Attribute-Augmented Semantic Hierarchy: Towards a Unified Framework for Content-Based Image Retrieval

HANWANG ZHANG, National University of Singapore
ZHENG-JUN ZHA, Chinese Academy of Sciences
YANG YANG, SHUICHENG YAN, YUE GAO, and TAT-SENG CHUA,
National University of Singapore

This article presents a novel attribute-augmented semantic hierarchy (A^2SH) and demonstrates its effectiveness in bridging both the semantic and intention gaps in content-based image retrieval (CBIR). A^2SH organizes semantic concepts into multiple semantic levels and augments each concept with a set of related attributes. The attributes are used to describe the multiple facets of the concept and act as the intermediate bridge connecting the concept and low-level visual content. An hierarchical semantic similarity function is learned to characterize the semantic similarities among images for retrieval. To better capture user search intent, a hybrid feedback mechanism is developed, which collects hybrid feedback on attributes and images. This feedback is then used to refine the search results based on A^2SH. We use A^2SH as a basis to develop a unified content-based image retrieval system. We conduct extensive experiments on a large-scale dataset of over one million Web images. Experimental results show that the proposed A^2SH can characterize the semantic affinities among images accurately and can shape user search intent quickly, leading to more accurate search results as compared to state-of-the-art CBIR solutions.

Categories and Subject Descriptors: H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval—*Retrieval models*

General Terms: Algorithms, Experimentation, Performance

Additional Key Words and Phrases: Image retrieval, attribute, semantic hierarchy

ACM Reference Format:
Hanwang Zhang, Zheng-Jun Zha, Yang Yang, Shuicheng Yan, Yue Gao, and Tat-Seng Chua. 2014. Attribute-augmented semantic hierarchy: Towards a unified framework for content-based image retrieval. ACM Trans. Multimedia Comput. Commun. Appl. 11, 1s, Article 21 (September 2014), 21 pages.
DOI: http://dx.doi.org/10.1145/2637291

1. INTRODUCTION

Content-based image retrieval (CBIR), a technique for retrieving images from a large database of digital images based on visual content, has been studied extensively since the early 1990s [Smith and Chang 1997]. It has gained increasing importance in both academia and industry because of the explosive growth of images shared in cyberspace and the compelling demands in various multimedia applications for Web and mobile clients [Lew et al. 2006; Datta et al. 2008]. In spite of the remarkable progress made in the last two decades, CBIR remains challenging mainly due to two critical scientific problems: (a) the "semantic gap" between the low-level visual features and high-level

This work is supported by the NUS-Tsinghua Extreme Search (NExT) project under grant #R-252-300-001-490.
Author's addresses: H. Zhang, National University of Singapore, Singapore; Z.-J. Zha (corresponding author), Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, China; email: junzzustc@gmail.com; Y. Yang, S. Yan, Y. Gao, and T.-S. Chua, National University of Singapore, Singapore.

DOI: http://dx.doi.org/10.1145/2637291

semantics [Smeulders et al. 2000], hindering the system interpretation of image content; and (b) the "intention gap" between user's search intent and the query at hand [Zha et al. 2009, 2010], hindering the system understanding of user's intent behind a query.

The cause of the semantic gap is that low-level visual features cannot correlate to high-level semantics accurately due to the large visual variance of image semantics [Zha et al. 2008]. Recent studies, especially those on TRECVID [Over et al. 2012; Zha et al. 2012], have shown that a promising route to narrowing the semantic gap is to exploit a set of concepts to form the semantic description of images. However, the performance of machine understanding of high-level concepts is still far from satisfactory. Besides the difficulty in modeling semantic concepts using visual features, a fundamental challenge is that a predefined concept lexicon with limited scale cannot generalize well to unseen domains. One may tackle this by increasing the size of the lexicon, but as Deng et al. [2010] have shown when they tried to classify 10K concepts, the accuracy drops to around 3.7% as compared to 77.1% on hundreds of concepts [Boureau et al. 2011]. A possible explanation is that the semantic gap may become much more significant as the scale of concepts increases.

The cause of the intention gap is much more difficult to delineate as it is dependent on subjective human interpretation. For example, even if a perfect vision system successfully detects the concepts of a query image of "car" and "people", it is still difficult for the system to know whether the user's intent is "car" or "people." Relevance feedback (RF) has been developed to address this problem [Crucianu et al. 2004; Rui et al. 1998]. In conventional RF, users are asked to label the top images returned by the search model as "relevant" or "irrelevant." Through iterative feedback and model refinement, RF attempts to capture users' information needs and improve the search results gradually. Although RF has shown encouraging potential in CBIR, its performance is usually unsatisfactory, because most of the RF solutions, in essence, rely on low-level visual features to infer high-level user intent, limiting their ineffectiveness in narrowing the search.

In this article, our goal is to develop a CBIR framework that addresses these two gaps. In particular, we build the framework based on a new type of semantics, named *attributes*. Here, attributes refer to semantic descriptions of concepts such as the visual appearances (e.g., "round" as shape, "metallic" as texture), subcomponents (e.g., "has wheel", "has leg"), and various discriminative properties (e.g., "properties that dog has but cat does not"). Different from the widely-used low-level features and high-level semantic concepts, attributes are intermediate-level feature representations of images. As compared to high-level concepts, attributes can be modeled more easily by a machine. For example, the visual variance of attribute "wheel" across many concepts like "car" and "carriage" is much smaller than that of the concepts. Compared to low-level features, attributes endow more semantic meanings and hence are more human-interpretable. For example, users can provide explicit negative feedback on the attribute "furry" given that most of the current search results are "furry dog." In our preliminary work [Zhang et al. 2012], we have shown that using attributes can successfully boost the performance of CBIR at a middle scale (e.g., 10K images). However, when the semantic diversity of the image repository scales up to real-world general-purpose CBIR, a list (or bag) of attributes becomes inefficient to represent image content as well as user intent. For example, the attribute "wing" is indispensable for the query image "bird," but it may also put irrelevant images which have "wing" closer, such as "airplane" and "insect." To the best of our knowledge, most attribute-based CBIR solutions are still limited in a particular domain where the attributes may not be ambiguous in describing concepts, such as fashion [Kovashka et al. 2012] or face [Kumar et al. 2011].

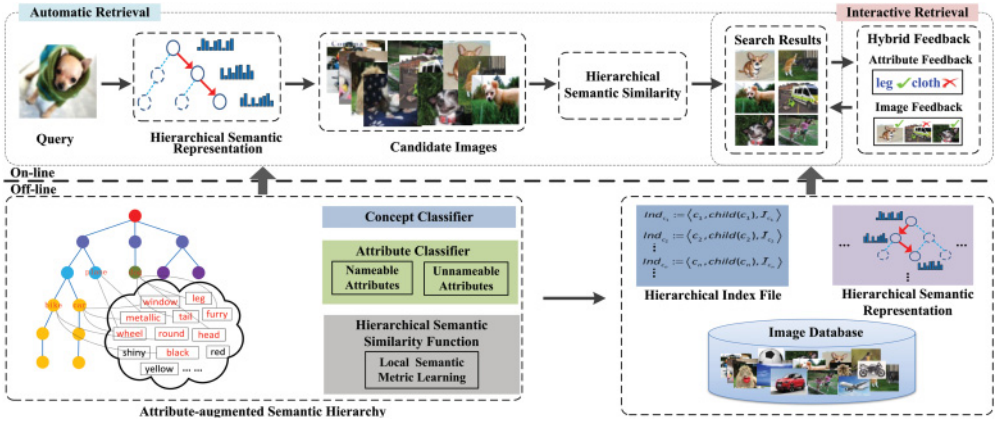


Fig. 1. The overview of the proposed attribute-augmented semantic hierarchy (A^2SH) and the unified image retrieval system developed based on A^2SH.

In order to build an attribute-based image retrieval system for general-purpose use, we propose a novel attributed-augmented semantic hierarchy (A^2SH), which is able to support developing a unified CBIR framework including both automatic and interactive CBIR. As shown in Figure 1, A^2SH is a semantic hierarchy of semantic concepts augmented by a pool of attributes. For example, "car" is augmented by the attributes "window" and "metallic", etc. At first glance, A^2SH is nothing new but a richer concept hierarchy decorated by attributes. However, we argue that A^2SH is essentially distinguishable from a traditional concept hierarchy as well as a bag of attributes in terms of the following two key characteristics.

—*Semantic Augmentation of Concepts*. Each concept in an existing semantic hierarchy can be delineated by a set of attributes. These attributes are explicitly descriptive semantics like some visual properties (e.g., "wheel", "metallic") or implicit discriminative semantics like subtle interconcept differences (e.g., "properties that cat has but dog does not"). By using such attributes, the semantics of concepts in an hierarchy are augmented. The reasons are twofold. First, the visual variance of attributes is much smaller and hence attributes can act as an intermediate bridge connecting concepts and low-level visual features. Second, as compared to the high dimensionality of low-level features, attributes serve as compact and detailed specifications for the multiple semantic facets of the corresponding concept.

