Learning Image Vicept Description via Mixed-Norm Regularization for Large Scale Semantic Image Search

Liang Li, Shuqiang Jiang
Key Lab of Intell. Info. Process.,
Inst. of Comput. Tech., CAS, China
{lli, sqjiang}@jdl.ac.cn

Qingming Huang
Graduate University
Chinese Academy of Sciences, China
qmhuang@jdl.ac.cn

Abstract

The paradox of visual polysemia and concept polymorphism has been a great challenge in the large scale semantic image search. To address this problem, our paper proposes a new method to generate image Vicept representation. Vicept characterizes the membership distribution between elementary visual appearances and semantic concepts, and forms a hierarchical representation of image semantic from local to global. To obtain discriminative Vicept descriptions with structural sparsity, we adopt mixed-norm regularization in the optimization problem for learning the concept membership distribution of visual word. Furthermore, considering the structure of BOV in images, visual descriptor is encoded as a weighted sum of dictionary elements using group sparse coding, which could obtain sparse representation at the image level. The wide applications of Vicept are validated in our experiments, including large scale semantic image search, image annotation, and semantic image reranking.

1. Introduction

Large scale semantic image analysis becomes a hot research topic recently for its wide applications in image search and mining, while the paradox between Visual Polysemia and Concept Polymorphism (VPCP) is still a great challenge in this area as illustrated in Fig. 1. Visual polysemia reveals that a certain visual appearance may have different semantic explanations. In Fig. 1-(a), we notice that the visual appearance v is shared by the elements in concept collection C={*leopard, clothes, shoes, bags, glove, belt, mouse*}. Without other context information, it is hard to assign v into one certain concept. From another point of view, concept polymorphism represents the fact that one concept may have many visual appearances under different instances. Fig. 1-(b) gives a tangible example of concept polymorphism for "*Skyscraper*". Even though the influence

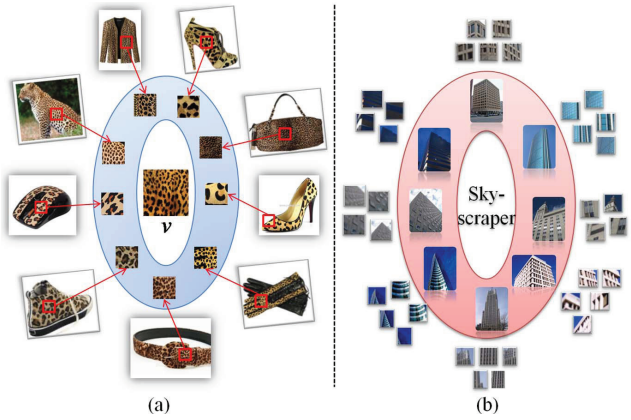

Figure 1. (a) Visual polysemia and (b) Concept polymorphism

of VPCP is slight in small image dataset, it is significant under large scale scenarios: on the side of VP, one visual appearance may correspond with thousands of concepts so that it is extremely difficult to infer its exact concept property; on the side of CP, one concept usually has hundreds of various instances and most of them have diverse visual appearances. In a word, there exists a potential connection between any visual appearance and any concept in large scale environment.

Though many significant works have been proposed to address the problem of large scale semantic image search, none of them solve the paradox of VPCP directly. The de-facto standard in these works is the *bag-of-visual-words* (BOV) model [27]. The BOV approach regards an image as a collection of visual appearance descriptors extracted from local patches, quantizes them into discrete "visual words", and then computes a compact histogram representation for semantic image search. BOV model has been extensively investigated for the following reasons: 1) Visual word is discriminative due to considering local salient and invariant information; 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 im-

ages; 3) The similarity between images can be measured quickly through simple operation, such as, dot-product.

The BOV approach has been refined in a number of ways: 1) *Local descriptor aggregation techniques*, such as DAISY configuration based descriptor learning [36], spatial-sensitive and affine-invariant image descriptor construction [2], and Fisher kernel representation based descriptor aggregating [11, 22, 23]; 2) *Discriminative codebook generation methods* [13, 17, 18, 30, 31, 32], such as scalable acceptance-radius based codebook clustering [13], online codebook learning scheme for sparse coding [17]; 3) *Efficient quantization techniques*, such as hierarchical k-means (HKM) [14] or approximate k-means (AKM) [24], hamming Embedding technique to provide binary signatures for visual word matching refinement [9] and small code to compress the BOV [29, 30, 33]. Recently, sparse coding is proposed to determine a small set of codeword from the dictionary to efficiently represent visual descriptors [14, 16, 17, 33]. Gemert et al. [32] 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*, such as query expansion [4] and geometrical re-ranking [10, 24, 25] etc.

