with heterogeneous appearance models for long-term learning of visual concepts from Web images. Zhu *et al.* [32] proposed to maximize relevancy and coverage to generate training images for visual concept learning from Web noisy images. Li *et al.* [11] proposed a multiple instance learning algorithm to learn mid-level visual concepts from Google and Bing image search

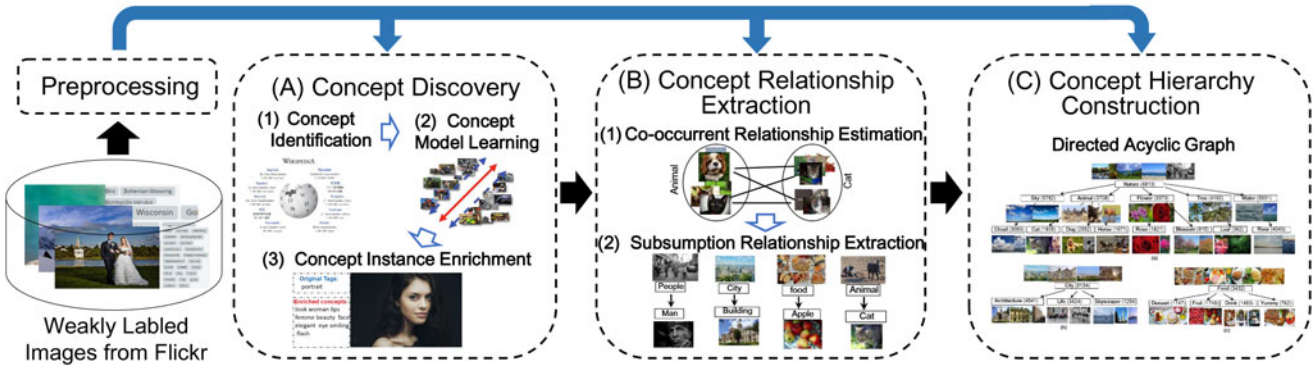


Fig. 2. Proposed framework for automatic construction of FBVO from weakly labeled Flickr images.

results. LEVAN [33] extracted keywords from Google Ngram to construct structured queries and retrieved the relevant image instances for each concept. Alternatively, there also exist approaches of collecting samples from user-generated weakly labeled images. [34] discovered concepts from images and the associated textual descriptions collected from a shopping website. Borth *et al.* [24] proposed SentiBank, a visual concept detector for visual sentiment analysis by exploiting Flickr images. Along with this research line, our approach also exploits the user-generated images with weakly-labeled tags. A series of techniques are developed to address the noisy tag issues and facilitate the concept model learning (CML) and concept instance enrichment. The images with associated refined tags are leveraged to extract co-occurrent and subsumption relationships between tags, and the derived concept models are utilized for automatic ontology update.

B. Visual Ontology Construction

In multimedia and computer vision fields, researchers have made extensive efforts to construct the visual ontology. Fan *et al.* [18] proposed a novel algorithm for mining multilevel image semantics via hierarchical classification. ImageNet [35] is an image database organized according to the WordNet hierarchy using Amazon Mechanical Turk,[1] in which each node of the hierarchy is depicted by hundreds and thousands of images. However, as discussed in the introduction, these human-based ontologies are labor-intensive and have limited scalability. The availability of huge amounts of web images provides the opportunity for automatic visual ontology construction. Recently, there have been growing interests in exploiting the returned images from image search engines to help visual ontology construction. Lu *et al.* [36] developed a novel framework to identify high-level concepts with small semantic gaps from a large-scale web image dataset. Wang *et al.* [16] proposed a bottom-up and top-down approach to build an image knowledge base called ImageKB from Bing image search engine. Zhang *et al.* [37] constructed an attribute-augmented semantic hierarchy from the web images associated with ImageNet for interactive image retrieval. NEIL [17] is a never ending learning system for visual ontology construction from image search engines, which iterates between concept relationship extraction, image instance recognition, and concept classifier/detector learning. With the popularity of social media and photo sharing folksonomies, users contribute to huge number of images with associated tags. We believe that the user-generated tags encode more compact visual-textual correspondences and provide an alternative solution for visual ontology construction. In this paper, several techniques are introduced to exert the potential of these user-generated tags in CML, concept relationship extraction, and concept hierarchy construction.

III. FBVO Construction

A. Concept Discovery

The first stage consists of three substages, i.e., concept identification, CML, and concept instance enrichment. The goal of concept identification is to determine the concept sets from Flickr tags. Concept models are then learned to enable automatic ontology update with consideration of the noisy issues of the tags. Based on the learned concept models, the raw image tags are refined and enriched to facilitate the following concept relationship extraction at the next stage.

1) Concept Identification: Preprocessing. The open contribution mechanism leads to the user-generated tag usage in a very causal way. For example, users may combine two or more words within a single tag, e.g., "bridge of lift", "Rocky-Mountains","black & white". These types of tags are usually created by bridging words that include prepositions of "at", "of", "in", conjunctions of "and", "or", or special characters such as '&', ':', '¿', ',', '-'. Therefore, a pre-filtering process is needed. We start by tokenizing the associated tags on these prepositions, conjunctions and special characters. White space is not tokenized to avoid breaking up proper names like "North America". The tags composed by only the non-alphanumeric characters or frequently-used common words like "the", 'and' and "myself" are also removed. The resultant tags are then normalized into the lower case.

Concept Identification. We call concept any tag that has a Wikipedia page. After pre-filtering, we match the remained tags with the entries in a Wikipedia thesaurus.[2] The tags that have a Wikipedia page are kept and constitute the final concept sets.

[1][Online]. Available: http://aws.amazon.com/mturk/

[2][Online]. Available: https://en.wikipedia.org/wiki/Main_Page

2) CML: The goal of this substage is to equip the concept nodes with a recognition capability, i.e., to recognize and import new images to enrich the concept descriptions and improve the performance of concept relationship extraction. We refer to concept model as a type of classifiers that has this recognition capability.

Positive Image Selection. Due to the imprecise issue of the user-contributed tags, it is impractical to directly use the images with associated concept tags as the positive training samples. To this end, we develop a neighborhood voting approach to select the reliable training images. Specifically, the classical Kernel Density Estimation (KDE)[38] is utilized to measure the relatedness between concept tags and images. Denote X_c as the set of images that originally contain concept tag c. The probability of image x_i belonging to concept c is defined as[3]

$$p(x_i|c) = \frac{1}{|X_c|} \sum_{x_j \in X_c} K_\sigma(x_i - x_j) \tag{1}$$

where image x_j belonging to concept c is from X_c and $|X_c|$ is the cardinality of X_c and K_σ is the Gaussian kernel function with the radius parameter σ, i.e.

$$K_\sigma(x_i - x_j) = \exp\left(-\frac{\| x_i - x_j \|^2}{2\sigma^2}\right) \tag{2}$$

where $\| \cdot \|$ denotes the l_2 norm, σ is the kernel radius and adaptively assigned as the median value of all pair-wise Euclidean distances between images. We can see that $p(x_i|c)$ measures the confidence that concept c is presented in image x_i. We select the images with highest $p(x_i|c)$ to construct the positive image set T_c for concept c.