—*Contextualization of Attributes*. As compared to an unstructured bag of attributes, the semantic hierarchy helps to define attributes with more specific semantic meanings in the context of concepts. Thus, attributes in the semantic hierarchy may have different semantics in the context of different concepts. For example, the attribute "*wing*" of concept "*bird*" refers to appendages that are feathered; while the same attribute refers to metallic appendages in the context of "*jet*." In this way, A^2SH can associate attributes to different concepts and reveal the heterogeneous meanings of the same attribute. This contextualization has two benefits for attribute-based CBIR. First, A^2SH interprets the semantics of image content with a hierarchical semantic representation, comprising multiple levels of semantic granulations. Second, user feedback on an attribute can refine the attribute representation of user intent only at related context, leading to more effective and efficient user feedback in terms of attributes.

The system based on A^2SH is illustrated in Figure 1. We expect A^2SH to be effective in narrowing both (a) the semantic gap by structurizing the semantics of image content with semantically meaningful hierarchical representations in terms of concepts and attributes and (b) the intention gap by shaping users' search intent accurately from the feedback. In the offline stage, we first learn concept classifiers, attribute classifiers, and a hierarchial semantic similarity function for A^2SH. We next use A^2SH to process images and obtain their hierarchical semantic representations. All images are indexed hierarchically based on their semantic paths in the hierarchy to facilitate efficient large-scale retrieval. In the online stage, the A^2SH first processes a given query image to derive its hierarchical semantic representation. A collection of candidate images are then returned from the database according to the index. Similar images are finally reranked from the candidate set based on their hierarchical semantic similarities to the query. After automatic retrieval, we present the results to solicit user feedback. We offer a broad channel of feedback mechanisms to help the user deliver search intent by providing hybrid feedback based on attributes and images. While the image feedback collects positive and negative samples of user intent, the feedback on attributes enable users to compose a clearer semantic description of their intent [Zhang et al. 2012], such as "*has head and legs, but not furry*." This hybrid feedback is analyzed by A^2SH, leading to a detailed semantic interpretation of user intent, which is then used to refine the search results. We expect the hybrid feedback to lead to better search results with less interaction effort. We evaluate the proposed system on a large-scale corpus of over one million Web images. The experimental results have demonstrated the superiority of the proposed system over state-of-the-art CBIR approaches. The main contributions of this article are summarized as follows.

—We propose a novel attribute-augmented semantic hierarchy (A^2SH) in which each concept is augmented by a set of related attributes. A^2SH models the semantics of images in the form of a hierarchical semantic representation, which is semantically meaningful.
—We learn a hierarchical semantic similarity function which is able to accurately characterize the semantic affinities among images. Moreover, we develop a hybrid feedback mechanism to collect feedback on both attributes and images, which can help to capture more details of users' search intent based on A^2SH.
—We develop a unified CBIR system based on the proposed A^2SH and demonstrate the effectiveness of the system in narrowing the semantic and intention gaps in image retrieval over a large-scale image dataset.

The rest of the article is organized as follows. Section 2 reviews related work. Section 3 describes the elementary building blocks of the proposed A^2SH, including concept classifiers, attribute classifiers, and the hierarchical semantic similarity function. Section 4 elaborates automatic and interactive image retrieval based on the proposed A^2SH. Experimental results and analysis are reported in Section 5, followed by conclusions and future work in Section 6.

2. RELATED WORK

In this section, we survey related work on the three key aspects of A^2SH.

2.1. Semantic Hierarchy

A semantic hierarchy is a formally defined taxonomy or ontology, where each node represents a semantic concept, such as WordNet [Fellbaum 2010], ImageNet [Deng et al. 2009], or LSCOM [Naphade et al. 2006], etc. It organizes semantic concepts from general to specific and has been shown to be effective in boosting visual recognition

[Marszalek and Schmid 2007] and retrieval [Deselaers and Ferrari 2011; Deng et al. 2011].

In A^2SH, we predict the semantic path of images by hierarchical concept classifiers. As compared to a bag of flat classifiers (e.g., a bag of one-vs.-all classifiers), hierarchical classifiers exploit the ontological relationship and thus significantly reduce the prediction time while producing comparatively good accuracy. For example, Griffin and Perona [2008] automatically built classification trees which achieve big speedup at a small performance drop. There are also studies which found that hierarchical top-down prediction using local classifiers can retain the same performance as one-vs.-all classifiers [Marszalek and Schmid 2007]. Here, "local" is a concept classifier training strategy that considers the images of the concept itself as positive and those of its siblings as negative. However, local classifiers may suffer severely from the error propagation problem. If a classifier at higher levels fails, it is hard for the successive classifiers at lower levels to stop the error (e.g., classify the error as negative), since they have never seen such an error as a negative sample. In order to alleviate this problem, Binder et al. [2012] developed a structure learning method that optimizes the performance of all local classifiers in a semantic hierarchy. However, the training cost of the method is prohibitive in the large-scale semantic hierarchy as employed in this work. For efficiency, we adopt an hierarchical one-vs.-all training strategy to learn local concept classifiers as in Song et al. [2010].

Most recent work that exploits semantic hierarchy for image retrieval focuses on designing a semantic similarity function that embeds hierarchical information. In the method proposed by Deselaers and Ferrari [2011], given two images, their visual nearest-neighbor images were first found, and then their semantic distance was computed as a distance between the concepts of their neighbors. Deng et al. [2011] developed an hierarchical bilinear similarity function for image retrieval. They first represented an image as a semantic vector $\mathbf{z}$ consisting of its relevances to a set of concepts and defined the bilinear similarity between any two images as $\mathbf{z}_i^T \mathbf{S} \mathbf{z}_j$, where $\mathbf{S}$ is a matrix encoding the pairwise semantic affinities among the concepts. They have shown that this method achieves state-of-the-art performance of image retrieval on ImageNet. Recently, Verma et al. [2012] proposed associating separated visual similarity metrics for every concept in an hierarchy and then learning the metrics jointly through an aggregated hierarchical metric. Differenting from existing work based on conventional semantic hierarchy, our work augments the semantic hierarchy with a pool of attributes. The augmented hierarchy has better capability in modeling the semantics of images. We characterize the semantic similarities among images by a hierarchical similarity function composed of a set of local semantic metrics learned in the semantic spaces spanned by attributes in the context of various concepts.

2.2. Attributes

Attributes have attracted increasing attention recently. Attributes refer to semantic descriptions of concepts, such as subcomponents (e.g., "has wheel"), visual appearances (e.g., "round" as shape), and various discriminative properties (e.g., "properties that dog has but cat does not"). Attributes are semantically meaningful as opposed to low-level visual features, and they share common features and thus are relatively easy to recognize automatically instead of the full concepts (e.g., "dog", "car") [Farhadi et al. 2009]. Attributes can be exploited as intermediate-level descriptors of images to boost visual recognition [Farhadi et al. 2009; Ma et al. 2012; Yu et al. 2013] and retrieval [Douze et al. 2011]. For example, Jaimes and Chang [2000] proposed a conceptual framework for indexing visual information based on semantic concepts and attributes. Douze et al. [2011] proposed representing an image by the concatenation of visual features and its response from attribute classifiers and have shown that such representation can

improve the performance of image retrieval significantly compared to using purely visual features. Scheirer et al. [2012] developed a probabilistic normalization method for normalizing the responses from attribute classifiers and have shown that the normalized representation is more effective. While most existing work represented images in the form of a flat semantic representations in terms of attributes, our work associates attributes to the concepts in a semantic hierarchy and represents images in the form of a hierarchical semantic representations which are more semantically meaningful. Moreover, as aforementioned, associating an attribute to concepts can reveal the heterogeneous meanings of the same attribute.

There are some recent studies on enhancing the semantic meanings of attributes. Parikh and Grauman [2011b] illustrated the idea of relative attributes towards endowing comparative semantics of attributes, for example, attributes like "shinier", "younger." However, the training of relative classifiers requires highly expensive annotation and computation of pairwise samples at certain relativity. Kumar et al. [2011] proposed defining a set of concept-dependent attributes, such as "Angelina Jolie's mouth," to discriminate between different "mouths" shared by people. Farhadi et al. [2009] introduced another kind of concept-dependent attributes. They exhaustively collected a huge number of attributes according to the combinatorial splits between any subsets of the concepts, for example, attributes that "cat" and "dog" have but "sheep" and "horse" do not. In our work, we call such attributes unnameable attributes that aim to discriminate between concepts. In particular, our unnameable discovery is inspired by Parikh and Grauman [2011a], where the unnameable attributes were discovered by discriminating images from the most confused concepts.

2.3. Relevance Feedback

Relevance feedback (RF) is a key technique for narrowing down the intention gap in image retrieval [Rui et al. 1999; Datta et al. 2008]. It attempts to capture users' search intent by iteratively collecting users' feedback on retrieved images and refining retrieval based on the feedback. A wealth of RF methods has been proposed to refine the original query or learn a relevance ranking function from the feedback, that is, "relevant" and "irrelevant" images labeled by users. For example, the Query Point Movement (QPM) method [Rui et al. 1998] gradually refines the query point by moving it towards the "relevant" images and away from the "irrelevant" images. Tong and Chang [2001] proposed learning a support vector machine (SVM) from the "relevant" and "irrelevant" images, and then ranking the images according to their responses from the SVM classifier. Tao et al. [2006] proposed an asymmetric bagging and random subspace SVM method for learning a robust classifier from user feedback. Yang et al. [2005] proposed providing semantic feedback on images with semantic concept labels, expecting to connect the high-level concept space to low-level feature space. However, this method requires database images with semantic labels and thus is not practical in content-based image retrieval.

