Fine-Grained Image Search

Lingxi Xie, Jingdong Wang, Bo Zhang, and Qi Tian, *Senior Member, IEEE*

***Abstract*—Large-scale image search has been attracting lots of attention from both academic and commercial fields. The conventional bag-of-visual-words (BoVW) model with inverted index is verified efficient at retrieving near-duplicate images, but it is less capable of discovering fine-grained concepts in the query and returning semantically matched search results. In this paper, we suggest that instance search should return not only near-duplicate images, but also fine-grained results, which is usually the actual intention of a user. We propose a new and interesting problem named fine-grained image search, which means that we prefer those images containing the same fine-grained concept with the query. We formulate the problem by constructing a hierarchical database and defining an evaluation method. We thereafter introduce a baseline system using fine-grained classification scores to represent and co-index images so that the semantic attributes are better incorporated in the online querying stage. Large-scale experiments reveal that promising search results are achieved with reasonable time and memory consumption. We hope this paper will be the foundation for future work on image search. We also expect more follow-up efforts along this research topic and look forward to commercial fine-grained image search engines.**

***Index Terms*—Applications, evaluation, fine-grained image search, problem formulation, semantic indexing.**

I. Introduction

RECENT years have witnessed the development of Web-scale content based image retrieval (CBIR). Based on the Bag-of-Visual-Words (BoVW) model and the inverted index structure, the state-of-the-art image search engines are capable of indexing billions of images and returning results in milliseconds. Although different approaches have been proposed to improve near-duplicate search performance, we still see few search engines which retrieve images based on the analysis of fine-grained concepts in the query.

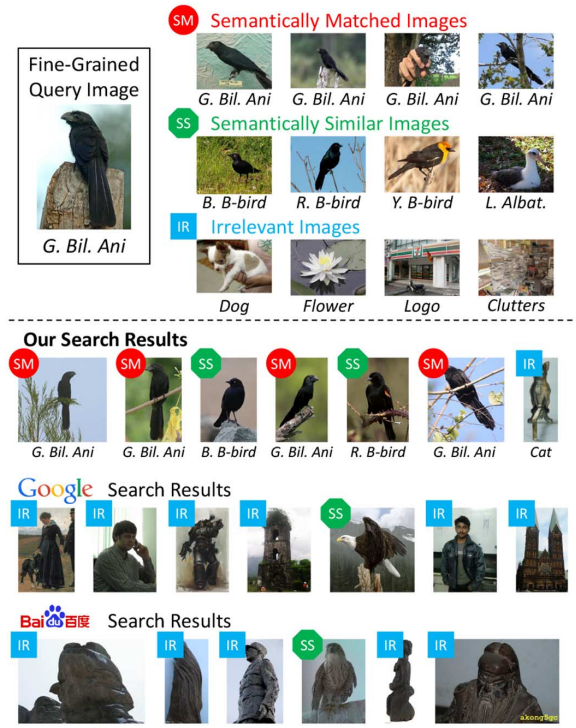


Fig. 1. (above) Typical fine-grained search query and three types of candidate images (semantically matched, semantically similar, and irrelevant). (below) The comparison between our results and those of Google/Baidu reveals that our approach might deal with fine-grained concepts in the query which is not captured by conventional search engines.

In fact, it is a common requirement that an image search engine should return *fine-grained* results, i.e., those candidates containing exactly the same semantic concept as the query image. For example, when a user uploads a *bird* or *flower* photo he/she just took, retrieving those images containing exactly matched *bird* or *flower* species helps a lot to his/her understanding; a user could also benefit from the search result of a *car* if it contains *cars* of the same brand/prototype with the query image.

Based on the observations above, we propose a new research topic, *fine-grained image search*, which differs from conventional cases in that candidate images might have multiple levels of relevance to the query. We take the query shown in Fig. 1 (*groove billed ani*, a *bird* species) as an example. We desire to retrieve other images also containing the same species, i.e., the *groove billed ani*. If we fail to find the exactly matched images, we still prefer those candidates with similar semantics, i.e., containing another *bird* species such as a *Brewer blackbird*, to the irrelevant ones containing, say, *dogs*, *flowers* or *clutters*.

Manuscript received August 13, 2014; revised December 07, 2014; accepted February 18, 2015. Date of publication March 04, 2015; date of current version April 15, 2015. This work was supported by the National Basic Research Program (973 Program) of China under Grant 2013CB329403, Grant 2012CB316301, and Grant 2014CB347600, by the National Natural Science Foundation of China under Grant 61332007, Grant 61273023, and Grant 61429201, and by the Tsinghua University Initiative Scientific Research Program under Grant 20121088071. The work of Q. Tian was supported in part by the NEC Laboratories of America Faculty Research Awards under ARO Grant W911NF-12-1-0057. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Vasileios Mezaris.

L. Xie and B. Zhang are with the State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing 100084, China, and also with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: 198808xc@gmail.com; dcszb@mail.tsinghua.edu.cn).

J. Wang is with Microsoft Research, Beijing 100080, China (e-mail: jingdw@microsoft.com).

Q. Tian is with the Department of Computer Science, University of Texas at San Antonio, San Antonio, TX 78249 USA (e-mail: qitian@cs.utsa.edu).

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

Digital Object Identifier 10.1109/TMM.2015.2408566

It is worth noting that fine-grained image search is very challenging especially in a large database. As shown in Fig. 1, commercial search engines such as Google and Baidu often fail in such cases, due to the lack of semantic analysis but simply parsing an image as a set of local or global features. Here, we are not claiming that our algorithm works better than Google or Baidu. Actually, the comparison is not fair since we only focus on a handful of basic-level categories, but real-world commercial search engines must deal with many other concepts in all the Web images. We are just demonstrating the difference between fine-grained and conventional image search tasks, which distinguishes this work from the previous ones not considering fine-grained concepts in the search process.

In this paper, we formulate the fine-grained image search problem by constructing a hierarchical database, and defining an evaluation method to judge the search quality. We propose a baseline algorithm which uses fine-grained classification scores to represent images' semantic attributes, and co-indexes images into an inverted file according to their distance in the attribute space. An efficient online querying stage is also designed to lookup the indexed structure in real-time. Experimental results reveal the promising performance of our approach. We will also release our benchmark database and expect more research efforts to be triggered in the future.

The main contribution of this paper could be summarized in three aspects.

- *New problem.* We formulate fine-grained image search, which is a new problem in the research field of multimedia information retrieval. To the best of our knowledge, this problem has not been well studied in the previous literatures. We expect it could become a new research direction which brings new insights and novel applications.
- *New dataset.* We construct a new database based on several basic image classification and object retrieval datasets. Both fine-grained and near-duplicate concepts are contained, making it difficult for conventional algorithms to return satisfying search results. We will make the database publicly available, and gradually add new images to enlarge the database.
- *New framework.* Based on state-of-the-art image classification and object retrieval techniques, we design a new framework for fine-grained image search. The flowcharts of both offline indexing and online querying are significantly different with conventional algorithms. Our framework serves as a very first trial towards this challenging problem, and also provides a baseline performance which is convenient to be compared with future works.

The remainder of this paper is organized as follows. The related works are briefly reviewed in Section II. We then formulate the fine-grained image search problem in Section III, and propose the search framework in Section IV. After experimental results and search examples are provided in Section V, we draw the conclusions in Section VI.

II. Related Works

Our work is closely related to several popular research topics, including, large-scale image search, fine-grained object recognition, semantic co-indexing and semantic ontology.

A. Large-Scale Image Search

Large-scale image search is aimed at retrieving query-relevant candidates from a large image corpus. One of the most popular and efficient approaches is based on the Bag-of-Visual-Words (BoVW) model and the inverted index structure [1]. Local descriptors such as SIFT [2], [3] and SURF [4] are extracted on the detected interest points [2]–[6]. A large visual vocabulary (codebook) is trained using the hierarchical [7] or approximate [8] versions of K-Means, and the descriptors are quantized onto the codebook as visual words. Codebook training-free approaches [9], [10] have also been proposed for quantization. The inverted index [1], [83] is then built as an efficient data structure representing the relationship between images and visual words.

