Attribute-augmented Semantic Hierarchy

Towards Bridging Semantic Gap and Intention Gap in Image Retrieval

Hanwang Zhang[†], Zheng-Jun Zha[†§], Yang Yang[†], Shuicheng Yan[‡], Yue Gao[†], Tat-Seng Chua[†]

[†]School of Computing, National University of Singapore
[§]Institute of Intelligent Machines, Chinese Academy of Sciences
[‡]Department of Electrical Computer Engineering, National University of Singapore
{hanwangzhang,junzzustc}@gmail.com; {yang.yang,eleyans,dcsgaoy,dcscts}@nus.edu.sg

ABSTRACT

This paper 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 the semantic concepts into multiple semantic levels and augments each concept with a set of related attributes, which describe the multiple facets of the concept and act as the intermediate bridge connecting the concept and low-level visual content. A hierarchical semantic similarity function is learnt to characterize the semantic similarities among images for retrieval. To better capture user search intent, a hybrid feedback mechanism is developed, which collects hybrid feedbacks on attributes and images. These feedbacks are then used to refine the search results based on A^2SH. We develop a content-based image retrieval system based on the proposed A^2SH. We conduct extensive experiments on a large-scale data set 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 precisely and 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**]: Retrieval Model

General Terms

Algorithms, Experimentation, Performance

Keywords

image retrieval, attribute, semantic hierarchy

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2502093.

Figure 1: Illustration of the semantic gap and intention gap in image retrieval.

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 [29, 15, 27, 23]. It has gained increasing importance in both the academia and industry, in the current era of social media, because of the explosive growth of images shared in cyberspace and the compelling demands in various multimedia applications for Web and mobile clients. In spite of the remarkable progress made in the last two decades, CBIR remains challenging mainly due to two critical scientific problems for image retrieval as shown in Figure 1: (a) the Semantic Gap between the low-level visual features and high-level semantics [28, 2]; and (b) the Intention Gap between user's search intent and the query [38, 11], which hinders the understanding of user's intent behind a query.

Recent studies, especially those on TRECVID [21], have shown that a promising route to narrowing the semantic gap is to exploit a set of concepts to form the semantic description of visual content [31, 19]. As the amount and scope of semantic concepts increase, semantic hierarchy, such as ImageNet [4] and LSCOM [20], has been developed to organize the semantic concepts from general to specific and essentially partition the semantic space hierarchically, towards better addressing the semantic gap problem. The semantic hierarchy has been found to be encouraging in improving the understanding of visual content recently [7, 18]. However, there still lacks the correspondences between visual features and semantics, due to the intra-concept variations and inter-concept similarities on visual properties. On the other hand, to address the intention gap problem, Relevance Feedback (RF) has been introduced into image retrieval. RF collects user feedbacks on candidate images, indicating them as "relevant" or "irrelevant" and lets the system to infer users' search intent from the labeled images [24, 1]. Due to the discrepancy between user's intent and low-level visual cues, RF is often ineffective in narrowing search to target in practice.

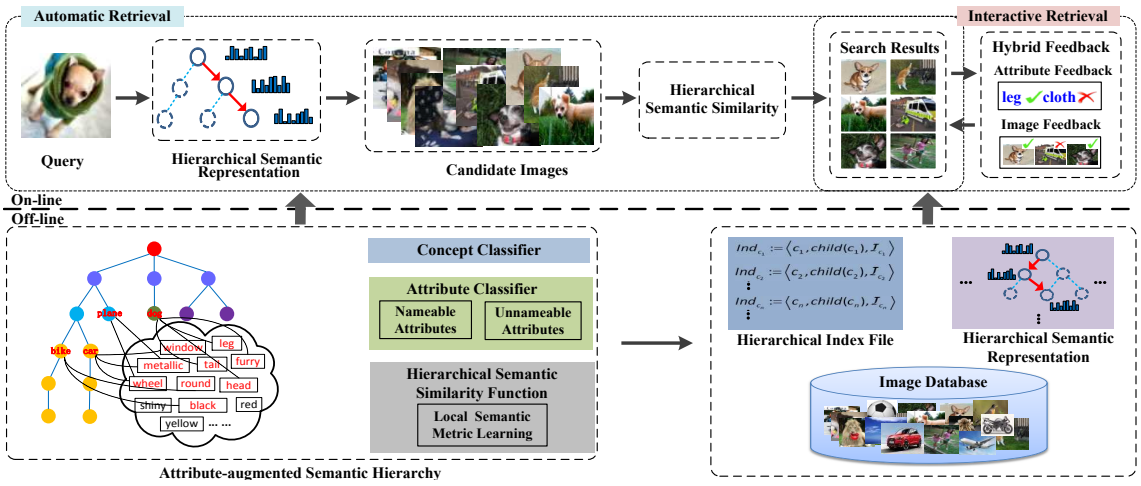


Figure 2: Illustration of the proposed Attribute-augmented Semantic Hierarchy (A^2SH) and the image retrieval system developed on A^2SH.