Recently, feedback on attribute-level semantics has been proposed to act as a bridge reliably connecting users' intent and visual features. Zhang et al. [2012] developed an attribute feedback scheme which collects user feedback on semantic attributes and ranks images according to the presence probabilities of attributes in the images. Kovashka et al. [2012] proposed collecting attribute feedback on relative attribute judgements, for example, "*show me shoes more formal than these and shinier than those.*" Their approach is for text-based image retrieval as it requires users to input text as relative feedback. Compared to these works, our work enables a broad channel of feedback to help users better delivery search intent by providing hybrid feedback on attributes and images. By modeling the feedback based on A^2SH, our work leads to a better understanding of user intent.

3. ATTRIBUTE-AUGMENTED SEMANTIC HIERARCHY

Attribute-augmented semantic hierarchy (A^2SH), denoted as $\mathcal{H} = (\mathcal{C}, \mathcal{A}, \mathcal{E}_C, \mathcal{E}_{CA})$, is a directed acyclic graph consisting of a set of concepts $\mathcal{C} = \{c\}$, a pool of attributes $\mathcal{A} = \{a\}$, a set of concept-concept edges $\mathcal{E}_C$, where an edge is an *ordered* pair of concepts in $\mathcal{C} \times \mathcal{C}$, and a set of concept-attribute edges $\mathcal{E}_{CA}$, where an edge is an *unordered* pair of a concept and an attribute in $\mathcal{C} \times \mathcal{A}$. The set of attributes linked to concept c is $\mathcal{A}_c$.

A^2SH organizes semantic concepts as well as associated attributes from general to specific. In order to represent and organize images using A^2SH, we need to automatically populate images into it. Therefore, we need to learn concept and attribute classifiers hierarchically. By doing this, a given image can be represented as the responses from the concept classifiers as well as the linked attribute classifiers, leading to an hierarchical semantic interpretation consisting of semantics at multiple levels. Furthermore, in order to facilitate image retrieval, we define a hierarchical similarity function for any two images. This similarity is an aggregation of local similarities in terms of attributes, along the common semantic paths of the two images.

3.1. Hierarchical Concept and Attribute Learning

3.1.1. Concept Classifiers. A concept classifier $f_c : \mathcal{X} \mapsto \{-1, +1\}$ predicts whether an image belongs to concept c, where $\mathcal{X}$ is a high-dimensional sparse visual feature that is widely used in many state-of-the-art visual modeling approaches (cf. Section 5.1.1). Given the concept classifiers in an hierarchy, the semantic path of an image can be efficiently predicted by the classifiers in a top-down fashion [Song et al. 2010; Marszalek and Schmid 2007]. Here, a semantic path $\mathcal{P}$ is a set of multilevel semantics $\mathcal{P} = (c_0 \rightarrow \cdots \rightarrow c_n)$ from the root c_0 and satisfies $\forall i > 0$, $f_{c_i}(x) = +1$. Next, we introduce the learning of concept classifiers. One way to learn each concept classifier is to use the conventional "one-vs.-all" strategy, that is, by training images from the concept as positive samples and images from the rest of concepts as negative samples [Wang et al. 2010]. However, this strategy neglects the hierarchical relation among concepts, resulting in classifiers ineffective for hierarchical classification. Another way for concept classifier learning is to locally train the classifier for a concept by using the images from its siblings as negative samples [Marszalek and Schmid 2007]. However, this local training strategy results in classifiers that suffer from the "error prorogation" problem. To address these problems, we adopt the "hierarchical one-vs.-all" strategy [Song et al. 2010] to learn concept classifiers by exploiting the hierarchical relationship among concepts and collecting training samples globally in the hierarchy. In particular, the positive training set $Pos(c)$ and the negative training set $Neg(c)$ are constructed as follows:

$$Pos(c) = \{I_i,\ s.t.\ \mathcal{L}(I_i) \cap (c \cup descend(c))\}, \quad Neg(c) = \{I_i,\ s.t.\ I_i \notin Pos(c)\}, \tag{1}$$

where $\mathcal{L}(I_i) \subseteq \mathcal{C}$ is the set of concept labels for sample I_i. For each concept c, the positive training set $Pos(c)$ consists of images labeled as either the concept itself or one of its descendant concepts, while the negative training set $Neg(c)$ contains images that are not in $Pos(c)$. Based on $Pos(c)$ and $Neg(c)$, we train a binary linear support vector machine (SVM) as the concept classifier f_c.

3.1.2. Attribute Classifiers. An attribute classifier under the context of concept c is denoted as $f_a^c : \mathcal{X}_a^c \mapsto \{-1, +1\}$, predicting the presence of attribute a of concept c in an image. Here, the feature space $\mathcal{X}_a^c$ is a discriminative visual feature space that is specially selected for attribute learning [Farhadi et al. 2009]. Note that the attributes are considered in the context of concepts, so the training images of attribute classifiers are constrained within the positive images, that is, $Pos(c)$. Specifically, given positive images where a certain attribute is present and negative images where the attribute

Fig. 2. Discovered unnameable attributes in the context of three concepts. From left to right, their ImageNet synsets are as follows: dog (n02084071), automotive vehicle (ID: n03791235), and seed plant (n13134947). For each concept, two sets of sample images (top 10 confidence score each) depicting the corresponding unnameable attributes, shown in two different rows separated by dashed lines. We may find some interesting interpretations. Left: dogs with longer fur and shorter fur; middle: hatchback cars and sedan cars; right: plants images taken from different angles, that is, perspective view and close-up view.

is absent, we can train a binary classifier for the attribute (e.g., SVM). Note that attributes normally correspond to partial visual cues of the whole image, so the whole feature used in concept learning may not characterize the attributes well. For example, a component attribute may only appear at one or more regions in the image, and an appearance attribute may correspond to only partial channels of visual descriptors. Therefore, we propose a hierarchical attribute feature selection strategy that can select the most informative features $\mathcal{X}_a^c$ for attributes from *specific to general*, gaining more visual cues. In particular, we propose selecting $\mathcal{X}_a^c$ in a bottom-up manner as follows. First, without loss of generality, suppose we have the selected features $\mathcal{X}_a^{c'}$ at hand, where $c' \in child(c)$ is a child concept of c. Second, we deploy the method of Farhadi et al. [2009] to select the most discriminative attribute features $\widetilde{\mathcal{X}}_a^{c'}$ within each concept c'. Third, we merge these features to obtain the resultant feature $\mathcal{X}_a^c$. Finally, we can repeat this procedure from bottom to top in the semantic hierarchy. For more details, please refer to Zhang et al. [2013].

3.1.3. Unnameable Attributes. Despite the expensive human labor in attribute labeling, the attributes we have discussed so far are only defined by domain experts or researchers of specific interest (e.g., like us), so we do not yet have a comprehensive set of them. This situation becomes more serious when we descend to lower specific concepts in the hierarchy. For example, dog breeds like "terrier" and "corgi" may share all the attributes such as "furry" and "tail." Therefore, we need auxiliary discriminative attributes. Figure 2 illustrates three such attributes: as we can see, unless we define specific terminologies for them, it is hard for casual users to name the semantics. In this article, we call such attributes *unnameable attributes*, since they are hard to be articulated explicitly by (casual) users. In order to distinguish these from predefined attributes, we call them *nameable attributes*, since they are straightforwardly named by humans. Note that both unnameable and nameable attributes offer a comprehensive description of the multiple facets of a concept. Inspired by Parikh and Grauman [2011a], we define unnameable attributes as the hypotheses that help in distinguishing each concept and its siblings in A^2SH. Algorithm 1 lists the detailed discovery procedure for any concept. Note that since unnameable attributes are automatically discovered to enhance interconcept separability, we use the same visual features as in concept learning.

3.2. Hierarchical Image Similarity Learning

From the concept and attribute classifiers previously learned, we can generate a hierarchical semantic representation of an image as $\{(c_0 \rightarrow \cdots \rightarrow c_n); (\mathbf{z}^{c_0}, \ldots, \mathbf{z}^{c_n})\}$, where $(c_0 \rightarrow \cdots \rightarrow c_n)$ is the semantic path predicted by concept classifiers, c_0 is the root of the hierarchy, and $(\mathbf{z}^{c_0}, \ldots, \mathbf{z}^{c_n})$ is the local semantic representation in terms of attributes along the path. Specifically, $\mathbf{z}^c$ is composed from the responses from the attribute classifiers of concept c as $\mathbf{z}^c = [f_{a_1}^c(\mathbf{x}), f_{a_2}^c(\mathbf{x}), \ldots, f_{a_{|\mathcal{A}_c|}}^c(\mathbf{x})]^T$, where f_a^c is the classifier for attribute $a \in \mathcal{A}_c$.

ALGORITHM 1: Unnameable Attribute Discovery

Input : Initial attributes $\mathcal{A} = \{a_i\}$ and the corresponding attribute models $\mathcal{M} = \{f_{a_i}\}$.
Images $\mathcal{I} = \{I_i\}$ labeled with any concept and its siblings, denoted as $\mathcal{Y}$.
Output: A comprehensive attributes $\mathcal{A}^*$ and its attribute models $\mathcal{M}^*$.

