Neurocomputing 174 (2016) 384–392

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Distributed image understanding with semantic dictionary and semantic expansion

Liang Li [a,1], Chenggang Clarence Yan [b,1], Xing Chen [c,d], Chunjie Zhang [a,*], Jian Yin [f], Baochen Jiang [f], Qingming Huang [a,e]

[a] Key Lab of Big Data Mining and Knowledge Management, University of Chinese Academy of Sciences, Beijing, China
[b] Department of Automation, Tsinghua University, Beijing, China
[c] Academy of Mathematics and Systems Science, Chinese Academy of Sciences (CAS), China
[d] National Center for Mathematics and Interdisciplinary Sciences, CAS, Beijing, China
[e] Key Lab of Intelligent Information Processing, Institute of Computing Technology, CAS, Beijing, China
[f] Department of Computer, Shandong University, Weihai, China

ARTICLE INFO

Article history:
Received 9 December 2014
Received in revised form
10 April 2015
Accepted 14 April 2015
Available online 1 September 2015

Keywords:
Image understanding
Semantic dictionary
Multi-task learning
Semantic expansion
Distributed systems

ABSTRACT

Web-scale image understanding is drawing more and more attention from the computer vision and multimedia domain. To solve the key problem of visual polysemia and concept polymorphism in the image understanding, this paper proposes a semantic dictionary to describe the images on the level of semantic. The semantic dictionary characterizes the probability distribution between visual appearances and semantic concepts, and the learning procedure of semantic dictionary is formulated into a minimization optimization problem. Mixed-norm regularization is adopted to solve the above optimization for learning the concept membership distribution of visual appearance. Furthermore, to improve the generalization ability of the semantic description, we propose the semantic expansion technology, where a concept transferring matrix is learnt to quantize the implicit relevancy among the concepts. Finally, the distributed framework on the basis of the semantic dictionary is constructed to speed up the large scale image understanding. The semantic dictionary is validated in the tasks of large scale semantic image search and image annotation.

1. Introduction

With the rapid development of mobile internet and multimedia technology, recently web scale image understanding becomes a hot research topic due to its wide applications in our daily life. However the phenomenon of Visual Polysemia and Concept Polymorphism (VPCP) has still been a great challenge in the image understanding. Visual polysemia depicts the fact that one certain visual appearance may have different semantic explanations, and concept polymorphism indicates the truth that one concept may have many visual appearances under the different examples. Particularly in web scale conditions, there exist the more complex connections between visual appearances and semantic concepts so that the VPCP problem becomes graver: on one hand of VP, one visual appearance may occur in thousands of web concepts so that it is extremely difficult to infer its exact semantic; on the other hand of CP, one concept has various instances, where there are diverse visual appearances. In one word, the VPCP problem is a complex and challenging issue in the large scale environment.

To solve the above problem, researchers mainly proposed their approaches at the perspective of the multimedia and computer vision, including image classification [1,2], image annotation [3,4], object and scene recognition [5], and image search [6], etc. In detail, Weinberger et al. [2] proposed the large margin nearest neighbor classification on the basis of distance metric learning model. Boiman et al. [1] introduced an Image-To-Class distance metric learning method for image classification by learning per-class Mahalanobis metric. Qi et al. [3] studied a technology for cross-category transfer learning for the classification task. Bucak et al. [5] introduced an algorithm for multi-label multiple kernel learning to recognize the objects. But none of them solve the problem of VPCP directly, either the VP problem or the CP problem. One main reason is that the relationship between image visual appearances and semantic information has not been individualized.

In the last few years, machine learning and distributed computing [7–16] are widely used in many hot domains, such as economics, biology, computer science philosophy, and big data.

* Corresponding author.
[1] L. Li and X. Chen have contributions.

http://dx.doi.org/10.1016/j.neucom.2015.04.108
0925-2312/

Inspired by the sparse learning and multi-task learning model, we learn the semantic dictionary (Fig. 1) to solve the problem of visual polysemia and concept polymorphism. In the viewpoint of the mathematics, this dictionary is a matrix where each column depicts the relationship between one concept with all the visual appearances while each row represents the relationship between one visual appearance with all the concepts. Further, inspired by the distributed computing theory, we design the distributed framework of semantic dictionary for the web-scale image understanding.