Concept Model Training. In addition to removing the noisy images, for the same concept, especially some general concepts, the selected positive images may reflect different aspects of the concept and exhibit significant visual variations. For example, the concept "vehicle" may involve with images of car, bicycle, bus, train, ship, aircraft, etc. In this case, treating all the positive images as a whole for training may result in models of weak discriminative capability. Therefore, we propose to learn several models corresponding to the subcategories of the same concept. For approach, for concept c and the positive image set T_c, we run Affinity Propagation [40] using the image visual features to identify the underlying clusters $\{X_m^c\}_{m=1}^M$. The number of clusters is adaptively determined by AP. The derived clusters represent different views or subcategories related to the concept c.

For each concept subcategory, we learn a classifier. For the m^{th} subcategory of concept c, the positive training set includes all the images from X_m^c. However, it is non-trivial to determine the negative training set as: 1) The number of negative samples is much larger than that of positive samples if we simply use all the images without concept c to construct the negative training set. This will lead to the imbalanced problem [41]. 2) Since the user-generated tags suffer from issue of incomplete, many images without concept c may serve as good candidates for the positive samples. Equally treating these images will deteriorate the training process and lead to inferior classifier. To address these problems, we utilize hard negative mining [22] to select the difficult instances as negative training samples during the model learning process. This method will iteratively seek the max-margin decision boundary that separates difficult negative samples and the positive training samples.

[3]In this work, for each Flickr image, we extract its deep feature for representation: 4096 dimensional feature vector from the Fully Connected Layer (FC) 7 layer of Caffe reference network [39].

Algorithm 1: CML

Input: Positive image clusters $\{X_m^c\}_{m=1}^M$ for concept c, images $\{S\}$ without concept c

Output: Concept models for concept c.

1: **for** each cluster m **do**
   Concept Model Training
2: $\mathcal{P}_c^{train} \Leftarrow \{I_i | I_i \in X_c, I_i \in \text{cluster } m\}$
3: $|\mathcal{N}_c^{train}| = \beta |\mathcal{P}_c^{train}|$
4: $\mathcal{N}_c^{train} \Leftarrow rand\_sample\ \{S\}$
5: **while** $\mathcal{N}_c^{train}$ is updated **do**
6: $\Psi_c \Leftarrow svm\_train\ (\mathcal{P}_c^{train}, \mathcal{N}_c^{train})$
7: $(\mathcal{N}_c^{\text{hard}}, \mathcal{N}_c^{easy}) \Leftarrow filter\ (\Psi_c, \mathcal{N}_c^{train})$
8: $\mathcal{N}_c^{train} \Leftarrow \mathcal{N}_c^{hard} \cup rand\_sample\ \{S - \mathcal{N}_c^{easy}\}$
9: **end while**
   Concept Subcategory Description
10: compute tag relevance score $s(t_j)$ to cluster m;
11: select the tags with highest $s(t_j)$ to construct the concept subcategory description $\mathcal{V}_m^c$.
12: **end for**

The detailed steps for CML are summarized in Algorithm 1. For a concept cluster, the algorithm starts with an initial cache of training instances, where the positive set is fixed as $X_m^c \in T_c$ and the negative set is generated by randomly sampling images without concept c. In each iteration, easy negative instances are removed from the cache and additional randomly selected negative images are added. LIBLINEAR SVM [42] is retrained on the new cache of training instances. β is the ratio of the number of negative samples to that of positive samples. To avoid the imbalanced training set problem, we keep $\beta = 1 \sim 5$. For each concept subcategory, we further assign relevant tags for semantic description. The relevance of tag t_j to the m^{th} subcategory of concept c is estimated as $s(t_j) = \sum_{x_k \in X_m^c} p(x_k|t_j)$, where $p(x_k|t_j)$ is calculated using Equation (1). The semantic description is represented as a tag set $\mathcal{V}_m^c = \{t_j\}_{j=1}^J$ constituted by the tags with highest relevance scores.

3) Concept Instance Enrichment: With the derived concept models and semantic descriptions, concept instance enrichment is conducted to facilitate the following concept relationship extraction. Since tag co-occurrence plays an important role in extracting the concept relationships, in this substage, we design approaches to enrich new tags as well as filtering out imprecise tags.

Given an image without raw tags, we leverage the learned models corresponding to the C most frequent concepts to score the image. Both the concepts with the highest confidence score,

Fig. 3. Examples of concept instance enrichment results.

and the semantic tags describing the concept subcategories that achieve the highest confidence scores are added as the new tags for the image. Fig. 3(a) shows such an example. The classifier for subcategory "bread breakfast" of the concept "food" obtains the highest confidence score on this image. Therefore, the subcategory descriptive tags as well as the concept are assigned as the new tags. Given an image with L raw tags, we leverage both the concept models corresponding to the raw tags and the learned models corresponding to the C most frequent concepts to score the image. The concept models with the highest confidence scores are recorded, and the corresponding subcategory tags and concepts are assigned. In Fig. 3(b), we can see the raw tag "hope" is removed as the concept models corresponding to the "hope" achieve low score. Therefore, concept instance enrichment is effective in both adding missing tags and filtering imprecise tags. In our experimental dataset, the average number of raw tags for each image is 12.1. After concept instance enrichment, the number increases to 17.7. The experimental evaluation results validate the effectiveness of concept instance enrichment in extracting concept relationships and thus in improving the quality of the derived FBVO.

B. Concept Relationship Extraction

To organize the concepts into a semantic hierarchy, the pairwise concept relationships need to be first extracted. In this stage, we introduce how to exploit the associations between images and concept tags to automatically extract the co-occurrent and subsumption relationships between concepts.

1) Co-Occurrent Concept Relationship Extraction: The co-occurrent relationship generally measures the semantic closeness between concepts. To construct the visual ontology, both visual and textual information are leveraged. We first compute the visual similarity between concepts. Recall that we have obtained image clusters of concepts in CML. The visual distance between a cluster $X_m^{c_i}$ of concept c_i and a cluster $X_m^{c_j}$ of concept c_j is estimated by aggregating the discrepancies between the included images

$$d(X_m^{c_i}, X_m^{c_j}) = \frac{1}{|X_m^{c_i}| \cdot |X_m^{c_j}|} \sum_{x \in X_m^{c_i}, y \in X_m^{c_j}} K_\sigma(x - y) \quad (3)$$

where $\{X_m^{c_i}\}_{m=1}^{M_{c_i}}$ and $\{X_m^{c_j}\}_{m=1}^{M_{c_j}}$ denote the image clusters of concept c_i and c_j, $K_\sigma(x-y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$ is the Gaussian kernel function similar to Equation (2). The visual similarity between cluster $X_m^{c_i}$ and cluster $X_m^{c_j}$ is defined as $f(d(X_m^{c_i}, X_m^{c_j}))$, where $f(\cdot)$ is standard sigmoid function, i.e., $f(x) = (1 + e^{-x})^{-1}$. Max-pooling is then conducted on the cluster similarities to obtain the visual similarity $\phi_v(c_i, c_j)$ between concept c_i and c_j.