Although above techniques can improve the search performance to some extent, the problem of VPCP has still been pendent. Recently, metric learning has received a lot of attention for face recognition [8], image annotation [3, 26] and classification [1, 34, 35]. Labeled pairs of bags are used for multiple instance logistic discriminant metric learning in [8]. A weighted similarity metric based on largest margin is learned by [3]. [1, 34] suggest an Image-To-Class (I2C) distance metric learning method by learning per-class Mahalanobis metrics. These techniques provide new insights on dealing with the paradox of VPCP; however, these works are in the primary 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 search. *Vicept* is introduced to characterize the membership distribution between each visual word and concepts. *Vicept* forms a hierarchical representation of image semantic from local to global and thus can directly deal with VPCP paradox. We adopt the idea of mixed-norm regularization in our optimization problem for learning the membership distribution, which is effective for obtaining a discriminative *Vicept* with structural sparsity. Furthermore, considering the structure of BOV in images, each visual descriptor is encoded as a weighted sum of dictionary elements using group sparse coding, which is possible to obtain sparse representation at image level.

The generating procedure of *Vicept* is illustrated in Fig. 2. First, a large image training dataset with concept labels is established based on a concept hierarchy covering frequently used concepts in the daily life. Second, SIFT descriptors are extracted from these images and an initial dictionary is obtained by clustering these descriptors with hierarchical k-means. Third, 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. Fourth, *Vicept* is obtained by learning with the mixed-norm regularization from above spare representations, and a *Vicept* description with a multi-level concept membership contribution is built based on a hierarchical concept structure [5]. The details will be discussed in Section 3. Finally, for one image, we can compute its global semantic description via vector product between *Vicept* and its group sparse code representation for further large scale applications.

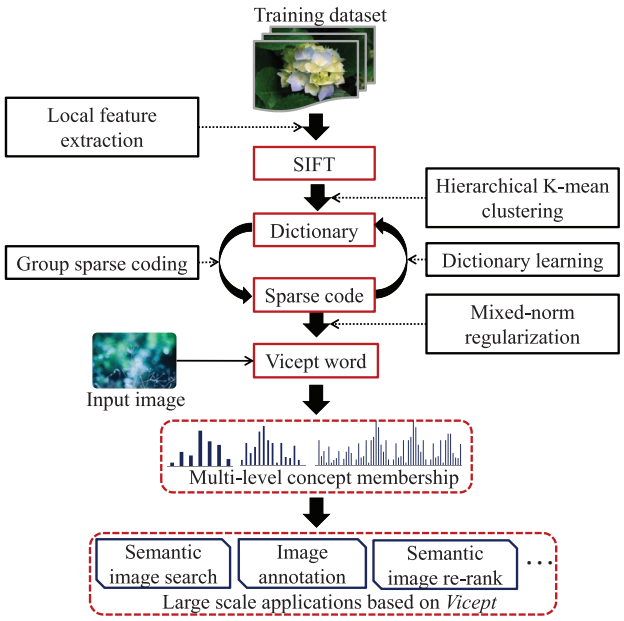

Figure 2. The proposed framework for Vicept generation and applications

For one given image, the proposed scheme quickly computes its semantic information without depending on specified training model so that it can be simply used for large scale applications. Experiments on large scale semantic image search tasks show strong semantic descriptive power of *Vicept*. Furthermore, in semantic image re-ranking and image annotation tasks, our method also shows promising performances.

The contributions of our work are summarized as follows:

1. The paradox between Visual Polysemia and Concept Polymorphism (VPCP) is discussed. Taking the para-

dox into account, a new method for generating image *Vicept* description is proposed for large scale semantic image search.

2. The idea of mixed norm regularization is adopted in our optimization problem to learn a discriminative *Vicept* with structural sparsity.
3. Group sparse coding is utilized to encode the images based on the visual words and a discriminative and compact dictionary is learned relying on this sparse representation.
4. The wide applications of *Vicept* are validated, including large scale semantic image search, hierarchical image annotation, and semantic image re-ranking.