In summary, this paper learns the semantic dictionary for web-scale image understanding, which is to characterize the membership distribution between each appearance of the visual set and each word of the concept set. With its help, the images can be represented into a description of the intuitive semantic, rather than the incomprehensible visual information. To learn the robust semantic dictionary, we introduce a mixed-norm regularization optimization algorithm to formulate the learning procedure, where the common visual patterns shared by the related concepts are learnt. The convergence guarantees the semantic dictionary to achieve the approximate global optimal solution. Furthermore, different from the visual description approach, where all the bins of the descriptor are independent, the semantic description has the implicit relevancy within it. Taking such a relevancy into account, we propose the semantic expansion technology to transfer the weights between related concepts and a concept transferring matrix is learned to expand the power of image semantic description. Finally, to speeding up the web-scale image understanding, we propose a distributed framework to integrate the semantic dictionary.

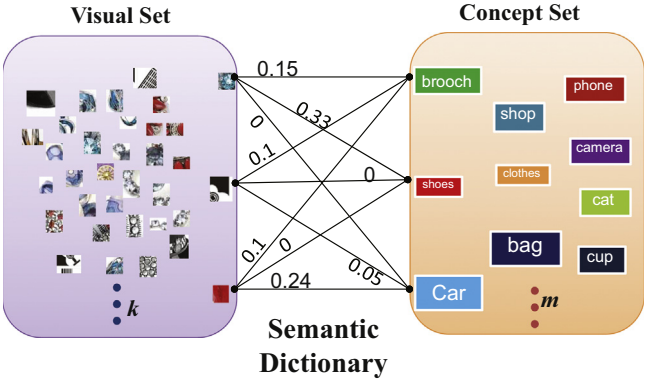

Fig. 1. The semantic dictionary built between visual set and concept set.

Fig. 2 shows the flowchart of our proposed scheme for image semantic representation. Our method can be divided into two independent learning procedure: one is the semantic dictionary learning (detailed in Section 3), and the other is the concept transferring matrix learning (detailed in Section 4). For the distributed system about the semantic dictionary, each visual appearance is treated as one node, which reveals the corresponding probability distribution of all the concepts. For the distributed system about the transferring matrix, each concept is treated as one node, which stores the relevancy with other concepts. Given an image, firstly, based on the Bag-of-visual-words model, the image visual representation is extracted. Secondly, each visual appearance is dispensed into the corresponding node of the distributed system, and its corresponding semantic representation can be obtained. The same operation is applied for the other visual appearance, and all the outputs of the nodes are summed up so that the original image semantic representation is obtained with the probability distribution about each concept. Thirdly, the probability of each concept is delivered into another distributed system, where each node is about the relevancy between one concept and other concepts in the practical environment. All the outputs of the nodes are integrated into the final image semantic representation, which can directly be used into the further applications, such as semantic image search, image annotation and so on.

The rest of this paper is organized as follows: Section 2 introduces the visual appearance representation method, and interprets the semantic dictionary. Section 3 details the learning procedure of semantic dictionary and represents the semantic description. Section 4 models the semantic expansion technology and learns the concept transferring matrix, which measures the perception relevancy of different concepts. Section 5 represents the distance metric based on semantic dictionary and concept transferring matrix. Section 6 shows the experimental results of different tasks on both the standard benchmark and the large scale image database. Finally, Section 7 concludes the ideas of this paper.

2. Semantic dictionary

As mentioned above, semantic dictionary is designed to bridge the image visual appearances and the semantic concepts. With its