Motivated by the above observations, in this paper, we propose a novel Attribute-augmented Semantic Hierarchy (A^2SH), towards narrowing both the semantic and intention gaps, and demonstrate its effectiveness in both automatic and interactive Content-based Image Retrieval. Here, attributes refer to semantic descriptions of concepts such as the visual appearances (*e.g.*, "round" as shape, "metallic" as texture), sub-components (*e.g.*, "has wheel", "has leg"), and various discriminative properties (*e.g.*, "properties that dog has but cat do not"). Figure 2 shows an illustration of A^2SH. A semantic hierarchy consisting of semantic concepts is augmented by a pool of attributes. Each semantic concept is linked to a set of related attributes. For example, "car" is augmented by the attributes "window" and "metallic", *etc.* These attributes are specifications of the multiple facets of the corresponding concept and can act as an intermediate bridge connecting the concept and low-level visual cues. Moreover, they span a local semantic space in the context of the concept. On the other hand, the same attribute 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*". Hence, associating an attribute to concepts can reveal the heterogeneous meanings of the same attribute. We equip A^2SH with a set of concept classifiers and attribute classifiers. The concept classifier is used to predict the presence of a concept in images, while the attribute classifier aims to predict the presence of the attributes in the context of its associated concept. As a result, A^2SH is able to interpret the semantics of image content with a hierarchial semantic representation. In particular, an image can be represented as the responses from the concept classifiers as well as the linked attribute classifiers, leading to a hierarchical interpretation consisting of multiple levels of semantic granulations. Based on such interpretation, we develop a hierarchical semantic similarity function to precisely characterize the semantic similarities between images. The semantic similarity between any two images is computed as a hierarchical aggregation of their similarities in the local semantic spaces of their common semantic concepts at multiple levels. In the local semantic space of each concept, a local semantic metric is learnt to capture the semantic affinities between images in the context of the concept.

Based on the above A^2SH, we develop a content-based image retrieval system, which supports both automatic retrieval and interactive retrieval with user feedbacks. We expect A^2SH to both effectively narrow the semantic gap by structurlizing the semantics of image content with semantically meaningful hierarchical representations in terms of concepts and attributes, as well as the intention gap by shaping user search intent accurately from the feedbacks. The system flowchart is illustrated in Figure 2. In the offline part, we first learn the concept classifiers, attribute classifiers, and hierarchial semantic similarity function to equip A^2SH. We next use A^2SH to process the database images and obtain their hierarchical semantic representations. All the images are indexed hierarchically based on their semantic paths in the hierarchy to enable efficient large-scale retrieval. In the online part, a given query image is first processed by A^2SH, getting its hierarchical semantic representation, based on which a collection of candidate images are returned from the database according to the indexing. Similar images are then retrieved from the candidate set based on their hierarchical semantic similarities to the query. After the automatic retrieval, we present the results to solicit user feedbacks. We enable a broad channel of feedback to help user deliver search intent by providing hybrid feedbacks on attributes and images. While the image feedbacks collect positive and negative samples of user intent, the feedbacks on attributes compose a clearer semantic description of the intent [39], such as "*has head and leg, not furry.*" These hybrid feedbacks are analyzed by A^2SH, leading to a precise semantic interpretation of user intent, and are used to refine the search results. We expect the hybrid feedbacks leading 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-arts CBIR approaches. The main contributions of this paper 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 seman-

tics of images in the form of a hierarchical semantic representation, which is semantically meaningful.

- We develop a CBIR system based on the proposed A^{2}SH and demonstrate the effectiveness of A^{2}SH in narrowing the semantic and intention gaps in image retrieval over a large-scale image data set.
- 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 feedbacks on both attributes and images, which can help capture users' search intent more precisely based on A^{2}SH.

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

2. RELATED WORK

2.1 Semantic Hierarchy

A semantic hierarchy is a formally defined taxonomy or ontology, where each node represents a semantic concept, such as WordNet [10], ImageNet [4], and LSCOM [20] etc. It organizes semantic concepts from general to specific and has been shown to be effective in boosting visual recognition [7, 18] and retrieval [30, 5, 3]. Most of the recent works of exploiting semantic hierarchy for image retrieval focus on designing a semantic similarity function that embeds hierarchical information. In the method proposed by Deselaers and Ferrari [5], given two images, their visual nearest neighbor images were first found, and their semantic distance was computed as a distance between the concepts of their neighbors. Deng *et al.* [3] developed a 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 the state-of-the-art performance of image retrieval on ImageNet. Recently, Verma *et al.* [35] proposed to associate separated visual similarity metrics for every concept in a hierarchy and then learn the metrics jointly through an aggregated hierarchical metric. Different from existing works based on conventional semantic hierarchy, our work augments semantic hierarchy with a pool of attributes. The augmented hierarchy has better capacity in modeling the semantics of images. We characterize the semantic similarities among images by a hierarchical similarity function composed by a set of local semantic metrics learnt in the semantic spaces spanned by attributes in the context of various concepts.

2.2 Attributes

Attributes have attracted increasing attentions recently. Attributes refer to semantic descriptions of concepts such as sub-components (*e.g.*, "has wheel"), the visual appearances (*e.g.*, "round" as shape), and various discriminative properties (*e.g.*, "properties that dog has but cat do not"). Attributes are semantically meaningful as opposed to low-level visual features, and they are relatively easier to recognize automatically instead of the full concepts (*e.g.*, "dog", "car") [8]. Attributes can be exploited as intermediate-level descriptors of images to boost visual recognition [8, 17] and retrieval [12, 6, 9]. For example, Jaimes *et al.* [12] proposed a conceptual framework for indexing visual information based on semantic concepts and attributes. Douze *et al.* [6] proposed to represent an image by the concatenation of visual features and its response from attribute classifiers, and have shown that such representation can improve image retrieval significantly compared to pure visual features. Scheirer *et al.* [26] developed a probabilistic normalization method to normalize the responses from attribute classifiers, and have shown that the normalized representation is more effective. While most existing works represented images in the form of a flat semantic representations in terms of attributes, our work associate attributes to the concepts in a semantic hierarchy and represent 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.