Initialization: $t = 1$, set $\mathcal{A}_t = \mathcal{A}$, $\mathcal{M}_t = \mathcal{M}$;
repeat
  Represent any image I_i as vector $[f_{a_1}(\mathbf{x}_i), \ldots, f_{a_{|\mathcal{A}_t|}}(\mathbf{x}_i)]$;
  Train a classifier to classify images $\mathcal{I}$ into $|\mathcal{Y}|$ concepts;
  Construct the confusion matrix $\mathbf{C}$ of the Y concepts and let $\mathbf{C} \leftarrow \frac{1}{2}(\mathbf{C} + \mathbf{C}^T)$;
  Apply Laplacian Eigenmap [Belkin and Niyogi 2003] for $\mathbf{C}$, collect V eigenvectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_V\}$ such that the sum of the V corresponding eigenvalues are no larger than 20% of the sum of all the eigenvalues;
  Apply K-means algorithm on matrix $[\mathbf{v}_1^T; \cdots; \mathbf{v}_V^T]$ into V clusters $\{\mathcal{V}_i\}$, each $\mathcal{V}_i$ is considered as a most confused set of concepts;
  for $i = 1$ **to** V **do**
    Split $\mathcal{V}_i$ into 2 groups by Max-Margin Clustering [Zhang et al. 2009];
    Train a binary SVM classifier f_{a_i} from the 2 groups, where a_i is the name of this binary membership, *i.e.*, a new discovered unnameable attribute;
    $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_t \cup a_i$, $\mathcal{M}_{t+1} \leftarrow \mathcal{M}_t \cup f_{a_i}$;
  end
  $t \leftarrow t + 1$;
until *no change to* $\mathcal{A}_t$ *or* $\mathcal{M}_t$;
Return: $\mathcal{A}^* = \mathcal{A}_t$, $\mathcal{M}^* = \mathcal{M}_t$;

With such a hierarchical semantic representation of images, we formulate a hierarchical semantic similarity function to characterize semantic similarities between images by aggregating their local similarities along their common semantic paths. The hierarchical semantic similarity between any two images is defined as follows:

$$S(I_i, I_j) = \sum_{c \in \mathcal{P}_{ij}} s(I_i, I_j; c), \tag{2}$$

where $\mathcal{P}_{ij}$ is the common semantic path of image I_i and I_j, $s(I_i, I_j; c)$ is the local similarity between I_i and I_j in the context of c along the path $\mathcal{P}_{ij}$.

There are two conventional ways to define $s(I_i, I_j; c)$. The first is to set $s(I_i, I_j; c)$ to 1, such that $S(I_i, I_j)$ is reduced to the length of the common path of I_i and I_j. This approach, however, lacks fine characterization of semantic affinities between the images along the path. The second is to calculate $s(I_i, I_j; c)$ as the visual similarity. This measurement suffers from the discrepancy between visual similarity and semantic similarity. Hence, they are both unable to characterize the semantic affinities between images well. In order to characterize the intrinsic semantic similarities between images, we propose learning a local semantic metric in the local semantic space of each concept.

We define the local semantic similarity between two images in the local semantic space of concept c in Eq. (2) as $s(I_i, I_j; c) = exp(-d(\mathbf{z}_i^c, \mathbf{z}_j^c; c))$, where $d(\mathbf{z}_i^c, \mathbf{z}_j^c; c) = \sqrt{(\mathbf{z}_i^c - \mathbf{z}_j^c)^T \mathbf{M}_c (\mathbf{z}_i^c - \mathbf{z}_j^c)}$, and $\mathbf{M}_c$ is a positive semi-definite symmetric matrix of size $|\mathcal{A}_c| \times |\mathcal{A}_c|$. $\mathbf{M}_c$ is the local semantic metric, which needs to be learned to bring together the images of the same concept as close as possible and separate the images of different concepts as far as possible. In particular, with $\mathbf{M}_c$, we expect the neighbor samples within the same semantic class c to be as close as possible towards preserving the fine neighborhood relation within the class, while the samples from the siblings of c to be separated away with a large margin. To achieve this, for image $I_i \in Pos(c)$, we require

the distance between I_i and its K-nearest neighbors $I_j \in Pos(c)$ to be as small as possible. $Pos(c)$ is the set of images belonging concept c. We denote $j \rightsquigarrow i$ as such a neighborhood. Moreover, the distance between I_i and I_j should be smaller than that between I_i and any image I_k from sibling concepts. Let $\mathcal{S}(c)$ denote the set of images of sibling concepts, we can have a set of training triples as $\mathcal{T} = \{(i, j, k) : j \rightsquigarrow i, I_i \in Pos(c), I_k \in \mathcal{S}(c)\}$, based on which we formulate the metric learning objective as follows:

$$\min_{\mathbf{M}_c} \sum_{j \rightsquigarrow i} d^2\left(\mathbf{z}_i^c, \mathbf{z}_j^c; c\right) + \lambda \sum_{(i,j,k)\in\mathcal{T}} \xi_{ijk},\ s.t.\ d^2\left(\mathbf{z}_i^c, \mathbf{z}_k^c; c\right) - d^2\left(\mathbf{z}_i^c, \mathbf{z}_j^c; c\right) \geq 1 - \xi_{ijk}, \xi_{ijk} \geq 0, \mathbf{M}_c \succeq \mathbf{0}, \tag{3}$$

where $\lambda > 0$ is the regularization constant. We employ the LMNN solver [Weinberger et al. 2006] modified with the previously defined training triplets $\mathcal{T}$ to solve the metric learning problem. Note that solving the preceding problem is very efficient, since the local semantic space is compact, that is, the dimension of $\mathbf{M}_c$ is low. With $\mathbf{M}_c$, we can compute the local semantic similarity between images as $s(I_i, I_j; c) = exp(-d(\mathbf{z}_i^c, \mathbf{z}_j^c; c))$, which is in turn used to compose the hierarchical semantic similarities in Eq. (2).

4. IMAGE RETRIEVAL WITH A^{2}SH

In this section, we develop a unified content-based image retrieval system based on A^{2}SH. The system enables efficient and effective automatic retrieval and interactive retrieval with hybrid feedback, aiming to bridge both semantic and intention gaps.

4.1. Automatic Retrieval with Hierarchical Indexing

A^{2}SH provides a much more efficient similarity search due to the aforementioned compact hierarchical semantic representations. However, the cost of linear scan of the entire database can be very high, even for such a compact representation and especially for large-scale databases. In order to support efficient large-scale image retrieval, we develop a hierarchical indexing strategy. All the images are indexed hierarchically based on their semantic paths in the hierarchy. We define an index file as $Ind_c = < c, child(c), \mathcal{I}_c >$, where $child(c)$ are children of the concept c, and $\mathcal{I}_c$ is the set of database images whose predicted semantic paths terminate at c.

Given a query image I_q, its retrieval using hierarchical indexing is carried out as follows. First, we generate the hierarchical semantic representation of I_q along its semantic path $c_0 \rightarrow \cdots \rightarrow c_n$ based on A^{2}SH. Second, we perform fast retrieval of candidate images by looking up the index file Ind_{c_n}. The candidate images consist of all images indexed by c_n along with its children $child(c_n)$. Note that the number of candidate images is significantly reduced compared to the size of the entire database. As the semantic path prediction may not be perfect, c_n may not be the correct semantic path terminal of the query image as the ground truth. To address this problem, we set a look-back level b ($b=3$ in the experiments as it obtains the best results) and retrieve more relevant candidate images by the index file $\text{Ind}_{c_{n-b}}$. Note that when $b = n$, it retrieves all the database images as candidates, degenerating the process into a linear scan. The basic assumption behind the choice of b is that semantic path prediction is more accurate if we only reach the node c_{n-b}. Also, note that thanks to the discriminative property of the semantic similarity defined in Eq. (3), we can rank the wrongly indexed candidate images lower, resulting in more accurate results.

Next, we analyze the time complexity of retrieval, including three major steps: (1) semantic path prediction, (2) looking up the index file to obtain candidate images, and (3) generating the top-K results according to the hierarchical semantic similarities of the candidate images. We denote the averaged fan-out (i.e., the averaged number of children of a concept in the hierarchy) of A^{2}SH as F, the averaged leaf depth (i.e., the

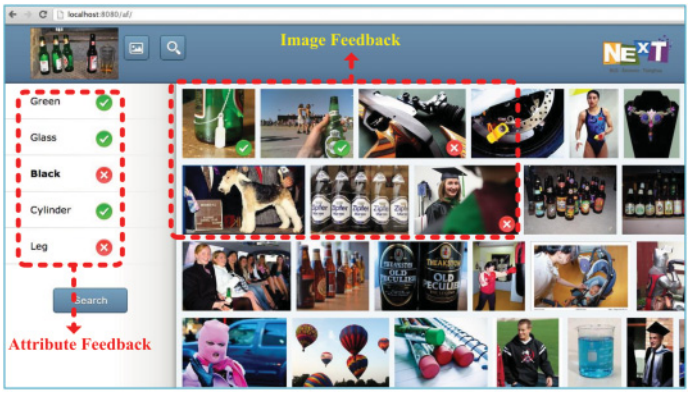


Fig. 3. The user interface of the A^{2}SH-based CBIR system.

averaged depth over all the leaves) as D, and the concept classifier prediction cost as C. Therefore, we can estimate the cost of semantic path prediction as $\mathcal{O}(nFC)$, where $n \leq D + b$ is the average depth of the predicted path, and the cost of candidate image retrieval as $\mathcal{O}(F^{D-n+b})$, which is the cost for subhierarchy traversal. Note that C is a small constant, D and F are, respectively, 6.3 and 3.8 in our ImageNet hierarchy, and thus the costs of prediction and candidate retrieval are very small. Consequently, the time cost for retrieval mainly comes from the third step, that is, the ranking of candidate images, which has a time complexity of $\mathcal{O}(ndN_c + N_c \log N_c)$, where d is the average dimension of local semantic spaces, and N_c is the number of candidate images, which is much smaller than the size of the entire database.