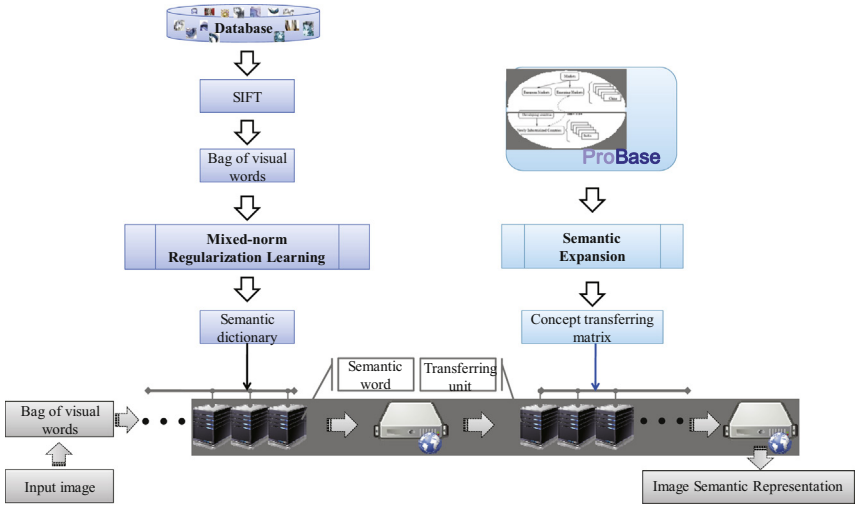

Fig. 2. The flowchart of our proposed scheme for image semantic representation.

help, the images can be represented into a description of the semantic, rather than the visual information. In this section, firstly, we introduce the popular image visual representation *bag-of-visual-words* (BOV) model [17], and then represent an intuitive interpretation about semantic dictionary, such as its structure and concept set source.

2.1. Image visual representation model

The de facto image visual representation method in the multimedia domain is based on the BOV model [17], which is motivated by the *bag-of-words* from the information retrieval domain. In this model, an image is characterized as a collection of visual appearance descriptors, which are extracted from local patches and quantized into discrete visual words, and then a compact histogram representation is computed for farther image applications.

The classical BOV model is based on the k-means algorithm. Let $\mathbf{X}=[\mathbf{x}1;\ldots;\mathbf{x}n]\in\Re n\times d$, depicts a collection of n visual descriptors, and each descriptor is a feature vector with d-dimensionality, e.g. the most popular local descriptor-SIFT [18]. The k-means algorithm is to minimize the reconstruction error:

$$\min_{B}\left(\sum_{i=1}^{n}\min_{j=1,\ldots,k}\|\mathbf{x}i-\mathbf{b}j\|2\right) \tag{1}$$

where $\mathbf{B}=[\mathbf{b}1;\ldots;\mathbf{b}k]\in\Re k\times d$ is called *dictionary*, where there are k clustering centers and every cluster center is a visual word. $\|\cdot\|$ is the ℓ_2-norm. The formulation of Eq. (1) can be transferred into a matrix factorization problem:

$$\min_{A,B}\sum_{i=1}^{n}\min\|\mathbf{x}i-\mathbf{a}i\mathbf{B}\|2 \tag{2}$$

$$\text{subject to}\quad \|\mathbf{a}i\|0=1,\quad \|\mathbf{a}i\|1=1,\quad \mathbf{a}i\geq 0,\quad \forall i$$

where $[\mathbf{a}1;\ldots;\mathbf{a}n]\in\Re n\times k$ is the cluster membership indictors. $\|\mathbf{a}i\|0=1$ is a cardinality constraint, meaning that only one element of $\mathbf{a}i$ is nonzero, $\mathbf{a}i>0$ means the nonnegative property of $\mathbf{a}i$, and $\|\mathbf{a}i\|1$ is the ℓ_1-norm, the sum of the absolute value of each element in $\mathbf{a}i$. After the optimization, the index of the only nonzero element in $\mathbf{a}i$ depicts the corresponding visual word of $\mathbf{x}i$.

The BOV model has become the most popular image representation model in the multimedia domain because of its advantages:

- As a result of the local salient and the invariant information of rotation and scale, the visual words are very discriminating.
- The BOV model can provide a compact and discriminative description with a collection of visual words, so that it is easy to be stored and searched.
- Based on the BOV, the distance among images can be computed quickly through some fast and simple operators, e.g. dot-product.

2.2. Semantic dictionary description

The semantic dictionary in Fig. 1 represents the between concept set and visual set, where k is the number of local visual appearances in the whole dictionary and m is the number of semantic concepts. The probability value in Fig. 1 denotes the degree of the relationship between the corresponding visual appearance and the corresponding concept. To efficiently make use of this structure, we detail the semantic dictionary:

1. *Visual set*: local visual descriptor is used to represent the image. In detail, SIFT [18], a robust local feature with invariant ration and scale, is located by Difference of Gaussian and consists of 128-dimensionality histogram. Finally, SIFT is quantized into visual words by hierarchical k-means algorithm.
2. *Concept set*: The concepts in real world are not independent but closely related. To provide the more discriminative semantic concepts and cover the wider objects and scenes, we follow the structure in [19], which is the most current image understanding database in the multimedia and computer vision domains. Here, we simplify the original concept structure with a flat representation.