The rest of this paper is organized as follows: Section 2 introduces the methods of visual appearance encoding. Section 3 details the learning procedure of *Vicept* under the mixed norm regularization. Section 4 presents experimental results on standard benchmarks and a large scale image database, showing the effective performance of our approach. Finally, Section 5 concludes the paper.

2. Visual Appearance Encoding

2.1. 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 local visual descriptors in a *P*-dimensional feature space, such as SIFT, i.e. $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M]^T \in \Re^{M \times P}$. The VQ method applies the k-means clustering algorithm to minimize the construction error:

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

where $\boldsymbol{D} = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_K]^T$ includes *K* cluster centers to be found, 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 $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M]^T$,

$$\min_{A,D} \sum_{m=1}^{M} \min_{k=1\ldots K} \|\boldsymbol{x}_m - \boldsymbol{a}_m \boldsymbol{D}\|^2 \tag{2}$$

$$\text{subject to } \|\boldsymbol{a}_m\|_0 = 1, \|\boldsymbol{a}_m\|_1 = 1, \boldsymbol{a}_m > 0, \forall m$$

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) [14, 16, 17, 38]:

$$\min_{A,D} \sum_{m=1}^{M} \|\boldsymbol{x}_m - \boldsymbol{a}_m \boldsymbol{D}\|^2 + \lambda \|\boldsymbol{a}_m\|_1 \tag{3}$$

$$\text{subject to } \boldsymbol{a}_m > 0, \forall m$$

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.

2.2. 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 application, 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 [6, 19, 21]:

$$\min_{A,D} \frac{1}{2} \sum_{m=1}^{M} \|\boldsymbol{x}_m - \boldsymbol{a}_m \cdot \boldsymbol{D}\|^2 + \lambda \sum_{k=1}^{K} \|\boldsymbol{a}_k\| \tag{4}$$

$$\text{subject to } \boldsymbol{a}_k^m \geq 0, \forall k, m$$

where $\boldsymbol{a}_k = (a_k^1, \ldots, a_k^M)$ and $\boldsymbol{a}^i = (a_1^i, \ldots, a_K^i)$ are nonnegative vectors and $\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K\} = \{\boldsymbol{a}^1; \ldots; \boldsymbol{a}^M\}^T$ is the reconstruction matrix, *M* is the total number of visual descriptors in the image.

Using the sparse coding with the norm ℓ_1/ℓ_2 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. First, a visual descriptor set extracted from a large image collection is used to solve Eq. 4 with respect to $\boldsymbol{A}$ and $\boldsymbol{D}$, where $\boldsymbol{D}$ is retained as the dictionary. In the coding phase, for each image represented as a descriptor set $\mathbf{X}$, the GSC code A is obtained by optimizing Eq. 4 with respect to $\boldsymbol{D}$ only. In the dictionary learning phase, the

optimization Eq. 4 is solved with respect to both $\boldsymbol{A}$ and $\boldsymbol{D}$. After several alternations between these two phases, we can obtain a discriminative and compact dictionary.

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.

3. 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 paradox into account. In this section, we first formulate the problem, and then introduce our approach for generating the *Vicept* and constructing a *Vicept* description with hierarchical semantic concepts.

3.1. Image Vicept Description

The observation of VPCP paradox motivates us that the relationship between concept collection and visual appearance set can be formalized as a bipartite graph. 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 [15] is detected and quantized into visual words [27] by group sparse coding.
2. *Semantic Concept Collection*: The concepts in real world are not independent but closely related. Following the structure in [5, 37], 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 Section 4.1.

Before learning the *Vicept*, a short interpretation is presented as follows. Suppose having a dictionary $\boldsymbol{D}$ with K visual words and a concept collection $\boldsymbol{C}$ with N concepts, a membership distribution can be learned between each visual word and concept collection. In a word, each visual word has a corresponding N-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 N-bin membership distribution histogram. Finally, we can obtain a *Vicept Dictionary* according to the dictionary $\boldsymbol{D}$.