Naive online querying stage ranks the candidates according to their number of feature matches to the query image. Due to the fact that the quantization loss greatly reduces the discriminative power of local features, the initial image search results usually suffer from low precision and/or recall. Various post-processing techniques are therefore designed to improve the initial search quality, including false match filtering [8], [11], Hamming embedding [12], soft assignment on visual words [13], query expansion [14], [15], query adaptation [16], prototype extraction [17], selecting high-quality features [18], [19], contextual weighting [20]–[22], sparsity-constrained measurement [23], distance learning [24], visual phrases [25], [26], feature similarity adaption [27], and diffusion-based methods [28]–[30]. It is also suggested to improve image retrieval with unlabeled training data [31].

B. Fine-Grained Object Recognition

Object recognition tasks often require capturing global property of images, therefore the Bag-of-Visual-Words (BoVW) model with feature pooling strategy [32] is often adopted. Local descriptors such as SIFT [2], [33] and HOG [34] are extracted on small patches, and a codebook is trained using K-Means [35] or GMM clustering [36]. Sparse coding [37], [35] or Fisher vector encoding [36] algorithms are used to quantize the descriptors onto the feature space, and spatial pooling strategies [38]–[40] are adopted to summarize local features into long vectors [41]. Generalized machine learning algorithms such as SVM are used for training and testing.

In the fine-grained classification tasks [42], [43], the main difficulty arises from the surprising inter-class similarity, therefore it becomes more important to uncover really useful patches on the objects [44]. One can see many state-of-the-art algorithms targeting at using visual attributes [45], pose estimation [46], template matching [47], pair-wise [48] and/or part-based [49], [50] pooling models, and geometric information [51], [52] for part segmentation.

C. Semantic Co-Indexing

In the fine-grained search tasks, it is not always true that the candidate images share a number of common visual words with the query. In this case, conventional BoVW model with inverted index might fail to retrieve really relevant candidates. Consequently, it requires other higher-level clues to help describing the visual properties of images. Intuitive solutions

come from summarizing local features into global representations [53], leveraging recognition scores [54] or referring to privileged information [55].

Semantic co-indexing is aimed at integrating other types of clues [56] into the inverted index structure, so that the online querying stage could benefit from these additional information to improve the search accuracy. It is suggested in [54] to use Hashing strategy for retrieving the approximate nearest neighbors in the attribute space efficiently, and [57] proposes to store the semantically similar pairs in the inverted index beforehand. Some authors also suggest to use feature fusion methods [58], [59] for integrating multiple types of information together.

D. Semantic Ontology

In computer science and information science, an ontology is a formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain of discourse. An ontology compartmentalizes the variables needed for some set of computations and establishes the relationships between them.

The field of computer vision creates ontologies to limit complexity and organize information. Especially, ontology is widely adopted to describe the category of a namable object. For example, the *ImageNet* [60] database is constructed based on the *WordNet* project [61] and its related works [62], [63]. Each leaf node of the WordNet concept tree correspond to a fine-grained object category, such as *groove-billed ani* (a *bird* species) or *golden retriever* (a *dog* breed), and basic-level concepts correspond to the nodes with higher levels.

Ontology can also apply to problem solving [64]. Semantic ontology could be used for bridging the semantic or intent gap between query and candidates, and many researchers have been adopting this idea to define and/or improve the performance of visual information retrieval [65]–[67]. A survey on this topic could be found in [68].

III. Problem Formulation

This section formulates *fine-grained image search*. Since this is a new problem which is less studied before, we first discuss the goal of the problem, and then construct a new database and suggest a hierarchical scoring function to evaluate different search engines.

A. Fine-Grained Image Search

Conventional image search problems usually require retrieving near-duplicate or partial-duplicate candidates, such as a *landmark building* (*Oxford* [8]), a *logo* (*FlickrLogo-32* [69]) or a *reoccurring object* (*UKBench* [7]). In those cases, a candidate image is either relevant or irrelevant to the query. Fine-grained image search, however, queries the database with an image containing a fine-grained concept, such as a *groove billed ani* (a *bird* species) or a *golden retriever* (a *dog* breed). In this case, there might exist multiple levels of relevance between query and candidate images.

Throughout this paper, we use three levels to model the relevance between query and candidates. Two images are named *semantically matched*, if they contain exactly the same semantic concept, e.g., the same species of *birds*, the same instance of *buildings*, or *logos* of the same brand. Image are considered *semantically similar* if the concepts are the same only at the basic level, e.g., a *groove billed ani* vs. a *Brewer blackbird* (both are *birds*), or a *Mona Lisa* vs. a *potato eater* (both are *paintings*). Image are defined *irrelevant* if the concepts are not the same even at the basic level, e.g., a *bird*, a *dog*, a *building* and a *painting* are irrelevant to each other.

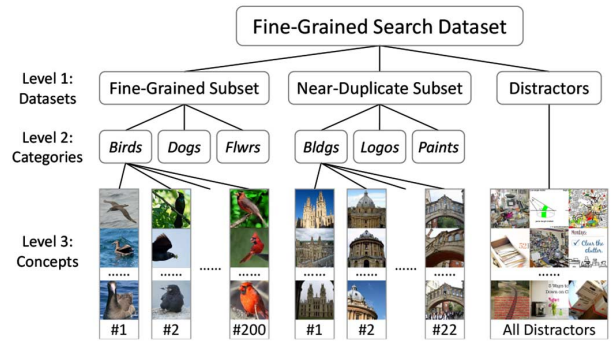


Fig. 2. Three-level hierarchical database for fine-grained image search.

B. Database

The database for fine-grained image search contains three levels, as illustrated in Fig. 2. In the first level, the database is partitioned into three subsets, i.e., fine-grained, near-duplicate, and distractor image sets. The composition of three subsets are summarized as follows.

1) *Three fine-grained concept groups: birds* (the Caltech-UCSD *Bird-200*-2011 dataset [42]), *dogs* (the Stanford *Dog-120* dataset [43]), and *flowers* (the Oxford *Flower-102* dataset [70]).
2) *Three near-duplicate instance groups: buildings* (the *Oxford* and *Paris Buildings* datasets with 22 named buildings [8]), *logos* (the *CarLogo-51* dataset [30]), and *paintings* (the *FamousPaint-26* dataset[1]).
3) *Distractors.* We crawl one million irrelevant images from the Web to test the scalability of the algorithm.

The basic-level visual concepts, e.g., *birds*, *dogs*, *buildings*, have the level index 2, and the fine-grained concepts, e.g., *groove billed ani* (a *bird*), *golden retriever* (a *dog*), *Oxford all souls* (a *building*), have the level index 3.

In all the fine-grained datasets, objects (*birds*, *dogs* and *flowers*) are labeled with their biology species. For example, the *Bird-200* dataset contains 200 *bird* species and each species contains up to 60 images. All these labels, such as *groove billed ani* (a *bird* species) or *golden retriever* (a *dog* breed), could be found in the *WordNet* [61], a large dictionary of semantic ontologies. In all the near-duplicate datasets, however, a category is defined as a repeatable object instance, such as a *building* picture taken in different conditions, a registered car brand *logo*, or a famous *paint* and its copies. The name of each instance (e.g., *Eiffel Tower*, *BMW logo* or *Mona Lisa paint*) is not ambiguous.

[1]This dataset is publicly available online. Type "FamousPaint-26 dataset" for searching.

TABLE I
COMPOSITION OF THE FINE-GRAINED IMAGE SEARCH DATABASE

Dataset Type	Dataset Name	Concept Groups	Training Images	Searching Images
Fine-Grained	*Bird*	200	5994	5794
	Dog	120	12000	8580
	Flower	102	2040	6149
	Subtotal	422	20034	20523
Near-Dup.	*Building*	22	2200	11455
	Logo	52	5200	11903
	Paint	26	2600	3148
	Subtotal	100	10000	26506
Distractors	*Web*	–	3000	1000000
	Total	–	33034	1047029

We take a small portion of the database as training images. For the fine-grained sets, we use the fixed training/testing splits provided by the authors, and each training image is labeled with its fine-grained category name, such as *groove billed ani*, a *bird* species. We also randomly select some non-query images (100 per instance) from the near-duplicate subsets, and 3000 images from the distractor set for training. These images are simply labeled as *not containing a fine-grained concept*. All the images, except for the fine-grained training cases, compose the search database. The query set consists of all the fine-grained testing images and author-specified near-duplicate querying cases. The detailed statistics of the database is summarized in Table I.

