Learning Hierarchical Semantic Description Via Mixed-Norm Regularization for Image Understanding

Liang Li, *Student Member, IEEE*, Shuqiang Jiang, *Senior Member, IEEE*, and Qingming Huang, *Senior Member, IEEE*

***Abstract*—This paper proposes a new perspective—Vicept representation to solve the problem of visual polysemia and concept polymorphism in the large-scale semantic image understanding. Vicept characterizes the membership probability distribution between visual appearances and semantic concepts, and forms a hierarchical representation of image semantic from local to global. In the implementation, incorporating group sparse coding, visual appearance is encoded as a weighted sum of dictionary elements, which could obtain more accurate image representation with sparsity at the image level. To obtain discriminative Vicept descriptions with structural sparsity, mixed-norm regularization is adopted in the optimization problem for learning the concept membership distribution of visual appearance. Furthermore, we introduce a novel image distance measurement based on the hierarchical Vicept description, where different levels of Vicept distance are fused together by multi-level separability analysis. Finally, the wide applications of Vicept description are validated in our experiments, including large-scale semantic image search, image annotation, and semantic image re-ranking.**

***Index Terms*—Image representation, large-scale systems, pattern analysis, semantic web, statistical learning.**

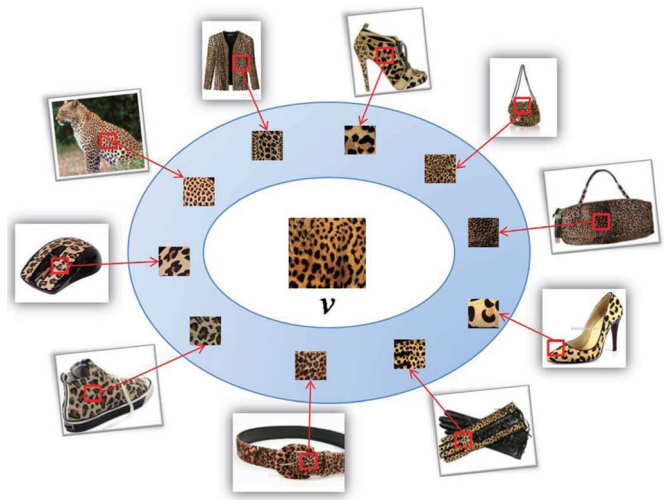


Fig. 1. Visual polysemia.

I. Introduction

LARGE scale semantic image analysis has recently become a hot research topic because of its wide applications in the fields of image understanding. To address this problem, researchers have proposed various approaches from different perspectives, including image classification [1]–[4], image annotation [5]–[7], object and scene recognition [8], [9], [11], [12], image search [13], [14], [16], [17], [27], etc. However, the problem of visual polysemia and concept polymorphism (VPCP) remains a great challenge for the task of large-scale semantic image understanding. Visual polysemia represents a fact that one certain visual appearance may have different semantic explanations as illustrated by Fig. 1. We observe that the visual appearance v is shared by the elements in concept collection $C = \{leopard, clothes, shoes, bags, glove, belt, mouse\}$. Without extra context information, it is difficult to assign v one certain concept. Concept polymorphism reveals another fact that one concept may have various visual appearances under different instances. Fig. 2 gives a concrete example of concept polymorphism for the concept *"skyscraper"*. The influence of VPCP may be slight in small image datasets; however, it becomes a significant problem under large-scale scenarios. On the side of VP, one visual appearance may occur in thousands of concepts so that it is extremely difficult to infer its exact concept. On the side of CP, one concept usually has thousands of various instances and most of them have diverse visual appearances. To sum up, the connections between visual appearances and semantic concepts are multi-aspect and complex in large-scale image environment.

Though many significant works have been proposed to address the large-scale image understanding, none of them solve the problem of VPCP directly. One main reason is that the relationship between image visual description and semantic information has not been specified. The de-facto image representation standard in these researches is based on the *bag-of-visual-words* (BOV) model [18]. The BOV approach regards an image as a collection of visual appearance descriptors extracted from local patches and quantized into discrete "visual words", and then computes a compact histogram representation for further image application. BOV has been extensively investigated for the following reasons: 1) visual words are discriminative due to the consideration of local salient and invariant information;

Manuscript received August 31, 2011; revised December 13, 2011, March 14, 2012; accepted March 22, 2012. Date of publication April 17, 2012; date of current version September 12, 2012. This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400, in part by National Natural Science Foundation of China: 61025011, 61035001, 61070108, and 60833006, and in part by Beijing Natural Science Foundation: 4111003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chong-Wah Ngo.

L. Li and S. Jiang are with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China (e-mail: lli@jdl.ac.cn; sqjiang@jdl.ac.cn).

Q. Huang is with the Graduate University of Chinese Academy of Sciences, and the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, (e-mail: qmhuang@jdl.ac.cn).

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.2012.2194993

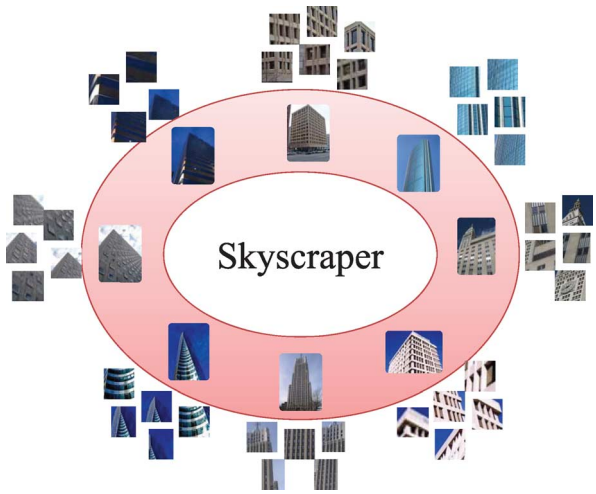


Fig. 2. Concept polymorphism.

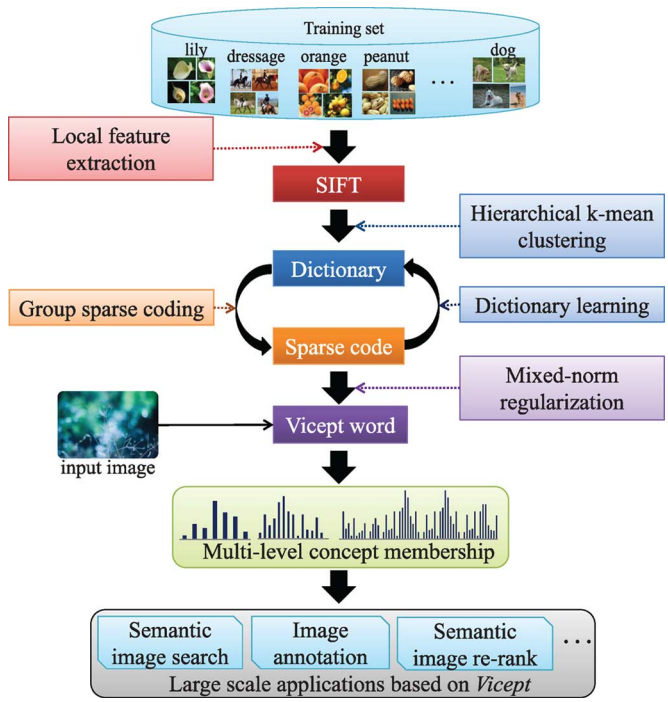


Fig. 3. Proposed framework for Vicept generation and applications.

2) similar to the word-document representation used in text retrieval, BOV provides a succinct and compact representation with a bag of visual words for images; 3) the similarity between images can be measured quickly through simple operations, such as dot-product.

Researchers have studied and improved the BOV approach for large-scale image application in the following four aspects: 1) aggregating local descriptors into more discriminative descriptions [2], [16], [19]; 2) generating compact codebook [3], [4], [8], [12], [20], [21], [22]; 3) efficient quantization techniques [3], [9], [11], [14], [15], [23]; 4) post-processing techniques on spatial information and image re-ranking [17], [24]. These above techniques can improve the performance to some extent, but the problem of VPCP has still been pendent.

Recently, researchers have paid more attention to machine learning for the applications of face recognition [6], image annotation [13], [25], image classification [1], [26], [28], scene and object recognition [29], [30], and so on. These works provide new insights on dealing with the problem of VPCP; however, they are still in the preliminary research stage and limited for wide applications on large-scale dataset.

In this paper, incorporating with the BOV model, we learn a Vicept (*vi*sual appearance-to-semantic con*cept*) image representation for large-scale semantic image understanding. Vicept is to characterize the membership distribution between each visual word and semantic concepts, and construct a hierarchical representation of image semantic. In details, Vicept is to present the probability relationship between visual appearances and hierarchical semantic concepts. Each Vicept word is a hierarchical concept membership distribution histogram about one visual appearance, which is illustrated in Figs. 4 and 5. Vicept can directly deal with the VPCP problem. For the visual polysemia problem, each Vicept word is an estimation of multiple concept possibilities of one visual appearance. It describes the concepts membership correlation with this visual appearance. For the concept polymorphism problem, in the Vicept dictionary, each visual appearance has a probability with one certain concept. In other words, one certain concept has the probability relationship with all the visual appearances.