Before learning the semantic dictionary, a short interpretation is represented as follows. Suppose having a visual dictionary **VD** with k visual words and a semantic concept collection **SC** with m concepts, a $k\times m$ membership distribution (semantic dictionary) can be jointly learned between each concept and each visual appearance. In other words, each concept in **SC** has the corresponding k-bin membership distribution histogram, and each visual appearance in **VD** has the corresponding m-bin membership distribution histogram.

3. Semantic dictionary learning with mixed-norm regularization learning

Semantic dictionary is to characterize the typical relationship between visual appearances and semantic concepts, where each column depicts the relationship between one concept with all the visual appearances while each row represents the relationship between one visual appearance with all the concepts. Here, we introduce the mixed-norm regularization learning algorithm to learn the semantic dictionary. Further, we impose the penalty term to restrict the number of common visual patterns in the related concepts, the objective function is defined as follows:

$$\Phi(\mathbf{X},\mathbf{Y},\mathbf{D})=\frac{1}{2}\sum_{i=1}^{N}\|\mathbf{y}i-\sum_{m=1}^{|xi|}|\sum_{j=1}^{K}x^{ij}{}_{m}\cdot\mathbf{d}j\|2+\lambda\sum_{j=1}^{K}\|\mathbf{d}j\|p \tag{3}$$

$$\text{subject to}\quad djk\geq 0,\quad \forall j,k$$

where $\mathbf{Y}=\{\mathbf{y}1,\ldots,\mathbf{y}N\}$ be the corresponding labels of images. $\mathbf{y}n=(y1,\ldots,yM)$ is relative to the whole concept collection with m concepts, and $yi\in[0,1]$ is the possibility that ith concept appears in the image. $x^{ij}{}_{m}$ is the description from the jth visual word for the mth descriptor of the ith image. $\mathbf{D}=\{\mathbf{u}1,\ldots,\mathbf{u}K\}$ denotes the *semantic dictionary* and K is the number of visual words in the image BOV representation. $\mathbf{d}j=(dj,1,\ldots,dj,M)$ indicates the relationship of jth visual word with semantic concept collection **SC**.

The first term in Eq. (3) measures the reconstruction quality. The mixed-norm regularization is the form of $\ell 1/\ell p$-norm, which is presented by the second term and measures the reconstruction complexity. It consists of two parts: one is ℓ_p-norm complexity of $\mathbf{d}j$, and the other is the ℓ_1-norm sum of $\mathbf{D}$. In the view of semantic dictionary, the mixed-norm regularization helps the dictionary to achieve the structural sparsity that all the images from the same concept have the similar sparse semantic representation, which is the essential difference with the ℓ_1 or ℓ_2 norm regularization. The parameter λ balances the effect of these two terms.

The problem of Eq. (3) can be solved by coordinate descent. Leaving all indices of $\mathbf{D}$ intact except for index t, omitting fixed argument of the objective, let φ be the term which does not rely on $\mathbf{d}t$ and $\sum^{m=1}{}^{|xi|}x^{ij}{}_{m}=s^{ij}$, we obtain the following reduced

objective function:

$$\Phi(\mathbf{d}t) = \frac{1}{2}\sum_{i=1}^{N} \| \mathbf{y}i - \sum_{j\neq t}^{K} s^{ji}\cdot \mathbf{d}j - s^{ti}\cdot \mathbf{d}t \|2 + \lambda \sum_{j=1}^{K} \| \mathbf{d}j \| p$$

$$= \sum_{i=1}^{N}\left(\sum_{j\neq t}^{K} s^{ji}\, s^{ti}\mathbf{d}j\cdot \mathbf{d}t - s^{ti}\mathbf{y}i\cdot \mathbf{d}t + \frac{1}{2}(s^{ti})2 \| \mathbf{d}t \|2 \right)$$

$$+\lambda \sum_{j=1}^{K} \| \mathbf{d}j \| p + \varphi$$

$$= \sum_{i=1}^{N}\left(\sum_{j\neq t}^{K} s^{ji}\, s^{ti} \sum_{x=1}^{M} \mathbf{d}t\chi \mathbf{d}j\chi - s^{ti} \sum_{x=1}^{M} \mathbf{y}^{xi}\mathbf{d}t\chi + \frac{1}{2}(s^{ti})2 \sum_{x=1}^{M} (\mathbf{d}t\chi)2 \right)$$