2.3 Relevance Feedback

Relevance Feedback (RF) is a key technique to narrow down the intention gap in image retrieval [23, 1, 33, 2]. It attempts to capture users' search intent by iteratively collecting users' feedbacks on the retrieved images and refining the retrieval based on the feedbacks. A wealth of RF methods has been proposed to refine the original query or learn a relevance ranking function from the feedbacks, *i.e.*, the "relevant" and "irrelevant" images labeled by users. For example, the Query Point Movement (QPM) method [24] gradually refines the query point by moving it towards the "relevant" images and away from the "irrelevant" images. Tong *et al.* [34] proposed to learn a Support Vector Machine (SVM) from the "relevant" and "irrelevant" images, and then rank the images according to their responses from the SVM classifier. Tao *et al.* [33] proposed a asymmetric bagging and random subspace SVM method to learn a robust classifier from user feedbacks. Recently, Zhang *et al.* [39] developed an attribute feedback scheme, which collects user feedbacks on semantic attributes and ranks images according to the presence probabilities of the attributes in the images. Kovashka *et al.* [13] proposed to collect the user's relative judgments on attributes, such as "show me shoes that are more formal than these and shinier than those," to improve text-based image retrieval. Compared to these works, our work enables a broad channel of feedback to help user better delivery search intent by providing hybrid feedbacks on attributes and images. By modeling the feedbacks based on A^{2}SH, our work leads to a better understanding of user intent.

3. ATTRIBUTE-AUGMENTED SEMANTIC HIERARCHY

In this section, we first present a definition of the proposed Attribute-augmented Semantic Hierarchy (A^{2}SH) as follows:

Definition 1. *Attribute-augmented Semantic Hierarchy is a directed acyclic graph* $\mathcal{H} = (\mathcal{C}, \mathcal{A}, \mathcal{E}_C, \mathcal{E}_{CA})$*, 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^{2}SH organizes semantic concepts from general to specific, where each concept is augmented with a set of related attributes. These attributes comprehensively describe the multiple semantic facets of the concept, and span a local semantic space tailored to the concept. Next, we equip A^{2}SH with concept classifiers, attribute classifiers, and a hierarchical semantic similarity function. In particular, each of the concept classifiers predicts the presence of a semantic concept c in images. The attribute classifiers for the attributes $\mathcal{A}_c$ linked to concept c predict the presence of the attributes in the context of c. With such hierarchy, a given image can be represented as the responses from the concept classifiers as well as the linked attribute classifiers, leading to a hierarchical semantic interpretation consisting of semantics at multiple levels. The similarity between two images is computed by the hierarchical semantic similarity function, which aggregates their local similarities in the context of their common semantic concepts at multiple levels. A local semantic metric is learnt in the local semantic space of each concept.

3.1 Concept Learning

A concept classifier $f_c : \mathcal{X} \mapsto \{-1, +1\}$ predicts whether an image belongs to concept c, where $\mathcal{X}$ is an arbitrary feature space. Generally, given the concept classifiers in a hierarchy, the semantic path of an image can be efficiently predicted by the classifiers in a top-down fashion [32, 18]. Here, a semantic path $\mathcal{P}$ is a set of multi-level semantics $\mathcal{P} = (c_0 \rightarrow ... \rightarrow c_n)$ from the root c_0 and satisfies $\forall i > 0,\ f_{c_i}(x) = +1$. Next, we introduce the learning of the concept classifiers.

One way to learn each concept classifier is to use the conventional "one-vs-all" strategy, that is, learning the classifier by the images are from the concept as positive samples and images from others as negative samples. 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 [18]. However, this local training strategy results in classifiers that suffer from the "error prorogation" problem. To address the above problems, we here use the "hierarchical one-vs-all" strategy [32] to learn concept classifiers by exploiting the hierarchical relation 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:

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

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 which 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.

3.2 Attribute Learning

As aforementioned, we augment the concepts in a semantic hierarchy using a pool of attributes. Here, we exploit two types of attributes, including nameable attributes and unnameable attributes. **Nameable attributes** refer to the attributes that are human-nameable, such as the visual appearances and sub-components of a concept [39]. Moreover, a bunch of discriminative properties among concepts, such as "properties that dog has but cat do not" are automatically discovered. These discriminative properties are termed as **unnameable attributes**, since they are hard to be articulated explicitly by human. Such unnameable attributes are important in depicting a concept especially when the concept shares most of the nameable attributes with many others concepts. For example, "dog" and "cat" may share many nameable attributes, such as "furry", "tail", *etc.* while unnameable attributes need to be learned to differentiate the fur of cats from that of dogs using image examples. Together they offer a comprehensive description of the multiple facets of a concept.

3.2.1 Nameable Attribute Learning

A nameable attribute classifier $f_a^c : \mathcal{V} \mapsto \{-1, +1\}$ predicts the presence of a nameable attribute a of concept c in an image, where $\mathcal{V}$ is an arbitrary visual feature space. This classifier is learnt in the context of c, *i.e.*, using the positive samples of c with ground truth labelings of attribute a. Note that attributes normally correspond to partial visual cues of the whole image. 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. Hence, the visual feature $\mathcal{V}$ describing the whole image may not characterize the attributes well. This motivates us to perform feature selection towards selecting the most informative feature $\mathcal{V}_a^c$ for learning the attribute classifier f_a^c.