In fact, the relationship between visual appearances and concepts is structural sparse in the practical situations: one visual appearance only has the correlation with limited concepts and one semantic concept only has the limited visual appearances. Taking this structural sparsity property of Vicept into account, we adopt the idea of mixed-norm regularization in our optimization problem for learning the Vicept, which is effective for obtaining a discriminative Vicept with structural sparsity. During visual appearance representation, considering the BOV structure, we encode visual descriptors as a weighted sum of dictionary elements by group sparse coding, which could obtain more accurate image representation with sparsity at the image level. Further, we formulate the image-level Vicept representation procedure from local to global. Besides we introduce the image distance measurement for the hierarchical Vicept description, where the different levels of Vicept distance are fused together by separability analysis.

The procedure of Vicept generation is illustrated in Fig. 3. Firstly, a large image training dataset with concept labels is established based on a hierarchical concept structure, which covers frequently used concepts in the daily life. Secondly, SIFT descriptors are extracted from these images and an initial dictionary is obtained by clustering these descriptors with hierarchical k-means. Thirdly, descriptors are encoded with group sparse coding based on the visual words, while a more discriminative and compact codebook can be learned from sparse representations of these descriptors. This is a procedure for visual appearance representation, and it is the crucial step before learning the Vicept. Fourthly, Vicept is obtained by learning from the above sparse representations with the mixed-norm regularization optimization method, and the Vicept description with hierarchical concept membership

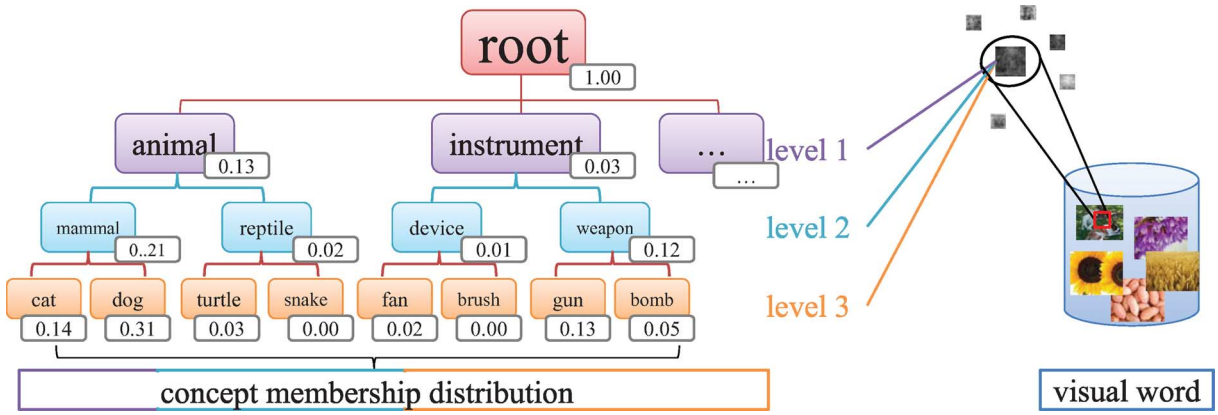


Fig. 4. One visualization of Vicept description.

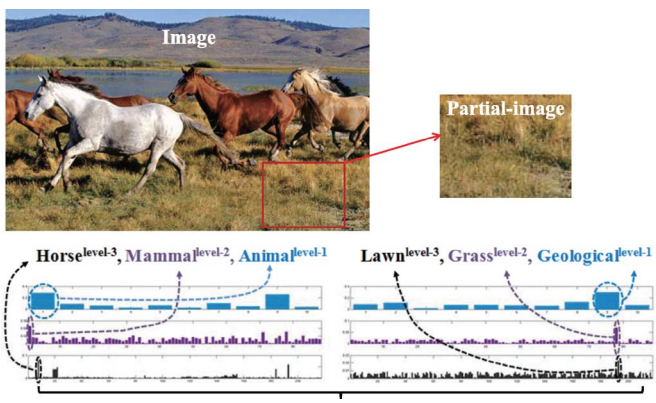


Fig. 5. Image representation based on Vicept description.

distribution is built based on the hierarchical concept structure. Finally, for one image, we can compute its global semantic description via vector product between Vicept and its group sparse code representation for further semantic applications.

The proposed scheme can quickly compute the semantic information of a given image without depending on specified training model so that it is easy to use under the large-scale applications. Experiments on large-scale semantic image search tasks show strong semantic descriptive power of Vicept. Furthermore, in the experiments of image annotation and semantic image re-ranking, our method also shows promising performances.

The contributions of our work are summarized as follows:

1) The problem of VPCP is discussed. Taking this paradox into account, a novel method for generating image hierarchical Vicept description is proposed for large-scale semantic image understanding.
2) Group sparse coding is utilized to encode the images for a more sparse representation. Meanwhile, a discriminative and compact dictionary is learned relying on this sparse representation.
3) The idea of mixed-norm regularization is adopted in our optimization problem to learn a discriminative Vicept with structural sparsity.
4) Aiming at the hierarchical structure of Vicept description, a distance metric is introduced to fuse the distance at different concept levels by separability analysis.

The research in this paper is an extension of our previous work "Learning image Vicept description via mixed-norm regularization for large scale semantic image search"(CVPR2011), where our main contribution focused on the local Vicept generation. However, in this paper, we focus on the whole Vicept framework in a global view. A summary of the main difference is as follows: 1) For the problem of visual polysemia and concept polymorphism, a clear motivation and intuition behind our proposed approach is described in Section I, and state-of-the-art related works is also added in Section II. 2) A deep analysis on the Vicept mechanism is done in Sections IV and V. Also the benefit of using mixed-norm regularization into the Vicept is presented in Section IV-C. 3) A new subsection (Section V-A) is added to formulate the image (or partial image) Vicept representation procedure, which provides a Vicept computing method from local to global. 4) A novel distance measurement for hierarchical Vicept representation is proposed based on multi-level separability analysis (Section V-B). 5) For each task in the experimental section, more state-of-the-art approaches are implemented as the baseline to validate the performance of our approach. Besides, more experimental analysis is presented in depth.

The rest of this paper is organized as follows: Section II gives an overview of previous related work. Section III introduces the visual appearance representation methods. Section IV presents the hierarchical Vicept description and details the learning procedure of Vicept under the mixed-norm regularization. Section V introduces the image representation based on the Vicept and its similarity measurement. Section VI presents the experimental results of different image applications on both the standard benchmark and a large-scale image database. Finally, Section VII concludes this paper.

II. Related Work

After presenting the problem of visual polysemia and concept polymorphism, we present the related work to provide a further comprehensive discussion about the large-scale image semantic understanding. In this section, we firstly discuss the state-of-the-art image representation model in large-scale image applications, and then introduce recent machine learning approaches on image understanding.

Traditional global features such as color and texture only capture parts of visual characteristics, and thus they usually cannot be directly correlated with image semantics. In most cases, semantic exists in the various scales and locations within an image. On noticing this, researchers try to detect image concepts based on local descriptions. Many significant works have been proposed to address the problem of large-scale image understanding. The de-facto standard in these works is based on the BOV model [18]. The BOV approach has been ameliorated in the following four ways: 1) *Local descriptor aggregation techniques*. Bronstein *et al.* [2] construct a spatially-sensitive image descriptor in which both the features and their relations are affine-invariant. Perronnin *et al.* [16] propose a simple yet efficient way of aggregating local descriptors into a vector of limited dimension inspired by the Fisher kernel representation. In all, the main aim of these works is to generate a discriminative and compact description with spatial information for further image applications. 2) *Discriminative codebook generation methods*. Sivic and Zisserman [18] originally propose to cluster the low-level features into codebooks with the k-means algorithm. Nister *et al.* [27] quantize the local region descriptors with hierarchical k-means in a vocabulary tree, which allows a larger and more discriminative vocabulary. Philbin *et al.* [17] improve the visual vocabulary with approximate k-means algorithm. Jurie and Triggs [8] propose a scalable acceptance-radius based codebook clustering method that can generate better codebooks. Marial *et al.* [31] raise an online codebook learning scheme for sparse coding. Germert *et al.* [22] learn the codebook with kernel method. Codebook generated by learning approaches has lower noise than that generated by clustering, and thus researchers are paying more attention to the related learning methods. 3) *Efficient quantization techniques*. Nearest neighbor algorithm is the most commonly used method for quantization. Besides this, Jégou *et al.* [14] propose a hamming Embedding technique to provide binary signatures for visual words matching refinement. Torralba *et al.* [9], [11] construct a small code to compress the BOV. Recently, sparse coding is applied to find a succinct set of codeword from the dictionary to efficiently represent visual descriptors [3], [23], [15], and sparse representations have obvious computational benefits of saving both processing time in handling visual descriptors and space in storing encoded images. Gemert *et al.* [22] demonstrate explicitly that modeling visual word assignment ambiguity improves search performance compared to the hard assignment of the traditional codebook model. 4) *Post-processing techniques*. Chum *et al.* [24] bring the query expansion idea into the text retrieval literature into the visual domain via spatial constraints and learn a latent feature model. Weak geometric consistency constraints [17] are also widely used in the post-processing phase. Post-processing plays an important role in improving the performance. Although the above techniques provide feasible solutions to some extent, how to find the semantic information from the local description is an important challenge due to the VPCP problem. In fact, local visual description such as SIFT [32] is able to keep elementary semantic information due to consideration of local salient and invariant information. Therefore, local description can be used as the basic visual unit to construct the visual-semantic relations under the large-scale environment.