$$+\lambda \sum_{j=1}^{K} \| \mathbf{d}j \| p + \varphi$$

Next we show how to find the optimum $\mathbf{d}t$. Let $\check{\Phi}$ be the first reconstruction term of the objective, and its partial derivatives with respect with to each $dt\chi$ are

$$\frac{\partial}{\partial dt\chi}\check{\Phi} = \sum_{i=1}^{N}\left(\sum_{j\neq t}^{K} s^{ji}\, s^{ti} dj\chi - s^{ti}y^{\chi i} + (s^{ti})2 dt\chi \right)$$

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

$$wx = \left| -\sum_{i=1}^{N}\left(\sum_{j\neq t}^{K} s^{ji}\, s^{ti} dj\chi - s^{ri}y^{\chi i} \right) \right| +$$

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:

$$0 \in -wx + \sum_{i=1}^{N} (s^{ti})2 dt\chi + \lambda \underbrace{\frac{\partial}{\partial dt\chi} \| dt\chi \|}_{1} \in [0,1]$$

$$\Rightarrow dt\chi \in \frac{wx - [0,\lambda]}{\sum_{i=1}^{N} (s^{ti})2} \tag{4}$$

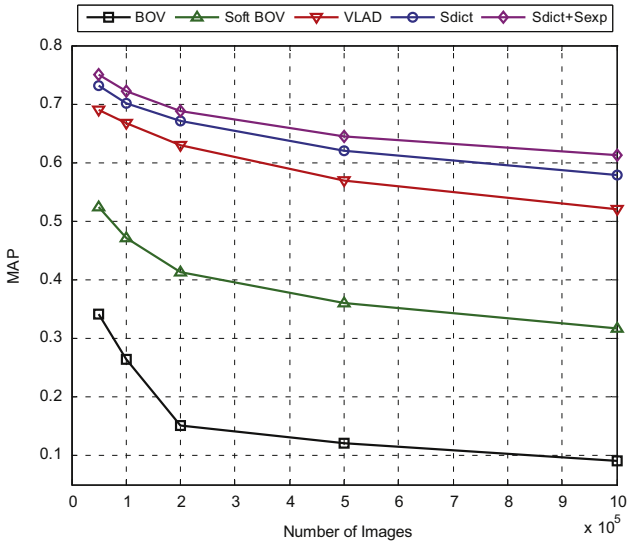

Fig. 3. Comparisons of different methods using MAP with different scale of image database.

Since $dt\chi \geq 0$, the above sub-gradient condition for optimality implies that $dt\chi = 0$ when $wx \leq \lambda$ and otherwise $dt\chi = (wx-\lambda)/\sum^{i=1N} (s^{ti})2$.

For $p=2$, indicating $\mathbf{w} = (w1,\ldots,wM)$, the gradient of $\Phi(\mathbf{d}t)$ with the ℓ_2-norm penalty is as follows:

$$\frac{\partial}{\partial \mathbf{d}t}\Phi = -\mathbf{w} + \sum_{i=1}^{N} (s^{ti})2\mathbf{d}t + \lambda \frac{\mathbf{d}t}{\| \mathbf{d}t \|} \tag{5}$$

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

$$\mathbf{d}t = \left(\sum_{i=1}^{N} (s^{ti})2 + \frac{\lambda}{\| \mathbf{d}t \|} \right) -1\mathbf{w} \tag{6}$$

Let $\mathbf{d}t = \tau\mathbf{w}$, τ is the scale. We can rewrite Eq. (6) as follows:

$$\tau\mathbf{w} = \left(\sum_{i=1}^{N} (s^{ti})2 + \frac{\lambda}{\| \tau\mathbf{w} \|} \right) -1\mathbf{w} \tag{7}$$

which infers that

$$\tau = \frac{1}{\sum^{i=1N}} (s^{ti})2 \left(1 - \frac{\lambda}{\| \mathbf{w} \|} \right) \tag{8}$$

Because τ should be a non-negative, we get that if $\| \mathbf{w} \| \leq \lambda$, $\mathbf{d}t = \mathbf{0}$; otherwise $\mathbf{d}t = \tau\mathbf{w}$ and τ is defined as Eq. (8).