There also exist some fine-grained datasets which are not included in the composed database, such as the Oxford **Pet-37** dataset [71] (37 *pet* breeds, 7390 images), the *Aircraft-100* dataset [72] (100 *aircraft* models, 10000 images), and the recently published *Food-101* dataset [73] (101 *food* categories, 101000 images). The number of images in these unused datasets is comparable with the used datasets and less than the distractor set. As illustrated in Section IV-E, the scalability of our model makes it easy to deal with a larger number of fine-grained concepts.

C. Evaluation

We start with defining the relevance between two images, say, $\mathbf{A}$ and $\mathbf{B}$, which is natural in the hierarchical database: $\mathrm{rel}(\mathbf{A},\mathbf{B}) = \max\{\mathrm{LCA}(\mathbf{A},\mathbf{B}) - 1, 0\}$, where $\mathrm{LCA}(\cdot,\cdot)$ is the level index of the least common ancestor of two images. Obviously, two images are *semantically matched* if their relevance value is 2, *semantically similar* if the value is 1, and *irrelevant* if the value is 0.

Given a query image $\mathbf{q}$, we can obtain a set of P retrieved images: $\mathcal{R}_\mathbf{q} = \{\mathbf{I}_{\mathbf{q},1}, \mathbf{I}_{\mathbf{q},2}, \ldots, \mathbf{I}_{\mathbf{q},P}\}$. Denote $\mathrm{rel}_{\mathbf{q},p} = \mathrm{rel}(\mathbf{q}, \mathbf{I}_{\mathbf{q},p})$ for $p = 1, 2, \ldots, P$, we can calculate the Discounted Cumulative Gain (DCG) [74] of $\mathcal{R}_\mathbf{q}$ as

$$\mathrm{DCG}\left(\mathcal{R}_\mathbf{q}\right) = \sum_{p=1}^{P} \frac{2^{\mathrm{rel}_{\mathbf{q},p}} - 1}{\log_2\left(p+1\right)} \tag{1}$$

DCG is then normalized into nDCG by dividing it by the ideal DCG value at this query. The calculation process is illustrated in Fig. 3. The nDCG values of all query images are averaged as the final score. It is proved that nDCG scores serve as a good measure for evaluating different ranking results [75].

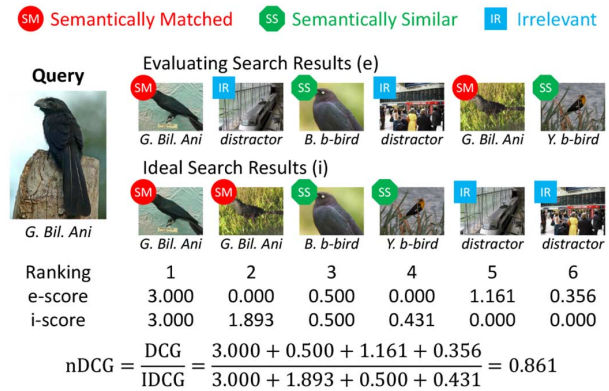


Fig. 3. Evaluation of a toy search results with two semantically matched, two semantically similar, and two distractor images.

IV. OUR APPROACH

This section presents our approach for fine-grained image search. The overall framework is illustrated in Fig. 4.

A. Fine-Grained Recognition

This module is aimed at training several classifiers for fine-grained concept discovery and recognition. We use the Bag-of-Visual-Words (BoVW) model with the SVM classifier.

The BoVW model consists of three stages, i.e., descriptor extraction, feature encoding and spatial pooling. We start from a raw image $\mathbf{I} = (a_{ij})_{W\times H}$, which is resized with the aspect ratio preserved so that the larger axis is 600 pixels. On a densely sampled set of patches, we use VLFeat [76] to extract 128-dimensional greyscale SIFT descriptors [2]. The spatial stride and window size are 8 and 16 for all images, respectively. The descriptors are reduced to 64 dimensions using PCA. Denote the set of reduced local descriptors with $\mathcal{D}_a = \{\mathbf{d}_{a,1}, \mathbf{d}_{a,2}, \ldots, \mathbf{d}_{a,M_a}\}$, where $\mathbf{d}_{a,m}$ denotes the description vector of the m-th descriptor, and M_a is the total number of descriptors. To quantize the descriptors onto the feature space, we train a Gaussian Mixture Model (GMM) with 256 components. The numbers of descriptors collected for PCA and GMM training are around 2 million. The Improved Fisher Vectors (IFV) [36] are calculated for efficient feature encoding. Given a GMM with 256 components, each patch descriptor $\mathbf{d}_{a,m}$ is assigned a $2 \times 256 \times 64$-dimensional feature vector $\mathbf{f}_{a,m}$. Denote $\mathcal{F}_a$ as the set of encoded features: $\mathcal{F}_a = \{\mathbf{f}_{a,1}, \mathbf{f}_{a,2}, \ldots, \mathbf{f}_{a,M_a}\}$. We aggregate the visual phrases with sum-pooling for image-level representation: $\mathbf{x}_a = \sum_{m=1}^{M_a} \mathbf{f}_{a,m}$. A 2-layer Spatial Pyramid Matching (SPM) [38] follows by dividing the image into 2×2 subregions, and concatenating the individual pooled vectors as a long vector $\widetilde{\mathbf{x}}_a$.

In practice, we also extract a set of 96-dimensional LCS descriptors [36] $\mathcal{D}_b$ on the densely sampled patches to capture color information. We follow the same flowchart to encode the descriptors into an image-level vector $\widetilde{\mathbf{x}}_b$, and concatenate $\widetilde{\mathbf{x}}_a$ and $\widetilde{\mathbf{x}}_b$ as a single vector $\widetilde{\mathbf{x}}$.

There also exist many advanced algorithms [47]–[52] aimed at detecting semantic object parts to capture more powerful features for fine-grained recognition. Most of them are especially designed for a small number of fine-grained concepts. We do

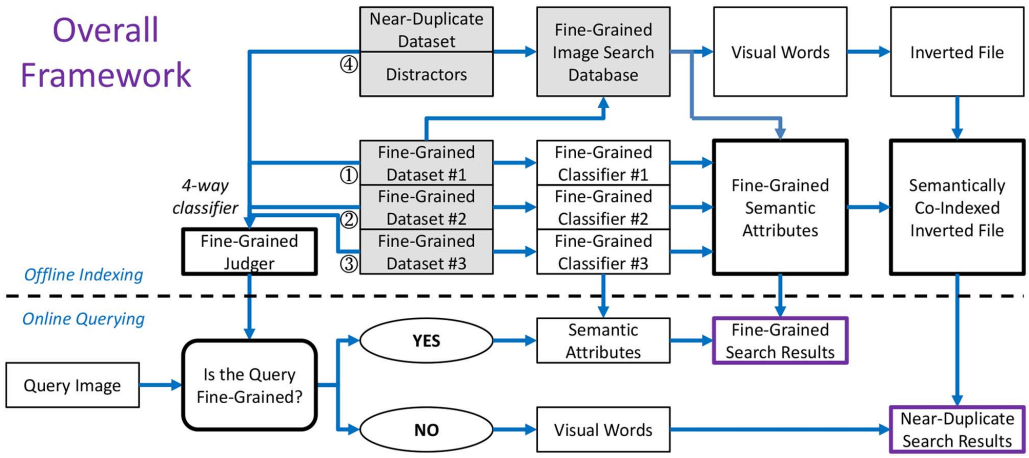


Fig. 4. Overall framework consisting of offline indexing and online querying stages.

not adopt them so that our framework is capable of dealing with many other fine-grained datasets.