Recently, machine learning approaches have received a lot of attention on the image understanding tasks, such as face recognition [6], image annotation [13], [25], image classification [1], [33], [28], object recognition, and scene understanding [12], [19], [29]. Weinberger and Saul [33] study how to improve large margin nearest neighbor classification by distance metric learning method. Boiman *et al.* [1] suggest an Image-To-Class distance metric learning method for image classification by learning per-class Mahalanobis metrics. Qi *et al.* [28] develop an approach for cross-category transfer learning for a visual classification task. Sadeghi and Farhadi [29] introduce visual phases as categories to recognize the object and understand the scene. Bucak *et al.* [30] develop an efficient algorithm for multi-label multiple kernel learning (ML-MKL) to recognize the visual object. These works provide insights on learning image semantics, but the problem of VPCP has not been directly resolved, especially under large-scale dataset. Some works focus on dealing with intra-class variance of images while ignoring the polysemia of visual appearances. Some works pay attention to learning the visual pattern of different concepts; however, they neglect the influence of semantic polymorphism.

To specifically solve the VPCP problem, this paper proposes a hierarchical semantic description—Vicept, which constructs the connection between visual appearance to semantic concepts. In the following sections, we will introduce the generating procedure of Vicept in detail: 1) visual appearance representation; 2) hierarchical Vicept learning.

III. Visual Appearance Representation

Visual appearance representation is one of the crucial steps in the BOV model and it is also the precondition of learning the Vicept. In this section, we will introduce several different methods, and explain the reasons of using the group sparse coding in our scheme.

A. Vector Quantization (VQ)

In the traditional BOV approach, every visual descriptor is encoded by k-means vector quantization. Let $\mathbf{X}$ be a set of T local visual descriptors in a P-dimensional feature space, such as SIFT [32], i.e., $\mathbf{X} = [\boldsymbol{x}_1; \ldots; \boldsymbol{x}_T] \in \Re^{T \times P}$. The VQ method applies the k-means clustering algorithm to minimize the construction error:

$$\min_{D} \sum_{m=1}^{T} \min_{k=1\ldots K} \|\boldsymbol{x}_m - \boldsymbol{d}_k\|^2 \tag{1}$$

where $\boldsymbol{D} = [\boldsymbol{d}_1; \ldots; \boldsymbol{d}_K] \in \Re^{K \times P}$ includes K cluster centers, which is called *dictionary*, and each cluster center is regarded as a visual word. $\|\cdot\|$ depicts the ℓ_2-norm of vector. The optimization problem can be re-formulated into a matrix factorization problem with cluster membership indictors $\boldsymbol{A} = [\boldsymbol{a}_1; \ldots; \boldsymbol{a}_T] \in \Re^{T \times K}$

$$\begin{aligned} &\min_{A,D} \sum_{m=1}^{T} \min_{k=1\ldots K} \|\boldsymbol{x}_m - \boldsymbol{a}_m \boldsymbol{D}\|^2 \\ &\text{subject to } \|\boldsymbol{a}_m\|_0 = 1, \|\boldsymbol{a}_m\|_1 = 1, \boldsymbol{a}_m \geq 0, \forall m \end{aligned} \tag{2}$$

where $\|\boldsymbol{a}_m\|_0 = 1$ is a cardinality constraint, meaning that only one element of $\boldsymbol{a}_m$ is nonzero, $\boldsymbol{a}_m > 0$ means that all the elements of $\boldsymbol{a}_m$ are nonnegative, and $\|\boldsymbol{a}_m\|_1$ is the ℓ_1-norm of the vector, the sum of the absolute value of each element in $\boldsymbol{a}_m$. After the optimization procedure, the index of the only nonzero element in $\boldsymbol{a}_m$ indicates which visual word the $\boldsymbol{x}_m$ belongs to.

However, the constraint $\|\boldsymbol{a}_m\|_0 = 1$ may be too rigorous, often giving rise to a coarse reconstruction of $\mathbf{X}$. We relax the constraint by putting a ℓ_1-norm regularization on $\boldsymbol{a}_m$, which enforces $\boldsymbol{a}_m$ to have a small number of nonzero elements. Then the VQ formulation is turned into another problem known as *sparse coding* (SC [3], [23], [31], [34]):

$$\min_{A,D} \sum_{m=1}^{T} \|\boldsymbol{x}_m - \boldsymbol{a}_m \boldsymbol{D}\|^2 + \lambda \|\boldsymbol{a}_m\|_1 \quad \text{subject to } \boldsymbol{a}_m > 0, \forall m. \tag{3}$$

The first term of the objective weighs the reconstruction error and the second term weighs the degree of sparsity. The larger the parameter λ is, the sparser the reconstruction coefficient is. Sparse representations have obvious benefits, by economizing both processing time in handling visual descriptors and the storage space in encoding image descriptors.

B. Group Sparse Coding (GSC)

The sparse code approaches based on ℓ_1-norm regularization consider each visual descriptor in the image as a separate coding problem and do not take the fact into account that descriptor coding is just an intermediate step in creating a BOV representation for the whole image. This might prevent the use of these methods in real large-scale image applications, which are constrained by either time or space resources. Thus, considering the structure of BOV in images, we encode jointly all the visual descriptors in an image by instead putting the ℓ_1/ℓ_2-norm regularizer [35]–[37]:

$$\min_{A,D} \frac{1}{2} \sum_{m=1}^{T} \|\boldsymbol{x}_m - \boldsymbol{a}_m \cdot \boldsymbol{D}\|^2 + \lambda \sum_{k=1}^{K} \|\boldsymbol{a}^k\|_2 \quad \text{subject to } a_j^i \geq 0, \forall j, i \tag{4}$$

where $\boldsymbol{a}_m = (a_m^1, \ldots, a_m^K)$ and $\boldsymbol{a}^k = (a_1^k; \ldots; a_T^k)$ are nonnegative vectors and $\boldsymbol{A} = [\boldsymbol{a}_1; \ldots; \boldsymbol{a}_T] = [\boldsymbol{a}^1, \ldots, \boldsymbol{a}^K]$ is the reconstruction matrix, T is the total number of visual descriptors in the image.

Using the sparse coding with the ℓ_1/ℓ_2-norm regularizer, we can for example specify an encoder that exploits the fact that once a visual word has been selected to help represent one of the visual descriptors of an image, it may as well be used to represent other visual descriptors of the same image without much additional regularization cost. Similar to SC, GSC has an encoding phase and a dictionary learning phase. In the encoding phase, for each image represented as a descriptor set $\mathbf{X}$, the GSC code $\boldsymbol{A}$ is obtained by optimizing (4) with respect to the dictionary $\boldsymbol{D}$ only. For improving the accuracy of encoding, the dictionary usually needs a further refinement, which is also called "dictionary learning". In the dictionary learning phase, the new $\boldsymbol{D}$ is obtained by current $\boldsymbol{D}$ and its corresponding code $\boldsymbol{A}$. After several alternations between these two phases, we can obtain a discriminative and compact dictionary.

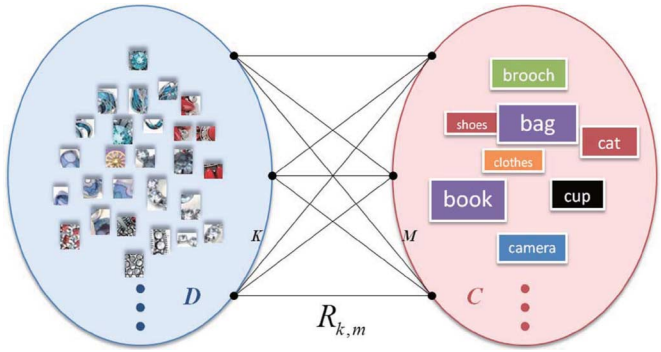


Fig. 6. Bipartite graph for local visual appearance set versus concept collection.

We choose GSC to derive image representations because it has a number of attractive properties: 1) Compared with the VQ method, GSC can achieve a much lower reconstruction error rate due to the less restrictive constraint; 2) Sparsity allows the representation to be special, and to capture salient properties of images; 3) Research in image statistics clearly reveals that image patches are sparse signals; 4) Compared with the SC coding, GSC can obtain the sparse representation at the level of image rather than descriptor.

IV. Vicept Generation

As mentioned above, Vicept builds the bridge between visual appearances and semantic concepts. In other words, we aim to provide a method which binds "visual word-semantic concept" together as well as takes the VPCP problem into account. In this section, we first formulate the problem, and then present a carefully prepared training image dataset. Finally, we detail the approach for generating the bottom-level Vicept and constructing the Vicept description with hierarchical semantic concepts.