The textual similarity between concepts is estimated by examining the co-occurrence of the concept tags. Specifically, the distance between concept c_i and c_j is calculated by Google distance [43]

$$d(c_i, c_j) = \frac{\max(\log N(c_i), \log N(c_j)) - \log N(c_i, c_j)}{\log N_{total} - \min(\log N(c_i), \log N(c_j))} \quad (4)$$

where $N(c_i)$ and $N(c_j)$ are the number of images containing tag c_i and tag c_j respectively, $N(c_i, c_j)$ is the number of images containing both c_i and c_j, and N_{total} is the total number of images. The textual similarity between concept c_i and c_j is then calculated as $\phi_t(c_i, c_j) = (1 + \exp(-d(c_i, c_j)))^{-1}$.

Finally, we linearly combine the visual and textual similarity to obtain the concept co-occurrent score as

$$s(c_i, c_j) = \lambda \phi_v(c_i, c_j) + (1 - \lambda)\phi_t(c_i, c_j) \quad (5)$$

where λ is the weighting parameter and $\lambda \in [0, 1]$.

2) Subsumption Concept Relationship Extraction: We further explore the directed subsumption relationships between concepts. Inspired by [12], the subsumption relationship between concept c_i and c_j is defined as follows: if c_i subsumes c_j, then wherever c_j is used, c_i can be used without ambiguity. The subsumption relation between c_i and c_j is denoted as $c_i \rightarrow_s c_j$. For example, *fruit*$\rightarrow_s$*apple* indicates that for any image annotated with *apple*, we can also annotate it with *fruit*. We discover the concept subsumption relations by estimating a conditional probability $p(c_i|c_j)$. The basic idea is that, if $c_i \rightarrow_s c_j$, c_i will have a wider coverage and higher usage frequency than c_j. In other words, if c_j is used, it will be very likely to see c_i as well, but not vice versa. Therefore, $p(c_i|c_j)$ is calculated via the concept co-occurrent score as

$$p(c_i|c_j) = \frac{s(c_i, c_j)}{\sum_z s(c_j, c_z)}. \quad (6)$$

We can see that $p(c_i|c_j)$ measures the extent that concept c_i subsumes concept c_j. With the subsumption probability, we can connect all the concepts into a directed graph $G = (V, E)$, where V is the set of concept nodes, and E is the set of directed edges. An edge e_{c_i,c_j} from c_i to c_j indicates the subsumption relationship $c_i \rightarrow_s c_j$. The weight of each edge $w(e_{c_i,c_j})$ is defined as the subsumption probability $p(c_i|c_j)$.

C. Concept Hierarchy Construction

The goal of concept hierarchy construction is to turn the directed graph G into a DAG. A DAG is a graph structure with no closed chains where a node can have multiple parents. It is appropriate to represent a folksonomy-based ontology, where the concepts are organized in a coarse-to-fine semantic hierarchy. To estimate the semantic broadness for each concept to construct the DAG, we further define the entropy for each concept c_i by utilizing the subsumption probability

$$H(c_i) = -\sum_z p(c_i|c_z) \log(p(c_i|c_z)). \quad (7)$$

The premise here is that, a concept with high entropy is expected to have broad semantics and wide out-links to other concepts.

Algorithm 2: DAG Construction Algorithm

Input: Concepts with entropy $O = \{H(c_i), c_i \in C\}$; weighted relations $R = \{c_i \rightarrow_s c_j, c_i \in C, c_j \in C, w_{c_i \rightarrow c_j} > 0\}$.
Output: A DAG of concepts $G^* = \{V^*, E^*\}$.

1: **for** $c_i \in C$ in descending concept entropy order **do**
2: $V^* \Leftarrow c_i$
3: choose the top-ranked subsumption relation pairs of $c_i, R_{c_i} = \{c_i \rightarrow_s c_j, c_j \in C\}$
4: **if** $H(c_i) \geq H(c_j)$ **then**
5: $E^* \Leftarrow c_i \rightarrow_s c_j, V^* \Leftarrow c_j$
6: **end if**
7: **end for**
8: output G^*

While, the concept with low entropy is more likely concentrated on specific semantics.

With the discovered subsumption relationships and the calculated entropy for each concept, we introduce a DAG construction algorithm to discover the concept hierarchy, as summarized in Algorithm 2. The algorithm traverses the concept nodes in the directed graph G and greedily adds the concept with the highest entropy and its selected subsumption relationships into the DAG. For each concept node in the derived DAG, we add the representative image instances discovered from AP-based clustering and the concept subcategory descriptions, to obtain the final FBVO.

In this section we have introduced the three stages in automatically constructing a FBVO from Flickr images with user-generated tags. Note that sequentially conducting the three stages only finishes one round of FBVO construction. The proposed framework is capable of updating by importing new images and self-enhancement in a never-end learning fashion: 1) Equipped with the concept models, the derived FBVO is ready to accept new images with or without user-generated tags. An online processing of the new images will contribute to improved extraction of concept relationships and dynamic update of the FBVO. 2) The derived concept hierarchy in the FBVO serves as high-level supervision to facilitate concept model enhancement in the next round. For example, knowing that "fruit" subsumes "apple" from the derived FBVO, we are confident to add the positive images of "apple" as the positive training samples when learning models for "fruit," and avoid using the positive images of "fruit" as the negative training samples when learning models for "apple." Since this paper focuses on introducing the basic FBVO construction framework, we leave the further implementation details, experimental evaluation and more discussions of FBVO update for future work.

IV. Experiments

A. Overview of the Constructed FBVO

We use Flickr as the example photo sharing folksonomy and the weakly labaled Flickr images to construct the dataset. We downloaded images with their associated tags from the Flickr groups via the public Flickr API. The selected Flickr groups cover a wide range of topics, including *photograph, people, nature, experience, world, travel, beautiful capture*, etc. The collected dataset contains 2 414 341 images with 1 173 730 unique tags.

TABLE I
Statistic of the Constructed FBVO

#concepts	#co-occurrent relation	#subsumption relation
139,825	12,433,209	1,545,854

Tag preprocessing is first conducted to filtered out the tags of non-alphanumeric characters, stop word, and usage frequency less than 10. This results in 499 532 tags. These tags are issued to Wikipedia and 139 825 concepts are identified to construct the concept set in the final FBVO. For each identified concept, the top 1000 images (if there has) with the highest relevance score as computed from Equation (1) are selected in the positive image set T_c, which are leveraged to identify the concept subcategories. To obtain reliable concept models trained with sufficient training images, we select concept subcategories that contain more than 50 positive images to learn the corresponding concept models. This leads to a total of 44 970 concept models. We select top 1000 concepts with the highest tag frequencies as the most frequent concepts and use the learned concept models corresponding to these concepts for concept instance enrichment.

For concept relationship extraction, calculating the pairwise distance between all the 139 825 concepts is not necessary and will lead to a high computational complexity. Instead, we extract the concept relationships only on the concepts that have tag co-occurrences in the collected dataset. by considering the existing concept tag co-occurrence pairs, which can significantly reduce the cost and guarantee the quality of concept relationships. This results in 12 433 209 co-occurrent relationships. In the derived FBVO, 1 545 854 subsumption relationships are remained and construct the final concept hierarchy. Table I summaries the basic statistics of the constructed FBVO. In Fig. 4, we show a part of the FBVO for illustration. Three subgraphs on "nature," "city" and "food," with some of their subsuming concept nodes and the corresponding representative images are presented. The numeric in the concept rectangle indicates the number of connected concepts. It is shown that the constructed FBVO is consistent with human perception and well captures the hierarchical structure among concepts. In the following, we conduct experiments to quantitatively evaluate the quality of the derived FBVO.