Based on the image-level vectors, we train two types of classification models. The first type, the *fine-grained judger*, is a 4-way classifier trained on 4 categories, i.e., *birds*, *dogs*, *flowers* and *others* (not containing any fine-grained concepts). We can use it to judge whether the query image contains a fine-grained concept, and which one it is. The second type is composed of three *fine-grained classifiers*, each trained on a fine-grained dataset. We denote the judger as $\mathbb{M}^{\mathrm{J}}$ (superscripts J for *judgement*), and other three classifiers as $\mathbb{M}^{\mathrm{C}}_{\mathrm{B}}$, $\mathbb{M}^{\mathrm{C}}_{\mathrm{D}}$ and $\mathbb{M}^{\mathrm{C}}_{\mathrm{F}}$ (superscript C for *classification*, subscript B, D and F for *birds*, *dogs* and *flowers*), respectively. We can train more classification models if more fine-grained concepts are introduced. All the SVM classifiers are trained with LibLINEAR [77] using the tradeoff parameter $C = 10$.

B. Large-Scale Feature Indexing

Feature indexing is a basic technique for large-scale image search, which is often conducted with the Bag-of-Visual-Words (BoVW) model and the inverted index structure. The images are also resized so that the larger axis size is 600 pixels. We extract SIFT descriptors [2] from the detected Hessian Affine regions [6]. The set of descriptors is denoted as $\mathcal{D}_c = \{\mathbf{d}_{c,1}, \mathbf{d}_{c,2}, \ldots, \mathbf{d}_{c,M_c}\}$. We train a large codebook containing one million codewords with the Approximate K-Means clustering [8]. Only descriptors on training images are collected for clustering. The descriptors are then assigned onto the nearest codeword, i.e., with hard quantization strategy. The set of visual words are denoted as $\mathcal{W}_c = \{\mathbf{w}_{c,1}, \mathbf{w}_{c,2}, \ldots, \mathbf{w}_{c,M_c}\}$. An inverted index [1] is constructed for efficient lookup. We filter the stop-words in the inverted index, defined as those occurring in more than 1% images. The visual words are then weighted by the ℓ_p-norm IDF [78].

The candidate images are first ranked by counting the weighted feature occurrences. We then construct ImageWeb [30], a graph-based structure to model the image-level relationship between images, and refine the initial search results with affinity propagation on the graph. There also exist many complicated object retrieval systems producing very high accuracy [11], [15]. However these methods could be time consuming at the online querying stage, therefore are not adopted in practise.

C. Semantic-Aware Co-Indexing

Semantic-aware co-indexing is aimed at incorporating semantic property of images into the inverted index, so that it is possible to lookup high-level semantic attributes at the online querying stage. We follow [57] to calculate semantic attributes from fine-grained object recognition scores, and perform two successive operations, i.e., semantic-isolated image deletion and semantic-nearest image insertion, to modify and expand the indexed image lists in the inverted file. The co-indexed structure is illustrated in Fig. 5.

Semantic attributes are those properties that help to describe semantic concepts of an image [79]. We use the fine-grained classifiers $\mathbb{M}^{\mathrm{C}}_{\mathrm{B}}$, $\mathbb{M}^{\mathrm{C}}_{\mathrm{D}}$ and $\mathbb{M}^{\mathrm{C}}_{\mathrm{F}}$ trained in Section IV-A to calculate the classification scores for each image. The output of $\mathbb{M}^{\mathrm{C}}_{\mathrm{B}}$, for example, is a 200-dimensional vector $\mathbf{s}_{\mathrm{B}}$, in which the element $s_{\mathrm{B},k}$ stands for the confident score of an image containing the k-th *bird* species. We normalize the scores using the soft-max function: $\widetilde{s}_{\mathrm{B},k} = \frac{\exp\{s_{\mathrm{B},k}\}}{\sum_k \exp\{s_{\mathrm{B},k}\}}$, so that we have $\sum_k \widetilde{s}_{\mathrm{B},k} = 1$. The normalized vectors, $\widetilde{\mathbf{s}}_{\mathrm{B}}$, $\widetilde{\mathbf{s}}_{\mathrm{D}}$ and $\widetilde{\mathbf{s}}_{\mathrm{F}}$, are concatenated as the semantic attribute vector $\widetilde{\mathbf{s}}$ with $200 + 120 + 102 = 422$ dimensions. The distance between two images $\mathbf{I}_a$ and $\mathbf{I}_b$ is then measured by the Total Variance Distance (TVD) score [57] between their attribute vectors

$$\mathrm{TVD}\left(\mathbf{I}_a, \mathbf{I}_b\right) = \mathrm{TVD}\left(\widetilde{\mathbf{s}}_a, \widetilde{\mathbf{s}}_b\right) = \sum_k \left|\widetilde{s}_{a,k} - \widetilde{s}_{b,k}\right|. \quad (2)$$

After the TVD value is calculated for each pair of images, we modify the inverted index structure by adopting two steps successively, i.e., semantic-isolated image deletion and semantic-nearest image insertion [57]. The semantic-isolated image deletion step traverses each entry in the inverted index and check the followed images. For those entries with no less than 3 images, it removes those images which are isolated from others, i.e., the minimum TVD from this image and others is larger than a threshold ρ. This process can effectively reduce the index size without impacting the retrieval precision in the search process.

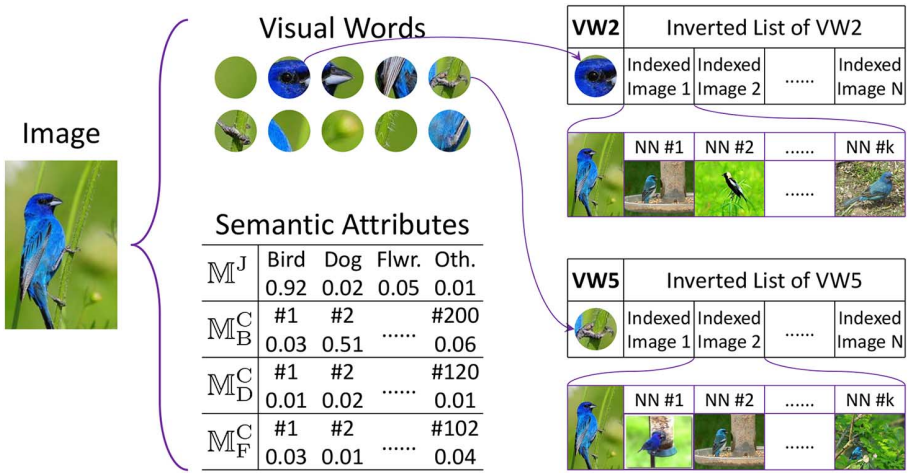


Fig. 5. Inverted file structure after semantic-aware co-indexing. For each image, we extract the visual words and calculate its semantic attributes. Visual words are then stored in the inverted index for efficient lookup. Additional nearest neighbor images in the attribute space are inserted into the inverted index for capturing the semantic similarity.

The semantic-nearest image insertion step, in opposite, adds images with large semantic similarity into the inverted index. It works by searching for a fixed number of nearest neighbors of each image according to the TVD scores, and add those nearest neighbors into the image's entry if they are not in the visual word's entry. In essential, this process could be considered as a query expansion operation based on the reliable candidate images.

It is obviously insufficient to represent an image's semantic attribute with only three fine-grained datasets (422 categories), but this serves as a baseline method to incorporate fine-grained information into the inverted index structure. As more fine-grained datasets, such as *cats* [71] or *aircrafts* [72], are introduced into the search task, we can extract longer attribute vectors to capture richer semantic properties.

D. Online Querying

When a query image comes, we first follow the previous sections to extract all the required features, i.e., an image-level representation vector, a set of visual words, and a semantic attribute vector. Then the flowchart of online querying is illustrated in the lower part of Fig. 4. We use the 4-way classifier $\mathbb{M}^{\mathrm{J}}$ to judge if a fine-grained concept, i.e., a *bird*, a *dog*, or a *flower*, is contained in the query image, or if the query does not contain any of these concepts. We then search the database according to the judgement.

If the prediction says that the query image contains a fine-grained concept, say, a *bird*, then we can calculate the 200-dimensional *bird* attribute vector with model $\mathbb{M}_{\mathrm{B}}^{\mathrm{C}}$, compare it to the images in the database, and sort them according to the TVD value $\mathrm{TVD}(\widetilde{\mathbf{s}}_{\mathbf{q},\mathrm{B}}, \cdot)$. Since we can index as many as one million images, calculating the TVD scores one-by-one can be very computationally expensive. We adopt the Locality-Sensitive Hashing (LSH) method [80] to index images into a small number of hashing tables and search only for those images sharing common bins with the query. In practise, LSH works much faster than bruteforce search and produces satisfying search results.