We propose a hierarchical feature selection mechanism to select features in a bottom-up fashion [8, 39]. Without loss of generality, we start with selecting visual features $\mathcal{V}_a^c$ for attribute a in the context of concept c. Suppose we have already selected the most informative features $\mathcal{V}_a^{c'}$ for attribute a in the context of $c' \in child(c)$. We merge these features to form the following base features $\widetilde{\mathcal{V}}_a^c$:

$$\widetilde{\mathcal{V}}_a^c = \bigcup_{c' \in child(c)} \mathcal{V}_a^{c'}. \tag{2}$$

The base features serve as a set of candidate features, from which we perform feature selection to discover the most informative features for learning a in the context of c. Intuitively, selecting features from the base features $\widetilde{\mathcal{V}}_a^c$ instead of the raw feature $\mathcal{V}$ leads to more informative features for learning attribute a. For example, it is more effective to select features for the attribute "head" of "animal" from the union of the features for 'head" of "dog", "cat", *etc.* Given the positive samples $\mathcal{I}_c$ of c labeled with/without attribute a, we train an ℓ_1-norm linear regressor to select $\mathcal{V}_a^c$ from $\widetilde{\mathcal{V}}_a^c$. For attributes appearing in most of $\mathcal{I}_c$, there are insufficient negative samples of a left. In order to learn an effective linear regressor, we instead access the images of c's ancestors until sufficient negative samples of a are collected. The regression results in a sparse set of model parameters where the nonzero elements correspond to feature dimensions that are selective for a. By selecting the nonzero dimensions of

$\widetilde{\mathcal{V}}_a^c$, we finally get the most informative feature $\mathcal{V}_a^c$, based on which we train a linear SVM as the attribute classifier f_a^c.

3.2.2 Unnameable Attribute Discovery

Since unnameable attributes are usually hard to be articulated by human, we cannot manually label images being positive/negative to an unnameable attribute and thus cannot obtain unnameable attributes classifiers by supervised learning. Hence, we propose to automatically discover unnameable attributes in an unsupervised fashion. Inspired by [22], we define unnameable attributes as the hypotheses that help in distinguishing a concept and its siblings. Next, we detail an iterative approach for discovering unnameable attributes in the context of concept c.

At each iteration t, we maintain an attribute set $\mathcal{A}_t$ containing nameable attributes of the concepts $\mathcal{C}_c = c \cup sibling(c)$, and unnameable attributes discovered thus far. We are concerned with the classification among the images of the concepts in $\mathcal{C}_c$, denoted as $\mathcal{I}$, represented by the responses from the attribute classifiers of $\mathcal{A}_t$. We use a nearest neighbor classifier to classify $\mathcal{I}$ into $\mathcal{C}_c$. Based on the classification result, we construct a symmetric confusion matrix[2], which can be viewed as a fully connected graph whose nodes correspond to $\mathcal{C}_c$. A strong edge weight indicates high confusion between the concepts linked by the edge. Next, we perform a spectral clustering algorithm on this graph to obtain several clusters. Each cluster is a subset of concepts that are most confused with each other. For images in the i-th cluster, we employ an unsupervised max-margin clustering algorithm [40] to generate a hyperplane separating them into two classes. The hyperplane serves as a hypothesis that helps in distinguishing the most confused concepts in the i-th cluster. We regard the hypothesis as an unnameable attribute a_i. Then, we learn a linear SVM classifier $f_{a_i}^c$ for a_i using the images of the two classes split by a_i. Finally, suppose that we have discovered m unnameable attributes at iteration t, we add them to the attribute set $\mathcal{A}_{t+1} \leftarrow \{\mathcal{A}_t, a_1, ..., a_m\}$, which is in turn used for the next iteration. The discovery process ends if no more new hypotheses can be found.

3.3 Hierarchical Semantic Similarity Learning

From the concept and attribute classifiers learnt above, we can generate a hierarchical semantic representation of an image as $\{(c_0 \rightarrow ... \rightarrow c_n); (\mathbf{z}^{c_0}, ..., \mathbf{z}^{c_n})\}$, where $(c_0 \rightarrow ... \rightarrow c_n)$ is the semantic path predicted by concept classifiers, c_0 is the root of the hierarchy and $(\mathbf{z}^{c_0}, ..., \mathbf{z}^{c_n})$ is the local semantic representations in terms of attributes along the path. Specifically, $\mathbf{z}^c$ is composed by the responses from the attribute classifiers of concept c as follows:

$$\mathbf{z}^c = \left[f_{a_1}^c(\mathbf{x}), f_{a_2}^c(\mathbf{x}), ..., f_{a_{|\mathcal{A}_c|}}^c(\mathbf{x}) \right]^T, \tag{3}$$

where f_a^c is the classifier for attribute $a \in \mathcal{A}_c$. We normalize f_a^c into the range of $[0, 1]$ by a probabilistic strategy [26].

With such hierarchical semantic representation of images, we formulate a hierarchical semantic similarity function to precisely characterize the 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:

[2]It is constructed by the sum of the original confusion matrix and its transposition.

$$S(I_i, I_j) = \sum_{c \in \mathcal{P}_{ij}} s(I_i, I_j; c), \tag{4}$$

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 lacks the fine characterization of the 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 precisely characterize the semantic similarities between images, we propose to learn a local semantic metric in the local semantic space of each concept.

3.3.1 Local Semantic Metric Learning

We define the local semantic distance between two images in the local semantic space of concept c as:

$$d\left(\mathbf{z}_i^c, \mathbf{z}_j^c; c\right) = \sqrt{(\mathbf{z}_i^c - \mathbf{z}_j^c)^T \mathbf{M}_c (\mathbf{z}_i^c - \mathbf{z}_j^c)}, \tag{5}$$

where $\mathbf{M}_c$ 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 learnt to bring together the images of the same concept as close as possible and sperate 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, and 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 $I_j \in Pos(c)$ as small as possible. $Pos(c)$ are the images belong to concept c. We denote $j \rightsquigarrow i$ as such 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:

$$\begin{aligned} \min_{\mathbf{M}_c} & \sum_{j \rightsquigarrow i} d^2(\mathbf{z}_i^c, \mathbf{z}_j^c; c) + \lambda \sum_{(i,j,k) \in \mathcal{T}} \xi_{ijk} \\ s.t. \;\; & \forall (i, j, k) \in \mathcal{T}, \\ & d^2(\mathbf{z}_i^c, \mathbf{z}_k^c; c) - d^2(\mathbf{z}_i^c, \mathbf{z}_j^c; c) \geq 1 - \xi_{ijk}, \\ & \xi_{ijk} \geq 0, \quad \mathbf{M}_c \succeq \mathbf{0}, \end{aligned} \tag{6}$$