B. Evaluation of Concept Models

We first evaluate the quality of the discovered nodes in the derived FBVO, i.e., the performance of the learned concept models in visual concept classification. Specifically, two evaluation sets are utilized, a subset of the collected Flickr images with user-generated tags (denoted as "WLTS"), and a well labeled image set MIRFlickr25K [44].

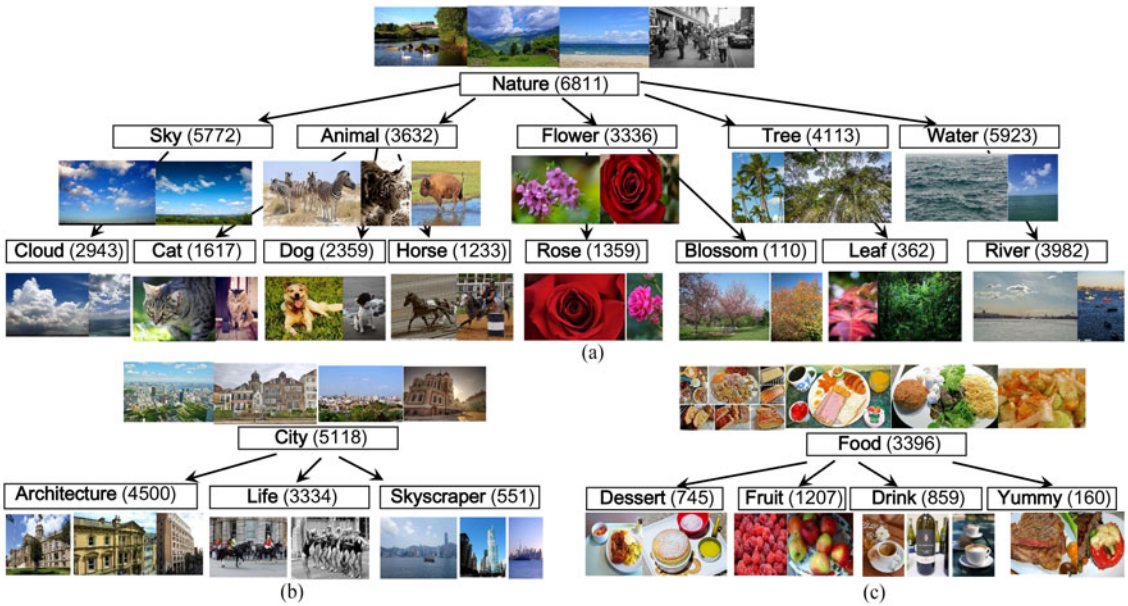


Fig. 4. Illustration for a part of the constructed FBVO.

TABLE II
ILLUSTRATION FOR PART OF THE TEST CONCEPTS ON WLTS

Concept list
portrait landscape nature travel people water sky street girl light sunset city white blue beautiful woman black clouds green sea beach asian woods food drink football interesting mono fire plant butterfly fish sexy young old festival concert dance kid hiking dusk man woman walking boy cafe sun flower architecture beach wildlife river asian fashion dog horse cat snow bird park tree ocean life love mountain face fruit cake ...

TABLE III
MEAN AP@20 OF THE DIFFERENT CONCEPT MODELS ON TWO DATASETS OF WLTS AND MIRFLICKR25K

	CRI	CPI	CML
WLTS	0.3643	0.5602	0.7235
MIRFlickr25K	0.3892	0.5642	0.7864

To evaluate the performance on WLTS, 1024 concepts are selected from FBVO with the highest tag frequencies. A portion of the 1024 testing concepts are listed in Table II for illustration. For training we randomly select 80% of positive images for each concept and twice as many negative images using the sampling scheme described in Algorithm 1. For testing, we use the remaining 20% of positive images as the positive test set and randomly sampled twice as many negative images as the negative test set (except those selected in the training set). The MIRFlickr25K dataset contains 24 concepts and 25 000 well labeled images. We locate the 24 concepts in the derived FBVO and directly use the learned concept models to evaluate the image classification performance on the 25 000 images.

We train concept models using Algorithm 1. Parameter tuning of SVM is performed by cross-validation optimizing Average Precision at rank 20 (AP@20), which is a evaluation metric concentrating on the top ranked samples. For the concepts with multiple concept subcategory models, we use all these models to score a test image and choose the maximum response as the confidence score for this concept. The performance of proposed CML is compared with the following baselines:

1) concept models learned from Random Images (CRI). Concept models are learned using a randomly chosen subset of images associated with the concept tag without positive image selection; and
2) concept models learned from all Positive Images (CPI). Subcategory separation is skipped. For each concept, we use all the selected positive images to train one concept model.

The evaluation results in terms of mean AP@20 on the two test sets are shown in Table III. AP@20 for each examined concept on MIRFlickr25K is shown in Fig. 5. We can see that by selecting positive images and discovering concept subcategories, the learned concept models obtain an obvious improvement in the recognition capability. On both test sets, *CPI* achieves about 50% improvement over *CRI*, while *CML* achieves another 30% improvement over *CPI*. This demonstrates the advantage of our introduced approaches to address the noisy issues of user-generated tags in learning concept models. Note that we do not quote and compare the performance on MIRFlickr25K with state-of-the-art approaches is that: We learn the concept models on the collected images with user-generated tags, which are only weakly labeled. This is different with the other approaches that use the well labeled training set. Moreover, the number of training images used in MIRFlickr25K is larger than that used in our approach. Simply listing the reported performances in literatures leads to unfair comparison.

C. Evaluation of Concept Relationships

In this subsection, we evaluate the quality of the identified subsumption relationships in the derived semantic hierarchy.

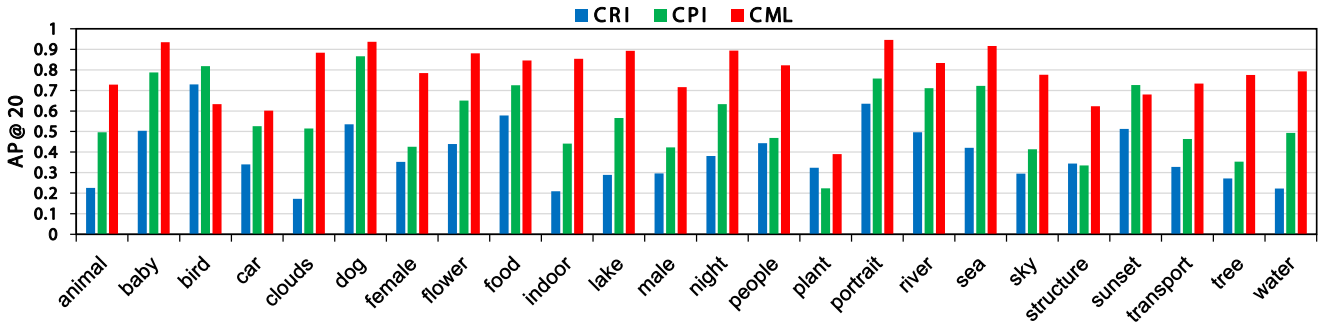


Fig. 5. Detailed performance for different concept learning approaches on MIRFlickr25K.