A. Image Vicept Description

The observation of VPCP problem motivates us that the relationship between concept collections C and visual appearance set D can be formalized as a bipartite graph, which is illustrated in Fig. 6, where K is the number of visual appearances and M is the number of concepts. The relationship $R_{k,m}$ indicates the relationship between the kth visual appearance and the mth concept. To efficiently make use of this structure, we design Vicept with the following details:

1) *Local Visual Appearance*: We adopt local descriptor to represent image. In our approach, SIFT [32] is detected and quantized into visual words by group sparse coding.
2) *Semantic Concept Collection*: The concepts in real world are not independent but closely related. Following the structure in [5] and [38], we simplify the concept modeling with a hierarchical representation and all the concepts are organized in a concept tree. We detail this concept collection in the next subsection.

Before learning the Vicept description, a short interpretation is presented as follows. Suppose having a dictionary $\boldsymbol{D}$ with K visual words and a hierarchical concept collection $\boldsymbol{C}$ with M concepts, a membership distribution can be learned between

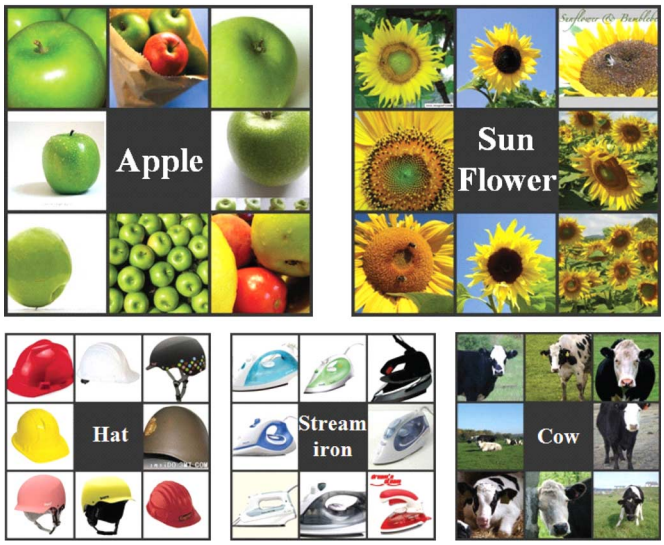


Fig. 7. Example images in the dataset.

each visual word and concept collection. In a Vicept word, each visual word has a corresponding M-bin membership distribution histogram with concept collection $\boldsymbol{C}$. Each Vicept consists of two parts: one is the original visual word, and the other is the corresponding M-bin membership distribution histogram. Finally, we can obtain a Vicept dictionary according to the visual dictionary $\boldsymbol{D}$.

B. Image Training Dataset and Concepts

As [5] declares, "*a large-scale ontology of images is a critical resource for developing advanced, large content-based image search and image understanding algorithms*." ImageNet [5] is the largest clean image dataset available to the vision research community. We use it as the source of our dataset by searching the concept names and downloading the returned images according to their URLs. Fig. 7 provides some example images in our dataset.

The concept collection in our dataset is organized by a semantic hierarchy which is used by ImageNet. However, to ensure enough training images and for the sake of limited computation capacity, we manually filter the concepts with less than 1 k images returned by ImageNet and select a frequently used collection with 217 concepts. In all, there are altogether 267 k images in our dataset and we use a simple 3-level concept structure: 10 concepts on level-1, 88 on level-2, and 217 on level-3. Note that we maintain an "Is-a" relationship between different concept levels and only the concepts at the same level are independent and comparable.

Although ImageNet offers clean image annotation, the generation of concept labeled local features is also "unclean". This is because the interest points are generated from both of the foreground and background. Supposing there is an image about a man using a cup to drink water, once the image is related with concept "*cup*", the local features generated from the man's area are also label as "*cup*" which is not expected in our approach.

On noticing this fact, we try to "purify" the dataset by manually segmenting the image and eliminate the irrelevant areas.

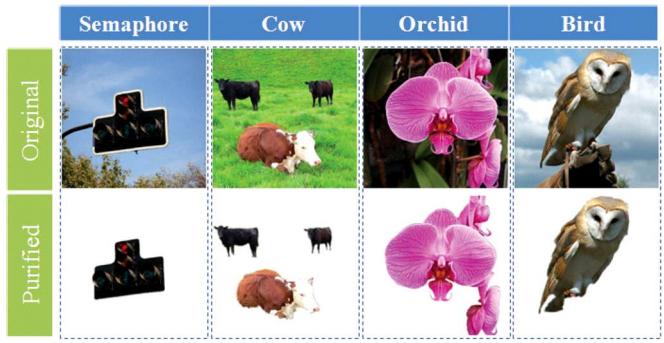


Fig. 8. Examples for image matting.

Balancing the workload for image matting and the data requirement in this task, we prepare a subset of the original 260 k image dataset with 120 "purified" images in each concept. This subset contains 120×217 one-concept-labeled images and it is only used for generating Vicept words. Fig. 8 illustrates the result for the "purification" of images, which bring the reliability and robustness for the Vicept learning.

C. Learning Vicept Word via Mixed Norm Regularization

Vicept is to present the characteristic/typical relationship between visual appearances and hierarchical semantic concepts. In fact, this relationship is structurally sparse in the practical situations: one visual appearance normally has the correlation with some limited concepts and one semantic concept usually has the limited visual appearances. Taking this structural sparsity property of Vicept into account, we adopt the idea of mixed-norm regularization in our optimization problem of learning the discriminative Vicept. Let $\boldsymbol{I} = \{I^1, \ldots, I^N\}$ be a group of images and $\boldsymbol{Y} = \{\boldsymbol{y}^1, \ldots, \boldsymbol{y}^N\}$ be the corresponding labels of images. $\boldsymbol{y}^n = (y_1, \ldots, y_M)$ is relative to the concept collection $\boldsymbol{C}$ with M concepts, and $y_i \in [0, 1]$ is the possibility that the ith concept appears in image $\boldsymbol{x}^n$. $\boldsymbol{A}^* = \{\boldsymbol{A}^1, \ldots, \boldsymbol{A}^N\}$ is the corresponding reconstruction coefficient related to dictionary $\boldsymbol{D}$. $\boldsymbol{U} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K\}$ denotes the *Vicept dictionary* and K is the number of visual words in $\boldsymbol{D}$. The following objective optimization is as follows:

$$\begin{aligned} \boldsymbol{J}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{A}, \boldsymbol{D}) = \frac{1}{2}\sum_{i=1}^{N} \left\| \boldsymbol{y}_i - \sum_{m=1}^{|x^i|}\sum_{j=1}^{K} a_j^{i,m} \cdot \boldsymbol{u}_j \right\|^2 \\ + \gamma \sum_{j=1}^{K} \|\boldsymbol{u}_j\|_p \\ \text{subject to } u_{j,k} \geq 0, \forall j, k \end{aligned} \tag{5}$$

where $a_j^{i,m}$ is the reconstruction coefficient of the jth visual word for the mth descriptor of the ith image, and the non-negative vector $\boldsymbol{u}_j = (u_{j,1}, \ldots, u_{j,M})$ indicates the relationship of the jth visual word with concept collection. The first term of the objective measures the reconstruction quality. The mixed-norm regularization is the form of ℓ_1/ℓ_p-norm [35], [36], which is presented by the second term of (5) and measures the reconstruction complesity. It consists of two parts: One is ℓ_p-norm complexity of $\boldsymbol{u}_j$, and the other is the ℓ_1-norm sum of $\boldsymbol{U}$. In the view of Vicept, the mixed-norm regularization helps the Vicept

achieve the structural sparsity that all the images from the same concept have the similar sparse Vicept representation, which is the essential difference with the ℓ_1- or ℓ_2-norm regularization. The parameter γ balances the effect of these two terms.

The problem of (5) can be solved by coordinate descent. Leaving all indices of $\boldsymbol{U}$ intact except for index r, omitting fixed argument of the objective, let φ be the term which does not rely on $\boldsymbol{u}_r$ and $\sum_{m=1}^{|x^i|} a_j^{i,m} = s_j^i$, we obtain the following reduced objective function:

$$\begin{aligned}\boldsymbol{J}(\boldsymbol{u}_r) &= \frac{1}{2}\sum_{i=1}^{N}\|\boldsymbol{y}^i - \sum_{j\neq r}^{K} s_j^i \cdot \boldsymbol{u}_j - s_r^i \cdot \boldsymbol{u}_r\|^2 \\ &\quad + \gamma\sum_{j=1}^{K}\|\boldsymbol{u}_j\|_p \\ &= \sum_{i=1}^{N}\left(\sum_{j\neq r}^{K} s_j^i s_r^i \boldsymbol{u}_j\cdot\boldsymbol{u}_r - s_r^i \boldsymbol{y}^i\cdot\boldsymbol{u}_r \right. \\ &\quad \left. + \frac{1}{2}(s_r^i)^2\|\boldsymbol{u}_r\|^2\right) \\ &\quad + \gamma\sum_{j=1}^{K}\|\boldsymbol{u}_j\|_p + \varphi \\ &= \sum_{i=1}^{N}\left(\sum_{j\neq r}^{K} s_j^i s_r^i \sum_{x=1}^{M}\boldsymbol{u}_{r,x}\boldsymbol{u}_{j,x} - s_r^i\sum_{x=1}^{M}\boldsymbol{y}_x^i\boldsymbol{u}_{r,x} \right. \\ &\quad \left. + \frac{1}{2}\left(s_r^i\right)^2\sum_{x=1}^{M}(\boldsymbol{u}_{r,x})^2\right) + \gamma\sum_{j=1}^{K}\|\boldsymbol{u}_j\|_p + \varphi.\end{aligned}$$