Finally, we can obtain the sematic dictionary $\mathbf{D} \in \Re k \times m$ via the above recursions. For any image i, we firstly describe it with the BOV model, and marked it by $xi \in \Re k$ where k is the dimensionality of the BOV representation. The semantic representation of the image i can be computed by the inner product,

$$SR(i) = xi \cdot \mathbf{D} \tag{9}$$

4. Semantic expansion

After obtaining the principal semantic representation of images, we can compute the semantic distance between images. However, different from the visual description approach, where all the bins of the descriptor are independent, the bins of the semantic description are the concepts, and they are not independent, i.e. there exists some implicit relevancy among the concepts. To capture these relevancies and provide the better and more comprehensive semantic description, we introduce the semantic expansion to transfer the weights between related concepts in the final semantic representation of images. In short, we learn a concept transferring matrix to expand the power of image semantic description.

Firstly, to mine the relevancy between the concepts in the semantic dictionary, we make use of a large scale probabilistic semantic network, known as Probase [20], which contains isA relations between subordinate concept (sub-concept) and super-ordinate concept (super-concept). The isA relationships in the Probase are harvested from 1.68 billion web pages and 2 year worth of Microsoft Bing search log using syntactic patterns [21]. For example, “Dog is an animal”, where “Dog” is a subordinate concept, and “animal” is a superordinate concept. Furthermore, Probase has the following significant properties:

Algorithm 1. Semantic expansion.

- Probase has a huge concept space with almost 2.7 million concepts, which completely covers the concept set of my above semantic dictionary.
- Probase is a network-structured taxonomy, i.e. a sub-concept may have several super-concepts.
- Each isA relation (x isA z) is associated with conditional probability $P(x|z)$ and $P(z|x)$ to measure the typicality, a.k.a.

Input: $\langle x, y\rangle$: a pair of concepts from the semantic dictionary (Sec. 3);

Φ_{isA}: the isA relationship of Probase;

Ψ: the synonym set in the Probase;

Output: The concept transferring matrix $\Omega \in \Re^{m\times m}$, Ω is a symmetric matrix, and m is the number of concept set **SC**in the semantic dictionary;

```
for τ = 1, ··· , m do
    x=SC(τ);
    Collect all the super-concepts of x from Φ_isA as the set ϒ^x;
    for μ = τ, ··· , m do
        y=SC(μ);
        Collect all the super-concepts of y from Φ_isA as the set ϒ^y;
        According to the synonym set Ψ, let ϒ_c = {ϒ^x ∩ ϒ^y} indicates the
        common super-concepts set;
        Ω⟨x, y⟩=max{P(x|ϒ_1^c) · P(y|ϒ_1^c), ··· , P(x|ϒ_n^c) · P(y|ϒ_n^c)}, here,
        n = |r^c|_0;
    end
end
```

typically scores, which derives from the co-occurrences:

$$P(x|z) = \frac{\text{occurrences of } (x,z) \text{in Hearst extraction}}{\text{occurrences of } z \text{ in Hearst extraction}}$$

Based on the Probase, we search all the super-concepts for each concept in our semantic dictionary (Section 3).

Then, for any two concept, we compute the common set of their super-concept set, according to the synonym set of the Probase. The transferring weight between two concepts is obtained by max pooling of probability.

Finally, the concept transferring matrix Ω is learned by the complete traversal of the above step. Algorithm 1 details the procedure of semantic expansion. The transferring matrix measures the perception relevancy of different concepts. On the other hand, the semantic expansion is a re-ranking model on the basis of semantic distance, which is different from the traditional re-ranking model, such as visual rank [22].

5. Image semantic distance metric

In this paragraph, firstly, we formulate the ultimate image semantic description based on the semantic dictionary and concept transferring matrix, and then discuss the image semantic distance metric.

According to Eq. (9), the principal image semantic description is obtained. Further, we implement the semantic expansion with the concept transferring matrix Ω. The final image semantic description is formulated as follows:

$$Des(i) = SR(i) + SR(i)\cdot\Omega, \tag{10}$$

where the first term is the principal semantic description of the image i based on the semantic dictionary from Section 3, and the second term is the context of semantic transferring.