If the prediction suggests that the query image does not contain a fine-grained concept, we can adopt the semantic-aware online querying algorithm [57] to lookup the inverted index. Here, the search process is very similar to the general case of near-duplicate image search, but we make full use of the co-indexed structure to improve the quality of retrieved images. In a word, we use the samples found in the semantic-nearest image insertion step to update the matching score of each candidate. This process is equivalent to expanding the query in the attribute space using the top-ranked candidates, and will be surprisingly effective when the query image turns out to contain a fine-grained concept (see the lower part of Fig. 6).

It is worth noting that we have made an early decision at the beginning of the querying stage. Although it limits the model flexibility and makes the search result highly unstable when the judgement is incorrect (see Fig. 6), the online querying time is significantly reduced. We expect more advanced approaches to refine this process in the future research.

E. Scalability Issues

Currently, our algorithm only considers a handful of basic-level categories, i.e., *birds*, *dogs* and *flowers*. Here we discuss the scalability issues when we a larger number of visual concepts are introduced into the database.

When the number of basic-level categories increases, we first need to calculate a larger number of semantic attributes and store them in the inverted index. Let us assume that we have 100 basic-level categories and each of them has 200 fine-grained classes, which leads to a total number of 20000 categories, comparable with the scale of *ImageNet* [60]. In such cases, extracting all the $100 + 20000$ attributes might be computationally expensive. We can adopt an approximate strategy, which first calculate 100 basic-level semantic attributes, then pick up the top-5 categories with the largest attribute scores, and calculate the fine-grained semantic attributes only for these 5 categories (5×200 fine-grained classes). For the remaining basic-level categories, we can simply set the same score for all the 200 fine-grained classes.

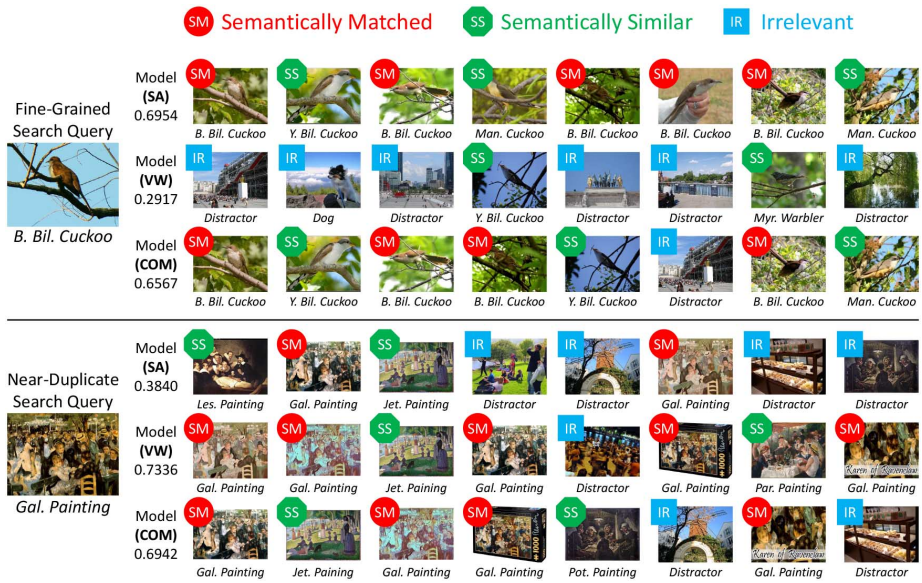


Fig. 6. Representative queries and search results. Above the dashed line are the queries correctly classified in the fine-grained concept judgement. Below, however, the two *birds*, one *logo*, and one *painting* images are wrongly recognized as: *not containing a fine-grained concept*, a *flower*, a *bird*, and a *flower*, respectively. The text below each image indicates its ground-truth label. One can see that our system produces satisfied results when the users' intention is correctly understood. Moreover, when a fine-grained query is misclassified as a near-duplicate query, we can also return fair quality search results thanks to the semantic-aware co-indexing; however, the search accuracy will become dramatically unstable when a near-duplicate query is misclassified.

The above method significantly reduces the time/memory costs in the offline indexing process. The only information loss happens when the actual basic-level category is not contained in the top-5 recognition results. According to the performance of our classification algorithm in the Caltech101 dataset [81], which contains 101 basic-level classes (not including the *background* category) and the number of training images is much smaller (30 per category), the top-5 classification accuracy over 101 basic-level categories is often higher than 90%. Therefore, the time/memory costs could be reduced to around 5% with less than 10% information loss.

At the online querying stage, the difference comes from seeking the approximate nearest neighbor in the semantic attribute space. Since the expanded space might have tens of thousands of dimensions, even LSH requires heavy computation. To cope with, we could adopt another approximate strategy, which considers the semantic attribute space of each basic-level category individually. When the basic-level classification scores of the query image are calculated, we find the top-5 categories and retrieve the top-1000 nearest neighbors from each of these 5 fine-grained attribute spaces. Finally, the retrieved images are combined into a single list according to their total distance to query in these 5 spaces. In this way we prevent our algorithm from calculating a large number of fine-grained concepts and becoming computationally expensive.

We will provide a more detailed scalability analysis in Section V-D. Meanwhile, we admit that many other approaches might be adopted to deal with this problem, such as adopting a hierarchical structure and a top-down approach to categorize an image from-coarse-to-fine. We aim at suggesting a possible solution to this new and challenging problem, and inspiring more research efforts in the future.

F. Generalization Issues

Of course, three levels of relevance are not enough for describing the similarity between two images. A more generalized approach is to use the natural structure of ontologies [61] to calculate a multi-level similarity score. For example, [54] uses hierarchical semantic indexing to improve the image retrieval performance, and [82] learns a fine-grained similarity with deep learning techniques. Both the above works inspire us to formulate fine-grained image search on a larger number of relevance levels.

Our two-stage framework could be easily modified to fit the new problem setting. We inherit the module of basic-level concept discovery, but add the hierarchical loss function into fine-grained image classification process. By hierarchical loss we mean that the classification score of an image is no longer either 1 (correct) or 0 (incorrect), but could be a fractional value between 0 and 1 indicating the relevance of the predicted label and the ground-truth label. Such technique, which is widely adopted in large-scale image classification, penalizes heavier on those basic-level errors (e.g., a *bird* is classified as a *building*) than those fine-grained errors (e.g., a *red-winged blackbird* is classified as a *brewer blackbird*).

V. Experimental Results

A. Recognition and Retrieval Accuracy

First, we report the accuracy on fine-grained concept judgement and object recognition. The results are listed in Table II and III, respectively. One can see the satisfied performance when judging the fine-grained concept in an image: 96.20% *birds*, 97.98% *dogs*, 96.62% *flowers* and 98.59% *others* (not containing a fine-grained concept) queries are correctly classified. The fine-grained recognition accuracy is also nearly

TABLE II
CONFUSION MATRIX OF FINE-GRAINED CONCEPT JUDGEMENT. TO BE CLEAR, 102 *BIRD* IMAGES ARE RECOGNIZED AS *DOGS*

	Bird	*Dog*	*Flower*	*Other*	Accuracy
Bird	5574	102	12	106	96.20%
Dog	45	8407	12	116	97.98%
Flower	83	102	5941	23	96.62%
Other	1	0	5	416	98.59%

TABLE III
FINE-GRAINED RECOGNITION ACCURACY

	Bird	*Dog*	*Flower*
Our accuracy	42.56%	35.29%	78.72%
Best known, w/o parts	$\approx 44\%$	$\approx 37\%$	$\approx 80\%$
Best known, with parts	$\approx 56\%$	$\approx 48\%$	$\approx 84\%$

TABLE IV
NEAR-DUPLICATE SEARCH ACCURACY (WITH ONE MILLION DISTRACTORS) OF BUILDING, *LOGO* AND *PAINTING* QUERIES

	Building	*Logo*	*Painting*
Our mAP score	0.7631	0.4842	0.5857
Best known, $t \leqslant 100$ms	≈ 0.78	≈ 0.49	≈ 0.60
Best known	≈ 0.87	≈ 0.53	≈ 0.66

the state-of-the-art without using human annotation and complicated part detection models [48]–[52]. It guarantees that in most cases, the users' intention could be correctly understood by the search engine.