Next we show how to find the optimum $\boldsymbol{u}_r$. Let $\widetilde{\boldsymbol{J}}$ be the first reconstruction term of the objective, and its partial derivatives with respect with to each $u_{r,x}$ are

$$\frac{\partial}{\partial u_{r,x}}\widetilde{\boldsymbol{J}} = \sum_{i=1}^{N}\left(\sum_{j\neq r}^{K} s_j^i s_r^i u_{j,x} - s_r^i y_x^i + (s_r^i)^2 u_{r,x}\right).$$

Let us make the following abbreviation for a given index γ,:

$$w_x = \left|-\sum_{i=1}^{N}\left(\sum_{j\neq r}^{K} s_j^i s_r^i u_{j,x} - s_r^i y_x^i\right)\right|_+$$

where $|x|_+ = \max(0, x)$. In this case of $p = 1$, the objective function is isolated and we can get the following sub-gradient condition for optimality:

$$\begin{aligned}0 \in -w_x + \sum_{i=1}^{N}(s_r^i)^2 u_{r,x} + \gamma\underbrace{\frac{\partial}{\partial u_{r,x}}\|u_{r,x}\|_1}_{\in[0,1]} \\ \Rightarrow u_{r,x} \in \frac{w_x - [0,\gamma]}{\sum_{i=1}^{N}(s_r^i)^2}. \quad (6)\end{aligned}$$

Since $u_{r,x} \geq 0$, the above sub-gradient condition for optimality implies that $u_{r,x} = 0$ when $w_x \leq \gamma$ and otherwise $u_{r,x} = (w_x - \gamma)/\sum_{i=1}^{N}(s_r^i)^2$.

For $p = 2$, indicating $\boldsymbol{w} = (w_1, \ldots, w_M)$, the gradient of $\boldsymbol{J}(\boldsymbol{u}_r)$ with the ℓ_2-norm penalty is as follows:

$$\frac{\partial}{\partial \boldsymbol{u}_r}\boldsymbol{J} = -\boldsymbol{w} + \sum_{i=1}^{N}\left(s_r^i\right)^2\boldsymbol{u}_r + \gamma\frac{\boldsymbol{u}_r}{\|\boldsymbol{u}_r\|}. \quad (7)$$

At the optimum, the value of the gradient should be equal to zero, thus we obtain

$$\boldsymbol{u}_r = \left(\sum_{i=1}^{N}\left(s_r^i\right)^2 + \frac{\gamma}{\|\boldsymbol{u}_r\|}\right)^{-1}\boldsymbol{w}. \quad (8)$$

Let $\boldsymbol{u}_r = h\boldsymbol{w}$, h is the scale. We can rewrite (8) as follows:

$$h\boldsymbol{w} = \left(\sum_{i=1}^{N}\left(s_r^i\right)^2 + \frac{\gamma}{\|h\boldsymbol{w}\|}\right)^{-1}\boldsymbol{w} \quad (9)$$

which infers that

$$h = \frac{1}{\sum_{i=1}^{N}\left(s_r^i\right)^2}\left(1 - \frac{\gamma}{\|\boldsymbol{w}\|}\right). \quad (10)$$

Because h should be a non-negative, we get that if $\|\boldsymbol{w}\| \leq \gamma, \boldsymbol{u}_r = \boldsymbol{0}$; otherwise $\boldsymbol{u}_r = h\boldsymbol{w}$ and h is defined as (10). Finally, we can obtain the Vicept dictionary $\boldsymbol{U}$ via above recursions.

D. Building Vicept With Hierarchical Concept Membership Distribution

Following the structure [5], concept collection is organized into a hierarchical tree. Based on the above learning, we obtain the bottom-level Vicept. In this paragraph we show how to establish the high-level concept memberships.

One basic approach is to bottom-up construct the multi-level histograms using "Is-a" relationship. That is, to sum the low-level concept membership distributions and obtain the higher one. However, as the density of the bottom-level concepts is not high enough (of course we cannot cover all the concepts in real world, in fact we only select 217 bottom-level concepts in this paper), the imbalance of the concept selection restricts the performance for high-level histogram generation. Fig. 9 details an example. According to the low-level histogram, the Vicept word is closely related with concept "*apple*" (0.4) and we further infer that the Vicept word may have a high possibility belonging to the concept "*fruit*" on high-level histogram. Nevertheless, because of the imbalance of the concept selection (6 concepts belong to animal but only 2 belong to fruit on low-level histogram), the Vicept word is more closely related to concept "*animal*" on high-level histogram and this error could be accumulated when calculating higher level histograms. In that case, rather than simply summing the bins that belong to the same high-level concept, we should assign large weights to the bins with large bin value. In fact, we implement this by adopting sigmoid function and normalize the sum of high-level histogram into 1. We implement this by adopting sigmoid function and normalize the sum of high-level histogram. The weight $w(i)$ for the ith bin is calculated by the value of itself:

$$w(i) = \frac{1}{1+\exp(v(i))}, v(i) \in [0,1]. \quad (11)$$

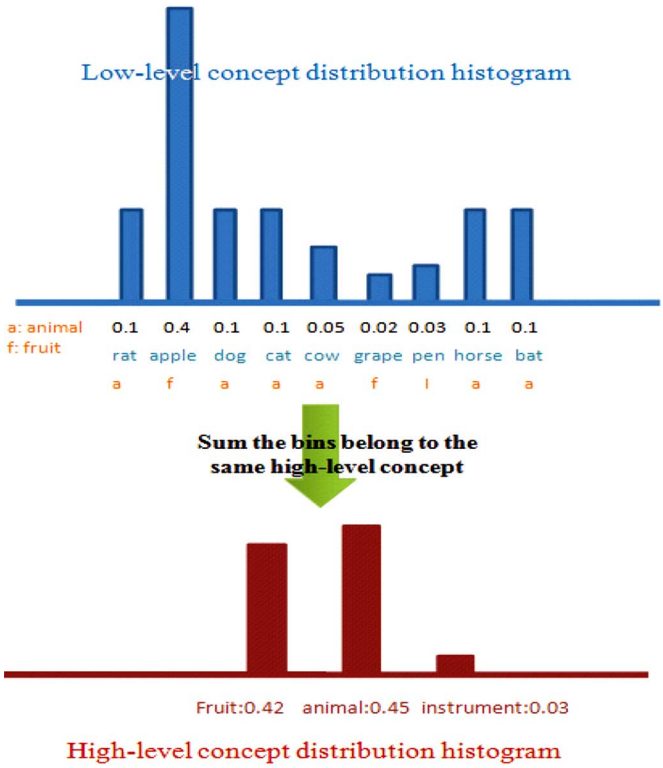


Fig. 9. Imbalance of the concept selection restricts the performance for high-level histogram generation.

Finally, we obtain a complete Vicept dictionary, where each Vicept word has a visual word with multi-level concept membership distribution histograms. Fig. 4 illustrates a typical Vicept word.

V. Image Vicept Representation and Similarity Measurement

Through the learning from the above section, we can obtain the Vicept description. In this section, we formulate the image-level Vicept representation procedure, which provide a Vicept computing method from local to global. Moreover, we introduce the distance metric for hierarchical Vicept representation based on multi-level separability analysis.

A. Image Vicept Representation

We aim to represent image (or part of the image for tasks such as partial annotation) into a concept membership distribution histogram in which the larger bin value denotes the higher probability of concept existence. In fact, we use the same multi-level concept structure with Vicept words as our image representation and this can be implemented by

$$\Pr(C_i|I) = \sum_{j=1}^{n} \Pr(C_i|Vw_j) \times \Pr(Vw_j|I). \qquad (12)$$

$Vw_j (j = 1, 2, \ldots, n)$ denotes the n Vicept words generated by Section IV-C. I is the input image (or partial image) and C_i is the ith concept. The multiplicand $\Pr(C_i|Vw_j)$ at the right side can be directly accessed according to the Vicept word j; the multiplicator $\Pr(Vw_j|I)$ can also be solved by counting visual words in the image. For a certain image, firstly we detect its interest points (SIFT), then encode them into sparse code according to the visual vocabulary generated in Section III-B. Thirdly we quantize them into Vicept words according to the Vicept vocabulary generated in Section IV-C. Finally, the image is represented as multi-level concept membership distribution histograms by (12). Note that our approach has already embedded semantic information into the Vicept description, after detecting the interest points, the multi-level histograms directly depict the existence probability for the semantic concepts. Fig. 5 illustrates a typical (partial) image represented by multi-level histograms according to Vicept. According to the bin values of the multi-level histograms at the left bottom, the image is likely to contain concept: "horse" (on level-3), "mammal" (on level-2), and "animal" (on level-1). Further, if the input is a partial image (areas in the red rectangle), we can also obtain the concept-level image representation by Vicept description.