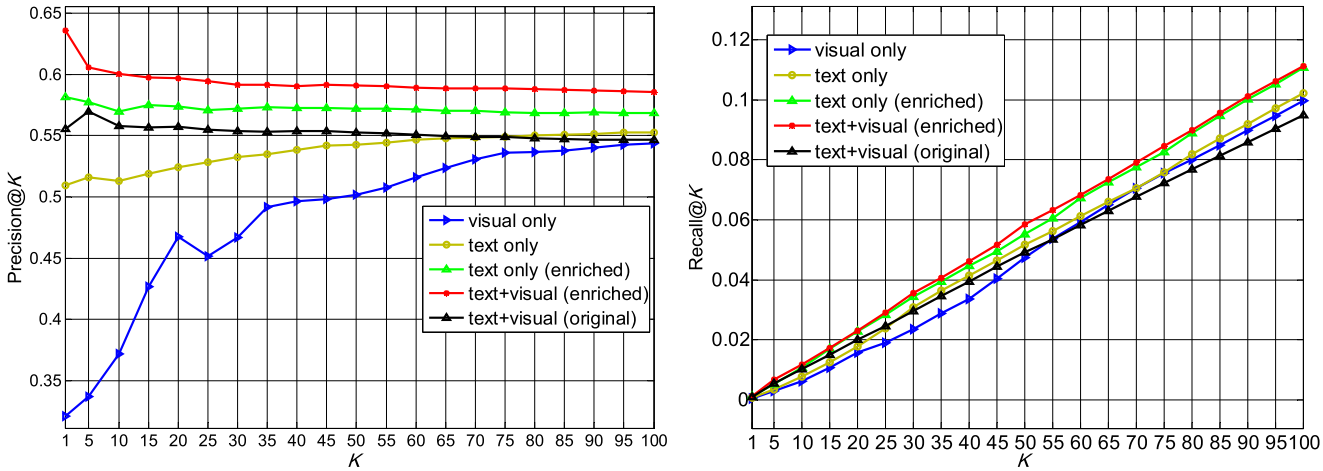


Fig. 6. Comparison of concept relationship extraction in terms of precision and recall.

WordNet is an ontology generated by well-trained experts. We take the concept relationship in WordNet as ground truth. We use a Python API for WordNet in NLTK [45], and selected 1000 shared concepts between our constructed FBVO and WordNet for quantitative evaluation. Specifically, we use NLTK API to obtain the similarity value of two concepts and then calculate their subsumption score using (6). For each concept, we remain the concept pairs whose subsumption scores are above the average level as the ground-truth subsumption relationship.

Recall@k and Precision@k are used as the evaluation metrics. Denote $S_k(c)$ as the top k identified subsumption relationships of concept c in the derived FBVO, $\mathcal{Q}$ as the set of test concepts, and $S_{truth}(c)$ as the ground-truth relationship set generated by WordNet. The evaluation metrics are calculated by

$$Recall@k = \frac{1}{|\mathcal{Q}|}\sum_{c\in\mathcal{Q}}\frac{|S_k(c)\cap S_{truth}(c)|}{|S_{truth}(c)|}$$

$$Precision@k = \frac{1}{|\mathcal{Q}|}\sum_{c\in\mathcal{Q}}\frac{|S_k(c)\cap S_{truth}(c)|}{k}. \quad (8)$$

The final performance is obtained by averaging the above metrics calculated from five independent trails. Fig. 6 illustrates the performance as k increases. *text+visual (enriched)* indicates the performance of the proposed approach. For comparison, we extracted the concept relationships based only on visual similarity (*visual only*) and tag co-occurrence (*text only*) as the baselines. To examine the advantage of concept instance enrichment, concept relationships extracted based on the raw tags are also evaluated, which is denoted as *text+visual (original)*.

From the results we can see the following. 1) The tag co-occurrence based method outperforms visual similarity based method. This is probably due to the semantic gap between low-level visual features and high-level semantic concepts. 2) Combining visual and textual information has achieved the best performance (the precision@10 and recall@10 are 0.6004 and 0.0118, respectively). This indicates that textual and visual information play complementary role for concept relationship extraction. 3) Without concept instance enrichment, even combined with visual information, *text+visual (original)* achieves inferior performance than *text only (enriched)*. This shows the necessity of filtering out imprecise tags and adding new tags before extracting concept relationships, and the advantage of the proposed concept instance enrichment solution.

We further conduct experiment to analyze the sensitivity of the proposed approach with respect to the weighting parameter λ in (5). The parameter λ is a tradeoff parameter to balance the contributions of tag co-occurrence and visual similarity. λ is set to range from 0 to 1.0. $\lambda = 0, 1$ reduces the approach to *text only* and *visual only*, respectively. Fig. 7 shows the results. It is observed that the performance remains steadily when $\lambda \in [0.1, 0.6]$, with the optimal performance achieved when $\lambda = 0.1$. This suggests that in practical applications, λ

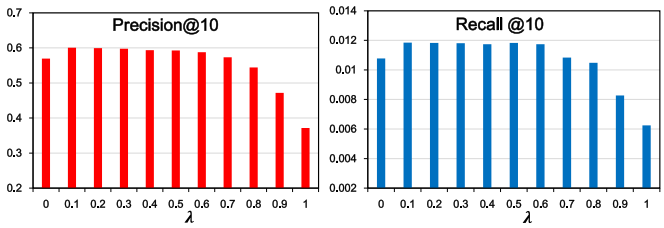


Fig. 7. Sensitivity of the concept relationship extraction performance to the weighting parameter λ.

may be set as around 0.3 to slightly emphasize the contribution from textual information.

V. Applications

In this section, we investigate the potential of the constructed FBVO in two applications as: 1) leveraging the discovered concept nodes as mid-level features for complex visual recognition task; and 2) leveraging the constructed concept hierarchy to expand user queries for serendipity and exploratory image search.

A. Concept Feature-Based Visual Recognition

The learned concept models in the FBVO can be directly used for visual recognition. Given an image with extracted deep feature, we leverage the concept models to obtain the response scores over all concepts. The concepts with high response scores are expected to present in the image. However, this approach suffers from two problems. 1) Each new image needs to be scored through all the concept models (44 970 models in our derived FBVO), which is very time consuming and impractical in real-world applications. 2) In the derived FBVO, both concept models and concept relationships are available. Conducting recognition without consideration of the concept relationship may lead to bad performance, especially for some complex visual recognition tasks, such as event concepts, location concepts, etc.

To address these problems, we propose to first find each complex concept its most S relevant concepts from the FBVO, then score the image with the S concept models and utilize these concept response scores as the mid-level feature. Representing images in a semantic space has shown promising performances in recent studies to tackle with visual recognition tasks [11], [46]. The constructed FBVO contains various types of concepts related to attributes, objects, scenes, etc. The extracted concept relationships can be used to map a specific concept onto the concept spaces for representation. Specifically, given a visual recognition task consisting of W categories, for each of the W categories, we find its S most relevant concepts in our constructed FBVO according to the concept co-occurrent relationships. Then we leverage the $W \times S$ concept models to score each image, and use the response scores to construct a $W \times S$ dimensional feature vector. With the concept response scores as mid-level features, we can train any supervised model as the visual recognition classifier.