3.2. Learning Vicept Word via Mixed Norm Regularization

Let $\boldsymbol{X} = \{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^N\}$ 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 $y_i \in [0, 1]$ the possibility that i-th 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 $\boldsymbol{D}$. We can learn the discriminative *Vicept* with structural sparsity by adopting the idea of mixed-norm regularization in the following objective optimization:

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

where $a_j^{i,m}$ is the reconstruction coefficient of the j-th visual word for the m-th descriptor of the i-th image, and the non-negative vector $\boldsymbol{u}_j = (u_{j,1}, \ldots, u_{j,M})$ indicates the relationship of j-th visual word with concept collection. The first term of the objective measures the reconstruction quality and the second term measures the reconstruction complexity. The parameter γ balances the effect of these two terms.

The problem of Eq. 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:

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

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 $u_{r,j}$ 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:

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

$$\Rightarrow u_{r,x} \in \frac{w_x - [0, \gamma]}{\sum_{i=1}^{N} (s_r^i)^2} \tag{6}$$

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 $u_{r,x} = (w_x - \gamma) / \sum_{i=1}^{N} (s_r^2)$.

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

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

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

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

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

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

which infers that:

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

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

3.3. Building Vicept with Multi-level Concept Membership Distribution

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

As we cannot cover all the concepts in real world, the imbalance of the concept selection restricts the performance for high-level *Vicept* generation. In this case, rather than simply summing the bins belonging to the same high-level concept, we assign large weights to the bins with large bin value. We implement this by adopting sigmoid function and normalize the sum of high-level histogram. The weight $w(i)$ for i-th bin is calculated by the value of itself:

$$w(i) = 1/(1 + \exp(v(i))), \; v(i) \in [0, 1] \tag{11}$$

Finally, we obtain a complete *Vicept Dictionary*, where each *Vicept* has a visual word with multi-level concept membership distribution histograms.

4. 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 reranking.

4.1. Database and Experimental Settings

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 267k 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. Because each image in ImageNet has only one label, we try to "purify" the dataset by manually segment the image and eliminate the irrelevant areas. Balancing the workload for image matting and the data requirement in this task, we prepare 120 "purified" images in each concept. This subset contains 120×217 one-concept-labeled images (later referred as ImageNet25K). Another standard benchmark (Corel5K) is used in our experiment, which consists of 5000 images. Furthermore, we used an additional set of 800k 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 is obtained from the hierarchical k-means cluster. A new dictionary with 473 visual words is generated through group sparse coding, which is detailed in Section 2.2.

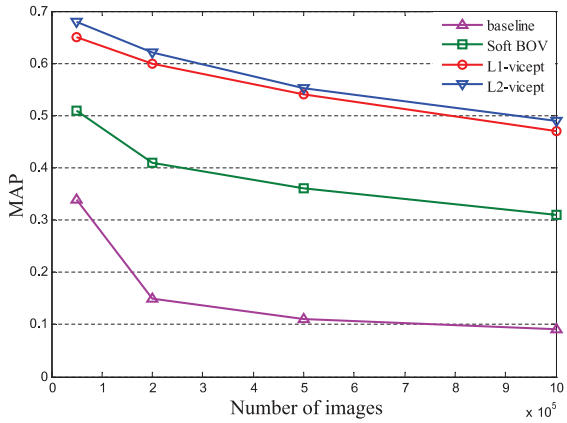

Figure 3. Comparisons of different methods using MAP with different scale of image dataset

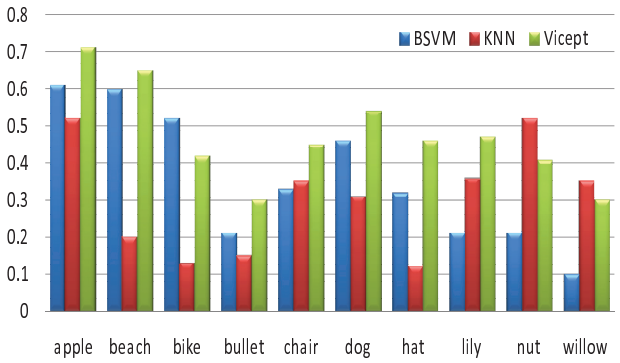

Figure 4. Comparisons of different annotation methods with AP over the ImageNet267K database: "BSVM" and "KNN" is the baseline approach; "Vicept" is the approach based the *Vicept*.

4.2. Large Scale Semantic Image Search

Incorporating with the BOV method, *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 ImageNet267K and Flickr800K.