4.2. Interactive Retrieval with Hybrid Feedback

Because of the presence of the intention gap that hinders understanding of users' search intent by the system, the results of automatic retrieval often do not satisfy users' information needs. We therefore perform interactive retrieval by involving users' interaction with the system. We propose a hybrid feedback (HF) mechanism to help users deliver their search intent by providing hybrid feedbacks on both the attributes and images. As shown in Figure 3, we allow users to indicate yes no feedback on attributes to state which attributes are present or not in his/her search intent. Meanwhile, we also allow users to provide relevance judgements on images to indicate which images are "relevant" or "irrelevant." In particular, the attributes on the left are dynamically suggested by the system based on mining the most informative attributes from the current search results [Zhang et al. 2012]. This hybrid feedback is then used to generate a semantic interpretation of users' search intent based on the proposed A^{2}SH. By iteratively collecting user feedback and refining retrieval, the system can shape users' intent more accurately and narrow the search to target gradually.

Suppose we are at the tth feedback iteration. The system records the "relevant" images as $\mathcal{R}_t$ and the "irrelevant" images as $\overline{\mathcal{R}}_t$, as well as the "yes" attributes as $\mathcal{B}_t$ and the "no" attributes as $\overline{\mathcal{B}}_t$. Suppose the hierarchical semantic representation of a query image is $\{\mathcal{Q} = (c_0 \rightarrow \cdots \rightarrow c_n); \mathcal{Z} = (\mathbf{z}^{c_0}, \ldots, \mathbf{z}^{c_n})\}$, where $\mathcal{Q}$ is the semantic path and $\mathcal{Z}$ is the set of local semantic representations along the path. We refine the query representation at iteration t, tailoring $\mathcal{Z}_t$ to user intent by incorporating semantic descriptions delivered by image feedback (i.e., $\mathcal{R}_t$ and $\overline{\mathcal{R}}_t$), and attribute feedback (i.e., $\mathcal{B}_t$ and $\overline{\mathcal{B}}_t$). More specifically, we refine the query $\mathcal{Z}_t$ to make it closer to the semantic representation of relevant images and away from that of irrelevant ones. This refinement

is carried out along the semantic path for every local semantic representation of the query, leading to an hierarchical semantic interpretation of user intent. Formally, $\forall c \in \mathcal{Q}$, we have

$$\mathbf{z}_{t+1}^c[a] = \mathbf{z}_t^c[a] + \beta \sum_{i \in \mathcal{R}_t} \big(\mathbf{z}_i^c[a] - \mathbf{z}_t^c[a]\big)/|\mathcal{R}_t| - \gamma \sum_{j \in \overline{\mathcal{R}}_t} \big(\mathbf{z}_j^c[a] - \mathbf{z}_t^c[a]\big)/|\overline{\mathcal{R}}_t|, \tag{4}$$

where β and γ are the trade-off parameters. Through image feedback, the semantic representation of the query is shaped closer towards the semantic representations of relevant images and farther away from those of irrelevant ones.

User feedback on attributes $\mathcal{B}_t$ and $\overline{\mathcal{B}}_t$ state the desired and undesired attributes, respectively. That is to say, the attributes in $\overline{\mathcal{B}}_t$ are expected be included in the query, while the attributes in $\overline{\mathcal{B}}_t$ are not. Hence, we refine the query $\mathcal{Z}_t$ by setting the values on the dimensions corresponding to $\mathcal{B}_t$ as 1 and those corresponding to $\overline{\mathcal{B}}_t$ as 0. $\forall c \in \mathcal{Q}$, we have

$$\forall a \in \mathcal{A}_c, \mathbf{z}_{t+1}^c[a] = \begin{cases} 1, & a \in \mathcal{B}_t, \\ 0, & a \in \overline{\mathcal{B}}_t, \\ \mathbf{z}_t^c[a], & otherwise. \end{cases} \tag{5}$$

The resultant query $\mathcal{Z}_{t+1}$ is then used to generate new search results based on the aforementioned hierarchical semantic similarity function, which leads to semantic reranking scores of the candidate images. Here, we emphasize the semantic dimensions corresponding to the attributes in $\mathcal{B}_t$ and $\overline{\mathcal{B}}_t$ to make them contribute more to the similarity, since they encapsulate users' clear intent on the attributes. Recall the distance function based on the local semantic metric $\mathbf{M}_c$ used in Eq. (2). We notice that emphasizing the dimensions is equivalent to giving large weights to the corresponding rows of the metric matrix $\mathbf{M}_c$ in similarity calculation. In our experiments, we set the weight to 0.7 for the rows corresponding to the attributes in $\mathcal{B}_t$ and $\overline{\mathcal{B}}_t$, and 0.3 for the rest.

In order to enhance the user intent delivered by attributes, we also incorporate the visual similarities between relevant images $\mathcal{R}_t$ and the other candidate images with respect to the desired attributes $\mathcal{B}_t$. For example, if a user specifies positive feedback on attribute "fur" and a relevant "dog" image, then the intention is most likely to be "dogs with fur like the relevant one should be also relevant." With the help of A^{2}SH, we further augment such similarities along the semantic hierarchy. Specifically, the visual reranking score of any candidate image I_i can be fused with the preceding semantic score. The visual score is calculated as

$$VisualScore(I_i) = \sum_{c \in \mathcal{Q}} \sum_{I_j \in \mathcal{R}_t} \sum_{a \in \mathcal{B}_t} \min\big(\mathbf{z}_i^c[a], \mathbf{z}_j^c[a]\big) exp(-\|\mathbf{x}_{a,c,i} - \mathbf{x}_{a,c,j}\|_2), \tag{6}$$

where $\mathbf{x}_{a,c,i}$ is the image I_i's visual feature of attribute a in the context of concept c (Section 3.1.2). In particular, we use the local attribute representation $\mathbf{z}_i^c[a]$ as the confidence of the presence of attributes to weight the visual similarities. Note that as compared to the time complexity of calculating the semantic scores, calculating the visual scores requires additional time cost (generally 100–300 ms in our experiments) at $\mathcal{O}(|\mathcal{Q}||\mathbf{R}_t||\mathcal{B}_t|N_c\bar{d})$, where N_c is the number of candidate images and $\bar{d}$ is the average number of feature dimensions of attributes along the semantic path.

5. EXPERIMENTS

In this section, we systematically evaluate the proposed attribute-augmented semantic hierarchy (A^{2}SH) in content-based image retrieval. We first evaluate the elementary

building blocks of A^{2}SH. We then investigate the effectiveness of A^{2}SH in automatic and interactive image retrieval.

5.1. Data and Methodology

5.1.1. Dataset and Features. We conducted experiments on ImageNet [Deng et al. 2009], which is a large-scale corpus of images organized according to the WordNet hierarchy. Each concept in the hierarchy is depicted by hundreds to thousands of images collected from the Web. We used a subset of ImageNet with 1,860 concepts and 1.27 million images, which were used for ILSVRC 2012[1]. This dataset contains a partial WordNet hierarchy and some isolated nodes outside WordNet. We used the WordNet hierarchy for evaluation. This hierarchy consists of 1.22 million images with 1,730 concepts, including 958 leaf concepts. Its maximum depth is 19. We merged the non-leaf nodes with no siblings into their parents since they are the sole heir to the semantics of their parents. This gives rise to a compressed hierarchy with a maximum depth of 11, consisting of 1,322 concepts and the original amount of leaf concepts and images. We augmented this hierarchy with a pool of attributes, including nameable and unnameable attributes. We defined 33 nameable attributes[2] based on the attribute pool used in Farhadi et al. [2009]. These attributes were linked to the concepts in a bottom-up manner. We first associated each leaf concept with its related attributes. Each non-leaf concept was then linked to the union of the attributes from its children. The unnameable attributes were automatically discovered for each concept as described in Section 3.1.3.

We randomly split the image set into a training subset with 50% of images for each concept, and a subset with the remaining images for testing. We generated ground truth on the images as follows. Based on the labeling of leaf concepts provided by ImageNet, we generated the labeling for each non-leaf concept in a bottom-up manner. A non-leaf concept was regarded as positive to an image if any child is positive and negative otherwise. For attributes, we conducted manual labeling. Since manual labeling is labor-intensive and time-consuming, we randomly selected 100 images of each leaf concept from the training subset for ground truth labeling. The attribute labeling in the context of non-leaf concepts was also generated in a bottom-up manner. We conducted image retrieval on the testing subset. We randomly selected 100 images from each of the 958 leaf concepts, giving rise to a total of 95,800 experimental queries.