Next, we report the object retrieval accuracy on near-duplicate datasets with one million distractors in Table IV. Once again, our search accuracy is nearly the state-of-the-art, given that we need to finish the search within 100 ms thus some complicated post-processing approaches [11], [15] are not used.

The good performance of separate modules helps us produce satisfying fine-grained search results. Please note that we are not aimed at comparing our classification or retrieval modules with the state-of-the-art algorithms. Most often, those complicated methods use some specified clues or tricks to boost the accuracy on one or two datasets. These algorithm are either too specific thus not generalizable onto a wide range of visual concepts, or very computationally expensive so that are not applicable onto large-scale image search.

B. Search Results

We report the fine-grained search accuracy using the evaluation method defined in Section III-C. For comparison, we test the proposed model combining both semantic attributes and visual words (*Model-COM*) illustrated in Section IV-D, as well as individual modules without performing fine-grained judgement on query images. *Model-SA* uses only semantic attributes from the 422 fine-grained categories $\widetilde{\mathbf{s}}$, and rank the candidates according to their Total Variance Distance (TVD) to the query. *Model-VW* uses only the quantized visual word set $\mathcal{W}_c$ to lookup the inverted index, and rank the candidates according to the feature occurrence weighted by the ℓ_p-norm IDF [78]. They are equivalent to regarding all query images as containing one of the fine-grained concepts or not so.

The results are summarized in Table V. One can observe that on each separate query set, either fine-grained or near-duplicate, *Model-COM* does not produce the best performance among all three models. Fig. 7 provides an intuitive comparison on both types of queries. When the query image contains a fine-grained concept, *Model-SA* better capture the image's global property; however when the query is a near-duplicate instance, *Model-VW* produces higher accuracy. *Model-COM* becomes the best choice only when the scores are averaged, i.e., there might be both types of queries, as in the real-world applications. This suggests that analysing the search intention is probably more important than achieving higher accuracy in separate tasks.

TABLE V
SEARCH PERFORMANCE OF THREE DIFFERENT MODELS

	Model-SA	**Model-VW**	**Model-COM**
Bird-200	**0.6741**	0.3215	0.6169
Dog-120	**0.7102**	0.3692	0.6727
Flower-102	**0.7961**	0.4250	0.7302
Building	0.5601	**0.9205**	0.8887
Logo	0.3406	**0.5940**	0.5432
Painting	0.4091	**0.6703**	0.6610
Fine-Grained	**0.7268**	0.3719	0.6733
Near-Duplicate	0.4366	**0.7283**	0.6976
Overall	0.5817	0.5501	**0.6855**

Once again, we shall emphasize that this paper is aimed at proposing a new problem and suggesting a new framework. Although we do not use those complicated classification/retrieval algorithm in practise, one might notice that the proposed framework is highly modular, implying that it is easy to replace each module with a more efficient algorithm designed in the future. We believe that baseline accuracy of *Model-SA* and *Model-VW* would be thereafter improved. Since the combined model (*Model-COM*) absorbs the advantages of both *Model-SA* and *Model-VW*, we can forecast that better search performance will also be produced by *Model-COM* in such cases.

C. Time and Memory Costs

We report the time and memory costs at the online querying stage with about one million images.

We need to store two parts of information (see Fig. 5). The first part, image ID for each occurrence of visual words, requires 4 bytes for each image-word pair. The average number of descriptors on each image is about 1000, thus we need 4 Kilobytes for a single image. According to [57], the memory consumption increases by about one half after semantic-aware co-indexing. The second part is the semantic attributes calculated from the fine-grained classification scores. For each image, we have to store $4 + 200 + 120 + 102 = 426$ floating numbers in total, requiring about 2 Kilobytes. Therefore, the total memory overhead is less than $4 \times 150\% + 2 = 8$ Kilobytes for one image, and about 8 Gigabytes for the whole database with one million images. ImageWeb [30] requires only about 160 Megabytes for one million images.

At the online querying stage, the time cost consists of three parts, i.e., descriptor extraction and quantization, concept judgement, and image search, which follows either fine-grained (using semantic attributes) or near-duplicate (using co-indexed file) flowchart, depending on the concept judgement result. The SIFT descriptor extraction takes 1000 ms (650 ms for dense sampling, 350 ms for interest point detection and description), quantization takes 400 ms (300 ms for Fisher vectors, and 100 ms for hard quantization). The concept judgement stage takes about 50 ms. If the query is considered as containing a

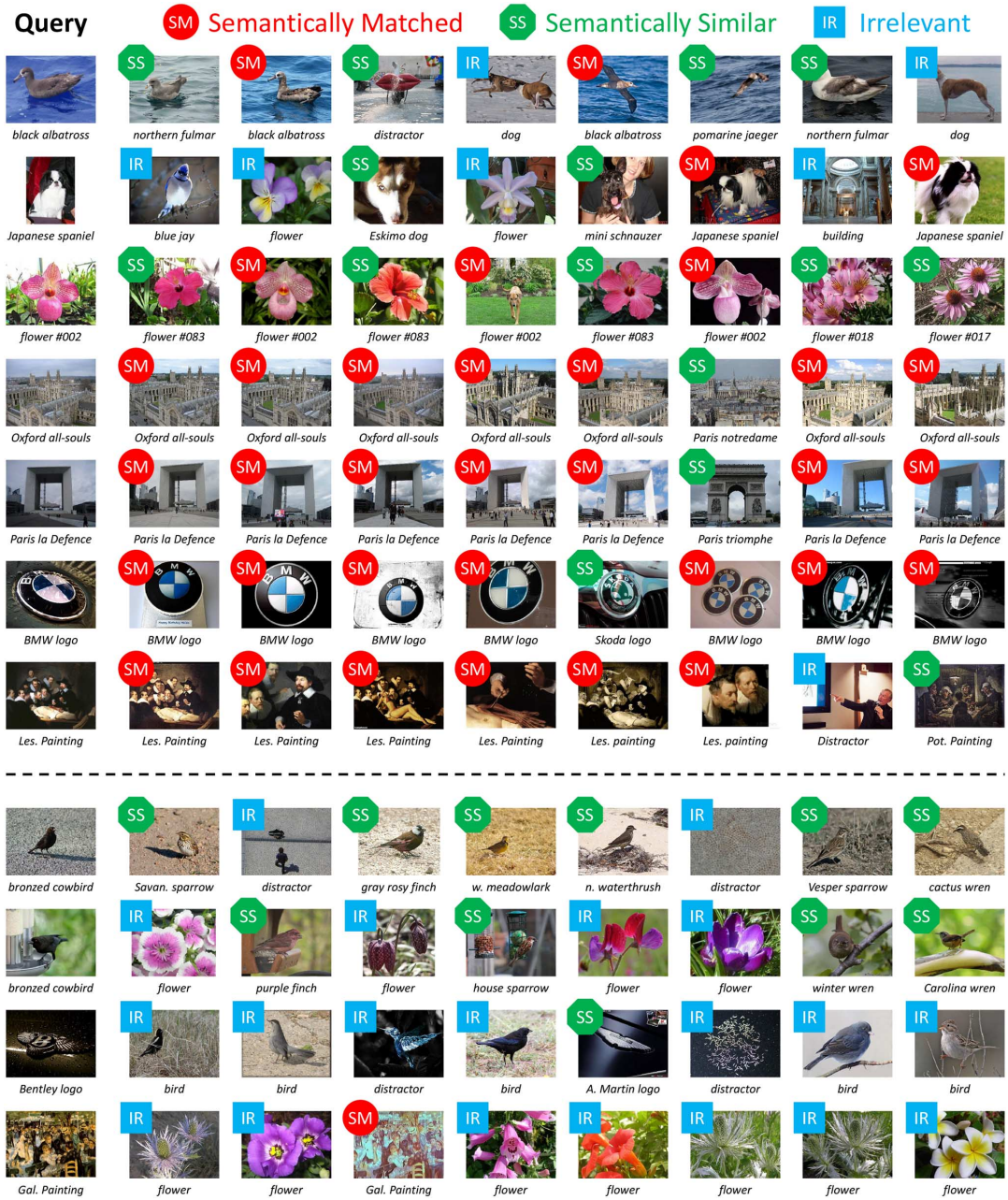


Fig. 7. Search results by three different models on two query images: the first one contains a fine-grained concept (*black billed cuckoo, a bird* species), while the other is a near-duplicate instance (*painting*). Numbers below model names indicate the nDCG scores of the corresponding search results.

fine-grained concept, then the fine-grained recognition takes about 100 ms, and LSH takes about 400 ms; otherwise the near-duplicate searching in the semantically co-indexed file takes about 550 ms. It takes less than 100 ms for ImageWeb to post-process the search results. Overall, the total time required for a single query would not exceed 2 seconds. All the time is recorded on a single 3.0 GHz CPU.