where $\lambda > 0$ is the regularization constant. We employ the LMNN solver [37] modified with the above defined training triplets $\mathcal{T}$ to solve the metric learning problem. Note that solving the above problem is very efficient since the local semantic space is compact, *i.e.*, the dimension of $\mathbf{M}_c$ is low. With $\mathbf{M}_c$ we can compute the local semantic similarity between images as:

$$s(I_i, I_j; c) = exp(-d(\mathbf{z}_i^c, \mathbf{z}_j^c; c)), \tag{7}$$

which is in turn used to compose the hierarchical semantic similarities in Eq. (4).

4. IMAGE RETRIEVAL WITH A^{2}SH

In this section, we develop a content-based image retrieval system based on A^{2}SH. The system enables efficient and

effective automatic retrieval and interactive retrieval with hybrid feedbacks.

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 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 follows:

$$\text{Ind}_c := \ < c, child(c), \mathcal{I}_c >, \tag{8}$$

where $child(c)$ is the 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, the retrieval with the hierarchical indexing is as follows. First, we generate the hierarchical semantic representation of I_q along its semantic path $c_0 \rightarrow ... \rightarrow c_n$ based on A^{2}SH. Next, we perform fast retrieval of candidate images by looking-up the index file Ind_{c_n}. The candidate images consist of the images indexed by c_n and its children $child(c_n)$. Note that the number of the candidate images is significantly reduced compared to the size of entire database. These candidate images are then ranked according to their hierarchical semantic similarities to the query as in Eq. (4). In practice, as the semantic path prediction may not be perfect, c_n may not be exactly the same as the ground truth semantic path terminal of the query image. To address this problem, we set a look-back level b (b=3 in the experiments) and retrieve more 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 to linear scan.

Next, we analyze the time complexity of the retrieval, including three major steps: 1) semantic path prediction, 2) looking-up the index file to obtain candidate images and 3) generating Top K results according to the hierarchical semantic similarities of the candidate images. Denote the averaged fan-out (*i.e.*, averaged number of the children of a concept in the hierarchy) of A^{2}SH as F, the averaged leaf depth (*i.e.*, averaged depth over all the leaves) as D, and the concept classifier prediction cost as C. Therefore, we can estimate cost of the semantic path prediction at $\mathcal{O}(nFC)$, where $n \leq D+b$ is the average depth of the predicted path, and the candidate images retrieval cost at $\mathcal{O}(F^{D-n+b})$, which is the cost for sub-hierarchy traversal. Note that C is a small constant, D and F are 6.3 and 3.8 in our ImageNet hierarchy, respectively. Hence, the prediction cost and the candidate retrieval cost are very small, and the time cost for for retrieval is mainly from the third step, *i.e.*, ranking candidate images, which has a time complexity of $\mathcal{O}(ndN_c+N_c \log N_c)$, where d is the average dimensions of local semantic spaces, and N_c is the number of the 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 intention gap that hinders the understanding of user search intent by the system, the results from automatic retrieval often do not satisfy users' information needs. We therefore execute interactive retrieval by involving users' interaction with the system. We propose a Hybrid Feedback (HF) mechanism to help user deliver search intent by providing hybrid feedbacks on both the attributes and images. In particular, we allow a user to give "yes"/"no" feedbacks on attributes to state which attributes are in or not in his/her search intent, as well as relevance judgements on images to indicate which images are "relevant" or "irrelevant" to the intent. These hybrid feedbacks are then used to generate a precise semantic interpretation of user intent based on the proposed A^{2}SH. By iteratively collecting user feedbacks and refining the retrieval, the system can shape user intent more accurately and narrow the search to target gradually.

Suppose we are at the t-th 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 $\{\mathcal{Q} = (c_0 \rightarrow ... \rightarrow c_n); \mathcal{Z} = (\mathbf{z}^{c_0}, ..., \mathbf{z}^{c_n})\}$, 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: $\mathcal{Z}_t$, tailoring it to user intent by incorporating semantic descriptions delivered by image feedbacks (*i.e.*, $\mathcal{R}_t$ and $\overline{\mathcal{R}}_t$), and attribute feedbacks (*i.e.*, $\mathcal{B}_t$ and $\overline{\mathcal{B}}_t$). More specifically, we refine the query $\mathcal{Z}_t$ to make it close to the semantic representation of relevant images while away from that of the irrelevant ones. This refinement is carried out along the semantic path for every local semantic representation of the query, leading to a hierarchical semantic interpretation of user intent. Formally, for $\forall c \in \mathcal{Q}$, we have

$$\begin{aligned} \mathbf{z}^c_{t+1}[a] =& \mathbf{z}^c_t[a] + \beta \sum_{i \in \mathcal{R}_t} (\mathbf{z}^c_i[a] - \mathbf{z}^c_t[a])/|\mathcal{R}_t| \\ & -\gamma \sum_{j \in \overline{\mathcal{R}}_t} (\mathbf{z}^c_j[a] - \mathbf{z}^c_t[a])/|\overline{\mathcal{R}}_t|, \end{aligned} \tag{9}$$

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

User feedbacks 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 the values on the dimensions for $\overline{\mathcal{B}}_t$ as 0. For $\forall c \in \mathcal{Q}$, we have

$$\forall a \in \mathcal{A}_c, \mathbf{z}^c_{t+1}[a] = \begin{cases} 1, & a \in \mathcal{B}_t, \\ 0, & a \in \overline{\mathcal{B}}_t, \\ \mathbf{z}^c_t[a], & otherwise. \end{cases} \tag{10}$$

The resultant query $\mathcal{Z}_{t+1}$ is then used to generate the new search results based on the aforementioned hierarchical semantic similarity function. 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 in Eq. (5). 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.