B. Image to Image Distance Based on Vicept

After generating the image-level Vicept representation, similarity measurement is another crucial question for further application. According to Vicept description, image is represented as multi-level concept membership distribution histograms (Fig. 5). Intuitively, we cannot concatenate the histograms into one and calculate the classical histogram distances (e.g., Euclidean distance, cosine distance, histogram intersection, chi-square distance, Minkowski-form) because the concepts at different levels are incomparable and with different discriminative powers. To measure the distance between two images, we first compute the distance of histograms on the same level and then fuse the results together by separability analysis.

We use weighted chi-square distance to measure the distance between histograms at the same level of Vicept. For m-bin histogram u and v, the distance is defined as

$$\psi(u, v) = \frac{1}{2}\sum_{i=1}^{m} \text{weight}\,(\text{bin}_i^u, \text{bin}_i^v) \times \frac{(\text{bin}_i^u - \text{bin}_i^v)^2}{\text{bin}_i^u + \text{bin}_i^v} \qquad (13)$$

where the weight reflects the importance of bin_i^u and bin_i^v. Intuitively, if both bin_i^u and bin_i^v are large, the two images are closely related with the concept in the ith bin, and we should assign a high weight value; otherwise, we assign a low weight. Here, $\min(\text{bin}_i^u, \text{bin}_i^v)$ is adopted as the weighting function.

In our approach, the weight for each level of Vicept is learned by Fisher linear discriminant [39] which aims to achieve high separability between different patterns. It is generally believed that the similarities of images from the same concept are higher than those from the different concept. Therefore we define and calculate the *inter-concept* and *intra-concept* statistics.

Suppose there are L levels histogram in the image Vicept representation. Without loss of generality, the ith level is composed by c_i concepts $(\text{Concept} = \{C_1, \ldots, C_{c_i}\})$. For concept C_{k_i}, we have N_{k_i} images in our dataset.

Intra-concept mean m_i^{intra} for the ith level:

$$m_i^{\text{intra}} = \frac{1}{c_i}\sum_{j=1}^{c_i}\frac{2}{N_j(N_j-1)}\sum_{u=2}^{N_j}\sum_{v=1}^{u-1}\psi(u,v).$$

Intra-concept variance v_i^{intra} for the ith level:

$$v_i^{\text{intra}} = \frac{1}{c_i}\sum_{j=1}^{c_i}\frac{2}{N_j(N_j-1)}\sum_{u=2}^{N_j}\sum_{v=1}^{u-1}\left[\psi(u,v)-m_i^{\text{intra}}\right]^2.$$

Inter-concept mean m_i^{extra} for the ith level:

$$m_i^{\text{extra}} = \frac{2}{c_i(c_i-1)}\sum_{f=2}^{c_i-1}\sum_{g=1}^{f-1}\frac{1}{N_fN_g}\sum_{u=1}^{N_f}\sum_{v=1}^{N_g}\psi(u,v).$$

Inter-concept variance v_i^{extra} for the ith level:

$$v_i^{\text{extra}} = \frac{2}{c_i(c_i-1)}\sum_{f=2}^{c_i-1}\sum_{g=1}^{f-1}\frac{1}{N_fN_g}\sum_{u=1}^{N_f}\sum_{v=1}^{N_g}\left[\psi(u,v)-m_i^{\text{extra}}\right]^2.$$

Finally, the weight for level i is defined as

$$w_i = \frac{\left(m_i^{\text{intra}}-m_i^{\text{extra}}\right)^2}{v_i^{\text{intra}}+v_i^{\text{extra}}}.$$

By learning the intra-concept and inter-concept distances for the images in our dataset, we find the best projection direction to fuse multi-level histograms of Vicept. The distance between image $\boldsymbol{x}$ and $\boldsymbol{y}$ is calculated with level weighting by the following equation:

$$D(x,y) = \sum_{i=1}^{|\text{level}|} w_i \cdot \psi(\boldsymbol{x}_i, \boldsymbol{y}_i)$$

where w_i is the weight for level i, $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are the Vicept description of $\boldsymbol{x}$ and $\boldsymbol{y}$ on the ith level, and $\psi(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is calculated by (13).

VI. Experiments

As a visual description closely integrated with semantic concepts, Vicept is recommended to be adopted in semantic related applications. In this section, we first introduce the experimental settings and then verify the validation of Vicept in three semantic related tasks: large-scale semantic image search, image annotation and semantic image re-ranking.

A. Database and Experimental Setting

Database: We use ImageNet [5] as the source of our training dataset, which is organized by a semantic hierarchy which is used by WordNet and ImageNet. We select a frequently-used collection with 217 low-level concepts and there are 267 k images (later referred to as ImageNet267K). We use a simple 3-level concept structure: 10 concepts on level-1, 88 concepts on level-2, and 217 concepts on level-3. There is a detailed introduction about the multi-level image dataset in Section IV-B. This subset contains 120×217 one-concept-labeled images (later referred as ImageNet25K). Another standard benchmark (Corel5K) is used in our experiment, which is based on the Corel image database [41] from 50 Corel Stock Photo CDs and consists of 5000 images. Furthermore, we use an additional set of 800 k distracter Flickr images (later referred to as Flickr800K).

Experimental settings: Vicept is learned from the ImageNet25K dataset. The SIFT description is extracted as the local visual appearance. The initial dictionary has 4056 visual words, which were obtained from the hierarchical k-means clustering. A new dictionary with 473 visual words is generated through group sparse coding, which is detailed in Section III-B.

B. Large-Scale Semantic Image Search

Incorporating with the BOV approach, Vicept builds the bridge between visual words and semantic concepts. In this paragraph, we validate its efficiency on a large-scale dataset, which consists of 1 M images (ImageNet267K and Flickr800K).

Baseline: We use a traditional BOV approach [27] as the baseline approach and a dictionary of 200 K visual words is used. We experimented with different size of visual word dictionary, and found 200 K dictionary to give the best performance.

Comparisons: We also enhance the baseline method with soft assignment [20], where the number of nearest neighbors is set to be 4. We call this method "Soft BOV". A state-of-the-art descriptor-VLAD (vector of locally aggregated descriptors) [10], which derived from both BOV and Fisher kernel, is compared with our method. Here, we use the same parameter as [10]: the cluster centroids k is 128, and the final dimension D is 16 384. The Vicept-based approach has two variants: 1) "ℓ_1-vicept", in which we set $p = 1$ as the norm penalty in (5). 2) "ℓ_1/ℓ_2-vicept", in which p is set to be 2 in (5). The parameter γ impacts the structural sparsity of Vicept. The larger it is, the sparser the Vicept is. This sparsity saves the processing time and storage space, but the performance decreases rapidly when the Vicept is too sparse. Thus, as the experiment setting of [40], we set $\gamma = 0.08$ for $p = 1$ and $\gamma = 0.01$ for $p = 2$. As to the similarity measurement between image representations based on Vicept, two methods are compared: 1) following [40], we simply compute a weighted histogram intersection for the level-3 membership distribution. We mark this similarity measurement method "HI"; 2) we measure the distance between images using the novel method, which is represented in Section V-B and is called "FisherWeight" for short.

In the evaluation, we select 250 representative images from the ImageNet267K as our queries. We use mean average precision (MAP) as our evaluation metric. For each query image, we compute its precision-recall curve and count the area below the curve. Finally, we take the mean value over all queries.

Fig. 10 compares the above seven approaches with MAP, leading to five observations. Firstly, our Vicept significantly improves the MAP, as can be seen by comparing the results with "baseline". On the 1 M image dataset, the methods based on the Vicept boost the MAP from 0.09 to 0.49, a 40% improvement. Secondly, soft assignment of visual words plays an important role in improving the performance (a 20% improvement

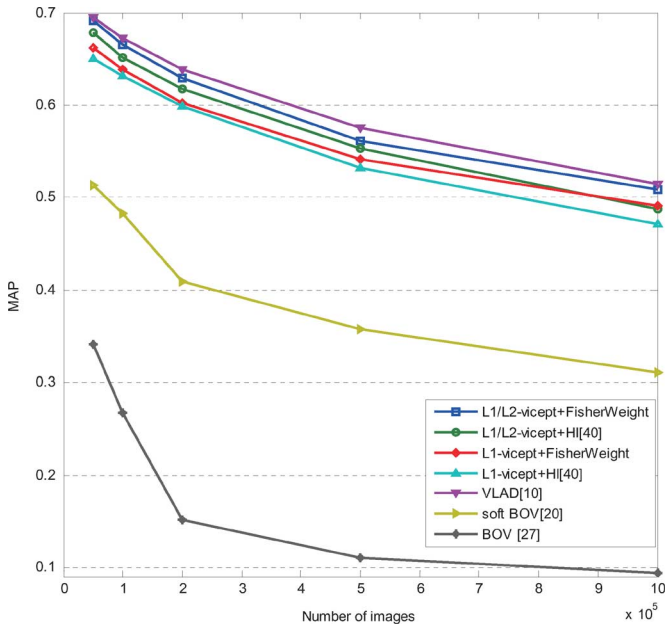


Fig. 10. Comparisons of different methods using MAP with different scale of image dataset.