D. Scalability Analysis

Let us still assume that there are 100 basic-level categories and each category has 200 fine-grained concepts. For each image, we first need to store the IDs and scores of the top-5 basic-level recognition results. Next, 5×200 fine-grained semantic attributes are calculated. In total, we need $5 \times 2 + 200 \times 5 = 1010$ floating numbers for each image, requiring nearly 4 Kilobytes. Considering the same amount of local descriptors, the memory overhead for one image is now $4 \times 150\% + 4 = 10$ Kilobytes. Real-world search engines such as Google and Baidu often deal with more than one billion images, therefore the required storage for the whole database is about 10 Terabytes.

The time cost of online querying goes up as the increasing number of basic-level categories and indexed images. If the query image is considered to have one of those fine-grained concepts, the LSH process is performed in 5 individual attribute spaces with 200 dimensions. It takes around 800 ms compared to 400 ms which is the time for searching in a single 400-dimensional space. The time used for combining the 5 retrieved lists

into one is around 100 ms. If the query is judged as a near-duplicate case, the search time complexity will grow sublinearly, as observed in ImageWeb [30] and semantic-aware co-indexing [78]. In either case, the time used to process a single query would be no more than 2.5 seconds.

To summarize, when the number of basic-level categories increases from 3 to 100, fine-grained classes from around 400 to 20000, and indexed images from 1 million to 1 billion, the estimated time and memory complexity (for storing and querying one image) only grows by 25.0%. Therefore we could expect that our algorithm scales up efficiently with reasonable approximation.

E. Sample Cases

We present some representative queries and corresponding search results in Fig. 6. One can see that, our approach works well for those correctly judged queries, suggesting that understanding the users' intention is crucial to improve the search quality. Meanwhile, we also present four wrongly judged cases, two fine-grained and two near-duplicate. For the first two cases, we can still obtain pretty good search results thanks to the semantic-aware co-indexing, however the latter cases show dramatically unstable results due to the ignorance of local features which are key to near-duplicate retrieval. This once again suggests to improve the concept judgement accuracy in the future research.

VI. Conclusions

This paper proposes *fine-grained image search*, a challenging topic which is less studied in the multimedia community. We argue that it is a common requirement that instance search returns not only near-duplicate but also fine-grained results. We formulate the fine-grained image search problem by constructing a new database and defining an evaluation method. A baseline framework is also proposed to incorporate semantic attributes into the inverted index, leading to an efficient search engine which produces promising search results in large-scale experiments. Since our framework is highly modular, it might cooperate well with many other classification/retrieval algorithms. The scalability of our algorithm also makes it easy to be transplanted onto commercial search engines.

In the future, we shall release the database used in the experiments, and gradually add more images, such as those fine-grained sets in *ImageNet* [60], into the database. The methods discussed in Section IV-E and IV-F will be implemented for scaling up and generalization. We hope this topic will reveal an exciting research direction which brings new insights and novel applications. We also expect more research efforts to be triggered on this interesting yet less explored problem.

References

[1] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, vol. 2, pp. 1470–1477.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[3] L. Xie, Q. Tian, and B. Zhang, "Max-SIFT: Flipping invariant descriptors for web logo search," in *Proc. Int. Conf. Image Process.*, 2014, pp. 5716–5720.

[4] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 404–417.

[5] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image Vis. Comput.*, vol. 22, no. 10, pp. 761–767, 2004.

[6] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *Int. J. Comput. Vis.*, vol. 60, no. 1, pp. 63–86, 2004.

[7] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 2, pp. 2161–2168.

[8] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–8.

[9] W. Zhou, Y. Lu, H. Li, and Q. Tian, "Scalar quantization for large scale image search," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 169–178.

[10] W. Zhou *et al.*, "Towards codebook-free: Scalable sascaded hashing for mobile image search," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 601–611, Apr. 2014.

[11] M. Perd'och, O. Chum, and J. Matas, "Efficient representation of local geometry for large scale object retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 9–16.

[12] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 304–317.

[13] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[14] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[15] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2911–2918.

[16] Y. Jiang, J. Wang, X. Xue, and S. Chang, "Query-adaptive image search with hash codes," *IEEE Trans. Multimedia*, vol. 15, no. 2, pp. 442–453, Feb. 2013.

[17] Y. Yang and A. Hanjalic, "Prototype-based image search reranking," *IEEE Trans. Multimedia*, vol. 14, no. 3, pt. 2, pp. 871–882 , Jun. 2012.

[18] H. Xie *et al.*, "Efficient feature detection and effective post-verification for large scale near-duplicate image search," *IEEE Trans. Multimedia*, vol. 13, no. 6, pp. 1319–1332, Dec. 2011.

[19] J. Huang, X. Yang, X. Fang, W. Lin, and R. Zhang, "Integrating visual saliency and consistency for re-ranking image search results," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 653–661, Aug. 2011.

[20] W. Lu, J. Wang, X. Hua, S. Wang, and S. Li, "Contextual image search," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 513–522.

[21] X. Wang *et al.*, "Contextual weighting for vocabulary tree based image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 209–216.

[22] Z. Liu, H. Li, W. Zhou, R. Zhao, and Q. Tian, "Contextual hashing for large-scale image search," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1606–1614, Apr. 2014.

[23] N. Morioka and J. Wang, "Robust visual reranking via sparsity and ranking constraints," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 533–542.

[24] Y. Jing, M. Covell, J. Rehg, and D. Tsai, "Learning query-specific distance functions for large-scale web image search," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 2022–2034, Dec. 2013.

[25] Y. Zhang, Z. Jia, and T. Chen, "Image retrieval with geometry-preserving visual phrases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 809–816.

[26] Y. Jiang, J. Meng, and J. Yuan, "Randomized visual phrases for object search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3100–3107.

[27] M. Wang, K. Yang, X. Hua, and H. Zhang, "Towards a relevant and diverse search of social images," *IEEE Trans. Multimedia*, vol. 12, no. 8, pp. 829–842 , Dec. 2010.

[28] Y. Lu, L. Zhang, J. Liu, and Q. Tian, "Constructing concept lexica with small semantic gaps," *IEEE Trans. Multimedia*, vol. 12, no. 4, pp. 288–299 , Jun. 2010.

[29] J. Wang and S. Li, "Query-driven iterated neighborhood graph search for scalable visual indexing," in *Proc. ACM Int. Conf. Multimedia*, 2012, pp. 179–188.

[30] L. Xie, Q. Tian, W. Zhou, and B. Zhang, "Fast and accurate near-duplicate image search with affinity propagation on the ImageWeb," in *Proc. Comput. Vis. Image Understand.*, 2014, vol. 124, pp. 31–41.

[31] Q. Tian, J. Yu, Q. Xue, and N. Sebe, "A new analysis of the value of unlabeled data in semi-supervised learning for image retrieval," in *Proc. 2004 IEEE Int. Conf. Multimedia Expo*, Jun. 2004, vol. 2, pp. 1019–1022.

[32] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop Statist. Learn. Comput. Vis.*, 2004, vol. 1, pp. 1–2.

[33] L. Xie, Q. Tian, J. Wang, and B. Zhang, "Image classification with max-sift descriptors," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2015, to be published.

[34] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 1, pp. 886–893.