We evaluate this application in the tasks of scene and event recognition. Two benchmark image sets, Scene-15 [48] and UIUC-Sport event [26] are utilized. Scene-15 has 15 natural scene categories, and UIUC-Sport has eight complex event classes. For each examined scene and event, we locate it in the concept space from the FBVO and find its most relevant 32 concepts. This results in 480 concept models for Scene-15 and 256 concept models for UIUC-Sport. In Fig. 8, we shows one category example for each of the image sets and several most relevant concepts in the constructed FBVO. We can see that, the categories to be recognized are very difficult. FBVO plays role of mapping the difficult scene and event concepts onto many simple and easy-for-recognition concepts.

For each training and testing image, the 4096 dimensional deep feature is extracted. The selected concept models are used to score the image into a 480 dimensional or 256 dimensional concept feature vector. Lib-linear SVM [42] is used for classifier training and testing where the penalty coefficient is set to 1. The results compared with other reported work are shown in Table IV. *CNN feature* indicates the classification performance directly using the 4096 dimensional deep feature. As shown, *CNN feature* shows impressive performance on the both image sets, which is consistent with the conclusion in recent work and validates our motivation to use deep feature for concept model training. We can see that, the proposed concept feature-based method has achieved comparable results on Scene-15 with state-of-the-art method in [11], which uses 14 200 concepts for mid-level representations. In contrast, we use a small subset of 256 concept models for image representation, which has the advantage of computational efficiency. on UIUC-Sport, the proposed concept feature outperforms all other methods. This demonstrates the effectiveness of the constructed FBVO, and its potential to serve as a concept vocabulary for complex visual recognition tasks.

B. Ontology-Based Exploratory Image Search

Exploratory search is recognized as an important way for information discovery, which indicates the activities users carry out when they have no specific search goals and do not yet know exactly what they want. Recent studies [49] show that image/multimedia search on the Web tends to be more "exploratory" than "deterministic" and requires heavy interactivity. A typical exploratory search starts when a user has interest in finding information on a topic with vague queries or broad search terms [50]. The search engine needs to understand the vague queries and help users clarify their intents via the interactive operations between search and browse.

Therefore, using the constructed FBVO, we design a solution for exploratory image search with two functions: 1) expanded query-based search, and 2) relevant query browse. Specifically, when a user inputs a concept query, we first locate the query in the FBVO and extract its top-k (in the experiment, we set k to 50) relevant concept nodes according the co-occurrent relationship score. The basic idea is to use the relevant concepts for query expansion. To determine the weight of each relevant concept in ranking the returned images, a lazy random walk with restart to the input concept is conducted in the concept hierarchy. Our method takes the relevant concept nodes as a graph, a self-loop probability β, and a start vector defined on the nodes of the graph. The random walk starts in the node corresponding to the query concept. At each step, it either remains in the current node

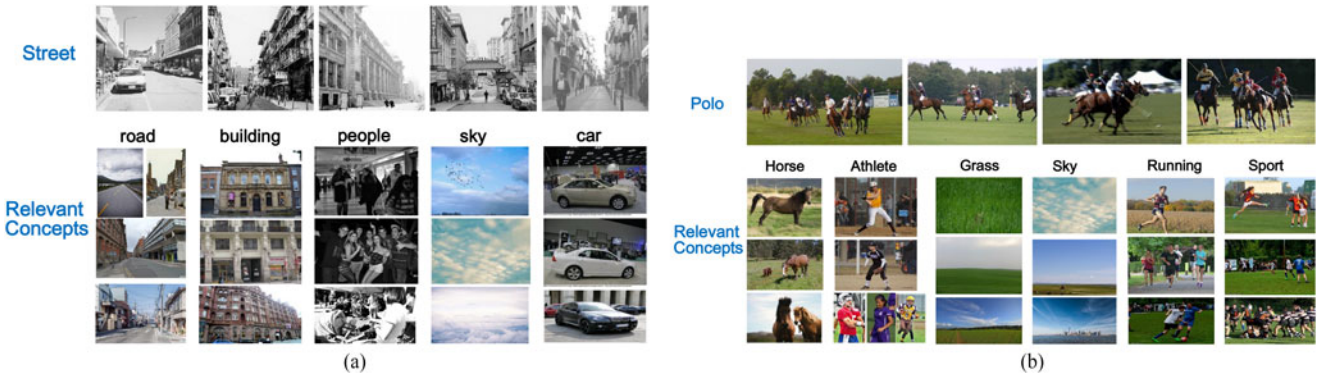


Fig. 8. Example categories of Scene-15 and UIUC Sport event and the relevant concepts in our FBVO. (a) Scene-15: Street. (b) UIUC-Sport: Polo.

TABLE IV
CLASSIFICATION ACCURACIES ON SCENE-15 AND UIUC-SPORT DATASETS

	Scene-15	UIUC-Sport
Object Bank [26]	80.9%	76.3%
LScSPM [47]	80.28%	82.74%
KSPM [48]	81.4%	–
Li *et al.* [11]	85.4%	88.4%
CNN feature	84.23%	92.83%
The proposed concept feature	85.05%	94.06%

Fig. 9. Examples of returned concepts and images for query "architecture" and "animal."

with probability β, or moves with one of the out-links with probability $1-\beta$. In the latter case, the links are followed with probability proportional to the weights of the edge w_e between the two concepts. Self-transitions are allowed to reinforce the importance of the starting node, by slowing diffusion to other nodes. The value of the self loop probability is set to $\beta = 0.9$, following the previous work [51]. The random walk is stopped after achieving a maximum iterations and we obtain the score $weight(c)$ for each relevant concept node c. To generate the returned images for the input query, we extract top-k relevant images for each concept node and calculate the relevant score for image I as $RW(I) = \Psi_c(\mathbf{x}_I) \cdot weight(c)$, where $\Psi_c(\mathbf{x}_I)$ is the response score of concept model Ψ_c on image I. Finally, the relevant concepts as well as the images are returned to the users in a descending order according to $weight(c)$ and $RW(I)$, respectively. Fig. 9 shows the examples to two concept queries. We can see that the returned images are both relevant to the query and having an adequate diversity. If the user clicks a presented relevant concept, the above search algorithm is to be conducted with the new concept as the input query and to return new relevant concepts and images. This process can be interactively performed along with the user's search-browse exploratory search activity.

For evaluation, ten common concepts are selected as the test queries, i.e., *architecture, animal, bird, flower, food, people, sports, sunset, travel, portrait*. To obtain the ground truth of the relevant orders of the returned images, we resort to a manual labeling procedure. Specifically, each image is labeled as three relevance levels with respect to the query: Highly Relevant (score 2), Relevant (score 1), and Irrelevant (score 0). We invite five subjects to manually label the relevance levels of the returned images. Each image is labeled by at least three subjects. The ground truth is obtained through the majority voting of subjects' labeling. For evaluation metric, we use normalized discounted cumulative gain (NDCG), where NDCG at position k is defined as

$$\text{NDCG@}k = \frac{1}{\text{Z}} \times \sum_{i=1}^{k} \frac{2^{r_i} - 1}{\log_2(i+1)} \tag{9}$$

where r_i is the relevance score of the sample at rank i. Z is a normalization term so that NDCG@$k \in [0, 1]$. We compare our lazy random walk based method with the three baseline methods: 1) Random search, searching images randomly from original images associated with the input concept tag; 2) Relevant search, returning images from the positive image set T_c; 3) Concept model based search, scoring relevant images using the responses from the learned concept models. The average NDCG results with different search depths K are illustrated in Fig. 10. It is shown that equipped with the constructed FBVO, we can understand the initial query with relevant concepts and return diversified images to meet the exploratory intents.