5. EXPERIMENTS

In this section, we systematically evaluate the proposed Attribute-augmented Semantic Hierarchy (A^2SH) in content-based image retrieval. We first evaluate the elementary building blocks of A^2SH. Then, we investigate the effectiveness of A^2SH in automatic and interactive image retrieval.

5.1 Data and Methodology

5.1.1 Dataset Description

We conducted experiments on ImageNet [4], 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[4]. This data set 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 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[5] based on the attribute pool used in [8]. 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 3.2.2.

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 its any child is positive, otherwise, negative. 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.

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

[5] We removed the concept-specific attributes in [8, 39], 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 [25]. 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.*

5.1.2 Visual Features

To represent image content, we extracted four types of visual features: edge, color, texture, and Scale Invariant Feature Transform (SIFT) descriptors [16]. While edge descriptor was extracted globally from the entire image, the other descriptors were extracted locally. In particular, edges were found using the standard canny detector and their orientations were quantized into 9 unsigned bins. This gives rise to a 9-D edge descriptor for each image. Color descriptors as the 3-channel LAB values were densely extracted from each pixel. Texture descriptors were computed for each pixel as the 48-D responses of texton filter banks [14]. SIFT descriptors were densely extracted from image patches at multiple scales of $\{8\times 8, 12\times 12, 16\times 16\}$-pixel size, with 4-pixel step. For the color, texture, and SIFT descriptors, we adopted the state-of-the-art locality-constrained linear sparse coding (LLC) [36] method with max-pooling strategy to generate the global representations of images. We used a 512-D codebook for color and texture, 4,096-D for SIFT. As a result, we had 5,129-D global feature representation for each image. Since attributes usually correspond to image regions but not the entire image. To get features better characterizing attributes, we split each image into 2×3 grids, and extracted the above features from each grids. Finally, we obtained a 35,903-D (*i.e.,* $5{,}129 \times 7$) feature vector for each image.

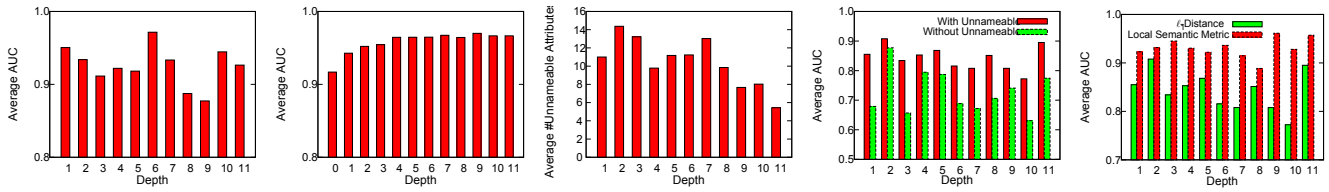
5.1.3 Experimental Setting

We learnt linear SVM classifiers for concepts and attributes by employing LIBLINEAR toolbox[1]. We learnt ℓ_1 linear logistic regressors to select informative features for learning attributes. For local metric learning, we deployed the LMNN toolbox [37] with training triplets configuration as described in Section 3.3.1. The algorithmic parameters of the above models were tuned through five-fold cross validation. We applied the Weibull distribution [26] 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, and e) **hBilinear** [3] retrieves images by the recently proposed bilinear semantic metric which was reported achieving the state-of-the-art performance on ImageNet dataset.

To evaluate the effectiveness of **A^2SH** in interactive retrieval, we compared it to the following three interactive retrieval methods. a) **QPM** [24]: Query Point Movement method updates the query based on image feedbacks, and refines search results using the new query; b) **SVM** [34]: this approach learns a SVM classifier from the "relevant" and "irrelevant" images and ranks images according to their

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



(a) (b) (c) (d) (e)

Figure 3: Performance at different depth levels measured by Average AUC: (a) concept classifers; (b) nameable attribute classifiers; (d) classification by exploiting unnameable attributes; (e) local semantic metrics. The average AUC at a depth level is obtained by averaging the AUC values of all the classifiers at that level. The average number of unnameable attribute discovered at different depth levels is shown in the subfigure (c). The depth of the root is 0.

responses from the classifiers; and c) **AF** [39]: the recently proposed Attribute Feedback approach collects user feedbacks on attributes and then ranks images according to the presence probabilities of the attributes in the images. Note that our approach enables hybrid feedbacks on attributes and images. For the sake of fair comparison, we incorporated image feedbacks into AF, such that it also uses hybrid feedbacks. Moreover, all the baseline methods were performed on the flat semantic representations of the images, rather than the low-level visual descriptors.

We conducted the evaluation in two settings with a fixed number of feedbacks and a fixed time limit, respectively. In the first setting, we conducted five feedback iterations with 20 feedbacks per iteration. For the QPM and SVM methods, 20 feedbacks on top 20 images were collected in each iteration. For the AF and our A^{2}SH methods, the same number of feedbacks were collected, including 5 attribute feedbacks and 15 image feedbacks. Five informative attributes were suggested in each iteration for soliciting attribute feedbacks. We here employed the suggestion strategy in [39]. 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 above 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 to 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 the above evaluations, we used the search results from the best automatic retrieval method, *i.e.*, the proposed A^{2}SH, as the initial results for interactive retrieval.

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

5.1.4 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 [21]. 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 j-th 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 overall performance metrics.