[35] J. Wang *et al.*, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 3360–3367 .

[36] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 143–156.

[37] H. Lee, A. Battle, R. Raina, and A. Ng, "Efficient sparse coding algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 801–808.

[38] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 2, pp. 2169–2178.

[39] L. Xie, Q. Tian, M. Wang, and B. Zhang, "Spatial pooling of heterogeneous features for image classification," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 1994–2008, May 2014.

[40] L. Xie, Q. Tian, and B. Zhang, "Generalized regular spatial pooling for image classification," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2015, to be published.

[41] L. Xie, Q. Tian, and B. Zhang, "Feature normalization for part-based image classification," in *Proc. Int. Conf. Image Process.*, Sep. 2013, pp. 2607–2611.

[42] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.

[43] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei, "Novel dataset for fine-grained image categorization," in *Proc. 1st Workshop Fine-Grained Vis. Categorization*, 2011.

[44] J. Deng, J. Krause, and L. Fei-Fei, "Fine-grained crowdsourcing for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 580–587.

[45] K. Duan, D. Parikh, D. Crandall, and K. Grauman, "Discovering localized attributes for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3474–3481.

[46] N. Zhang, R. Farrell, and T. Darrell, "Pose pooling kernels for sub-category recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3665–3672 .

[47] S. Yang, L. Bo, J. Wang, and L. Shapiro, "Unsupervised template learning for fine-grained object recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 3122–3130.

[48] T. Berg and P. Belhumeur, "POOF: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 955–962.

[49] L. Xie, Q. Tian, and B. Zhang, "Hierarchical part matching for fine-grained visual categorization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1641–1648 .

[50] L. Xie, J. Wang, B. Guo, B. Zhang, and Q. Tian, "Orientational pyramid matching for recognizing indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 3734 –3741.

[51] Y. Chai, V. Lempitsky, and A. Zisserman, "Symbiotic segmentation and part localization for fine-grained categorization," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2013, pp. 321–328.

[52] E. Gavves, B. Fernando, C. Snoek, A. Smeulders, and T. Tuytelaars, "Fine-grained categorization by alignments," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1713–1720.

[53] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[54] J. Deng, A. Berg, and L. Fei-Fei, "Hierarchical semantic indexing for large scale image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 785–792.

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

[56] F. Yu, R. Ji, M. Tsai, G. Ye, and S. Chang, "Weak attributes for large-scale image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2949–2956.

[57] S. Zhang, M. Yang, X. Wang, Y. Lin, and Q. Tian, "Semantic-aware co-indexing for near-duplicate image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1673–1680.

[58] M. Douze, A. Ramisa, and C. Schmid, "Combining attributes and fisher vectors for efficient image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 745–752.

[59] S. Zhang, M. Yang, T. Cour, K. Yu, and D. N. Metaxas, "Query specific fusion for image retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 660–673.

[60] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 248–255.

[61] C. Fellbaum, *WordNet*. New York, NY, USA: Wiley, 1998.

[62] G. Miller and C. Fellbaum, *WordNet: An Electronic Lexical Database*. Cambridge, MA, USA: MIT Press, 1998.

[63] C. Fellbaum, "WordNet and WordNets," in *Encyclopedia of Language and Linguistics*, 2nd ed. New York, NY, USA: Oxford, 2005.

[64] S. Chang, J. Smith, M. Beigi, and A. Benitez, "Visual information retrieval from large distributed online repositories," *Commun. ACM*, vol. 40, no. 12, pp. 63–71, 1997.

[65] E. Hyvonen, S. Saarela, A. Styrman, and K. Viljanen, "Ontology-based image retrieval," Helsinki Inst. Inf. Technol., Helsinki, Finland, Rep. 2002-03, 2003.

[66] V. Mezaris, I. Kompatsiaris, and M. Strintzis, "An ontology approach to object-based image retrieval," in *Proc. Int. Conf. Image Process.*, Sep. 2003, vol. 2, pp. II-511–II-514.

[67] P. Stanchev, D. Green, Jr., and B. Dimitrov, "High level color similarity retrieval," *Int. J. Inst. Inf. Theories Appl.*, vol. 10, pp. 283–287, 2003.

[68] Y. Liu, D. Zhang, G. Lu, and W. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recog.*, vol. 40, no. 1, pp. 262–282, 2007.

[69] S. Romberg, L. G. Pueyo, R. Lienhart, and R. van Zwol, "Scalable logo recognition in real-world images," in *Proc. 1st ACM Int. Conf. Multimedia Retrieval*, 2011, pp. 25–32.

[70] M. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2008, pp. 722–729.

[71] O. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3498–3505.

[72] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *CoRR* vol. abs/1306.5151, 2013.

[73] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–Mining discriminative components with random forests," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 446–461.

[74] K. Jarvelin and J. Kekalainen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, 2002.

[75] Y. Wang *et al.*, "A theoretical analysis of NDCG ranking measures," in *Proc. 26th Annu. Conf. Learn. Theory*, 2013, pp. 1–26.

[76] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1469–1472.

[77] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008.

[78] L. Zheng, S. Wang, and Q. Tian, "Lp-norm IDF for scalable image retrieval," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3604–3617, Aug. 2014.

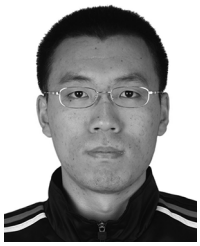
[79] D. Parikh and K. Grauman, "Relative attributes," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 503–510.

[80] M. Datar, N. Immorlica, P. Indyk, and V. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proc. 20th Annu. Symp. Comput. Geom.*, 2004, pp. 253–262.

[81] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," *Comput. Vis. Image Understand.*, vol. 106, no. 1, pp. 59–70, 2007.

[82] J. Wang *et al.*, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 1386–1393.

[83] J. Wang *et al.*, "Trinary-projection trees for approximate nearest neighbor search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 388–403, Feb. 2014.



Lingxi Xie received the B.E. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2010. He is currently working toward the Ph.D. degree at the same university.

He was a Research Intern with Microsoft Research Asia, Beijing, China, from August 2013 to January 2014, and from September 2014 to present. He was a Visiting Researcher with the Department of Computer Science, University of Texas at San Antonio, San Antonio, TX, USA, from February 2014 to July 2014. His research interests include computer vision, multimedia information retrieval, and machine learning.

Jingdong Wang received the B.Sc. and M.Sc. degrees from the Department of Automation, Tsinghua University, Beijing, China, in 2001 and 2004, respectively, and the Ph.D. degree from the Department of Computer Science, Hong Kong University of Science and Technology, Hong Kong, in 2007.

He is currently a Lead Researcher with the Visual Computing Group, Microsoft Research Asia, Beijing, China. His areas of interest include computer vision, machine learning, and multimedia search. He is currently working on the Big Media project, including large-scale indexing and clustering, and Web image search and mining.

Dr. Wang is an Editorial Board Member of *Multimedia Tools and Applications*.

Bo Zhang received the B.E. degree from the Department of Automatic Control, Tsinghua University, Beijing, China, in 1958.

From 1980 to 1982, he was a Visiting Researcher with the University of Illinois at Urbana-Champaign, Champaign, IL, USA. He is currently a Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His research interests include artificial intelligence, machine learning, pattern recognition, knowledge engineering, intelligent robotics, and intelligent control.

Prof. Zhang was the recipient of the Prize of European Artificial Intelligence in 1984. He was the recipient of the First Class of Science and Technology Progress Prize three times (in 1987, 1993, and 1998) and the Third Class Prize twice (in 1995 and 1999). He is a Member of the Chinese Academy of Sciences.

Qi Tian (S'95–M'96–SM'03) received the B.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 1992, the M.S. degree in electrical and computer engineering from Drexel University, Philadelphia, PA, USA, in 1996, and the Ph.D. degree in electrical and computer engineering from the University of Illinois, Urbana-Champaign, Champaign, IL, USA, in 2002.

He is currently a Professor with the Department of Computer Science, University of Texas at San Antonio, San Antonio, TX, USA. He took a one-year faculty leave with Microsoft Research Asia, Beijing, China, from 2008 to 2009. His research interests include multimedia information retrieval and computer vision.