After obtaining the ultimate semantic description, next we discuss the distance metric about semantic description. There are three classical metrics in the text retrieval domain as follows:

- Jaccard coefficient:

$$SimJaccard(A,B) = \frac{|A\bigcap B|}{|A\bigcup B|} \tag{11}$$

- Dice's coefficient:

$$SimDice(A,B) = \frac{2|A\bigcap B|}{|A|+|B|} \tag{12}$$

- Cosine similarity:

$$\frac{Simcosine(A,B) = \frac{|A\cdot B|}{\|A\|\cdot\|B\|} = \sum^{i=1m} Ai \times Bi}{\sqrt{\sum}^{i=1m}(Ai)2 \times \sqrt{\sum}^{i=1m}(Bi)2} \tag{13}$$

The first two methods derived from the Set system, where the specific value of each bin from the description is ignored. These methods are more suitable for the representation with huge concepts, i.e.10+ million. The aim of our semantic representation is to describe the image with the probability distribution of a series of concepts, and the cosine similarity can make full use of the probability paradigm and serve as the image semantic distance metric.

Table 1
Average time by different methods for search an image.

	BOV [24]	SoftBOV [25]	VLAD [26]	Sdict	Sdict+Sexp
Average time (ms)	313.3	521.6	285.4	323.3	402.7

6. Experiments

In this section, we first introduce the database and the relevant experimental settings, and then validate the effectiveness of the proposed semantic dictionary on the public test sets for the common image understanding tasks: web-scale semantic image search and image annotation.

6.1. Database and experimental setting

Database: The most popular image database-ImageNet [19] is used as our training data, which is organized by a semantic hierarchy. The original data has over one thousand categories and over one million images, and we simplify the data and choose frequently-used 217 concepts and there are 267k images together (named by ImageNet267K). Furthermore, we re-label 120×217 images from the Image267K as the training set (named by ImageNet25K). Another standard image database is the (Corel5K) [23], which is from 50 Corel Stock Photo CDs and consists of 5000 images. Besides, we collect 0.8 million Flickr image as the distracter images (named by Flickr800K).

Experimental settings: Semantic dictionary is learnt from the ImageNet25K. The visual set consists of 131,072 visual words, which are obtained by the hierarchical k-means clustering, and the concept set includes 217 concept from the ImageNet.

6.2. Web-scale semantic image search

In this paragraph, we validate its efficiency of semantic dictionary on large scale image database, where there are one million images (ImageNet267K+Flickr800K).

Comparisons methods: (1) The classical BOV approach [24] is compared as the baseline approach, where the size of visual words is 0.2 million. (2) SoftBOV [25], an improvement version of BOV, where each descriptor is encoded with soft assignment of 4 nearest neighbors. (3) VLAD [26] (vector of locally aggregated descriptors), which is a state-of-the-art method and derive from the Fisher kernel. The parameter is the same with that in [26]. (4) Our semantic dictionary approach [10], called by 'Sdict', $\lambda = 0.78 \times 10-2$, which is obtained by the cross-validation. λ restricts the sparsity of semantic dictionary. θ is a titleholding parameter, and it controls the inter-impact of the semantic dictionary. (5) Semantic dictionary+semantic expansion approach (called by 'Sdict+Sexp'), after obtaining the principal semantic description by the proposed semantic dictionary, we add the semantic expansion stage, i.e. the original semantic description does the dot product with the concept transferring matrix (detailed in Section 4).

For the image retrieval task, we follow [24–26] and take the mean average precision (MAP) as the evaluation metric. In detail, 250 representative images are chosen from the ImageNet267K as the query images. For each query result, the precision-recall curve is computed and the average precision (AP) is obtained by summing the area below the precision-recall curve. Finally, MAP is the meaning of AP from all the query images.

Fig. 3 shows the comparisons of the above approaches on the ImageNet dataset with a different number of images. We can find: firstly, our two methods (Sdict and Sdict+Sexp) improve the MAP sharply than the classical BOV and SoftBOV models, and compared with the state-of-the-art method VLAD, our method also boosts the MAP with 5.7% and 6.8%, the good performance of our method benefits from the consideration of the VPCP problem when we design the semantic dictionary model. Secondly, the semantic expansion can improve the performance, which demonstrates that the relevancy between concepts exists and it has a significant influence. Thirdly, as the number of images increases, the retrieval performance decreases. But the decreasing rate of our approaches is slow, which demonstrates that our method is less sensitive to the increment of the data scale.