Baseline: We use a traditional BOV approach [20] as the "baseline" approach and a dictionary of 200K visual words is used. We experimented with different size of visual word dictionary, and found 200K dictionary to give the best performance.

Comparisons: we also enhance the baseline method with soft assignment [25], where the number of nearest neighbors is set to be 4. We call this method "Soft BOV". Our *Vicept* based approach has two variants: 1) "L1-Vicept", in which we set p=1 as the norm penalty in Eq. 5 and γ is 0.08. 2) "L2-Vicept", in which p is set to be 2 in Eq. 5 and γ is 0.01.

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. 3 compares the above four approaches with MAP, leading to three observations. First, our *Vicept* significantly improves the MAP, as can be seen by comparing the results with "baseline". On the 1M image dataset, the methods based on the *Vicept* boost the MAP from 0.09 to 0.48, a 37% improvement. Second, soft assignment of visual words plays an important role in improving the performance (a 20% improvement on average). This point is also demonstrated in [32]. Three, the "L2-Vicept" method reaches a higher MAP than the "L1-Vicept". One main reason is that the *Vicept* learned via ℓ_1/ℓ_2 norm regularization is sparser, which allows the representation to be special and to capture salient properties of relations between visual words and semantic concepts.

4.3. Image Annotation

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

4.3.1 Image Annotation on ImageNet267K

Two basic approaches are implemented as baseline for Image Annotation, (1) Binary SVM; (2) KNN based voting. 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. For the KNN, we replicate the experiment described in [28]. Firstly, the query image and the images in 267K 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.

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. 4 illustrates the average precision for three approaches. We can find that *Vicept* provides a better concept annotation result than BSVM and KNN for most of the query images. Besides, the fluctuations on concept "nut" and "willow" are likely to be influenced by small number of training data for *Vicept* learning. In a way, the 47.1% mean AP of *Vicept* seems to be satisfactory in this annotation task.

4.3.2 Image Annotation on Corel5K

Corel5K [7] has become the benchmark for image annotation, which contains 5000 images with 260 concepts. We

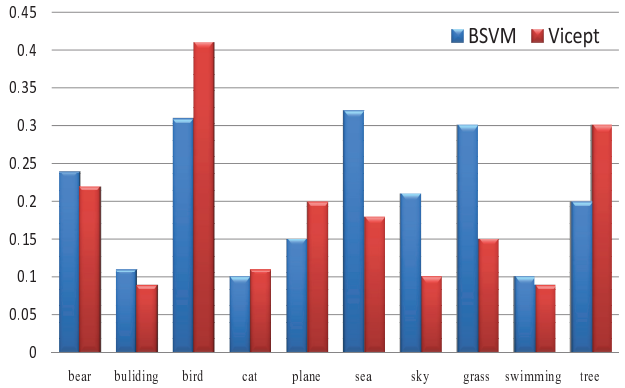

Figure 5. Comparisons of two annotation methods on Corel5K dataset

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; Similar to above procedure, we train 260 binary SVM classifiers with an output of classification probability. In the training stage, we split the Corel5K into 4500 training and 500 test examples.

Average Precision (AP) is used as the evaluation metric. To have a fair evaluation, we select 10 keywords concepts, which are included by the concept 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 with the "BSVM". The result in Fig. 5 shows that *Vicept* has potential to be used without relying on any outside information.

4.4. 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 [12], 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 (nDCG@k) is adopted as the evaluation metric. nDCG is a normalized version of DCG metric. Two assumptions of DCG metric are that highly relevant results are more useful when appearing earlier in a result list and that highly relevant results are more useful than marginally relevant ones, which are in turn more useful than irrelevant results. nDCG@k is calculated by,

$$\text{nDCG@}k = \frac{1}{Z}\sum_{n=1}^{k}\frac{2^{s(p)}-1}{\log(1+p)} \tag{12}$$

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.