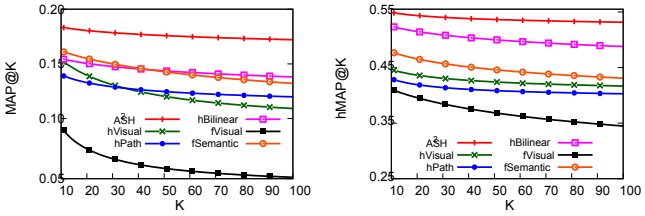
5.2 Experimental Results

5.2.1 Evaluations of Concept Classifiers, Attribute Classifiers, and Local Semantic Metrics

Figure 3(a) shows the average AUC values of the concept classifiers at different depth levels in the hierarchy [4]. From these results, we can see that the concept classifiers at most levels achieve an average AUC above 0.9 except those at depth 8 and 9 with average AUC of 0.89 and 0.88, respectively. The performance of attribute classifiers is illustrated in Figure 3(b), from which we can see that the attribute classifiers at every level obtain 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 precise hierarchial semantic representations of images.

A bunch of unnameable attributes are discovered to complement the nameable attributes for better characterizing a concept. Figure 3(c) shows the average number of unnameable attributes discovered for the concepts at different depth levels. As aforementioned, we have no ground truth of the unnameable attributes on images and thus cannot evaluate the performance of their classifiers directly. Alternatively, 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 unnameable attributes. The classification performance is illustrated in Figure 3(d). From the results, we can see that the discovered unnameable attributes can improve the classification performance significantly. This indicates that the unnameable attributes can help to provide a more comprehensive and discriminative description of the multiple facets of a concept.

We evaluated the effectiveness of the local semantic metrics as follows. We used the local semantic metric of each



(a) (b)

Figure 4: Performance of automatic image retrieval over the 95,800 queries.

concept to help classifying the images of the concept from those of all its siblings by using the 5-nearest neighbor classifier [37]. We compared the local semantic metrics again the widely used ℓ_1 distance. The classification performance comparison is illustrated in Figure 3(e). We can see that the proposed local semantic metric outperforms the ℓ_1 distance significantly at every depth level. It achieves relative improvement 10.7% at various depth levels on average. This demonstrates the effectiveness of the local semantic metric and its capacity in composing an effective hierarchical semantic similarity to precisely characterize the semantic affinities among images.

5.2.2 *Evaluations of Automatic Retrieval*

Figure 4 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 performance by 22.4%, 23.1%, 41.5%, and 46.0% relatively in terms of 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 A^{2}SH in image retrieval. The superiority of A^{2}SH to the other methods arises from the following aspects: a) A^{2}SH models the semantics of images in the form of a 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 precise interpretation of image semantics; and b) The hierarchical similarity function in A^{2}SH more accurately characterizes the semantic similarities among images by ensembling the local semantic metrics in the context of various concepts.

Table 1: Average retrieval time per query of automatic image retrieval over the 95,800 queries.

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

Table 1 lists the average retrieval time per query of the five automatic retrieval over the 95,800 queries by the five approaches. We can observe that A^{2}SH provides highly efficient retrieval. It significantly reduces the retrieval time by several orders of magnitude compared to the other methods. The reasons are two folds. First, A^{2}SH represents images in the form of a compact hierarchical semantic representation, enabling to fast similarity computation. Second, the hierarchical indexing in A^{2}SH 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^{2}SH system. hBilinear was accelerated using the indexing technique in [3]. Here, we do not list the time cost of the hPath method, since it is a sub-procedure of A^{2}SH and hVisual, *i.e.*, 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 as described in Section 5.1.2, by inverted files to accelerate the visual retrieval in fVisual and hVisual.

Table 2: 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^{2}SH	**24.67**	**22.80**	**22.03**	**68.37**	**66.04**	**64.08**
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

5.2.3 *Evaluations of Interactive Retrieval with Hybrid Feedbacks*

Figure 5 illustrates the performance of interactive retrieval with five feedback iterations in terms of MAP and hMAP at the top 20, 50, and 100 search results, respectively. From these results, the following observations can be obtained: a) The proposed A^{2}SH based interactive retrieval approach outperforms the other three methods at every iteration and all the top 20, 50, and 100 results; b) A^{2}SH significantly reduces the interaction efforts while it achieves comparable performance to the other three methods. For example, consider the MAP at top 20 results, A^{2}SH obtains a comparable performance at the 3rd, 2nd, and 2nd iteration as compared to AF, QPM and SVM at the last round, respectively. In other words, A^{2}SH can reduce labeling efforts by about 40%, 60%, and 60% as compared to the three methods, respectively; and c) The performance improvements of A^{2}SH and AF over QPM and SVM indicate the effectiveness of attribute feedbacks in delivering user search intent. The superiority of A^{2}SH to AF demonstrates that A^{2}SH can infer user intent more accurately from the feedbacks.

Table 2 lists the performance of the four interactive retrieval approaches with a fixed time limit of 2 minutes. From these results, we can see that the proposed A^{2}SH achieves the best performance in terms of both MAP and hMAP at all the top 20, 50, and 100 results. This demonstrates that A^{2}SH shapes user intent more precisely and quickly within the same interaction time and can generate more accurate search results as compared to the other methods.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a novel Attribute-augmented Semantic Hierarchy (A^{2}SH) 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 the concept classifiers, attribute classifiers, and hierarchical similarity function to equip A^{2}SH.