Moreover, users are allowed to browse and move to other concepts along the constructed concept hierarchy to explore interesting images. We further conduct a small-scale user study to evaluate the user experience of the proposed exploratory image search solution. The five labeling subjects are asked to assess the relevance, diversity of the search results, and the exploratory usability of the search system. The rate has a scale of 1 to 5, where 1 is the worst and 5 is the best. The averaged results are shown in Table V. It is shown that the five subjects generally gave positive feedbacks to the proposed exploratory search solution. This validates the potential of the constructed FBVO in advanced image search tasks.

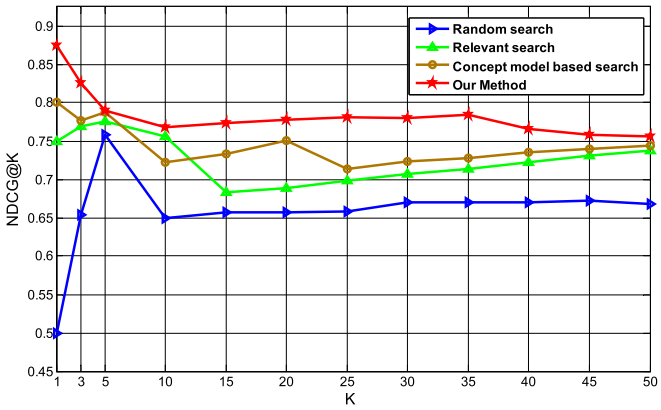


Fig. 10. Comparison of NDCG@K using different image search methods.

TABLE V
USER STUDY RESULTS FOR EXPLORATORY IMAGE SEARCH

Search Relevance	Search Diversity	Exploratory Usability
3.8	4.2	4.3

VI. CONCLUSION

In this paper, we have introduced a systematic solution for automatically constructing a visual ontology from folksonomy-based images. The constructed visual ontology consists of rich concept set, pairwise concept relationships and hierarchical structure of concepts. The learned models corresponding to the concept nodes are capable of recognizing new images to update the ontology. The extracted concept relationships and semantic hierarchy is demonstrated to be consistent with human cognition. The utility and potential of the constructed visual ontology are further examined and validated through two carefully designed applications. Our future work includes: 1) implementing an online never-end version of the proposed FBVO framework; 2) integrating other data source to form a multimodal knowledge base; and 3) applying FBVO in more applications such as multimedia question-answering.

REFERENCES

[1] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41. 1995.
[2] M. R. Naphade *et al.*, "Large-scale concept ontology for multimedia," *IEEE MultiMedia*, vol. 13, no. 3, pp. 86–91, Jul.-Sep. 2006.
[3] D. Fensel, *Ontologies*. New York, NY, USA: Springer, 2001.
[4] M. Marszalek and C. Schmid, "Semantic hierarchies for visual object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–7.
[5] J. Fan, H. Luo, Y. Gao, and R. Jain, "Incorporating concept ontology for hierarchical video classification, annotation, and visualization," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 939–957, Aug. 2007.
[6] J. Fan, Y. Gao, and H. Luo, "Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation," *IEEE Trans. Image Process.*, vol. 17, no. 3, pp. 407–426, Mar. 2008.
[7] X. Wei, C. Ngo, and Y. Jiang, "Selection of concept detectors for video search by ontology-enriched semantic spaces," *IEEE Trans. Multimedia*, vol. 10, no. 6, pp. 1085–1096, Oct. 2008.
[8] C. G. M. Snoek *et al.*, "Adding semantics to detectors for video retrieval," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 975–986, Aug. 2007.
[9] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 2, no. 1, pp. 1–19, 2006.
[10] D. Oberle, "How ontologies benefit enterprise applications," *Semantic Web*, vol. 5, no. 6, pp. 473–491, 2014.
[11] Q. Li, J. Wu, and Z. Tu, "Harvesting mid-level visual concepts from large-scale internet images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 851–858.
[12] M. Sanderson and B. Croft, "Deriving concept hierarchies from text," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1999, pp. 206–213.
[13] H. Alani *et al.*, "Automatic ontology-based knowledge extraction from web documents," *IEEE Intell. Syst.*, vol. 18, no. 1, pp. 14–21, Jan./Feb. 2003.
[14] A. Plangprasopchok, K. Lerman, and L. Getoor, "Growing a tree in the forest: Constructing folksonomies by integrating structured metadata," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 949–958.
[15] L. Wu, X. Hua, N. Yu, W. Ma, and S. Li, "Flickr distance: A relationship measure for visual concepts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 863–875, May 2012.
[16] X. Wang, Z. Xu, L. Zhang, C. Liu, and Y. Rui, "Towards indexing representative images on the web," in *Proc. ACM Multimedia*, 2012, pp. 1229–1238.
[17] X. Chen, A. Shrivastava, and A. Gupta, "NEIL: Extracting visual knowledge from web data," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1409–1416.
[18] J. Fan, Y. Gao, H. Luo, and R. Jain, "Mining multilevel image semantics via hierarchical classification," *IEEE Trans. Multimedia*, vol. 10, no. 2, pp. 167–187, Feb. 2008.
[19] M. Koskela, A. F. Smeaton, and J. Laaksonen, "Measuring concept similarities in multimedia ontologies: Analysis and evaluations," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 912–922, Aug. 2007.
[20] P. F. Felzenszwalb, D. A. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.
[21] V. Ferrari and A. Zisserman, "Learning visual attributes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 433–440.
[22] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[23] G. Patterson, C. Xu, H. Su, and J. Hays, "The SUN attribute database: Beyond categories for deeper scene understanding," *Int. J. Comput. Vis.*, vol. 108, no. 1/2, pp. 59–81, 2014.
[24] D. Borth, T. Chen, R. Ji, and S. Chang, "Sentibank: Large-scale ontology and classifiers for detecting sentiment and emotions in visual content," in *Proc. ACM Multimedia*, 2013, pp. 459–460.
[25] Q. Fang, J. Sang, and C. Xu, "Discovering geo-informative attributes for location recognition and exploration," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 11 no. 1s. pp. 19:1–19:23, Oct. 2014.
[26] L. Li and F. Li, "What, where and who? Classifying events by scene and object recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
[27] P. Over *et al.*, "Trecvid 2014—An overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proc. TRECVID*, 2014, pp. 52.
[28] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge" *CoRR*. 2014. [Online]. Available: http://arxiv.org/abs/1409.0575
[29] P. Mylonas, E. Spyrou, Y. S. Avrithis, and S. D. Kollias, "Using visual context and region semantics for high-level concept detection," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 229–243, Feb. 2009.
[30] J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha, "Real-time visual concept classification," *IEEE Trans. Multimedia*, vol. 12, no. 7, pp. 665–681, Nov. 2010.
[31] R. Ewerth, K. Ballafkir, M. Mühling, D. Seiler, and B. Freisleben, "Long-term incremental web-supervised learning of visual concepts via random savannas," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1008–1020, Aug. 2012.
[32] S. Zhu, C. Ngo, and Y. Jiang, "Sampling and ontologically pooling web images for visual concept learning," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1068–1078, Aug. 2012.
[33] S. K. Divvala, A. Farhadi, and C. Guestrin, "Learning everything about anything: Webly-supervised visual concept learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 3270–3277.