For low-level visual features of images, we used color, texture, edge, and shape visual cues in a two-level spatial pyramid fashion, resulting in a 35,903-D sparse feature vector (cf. [Zhang et al. 2013] in detail). These visual cues are shown to be effective in attribute and concept learning [Farhadi et al. 2009]. Note that even the 35,903-D feature is high-dimensional, it is very sparse (generally 90% zeros) and has been shown to be very effective in modeling with simple linear classifiers such as linear SVM as we did in this article [Wang et al. 2010].

5.1.2. Experimental Setting. We learned linear SVM classifiers for concepts and attributes by employing the LIBLINEAR toolbox[3]. We learned ℓ_1 linear logistic regressors to select informative features for learning attributes. For the MMC clustering [Zhang et al. 2009] used in unnameable attribute discovery, we used the code provided in the

[1] `http://www.image-net.org/challenges/LSVRC/2012/index.`

[2] We removed the concept-specific attributes in Farhadi et al. [2009] and Zhang et al. [2012], such as "*jet-engine*," since in our work, we have such concept-specific descriptions by linking the attributes (e.g., "*wing*") to concepts (e.g., "*jet*"). We also added seven color attributes because of their effectiveness in image retrieval [Russakovsky and Li 2010]. As a result, we have 33 attributes as follows: *black, blue, brown, cylinder, furry, glass, gray, green, handle, head, leg, metallic, plastic, rectangular, red, round, scale, screen, shiny, skin, smooth, spotted, stripped, tail, triangle, vegetation, wet, wheel, white, window, wing, wooden, and yellow.*

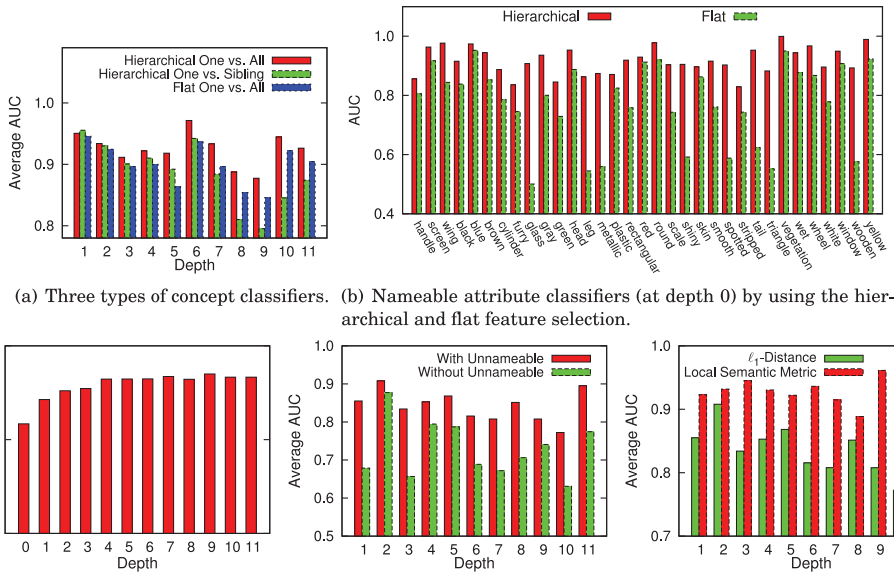
[3] `http://www.csie.ntu.edu.tw/cjlin/liblinear/.`

author's website. For local metric learning, we deployed the LMNN toolbox [Weinberger et al. 2006] with training triplets configuration as described in Eq (3). The algorithmic parameters of these models were tuned through fivefold cross validation. We applied the Weibull distribution [Scheirer et al. 2012] to normalize the responses from attribute classifiers.

To evaluate the effectiveness of the proposed A^2SH in automatic retrieval, we compared it against the following five representative retrieval solutions, including two flat methods and three hierarchical ones. (a) fVisual retrieves images based on visual similarities with Eculidean metric. (b) fSemantic represents each image into a flat semantic representation composed by the responses from the 1,322 concept classifiers and the 33 attribute classifiers of the root concept. It retrieves images based on such representation using the ℓ_1 distance. (c) hPath performs retrieval based on the length of the common semantic path of an image and the query. (d) hVisual computes the similarities between any two images by aggregating their visual similarities along their common semantic path, then conducts retrieval based on such similarity. (e) hBilinear [Deng et al. 2011] retrieves images by the recently proposed bilinear semantic metric which was reported achieving state-of-the-art performance on the ImageNet dataset.

To evaluate the effectiveness of A^2SH in interactive retrieval, we compared it to the following three interactive retrieval methods. (a) QPM [Rui et al. 1998]: Query Point Movement method updates the query based on image feedback and refines search results using the new query; (b) SVM [Tong and Chang 2001]: this approach learns an SVM classifier from the "relevant" and "irrelevant" images and ranks images according to their responses from the classifiers; and (c) AF [Zhang et al. 2012]: the recently proposed Attribute Feedback approach collects user feedback on attributes and then ranks images according to the presence probabilities of the attributes in the images. Note that our approach enables hybrid feedback on attributes and images. For the sake of fair comparison, we incorporated image feedback into AF such that it also uses hybrid feedback. Moreover, all the baseline methods were performed on the flat semantic representations of the images, rather than the low-level visual descriptors. For our A^2SH methods, we denote the method that uses only the semantic ranking score in Eqs. (4) and (5) as A^2SH and the method that uses both semantic and visual ranking score in Eq. (6) as A^2SH^*.

We conducted the evaluation in two settings with a fixed number of feedback interations and a fixed time limit, respectively. In the first setting, we conducted five feedback iterations with 20 feedback per iteration. For the QPM and SVM methods, 20 feedbacks on the top 20 images were collected in each iteration. For the AF and our A^2SH methods, the same number of feedback iterations were collected, including 5 attribute feedbacks and 15 image feedbacks. Five informative attributes were suggested in each iteration for soliciting attribute feedback. We here employed the suggestion strategy in Zhang et al. [2012]. Given a query, the feedback process was simulated by the computer according to the ground truth of the query category on the images and the association between the attributes and the category. In the setting of fixed time limit, we invited 25 novice users to interact with the system through the preceding four feedback methods, respectively. We did not constrain the numbers of feedbacks and iterations and allowed the users to interact with the system in a free way. Since it is time-consuming and labor-intensive for users to evaluate a large number of queries. We randomly selected 10 images from each leaf concept as queries, giving rise to 9,580 queries in total, and assigned these queries to the users approximately evenly with no overlap between them. We set the time limit to 2 minutes in the experiments. For a given query in all these preceding evaluations, we used the search results from the best automatic retrieval method, that is, the proposed A^2SH, as the initial results for



(a) Three types of concept classifiers. (b) Nameable attribute classifiers (at depth 0) by using the hierarchical and flat feature selection.

(c) Nameable attribute classifiers. (d) Concept classification by exploiting unnameable attributes. (e) Local semantic metric.

Fig. 4. Performance at different depth levels measured by average AUC. The depth of the root is 0.

interactive retrieval. In particular, the feedback is done on a small fraction of the retrieved images (e.g., top 1,000 results in our experiments).

All the experiments were conducted on a server with Intel(R) Xeon(R) CPU X5650 at 2.67GHz on 24 cores, 48GB RAM and 64-bit Centos 5.4 operating system.

5.1.3. Performance Metric. We adopted the widely used metric AUC (area under ROC curve) value for classification performance evaluation. We adopted average precision at top-K retrieved images (AP@K) for retrieval performance evaluation [Over et al. 2012]. We denote R as the number of relevant images in the database. At any ranked position j ($1 \leq j \leq K$), let R_j be the number of relevant images in the top j results and let $I_j = 1$ if the jth image is relevant and 0 otherwise, then AP@K is defined as $\frac{1}{min(R,K)} \sum_{j=1}^{K} \frac{R_j}{j} \times I_j$. Moreover, we also used the following hierarchical average precision at top K for retrieval performance evaluation. The hMAP@K is defined as $\frac{1}{min(R,K)} \sum_{j=1}^{K} \frac{D^*_{jq}/D^*_q}{j}$, where D^*_{jq} is the depth of the lowest common ground truth ancestor of ground-truth concept of the image ranked at position j and the query, D^*_q is the depth of the ground truth concept of the query. The intuition of hAP@K is that if a returned image does not exactly match the query, it is expected to be as semantically close to the query as possible in order for a better user experience. We averaged the AP@K and hAP@K over all the queries to compute the MAP@K and hMAP@K, which are used as the overall metrics.

5.2. Experimental Results

5.2.1. Evaluations of Concept Classifiers, Attribute Classifiers, and Local Semantic Metrics. Figure 4(a) shows the average AUC values of the concept classifiers at different depth levels in the hierarchy [Deng et al. 2009]. We compared the deployed Hierarchical One vs. All classifiers with the other two state-of-the-art concept classifiers in the hierarchy:

Hierarchical One vs. Sibling [Marszalek and Schmid 2007] and Flat One vs. All [Wang et al. 2010], which are discussed in Section 3.1.1. From these results, we can see that Hierarchical One vs. All classifiers achieved the best performance at almost all the depths. This is because the deployed method effectively leverages the hierarchical information (compared to Flat One vs. All) and alleviates the error propagation problem (compared to Hierarchical One vs. Sibling). In particular, the concept classifiers at most levels achieve an average AUC of above 0.9, except for those at depth 8 and 9, with an average AUC of 0.89 and 0.88, respectively. One possible explanation for the relatively lower performance of depth 8 and 9 is that the concept taxonomy (generally for animals) at this level in ImageNet leads to confusing visual cues for visual classifiers for example "hunting dogs" and "watch dog".