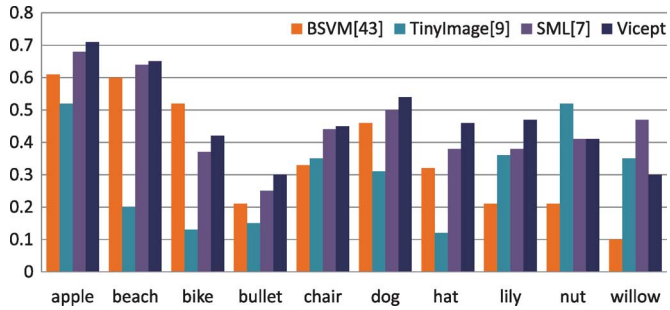


Fig. 11. Comparisons of different annotation methods with AP over the ImageNet267 K database: "BSVM", "TinyImage", and "SML" is the baseline approach; "Vicept" is the approach based on the ℓ_1/ℓ_2-vicept.

on average). This point is also demonstrated in [22]. Thirdly, when the number of images is under 0.5 M, the MAP of our approach is comparable with that of VLAD [10]. The good performance of our method attributes to the reason that we take the VPCP problem into account when designing the Vicept schema. However, when the number of images is about 1 M, the VLAD performance descends rapidly with a 10.43% rate while our "ℓ_1/ℓ_2-vicept" method has only a 9.27% descending rate. The slow performance descent of our method demonstrates that our method is less sensitive to the increment of the dataset scale compared with the VLAD method. Fourthly, the "ℓ_1/ℓ_2-vicept" method reaches a higher MAP than the "ℓ_1-vicept". One main reason is that the Vicept learned via ℓ_1/ℓ_2-norm regularization is structural sparser, which allows the representation to capture salient properties of relations between visual appearances and semantic concepts. Fifthly, for both ℓ_1/ℓ_2-vicept and ℓ_1-vicept, the new distance metric has a better performance, which attribute to the fact that the learned weights by Fisher linear method can capture the discriminative power of concepts at the different levels.

C. Image Annotation

To evaluate the performance, Vicept is evaluated on the ImageNet267K and one standard benchmark (Corel5K).

1) Image Annotation on Imagenet267K: Three different approaches are implemented as baseline for Image Annotation, 1) Binary SVM [43]: For Binary SVM, we prepare 217 binary SVM classifiers with an output of classification probability. In the training phase, for every SVM concept classifier, we pick 100 positive samples and 200 negative samples from ImageNet25K. 2) Tiny Image voting [9]: For the Tiny Image voting, we replicate the experiment described in [9]. Firstly, the query image and the images in 267 K dataset are down sampled to 32×32. Then 100 nearest neighbors for the query are returned based on SSD pixel distance. Finally, we obtain the concept by aggregating the votes. 3) Supervised Multiclass Labeling (SML) method [7]: Following [7], we represented images as bags of localized features and Gaussian mixture model (GMM) of 64 components is learned from each mixture.

The accuracy is measured as the average precision (AP) averaged over the 100 queries from 10 concepts. For an image, each approach provides a top-5 annotation. If one of the five labels is correct, this annotation is valid.

Fig. 11 illustrates the average precision for four approaches. We can find that Vicept provides a better concept annotation result than BSVM, TinyImage, and SML for most of the query images. However, TinyImage has the best performance on the "nut", because the diversity of images in "nut" is slight and in this situation the nearest neighbor is reilable. Besides, benefitting from the GMM, SML has the better annotation result than Vicept on the "willow". In the view of Vicept, the fluctuations on concept "nut" and "willow" are likely to be influenced by small number of training data for Vicept learning. But the 47.1% mean AP of Vicept seems to be satisfactory in this annotation task.

2) Image Annotation on Corel5k: Corel5K [41] has become the benchmark for image annotation, which contains 5000 images with 260 concepts. We find that the concept collection from Vicept covers most of the major keywords. In this paragraph, we complement the annotation task with the Vicept learned from ImageNet25K.

The baseline is Binary SVM [43]. Similar to the above procedure, we train 260 binary SVM classifiers with an output of classification probability. In the training stage, we split the Corel5K into a training set of 4000 images, a validation set of 500 images, and a test set of 500 images. Another comparison approach is the supervised multiclass labeling (SML) method [7]. Following [7], images are represented as bags of localized features and GMM of 64 components is learned from each mixture.

AP is used as the evaluation metric. To have a fair evaluation, we select 10 keyword concepts, which are included by the concept set of Vicept. For each keyword, 8 representative images are picked from the test data. During the annotation, we judge the validation of each annotation if one of its top-3 labels is correct.

Although our Vicept was not trained on the Corel5K dataset, the mean average precision of our proposed method is comparable to the "BSVM" [43] and SML [7], with a 1.8% and 2.1%

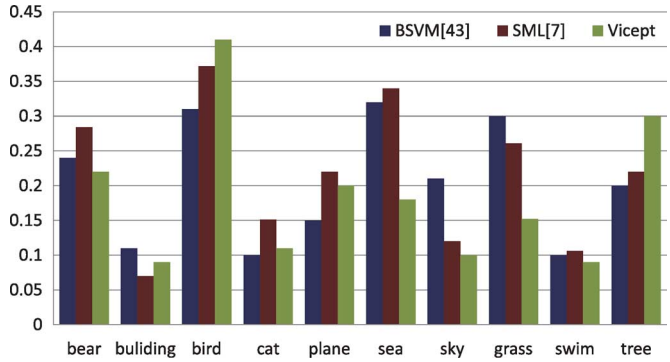


Fig. 12. Average precision comparisons of three annotations on Corel5K.

difference. For the "bear", "cat", and "grass" class, the annotation performance of our method has a degradation due to the fact that the test images from the Corel5K have a great difference with our training images from ImageNet25K. The feasible solutions of this problem are 1) to increase the number of training data and choose diverse training images for one class; 2) to enlarge the concept set in the Vicept learning procedure. The result in Fig. 12 shows that Vicept has potential to be used without relying on any outside information. Thus we could have a coarse annotation for large/web scale image datasets, such as Flickr and ImageNet, when the concept set of Vicept covered most concepts in our daily life. Benefiting from this point, we would have some significant expansions on these dataset. For users, they can search interesting images with mature text retrieval technology. For researchers, they can study image re-tagging technology [42], [44] to obtain better label results.

D. Semantic Image Re-Ranking

Image re-ranking is to re-rank the images returned by text-based search engines according to their visual appearances to make the top-ranked images more relevant to the query. Based on the Vicept, we propose a novel image re-ranking model: ViceptRank, which can be considered as distinguishing the semantic concept of the returned images from search engines and re-ranking the images based on the semantic relevance with the identified concept.

In our experiment, we submit 50 text queries to Google Image Search; we crawl 1000 images for each query and score the graded relevance of the returned results with the query text. The first image of each category from Google Image is regarded as the query image. Our baseline is VisualRank [45], which computes the visual similarities between images and leverages the algorithm similar to PageRank to re-rank the images.

Normalized discounted cumulative gain at top k ($\mathrm{nDCG@}k$) is adopted as the evaluation metric. nDCG is a normalized version of DCG metric. Two assumptions of DCG metric are: 1) highly relevant results are more useful when appearing earlier in a result list; 2) highly relevant results are more useful than marginally relevant ones, which are in turn more useful than irrelevant results. $\mathrm{nDCG@}k$ is calculated by

$$\mathrm{nDCG@}k = \frac{1}{Z}\sum_{n=1}^{k}\frac{2^{s(p)}-1}{\log(1+p)} \tag{14}$$

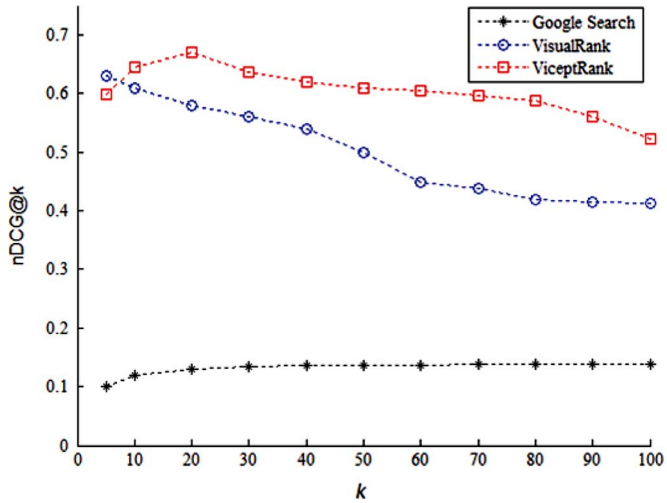


Fig. 13. Performance of semantic image re-ranking with $\mathrm{nDCG@}k$.

TABLE I
AVERAGE TIME BY VISUALRANK AND VICEPTRANK FOR RE-RANKING AN IMAGE

	VisualRank [45]	ViceptRank
Average Time	0.914s	0.142s

where $s(p)$ is the score that represents the relevance given to the retrieved image at position p, Z is a normalization term derived from the perfect ranking of top k images.