[34] T. L. Berg, A. C. Berg, and J. Shih, "Automatic attribute discovery and characterization from noisy web data," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 663–676.
[35] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 248–255.
[36] Y. Lu, L. Zhang, J. Liu, and Q. Tian, "Constructing concept lexica with small semantic gaps," *IEEE Trans. Multimedia*, vol. 12, no. 4, pp. 288–299, Jun. 2010.
[37] H. Zhang *et al.*, "Attribute-augmented semantic hierarchy: Towards bridging semantic gap and intention gap in image retrieval," in *Proc. ACM Multimedia*, 2013, pp. 33–42.
[38] C. Bishop, *Pattern Recognition And Machine Learning*, ser. Inform. Sci. and Statist. New York, NY, USA: Springer, 2006. [Online]. Available: http://books.google.com.sg/books?id=kTNoQgAACAAJ
[39] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Multimedia*, 2014, pp. 675–678.
[40] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, 2007.
[41] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Proc. Mach. Learn*, 2004, pp. 39–50.
[42] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learning Res*, vol. 9, pp. 1871–1874, 2008.
[43] R. Cilibrasi and P. M. B. Vitányi, "The Google similarity distance," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 370–383, Mar. 2007.
[44] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proc. ACM Multimedia Inf. Retrieval*, 2008, pp. 39–43.
[45] S. Bird, "NLTK: The natural language toolkit," in *Proc. 44th Annu. Meeting Assoc. Comput. Linguistics*, 2006, pp. 69–72.
[46] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 776–789.
[47] J. Yang, K. Yu, Y. Gong, and T. S. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 1794–1801.
[48] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 2, pp. 2169–2178.
[49] N. O'Hare, J. Y. Park, R. Schifanella, A. Jaimes, and C.-W. Chung, "A large-scale study of user image search behavior on the web," in *Proc. 33rd Annu. ACM Conf. Human Factors Comput. Syst.*, 2015, pp. 985–994.
[50] R. W. White and R. A. Roth, *Exploratory Search: Beyond the Query-Response Paradigm*. ser. Synthesis Lectures on Inform. Concepts, Retrieval, and Services. San Rafael, CA, USA: Morgan & Claypool, 2009. [Online]. Available: http://dx.doi.org/10.2200/S00174ED1V01Y200901ICR003
[51] N. Craswell and M. Szummer, "Random walks on the click graph," in *Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2007, pp. 239–246.

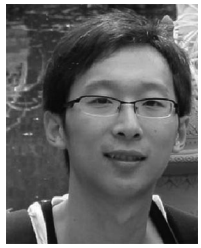
Quan Fang received the B.E. degree from Beihang University, Beijing, China, in 2010, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2015.

His research interests include georeferenced social media mining and application, multimedia content analysis, knowledge mining, computer vision, and pattern recognition.

Mr. Fang was the recipient of the 2013 Microsoft Research Asia Fellowship. He was the recipient of the Best Student Paper in Internet Multimedia Modeling 2013 and was the Best Paper Finalist in ACM Multimedia 2013.

Changsheng Xu (M'97–SM'99–F'14) is a Professor with the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences and the Executive Director of China-Singapore Institute of Digital Media, Singapore. He holds 30 granted/pending patents and has authored or coauthored more than 200 refereed research papers. His research interests include multimedia content anaysis/indexing/retrieval, pattern recognition, and computer vision.

Prof. Xu is a Fellow of the IAPR and an ACM Distinguished Scientist. He is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, *ACM Transactions on Multimedia Computing*, *Communications and Applications*, and *ACM/Springer Multimedia Systems Journal*. He was the recipient of the Best Associate Editor Award of *ACM Transactions on Multimedia Computing, Communications and Applications* in 2012 and the Best Editorial Member Award of *ACM/Springer Multimedia Systems Journal* in 2008. He served as the Program Chair of *ACM Multimedia* 2009. He has served as an Associate Editor, Guest Editor, General Chair, Program Chair, Area/Track Chair, Special Session Organizer, Session Chair, and TPC Member for more than 20 IEEE and ACM multimedia journals, conferences, and workshops.



Jitao Sang received the B.E. degree from the South-East University, Nanjing, China, in 2007, and the Ph.D. degree from the Chinese Academy of Sciences (CASIA), Beijing, China, in 2012.

He is an Assistant Professor with the National Laboratory of Pattern Recognition, Institute of Automation, CASIA. He has authored or coauthored several refereed research papers. His research interests include multimedia content analysis, social media mining, and social network analysis.

Prof. Sang has served as a Special Session Organizer in MMM 2013 and Publication Chair in ACM ICIMCS 2013 and 2014. He was the recipient of the Special Prize of President Scholarship by Chinese Academy of Sciences. He was the coauthor of the recipient of the Best Student Paper in Internet Multimedia Modeling 2013 and the Best Paper Candidate in ACM Multimedia 2012 and 2013.

M. Shamim Hossain (S'03–M'07–SM'09) received the Ph.D. degree in electrical and computer engineering from the University of Ottawa, Ottawa, ON, Canada, in 2009.

He is an Associate Professor with the Department Software Engineering, King Saud University, Riyadh, Saudi Arabia. He has authored or coauthored more than 100 publications, including refereed IEEE/ACM/Springer/Elsevier journals, conference papers, books, and book chapters. His research interests include serious games, social media, cloud and multimedia for health care, resource provisioning for big data processing on media clouds, and biologically inspired approaches for multimedia and software system.

Prof. Shamim Hossain has served as a Co-Chair, the General Chair, the Workshop Chair, the Publication Chair, and the Technical Program Committee Chair for more than 12 IEEE and ACM conferences and workshops. He served as a Guest Editor for the IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, *International Journal of Distributed Sensor Networks*, and the *International Journal of Multimedia tools and Applications* (Springer). He currently serves on the editorial board of *International Journal of Multimedia Tools and Applications*, and as a Lead Guest Editor for the IEEE TRANSACTIONS ON CLOUD COMPUTING, *Future Generation Computer Systems* (Elsevier), *Computers and Electrical Engineering* (Elsevier), and *Cluster Computing* (Springer).

Ahmed Ghoneim (M'10) received the M.Sc. degree in software modeling from the University of Menoufia, Shebeen El-Kom, Egypt, in 1999, and the Ph.D. degree in software engineering from the University of Magdeburg, Magdeburg, Germany, in 2007.

He is currently an Assistant Professor with the Department of Software Engineering, King Saud University, Riyadh, Saudi Arabia. His research activities include software evolution, service oriented engineering, software development methodologies, quality of services, net-centric computing, and human–computer interaction.