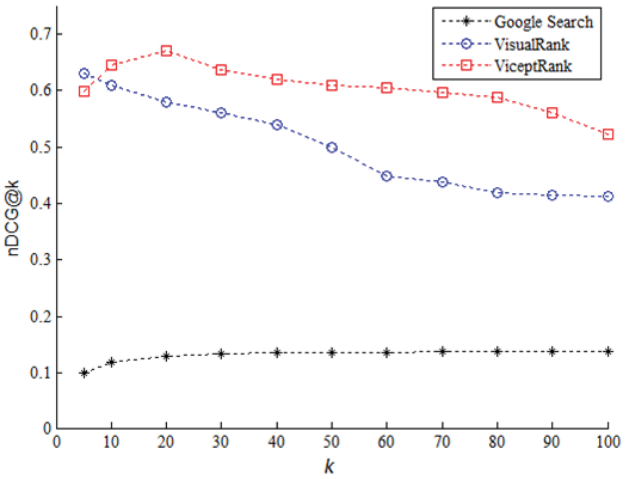

Figure 6. The performance of semantic image re-ranking using nDCG@k

Fig. 6 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 bases upon the concept membership distribution.

5. 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" (concepts). 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. *Vicept* approach provides fast computation, compact expression, and local-to-global description, and thus can be implemented for large scale web applications. In the future, we will focus on learning more powerful *Vicept* descriptions based on multiple features with a web-scale image dataset and concept collection.

Acknowledgment

We thank the anonymous reviewers for their valuable comments. This research was supported in part by National Natural Science Foundation of China: 61025011, 60833006, 61035001 and 61070108, and in part by National Basic Research Program of China (973 Program): 2009CB320906.

References

[1] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *CVPR*, 2008.

[2] A. M. Bronstein and M. M. Bronstein. Spatially-sensitive affine-invariant image descriptors. In *ECCV*, 2010.

[3] H. Cai, F. Yan, and K. Mikolajczyk. Learning weights for codebook in image classification and retrieval. In *CVPR*, 2010.

[4] 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 *ICCV*, 2007.

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

[6] J. Duchi and Y. Singer. Boosting with structural sparsity. In *ICML*, 2009.

[7] P. Duygulu, K. Barnard, J. Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002.

[8] M. Guillaumin, J. Verbeek, and C. Schmid. Multiple instance metric learning from automatically labeled bags of faces. In *ECCV*, 2010.

[9] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, 2008.

[10] H. Jégou, M. Douze, and C. Schmid. Improving bag-of-features for large scale image search. *IJCV*, 2010.

[11] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *CVPR*, 2010.

[12] Y. Jing and S. Baluja. Visualrank: applying page-rank to large-scale image search. *PAMI*, 2008.

[13] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *ICCV*, 2005.

[14] H. Lee, A. Battle, R. Raina, and A. Ng. Efficient sparse coding algorithms. In *NIPS*, 2007.

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

[16] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *ICML*, 2009.

[17] J. Mairal, M. Leordeanu, F. Bach, M. Hebert, and J. Ponce. Discriminative sparse image models for class-specific edge detection and image interpretation. In *ECCV*, 2008.

[18] F. Moosmann, E. Nowak, and F. Jurie. Randomized clustering forests for image classification. *PAMI*, 2008.

[19] S. Negahban and M. Wainwright. Phase transitions for high-dimensional joint support recovery. In *NIPS*, 2008.

[20] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006.

[21] G. Obozinski, B. Taskar, and M. Jordan. Joint covariate selection for grouped classification. *Technical Report 743, University of California Berkeley*, 2007.

[22] F. Perronnin and C. R. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

[23] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *CVPR*, 2010.

[24] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[25] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 2008.

[26] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In *NIPS*, 2004.

[27] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[28] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: a large dataset for nonparametric object and scene recognition. In *PAMI*, 2008.

[29] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large databases for recognition. In *CVPR*, 2008.

[30] T. Tuytelaars and C. Schmid. Vector quantizing feature space with a regular lattice. In *ICCV*, 2007.

[31] J. van Gemert, J. Geusebroek, C. Veenman, and A. Smeulders. Kernel codebooks for scene categorization. In *ECCV*, 2008.

[32] J. van Gemert, C. Veenman, A. Smeulders, and J. Geusebroek. Visual word ambiguity. *PAMI*, 2010.

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

[34] Z. Wang, Y. Hu, and L. Chia. Image-to-class distance metric learning for image classification. In *ECCV*, 2010.

[35] K. Weinberger and L. Saul. Fast solvers and efficient implementations for distance metric learning. In *ICML*, 2008.

[36] S. Winder, G. Hua, and M. Brown. Picking the best daisy. In *CVPR*, 2009.

[37] 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 *ACM MM*, 2010.

[38] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.