Figure 4(b) shows the AUC of the 33 attribute classifiers (Depth 0) by the proposed hierarchical attribute learning (Section 3.1.2) and flat attribute learning [Farhadi et al. 2009]. We can see that the hierarchical attribute classifiers significantly outperform the flat ones by 19.5% in relative improvement. The performance of attribute classifiers is illustrated in Figure 4(c), from which we can see that the attribute classifiers at every level achieve an average AUC higher than 0.9. These results demonstrate the effectiveness of our concept and attribute classifiers in capturing the semantics of image content, leading to hierarchial semantic representations of images.

A collection of unnameable attributes are discovered to complement the nameable attributes to better characterize a concept. In particular, the average number of unnameable attributes discovered for the concepts at different depth levels is generally from 5 to 15. As aforementioned, we have no ground truth of the unnameable attributes on images and thus cannot evaluate the performance of their classifiers directly. Instead, we evaluated the effectiveness of unnameable attributes in improving the accuracy of distinguishing sibling concepts. In particular, for each set of sibling concepts in the hierarchy, we used the nearest-neighbor classifier to classify their images based on the local semantic representations with or without using unnameable attributes. The classification performance is illustrated in Figure 4(d). From the results, we can see that the discovered unnameable attributes can improve classification performance significantly. This indicates that the unnameable attributes can help to provide a more comprehensive and discriminative description of a concept.

We evaluated the effectiveness of the local semantic metrics as follows. We used the local semantic metric of each concept to help classify the images of the concept from that of all its siblings by using the 5-nearest-neighbor classifier [Weinberger et al. 2006]. We compared the local semantic metrics against the widely used ℓ_1 distance. The classification performance comparison is illustrated in Figure 4(e). We can see that the proposed local semantic metric outperforms the ℓ_1 distance significantly at all depth levels. On average, it achieves a relative improvement of 10.7% at various depth levels. This demonstrates the effectiveness of the local semantic metric and its capacity in composing an effective hierarchical semantic similarity to characterize the semantic affinities among images.

5.2.2. Evaluations of Automatic Retrieval. Figure 5 illustrates the performance comparison between the proposed A^{2}SH and the other five automatic retrieval methods. We can see that A^{2}SH achieves the best retrieval performance in terms of both MAP and hMAP at all the top K results as compared to the other methods. The performance improvements of A^{2}SH over the other methods are significant. For example, A^{2}SH improves the relative performance by 22.4%, 23.1%, 41.5%, and 46.0% in MAP at the top 50 results as compared to the hBilinear, fSemantic, hPath, and hVisual methods, respectively. The corresponding performance improvements in terms of hMAP are 7.3%, 20.2%, 31.3%, and 26.5%, respectively. These results demonstrate the effectiveness of

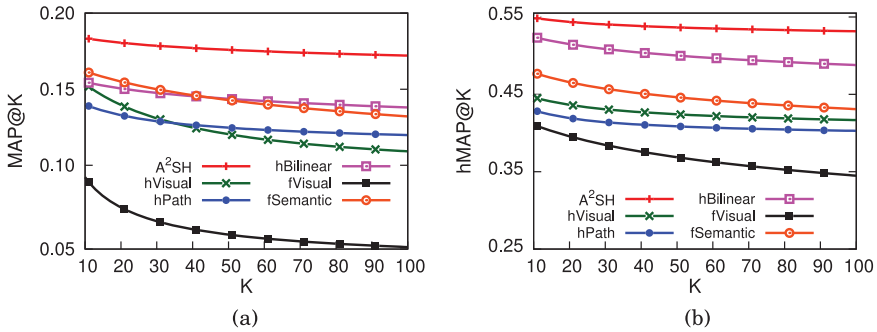


Fig. 5. Performance of automatic image retrieval over the 95,800 queries.

Table I. Average Retrieval Time per Query of Automatic Image Retrieval over the 95,800 Queries

Method	fVisual	fSemantic	hVisual	hBilinear	A^2SH
Time (ms)	1.18×10^4	3.62×10^3	7.42×10^2	4.47×10^2	**70.6**

A^2SH in image retrieval. The superiority of A^2SH to the other methods arises from the following aspects: (a) A^2SH models the semantics of images in the form of an hierarchical semantic representation consisting of multiple levels of concepts, each of which is associated with a local semantic representation in terms of related attributes. Such hierarchical semantic representation provides a more comprehensive and more accurate interpretation of image semantics. (b) The hierarchical similarity function in A^2SH more accurately characterizes the semantic similarities among images by ensembling the local semantic metrics in the context of various concepts. (c) As compared to the state-of-the-art hBilinear that only exploits 958 leaf node concepts, A^2SH encodes much more semantics, for example, 1,322 semantic concepts and about 58,000 attributes. Moreover, hBilinear only incorporates the semantic relations (e.g., depth of the common ancestor) of the concepts, however, A^2SH not only uses the semantic relations (i.e., the common semantic path) but also the much richer semantic similarities in terms of attributes in context.

Table I lists the average retrieval time per query of automatic retrieval over the 95,800 queries by the five approaches. We can observe that A^2SH offers highly efficient retrieval. It significantly reduces the retrieval time by several orders of magnitude compared to the other methods. The reasons are twofold. First, A^2SH represents images in the form of a compact hierarchical semantic representation, thus enabling fast similarity computation to be carried out. Second, the hierarchical indexing in A^2SH significantly reduces the size of the search space. For the sake of fair comparison, we also accelerated the other four retrieval methods using indexing techniques. In particular, hVisual was carried out based on the hierarchical indexing in our A^2SH system. hBilinear was accelerated using the indexing technique in Deng et al. [2011]. Here, we do not list the time cost of the hPath method, since it is a sub-procedure of A^2SH and hVisual, that is, retrieving candidate images from the hierarchical index files. The fSemantic method was accelerated by indexing the semantic concepts and attributes using inverted files. We also indexed the low-level visual features which are high-dimensional and sparse by using inverted files to accelerate the visual retrieval in fVisual and hVisual.

5.2.3. Evaluations of Interactive Retrieval with Hybrid Feedbacks. Figure 6 illustrates the performance of interactive retrieval with five feedback iterations in terms of MAP and hMAP at the top 20, 50, and 100 search positions, respectively. From these results, the

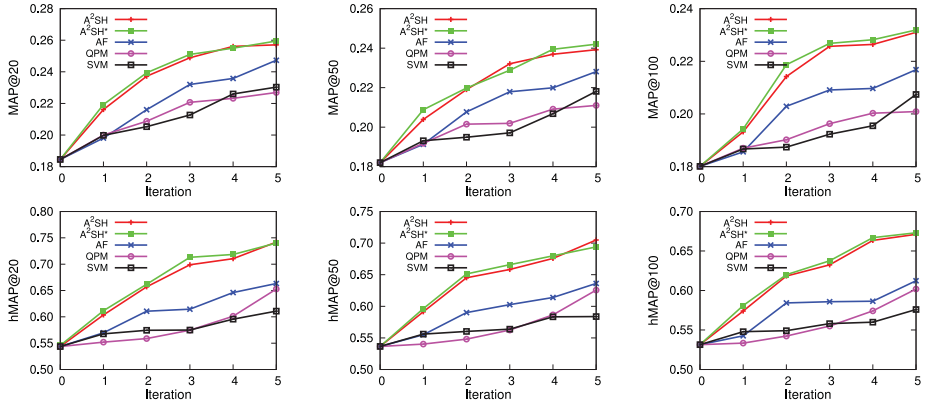


Fig. 6. Performance of interactive retrieval with five feedback iterations over the 95,800 queries.

Fig. 7. Illustrative examples of retrieval results by different methods. The query is "beer bottle" (ImageNet ID: n02823428). The red rectangular represents exact matched results (i.e., in the same category of n02823428) and the yellow rectangular denotes semantic similar results (i.e., in different categories of the same parent node n02876657 named "bottle"). Left three rows: automatic retrieval. Right three rows: interactive retrieval after five feedback iterations.

following observations can be made. (a) The proposed A^{2}SH/A^{2}SH*-based interactive retrieval approaches outperform the other three methods at every iteration and all the top 20, 50, and 100 results. (b) A^{2}SH/A^{2}SH*-based approaches significantly reduce the interaction efforts while achieving comparable performance to the other three methods. For example, consider the MAP at top 20 results: A^{2}SH/A^{2}SH*-based approaches obtain a comparable performance at their 3rd, 2nd, and 2nd iteration, as compared to AF, QPM, and SVM at the last round, respectively. In other words, the A^{2}SH framework can reduce labeling efforts by about 40%, 60%, and 60%, respectively, as compared to the three methods. (c) The performance improvements of A^{2}SH/A^{2}SH*-based approaches and AF over QPM and SVM indicate the effectiveness of attribute feedback in delivering user search intent. The superiority of A^{2}SH/A^{2}SH*-based approaches to AF demonstrates that A^{2}SH can infer user intent more accurately from the feedback, improving retrieval performance over A^{2}SH without feedback. In particular, A^{2}SH can interpret more accurate meanings of attributes by automatically inferring the context of them. (d) Finally, A^{2}SH* that also incorporates visual similarities of attributes can further boost the overall performance, as compared to A^{2}SH that only uses semantic similarities. Figure 7 illustrates an example of three interactive retrieval methods with attributes.