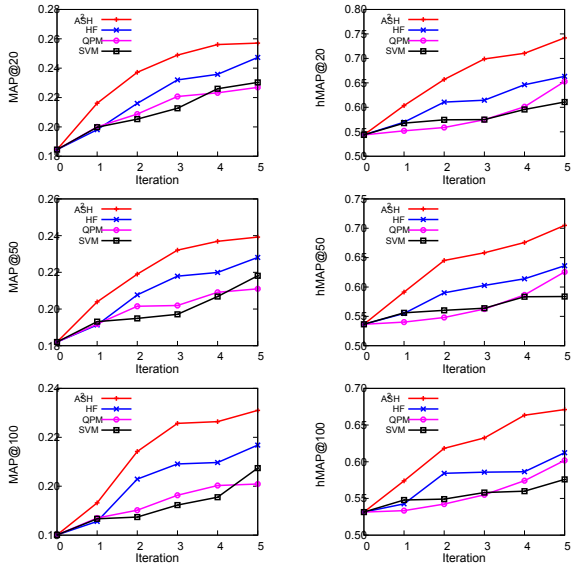


Figure 5: Performance of interactive retrieval with five feedback iterations over the $95,800$ queries.

Based on the proposed A^{2}SH, we developed a content-based image retrieval system supporting both automatic retrieval and interactive retrieval with user feedbacks. A hybrid feedback mechanism was developed to collect broad array of feedbacks on attributes and images. These feedbacks were then utilized to improve the retrieval based on A^{2}SH. We systematically evaluated the A^{2}SH based image retrieval system on a large-scale corpus of over one million Web images. The experimental results demonstrated the effectiveness of A^{2}SH in bridging the semantic and intention gaps, leading to more accurate results compared to the state-of-the-arts CBIR approaches.

We will continue our future works in two directions. First, we will study how to build or refine the A^{2}SH automatically by mining concepts, attributes, and their intrinsic relations from online user-generated content. Second, we will apply the proposed A^{2}SH to other applications, such as the user-generated content organization and web video retrieval.

7. ACKNOWLEDGMENTS

This work was supported by NUS-Tsinghua Extreme Search (NExT) project under the grant No.: R-252-300-001-490.

8. REFERENCES

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

[2] R. Datta, D. Joshi, J. Li, and J. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 2008.

[3] J. Deng, A. C. Berg, and L. Fei-Fei. Hierarchical semantic indexing for large scale image retrieval. In *CVPR*, 2011.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[5] T. Deselaers and V. Ferrari. Visual and semantic similarity in imagenet. In *CVPR*, 2011.

[6] M. Douze, A. Ramisa, and C. Schmid. Combining attributes and fisher vectors for efficient image retrieval. In *CVPR*, 2011.

[7] J. Fan, Y. Gao, and H. Luo. Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation. *TIP*, 2008.

[8] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[9] X. Felix, R. Ji, M. Tsai, G. Ye, and S. Chang. Weak attributes for large-scale image retrieval. In *CVPR*, 2012.

[10] C. Fellbaum. Wordnet. *Theory and Applications of Ontology: Computer Applications*, 2010.

[11] A. Hanjalic, C. Kofler, and M. Larson. Intent and its discontents: the user at the wheel of the online video search engine. In *MM*, 2012.

[12] A. Jaimes and S. fu Chang. A conceptual framework for indexing visual information at multiple levels. In *SPIE Internet Imaging*, 2000.

[13] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *CVPR*, 2012.

[14] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 2001.

[15] M. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *TOMCCAP*, 2006.

[16] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[17] Z. Ma, Y. Yang, Z. Xu, S. Yan, N. Sebe, and A. G. Hauptmann. Complex event detection via multi-source video attributes. In *CVPR*, 2012.

[18] M. Marszalek and C. Schmid. Semantic hierarchies for visual object recognition. In *CVPR*, 2007.

[19] F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. In *MM*, 2003.

[20] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *Multimedia, IEEE*, 2006.

[21] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A. F. Smeaton, and G. Quénot. Trecvid 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID*, 2012.

[22] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*, 2011.

[23] Y. Rui, T. S. Huang, and S.-F. Chang. Image retrieval: Current techniques, promising directions, and open issues. *JVCIR*, 1999.

[24] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *TCSVT*, 1998.

[25] O. Russakovsky and L. Fei-Fei. Attribute learning in large-scale datasets. In *ECCV*, 2010.

[26] W. J. Scheirer, N. Kumar, P. N. Belhumeur, and T. E. Boult. Multi-attribute spaces: Calibration for attribute fusion and similarity search. In *CVPR*, 2012.

[27] N. Sebe, M. S. Lew, X. Zhou, T. S. Huang, and E. M. Bakker. The state of the art in image and video retrieval. In *Image and Video Retrieval*. 2003.

[28] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *TPAMI*, 2000.

[29] J. R. Smith and S.-F. Chang. Visualseek: a fully automated content-based image query system. In *MM*, 1997.

[30] C. G. Snoek, B. Huurnink, L. Hollink, M. De Rijke, G. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. *TMM*, 2007.

[31] C. G. Snoek and M. Worring. Concept-based video retrieval. *FTIR*, 2008.

[32] Y. Song, M. Zhao, J. Yagnik, and X. Wu. Taxonomic classification for web-based videos. In *CVPR*, 2010.

[33] D. Tao, X. Tang, X. Li, and X. Wu. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *TPAMI*, 2006.

[34] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *MM*, 2001.

[35] N. Verma, D. Mahajan, S. Sellamanickam, and V. Nair. Learning hierarchical similarity metrics. In *CVPR*, 2012.

[36] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[37] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2006.

[38] Z.-J. Zha, L. Yang, T. Mei, M. Wang, and Z. Wang. Visual query suggestion. In *MM*, 2009.

[39] H. Zhang, Z.-J. Zha, S. Yan, J. Bian, and T.-S. Chua. Attribute feedback. In *MM*, 2012.

[40] K. Zhang, I. W. Tsang, and J. T. Kwok. Maximum margin clustering made practical. *TNN*, 2009.