Fig. 13 shows the experimental results. We find that both VisualRank and ViceptRank outperform the Google search by 40%, which demonstrates the fact that image re-ranking technique can substantially improve the performance. Although our method is not as good as VisualRank in the performance of top-10 in the re-ranked images, our proposed ViceptRank outperforms the VisualRank by 10.1% in the overall performance. The imperfection lies in the fact that human are instinctive to score higher to visual similarity than semantic similarity while the similarity in our approach is measured based upon the concept membership distribution.

Besides the obvious improvements on $\mathrm{nDCG@}k$, it is necessary to point out that ViceptRank approach is also efficient. Table I shows the average time of re-ranking an image by VisualRank and ViceptRank. The above experiments are carried out on a laptop with 2-GB memory and 2-core 2.10-Ghz processor. Our method is faster than "VisualRank" with about 6 times. The low efficiency of VisualRank is mainly rooted in the expensive image pair similarity computation based on SIFT and LSH. However, in ViceptRank, images are represented with multi-level concept membership distribution, which is sparse representation and easy to compute. In short, the ViceptRank approach shows significant advantages on both accuracy and efficiency over the VisualRank.

VII. CONCLUSION

There is a saying "A picture is worth a thousand words." In this paper, we propose a new perspective to interpret an image into its "semantic words" (concept). A Vicept description is introduced to characterize the membership distribution between visual appearance and concepts. The mixed norm regularization

is adopted in our optimization problem for learning the membership distribution, which is effective for obtaining a discriminative Vicept with structural sparsity. Further, to aim at the hierarchal structure of Vicept, a distance metric is introduced to measure the similarity. Vicept approach provides fast computation, compact expression, and local-to-global description, and thus can be implemented for large-scale web applications.

The scalability of Vicept is restrained on the web-scale image dataset because our Vicept learning algorithm works well on the purified image dataset. Although our training set has 217 frequently-used concepts in our daily life, web-scale image dataset has thousands of concepts. For such large scale, the work of preparing the training images will be a tremendous task. In the future, we will focus on two tasks: 1) to improve the learning algorithm to obtain more powerful Vicept descriptions; 2) by making use of multi-features co-occurrence, to introduce novel approaches of learning Vicept, which do not depend on this purified image training dataset.

REFERENCES

[1] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proc. IEEE CVPR*, 2008, pp. 1–8.

[2] A. M. Bronstein and M. M. Bronstein, "Spatially-sensitive affine-invariant image descriptors," in *Proc. ECCV*, 2010, pp. 197–208.

[3] J. Mairal, M. Leordeanu, F. Bach, M. Hebert, and J. Ponce, "Discriminative sparse image models for class-specific edge detection and image interpretation," in *Proc. ECCV*, 2008, pp. 43–56.

[4] F. Moosmann, E. Nowak, and F. Jurie, "Randomized clustering forests for image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1632–1646, Sep. 2008.

[5] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE CVPR*, 2009, pp. 248–255.

[6] M. Guillaumin, J. Verbeek, and C. Schmid, "Multiple instance metric learning from automatically labeled bags of faces," in *Proc. ECCV*, 2010, pp. 793–806.

[7] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 394–410, Mar. 2007.

[8] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *Proc. IEEE ICCV*, 2005, pp. 604–610.

[9] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large dataset for nonparametric object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1958–1970, Nov. 2008.

[10] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE CVPR*, 2010, pp. 3304–3311.

[11] A. Torralba, R. Fergus, and Y. Weiss, "Small codes and large databases for recognition," in *Proc. IEEE CVPR*, 2008, pp. 1–8.

[12] J. van Gemert, C. Snoek, C. Veenman, A. Smeulders, and J. Geusebroek, "Comparing compact codebooks for visual categorization," *Comput. Vis. Image Understand.*, vol. 114, no. 4, pp. 450–462, 2010.

[13] H. Cai, F. Yan, and K. Mikolajczyk, "Learning weights for codebook in image classification and retrieval," in *Proc. IEEE CVPR*, 2010, pp. 2320–2327.

[14] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *IJCV*, vol. 87, no. 3, pp. 316–336, Feb. 2010.

[15] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE CVPR*, 2010, pp. 3360–3367.

[16] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier, "Large-scale image retrieval with compressed Fisher vectors," in *Proc. IEEE CVPR*, 2010, pp. 3384–3391.

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

[18] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE ICCV*, 2003, pp. 1470–1477.

[19] F. Perronnin and C. R. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE CVPR*, 2007, pp. 1–8.

[20] 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 CVPR*, 2008, pp. 1–8.

[21] T. Tuytelaars and C. Schmid, "Vector quantizing feature space with a regular lattice," in *Proc. IEEE ICCV*, 2007, pp. 1–8.

[22] J. van Gemert, C. Veenman, A. Smeulders, and J. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Jul. 2010.

[23] H. Lee, A. Battle, R. Raina, and A. Ng, "Efficient sparse coding algorithms," in *Proc. NIPS*, 2007, pp. 801–808.

[24] 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. IEEE ICCV*, 2007, pp. 1–8.

[25] M. Schultz and T. Joachims, "Learning a distance metric from relative comparisons," in *Proc. NIPS*, 2003.

[26] Z. Wang, Y. Hu, and L. Chia, "Image-to-class distance metric learning for image classification," in *Proc. ECCV*, 2010, pp. 706–719.

[27] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE CVPR*, 2006, pp. 2161–2168.

[28] G. Qi, C. Aggarwal, Y. Rui, Q. Tian, S. Chang, and T. Huang, "Towards cross-category knowledge propagation for learning visual concepts," in *Proc. IEEE CVPR*, 2011, pp. 897–904.

[29] M. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *Proc. IEEE CVPR*, 2011, pp. 1745–1752.

[30] S. Bucak, R. Jin, and A. Jain, "Multi-label multiple kernel learning by stochastic approximation: Application to visual object recognition," in *Proc. NIPS*, 2010, pp. 1145–1154.

[31] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, Jan. 2010.

[32] D. G. Lowe, "Distinctive image features from scale invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[33] K. Weinberger and L. Saul, "Fast solvers and efficient implementations for distance metric learning," in *Proc. ICML*, 2008, pp. 1160–1167.

[34] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE CVPR*, 2009, pp. 1794–1801.

[35] J. Duchi and Y. Singer, "Boosting with structural sparsity," in *Proc. ICML*, 2009, pp. 297–304.

[36] S. Negahban and M. Wainwright, "Phase transitions for high-dimensional joint support recovery," in *Proc. NIPS*, 2008, pp. 1161–1168.

[37] G. Obozinski, B. Taskar, and M. Jordan, Joint Covariate Selection for Grouped Classification University of California, Berkeley, Tech. Rep. 743, 2007.

[38] Z. Wu, S. Jiang, L. Li, P. Cui, Q. Huang, and W. Gao, "Vicept: Link visual features to concepts for large-scale image understanding," in *Proc. ACM Multimedia*, 2010, pp. 711–714.

[39] R. Duda, P. Hart, and D. Stork, Pattern Classification, Citeseer.

[40] L. Li, S. Jiang, and Q. Huang, "Learning image Vicept description via mixed-norm regularization for large scale semantic image search," in *Proc. IEEE CVPR*, 2011, pp. 825–832.

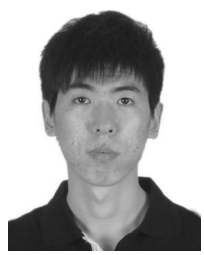
[41] P. Duygulu, K. Barnard, J. Freitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Proc. ECCV*, 2002, pp. 97–122.

[42] D. Liu, X. Hua, M. Wang, and H. Zhang, "Image retagging," in *Proc. ACM Multimedia*, 2010, pp. 491–500.

[43] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.* vol. 2, pp. 27:1–27:27, 2011 [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm (software).

[44] J. Tang, S. Yan, R. Hong, G. Qi, and T. Chua, "Inferring semantic concepts from community-contributed image and noisy tags," in *Proc. ACM Multimedia*, 2009, pp. 223–232.

[45] Y. Jing and S. Baluja, "Visualrank: Applying page-rank to large-scale image search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1877–1890, Nov. 2008.



Liang Li (S'10) is currently pursuing the Ph.D. degree at the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China.

His research interests include image processing, large-scale image retrieval, image semantic understanding, multimedia content analysis, computer vision, and pattern recognition.

Shuqiang Jiang (SM'08) received the M.S. degree from the College of Information Science and Engineering, Shandong University of Science and Technology, Shandong, China, in 2000, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2005.

He is currently an Associate Professor with the Institute of Computing Technology, Chinese Academy of Sciences. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include multimedia processing and semantic understanding, pattern recognition, and computer vision. He has authored or coauthored more than 90 papers on the related research topics.

Qingming Huang (SM'08) received the B.S. degree in computer science and the Ph.D. degree in computer engineering from Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively.

He is currently a Professor with the Graduate University of the Chinese Academy of Sciences (CAS), Beijing, China, and an Adjunct Research Professor with the Institute of Computing Technology, CAS. He has authored or coauthored nearly 200 academic papers in prestigious international journals and conferences. His research areas include multimedia video analysis, video adaptation, image processing, computer vision, and pattern recognition.

Dr. Huang is a reviewer for the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON COMMUNICATIONS. He has served as program chair, track chair, and TPC member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, PSIVT, etc.