Table II lists the performance of the five interactive retrieval approaches with a fixed time limit of two minutes. From these results, we can see that the proposed A^{2}SH/A^{2}SH*-based approaches with attribute visual scores are able to achieve the best performance in terms of both MAP and hMAP. This demonstrates that A^{2}SH/A^{2}SH*-based approaches are able to shape user intent more quickly within the same interaction time and can generate more accurate search results as compared to the other

Table II. Performance of Interactive Retrieval with 2-Minute Time Limit over the 9,580 Queries

	MAP(%)			hMAP(%)		
RF Methods	@20	@50	@100	@20	@50	@100
A^2SH	24.67	**22.80**	22.03	68.37	66.04	64.08
A^2SH^*	**25.72**	22.03	**22.84**	**69.95**	**67.12**	**64.40**
AF	22.59	21.38	20.63	62.84	60.20	58.54
QPM	21.24	20.53	19.52	58.00	56.73	55.83
SVM	21.56	20.08	19.15	58.50	57.18	55.45

methods. Moreover, A^2SH^* further exploits the visual similarities of attributes and thus generally outperforms A^2SH.

6. CONCLUSIONS AND FUTURE WORK

In this article, we proposed a novel attribute-augmented semantic hierarchy (A^2SH) which organizes semantic concepts from general to specific and augments each semantic concept with a set of related attributes, which are specifications of the multiple facets of the concept and act as an intermediate bridge connecting the concept and low-level visual features. We learned concept classifiers, attribute classifiers, and an hierarchical similarity function under the framework of A^2SH. Based on the proposed A^2SH, we developed a unified content-based image retrieval system that supports both automatic retrieval and interactive retrieval with user feedback. A hybrid feedback mechanism was developed to collect a broad array of feedback based on both attributes and images. This feedback was then utilized to improve the retrieval performance based on A^2SH. We systematically evaluated the A^2SH-based image retrieval system on a large-scale corpus of over one million Web images. The experimental results demonstrated the effectiveness of A^2SH in bridging the semantic and intention gaps, leading to more accurate results as compared to state-of-the-art CBIR approaches.

The essence of A^2SH is to enrich an existing knowledge base with sublevel semantics such as attributes. In this way, we are able to deploy a unified and computable knowledge base with recent advances in semantic concepts and attributes. Therefore, part of our future research may focus on integrating FrameNet (a hierarchy defines the interactions of concepts) [Baker et al. 1998], WordNet (ImageNet), and attributes, towards a more challenging task such as multimedia event detection (MED). However, defining a complete set of concepts and attributes only by domain experts is a challenging and expensive task. It may be limited to semantic scale and general user interest (although the proposed automatic discovery of unnameable attributes is a practical remedy). Hence, other future work may focus on automatically mining semantics and their intrinsic relations from the inexhaustible amount of online user-generated content.

REFERENCES

C. F. Baker, C. J. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*.

M. Belkin and P. Niyogi. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computat*. 15, 6, 1373–1396.

A. Binder, K.-R. Müller, and M. Kawanabe. 2012. On taxonomies for multi-class image categorization. *Int. J. Comput. Vision* 99, 3, 281–301.

Y. Boureau, N. Le Roux, F. Bach, J. Ponce, and Y. LeCun. 2011. Ask the locals: Multi-way local pooling for image recognition. In *Proceedings of the International Conference on Computer Vision*.

M. Crucianu, M. Ferecatu, and N. Boujemaa. 2004. Relevance feedback for image retrieval: A short survey. *DELOS2 Report*.

R. Datta, D. Joshi, J. Li, and J. Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.* 40, 2, Article 50.

J. Deng, A. C. Berg, and F.-F. Li. 2011. Hierarchical semantic indexing for large scale image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

J. Deng, A. C. Berg, K. Li, and F.-F. Li. 2010. What does classifying more than 10,000 image categories tell us? In *Proceedings of the European Conference on Computer Vision*.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*.

T. Deselaers and V. Ferrari. 2011. Visual and semantic similarity in ImageNet. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*.

M. Douze, A. Ramisa, and C. Schmid. 2011. Combining attributes and fisher vectors for efficient image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. 2009. Describing objects by their attributes. In *Proceedings of the IEEE Conference on Computer Vision and Patter Recognition*.

C. Fellbaum. 2010. WordNet. In *Theory and Applications of Ontology: Computer Applications*. Springer.

G. Griffin and P. Perona. 2008. Learning and using taxonomies for fast visual categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

A. Jaimes and S.-F. Chang. 2000. A conceptual framework for indexing visual information at multiple levels. *Proc. SPIE 3964*.

A. Kovashka, D. Parikh, and K. Grauman. 2012. WhittleSearch: Image search with relative attribute feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. 2011. Describable visual attributes for face verification and image search. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 10, 1962–1977.

M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. 2006. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.* 2, 1–90.

Z. Ma, Y. Yang, Z. Xu, S. Yan, N. Sebe, and A. G. Hauptmann. 2012. Complex event detection via multi-source video attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

M. Marszalek and C. Schmid. 2007. Semantic hierarchies for visual object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. 2006. Large-scale concept ontology for multimedia. *IEEE Multimedia* 13, 3, 86–91.

P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A. F. Smeaton, and G. Quenot. 2012. TRECVID 2012 – An overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of the TRECVID Conference*.

D. Parikh and K. Grauman. 2011a. Interactively building a discriminative vocabulary of nameable attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

D. Parikh and K. Grauman. 2011b. Relative attributes. In *Proceedings of the IEEE International Conference on Computer Vision*.

Y. Rui, T. S. Huang, and S.-F. Chang. 1999. Image retrieval: Current techniques, promising directions, and open issues. *J. Visual Commun. Image Represent.* 10, 1, 39–62.

Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. 1998. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Trans. Circuits Syst. Video Techno.* 8, 5, 644–655.

O. Russakovsky and F.-F. Li. 2010. Attribute learning in large-scale datasets. In *Trends and Topics in Computer Vision*. Lecture Notes in Computer Science, vol. 6553. Springer.

W. J. Scheirer, N. Kumar, P. N. Belhumeur, and T. E. Boult. 2012. Multi-attribute spaces: Calibration for attribute fusion and similarity search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 12, 1349–1380.

J. R. Smith and S.-F. Chang. 1997. VisualSeek: A fully automated content-based image query system. In *Proceedings of the ACM International Conference on Multimedia*.

Y. Song, M. Zhao, J. Yagnik, and X. Wu. 2010. Taxonomic classification for web-based videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

D. Tao, X. Tang, X. Li, and X. Wu. 2006. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* 28, 7, 1088–1099.

S. Tong and E. Chang. 2001. Support vector machine active learning for image retrieval. In *Proceedings of the ACM International Conference on Multimedia*.

N. Verma, D. Mahajan, S. Sellamanickam, and V. Nair. 2012. Learning hierarchical similarity metrics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. 2010. Locality-constrained linear coding for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

K. Q. Weinberger, J. Blitzer, and L. K. Saul. 2006. Distance metric learning for large margin nearest neighbor classification. In *Proceedings of the 20th Annual Conference on Neural Information Processing Systems*.

C. Yang, M. Dong, and F. Fotouhi. 2005. Semantic feedback for interactive image retrieval. In *Proceedings of the ACM International Conference on Multimedia*.

F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang. 2013. Designing category-level attributes for discriminative visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang. 2008. Joint multi-label multi-instance learning for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Z.-J. Zha, W. Meng, Y.-T. Zheng, Y. Yang, R. Hong, and T.-S. Chua. 2012. Interactive video indexing with statistical active learning. *IEEE Trans. Multimedia* 14, 1.

Z.-J. Zha, L. Yang, T. Mei, M. Wang, and Z. Wang. 2009. Visual query suggestion. In *Proceedings of the ACM International Conference on Multimedia*.

Z.-J. Zha, L. Yang, T. Mei, M. Wang, Z. Wang, T.-S. Chua, and X.-S. Hua. 2010. Visual query suggestion: Towards capturing user intent in internet image search. *ACM Trans. Multimedia Comput. Commun. Appl.* 6, 3.

H. Zhang, Z.-J. Zha, S. Yan, J. Bian, and T.-S. Chua. 2012. Attribute feedback. In *Proceedings of the ACM International Conference on Multimedia*.

H. Zhang, Z.-J. Zha, Y. Yang, S. Yan, Y. Gao, and T.-S. Chua. 2013. Attribute-augmented semantic hierarchy: Towards bridging semantic gap and intention gap in image retrieval. In *Proceedings of the ACM International Conference on Multimedia*.

K. Zhang, I. W. Tsang, and J. T. Kwok. 2009. Maximum margin clustering made practical. *IEEE Trans. Neural Netw.* 20, 4, 583–596.

Received February 2014; revised May, June 2014; accepted June 2014