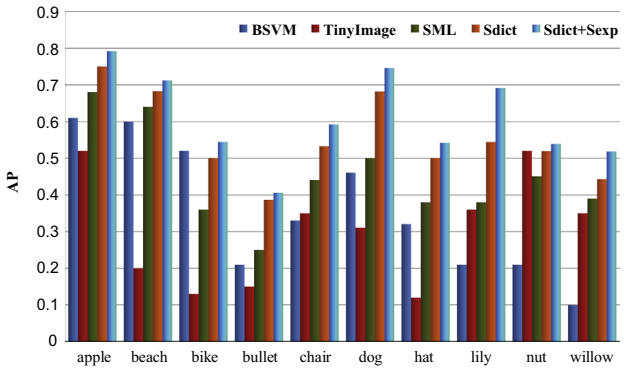

Fig. 4. Comparisons of different annotation methods with AP on the ImageNet267K database.

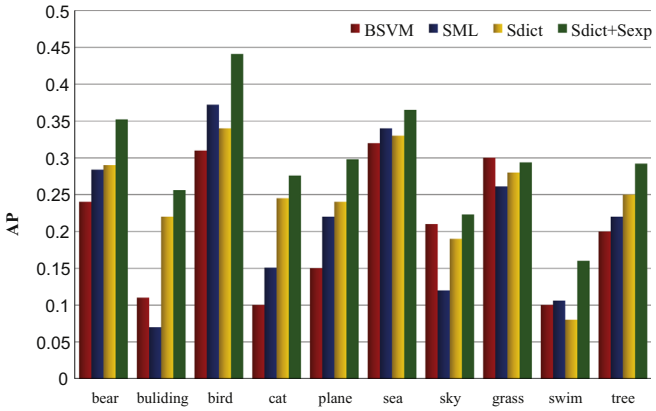

Fig. 5. Average precision comparisons of three annotation methods on Corel5K database.

Besides the obvious improvements of MAP, it is necessary to point out that our approaches (Sdict and Sdict+Sexp) are also efficient. Table 1 shows the average time of search an image by different methods. Experimental results show that our methods are comparable and satisfied for the real application. The lowness of SoftBOV is mainly rooted in the expensive image soft expansion. The above experiments are carried out on a server with 128 GB memory and 16-core 2.13 GHz processor with nVIDIA GeForce GTX 780. In short, the semantic dictionary and semantic expansion approaches show significant advantages on both accuracy and efficiency.

6.3. Image annotation on ImageNet267K

Image annotation is one of the most important tasks in the image understanding, and effective image annotation approach can help the images from the Internet to obtain more accurate label result, which can be widely used for many kinds of applications, such as the data mining, text-based retrieval and so on. To evaluate the performance of image annotation, semantic dictionary is tested and compared on the real web data-ImageNet.

Comparisons methods: (1) Binary SVM (BSVM) [27], as a one-vs.-all classification model, we have to separately train the SVM classifiers for each of 217 concepts. One hundred positive images and two hundred negative images are selected from the ImageNet25K. (2) TinyImage [28], this is one idea of nearest neighbor voting. Following the work [28], firstly, the images in ImageNet267K database are down sampled to 32×32, then the top-100 nearest neighbors for the query are searched based on SSD pixel distance, and at last the label of query image is replaced by the concept with the most votes. (3) Supervised Multi-class Labeling (SML) [29], as the setting of [29], images are represented by the bags of localized features, and Gaussian mixture model (GMM) consists of 64 components, which are obtained by the separate training. (4) Semantic dictionary (Sdict), the parameters are the same with those in Section 6.2. (5) Semantic dictionary + semantic expansion model (Sdict + Sexp) (detailed in Section 4).

Evaluation rule: the average precision (AP) is employed as the evaluation metric. Following the rule in the ImageNet [19], every method provides a top-5 annotation result for the query image, and this annotation is valid if one of the top-5 results is the same with the benchmark. The final AP is averaged over 100 query images from the ImageNet database.

The average precision of different approaches is shown in Fig. 4, where we can observe that our method, which combines the semantic dictionary with the semantic expansion technology, provides the best results of image annotation for most of the query images. In detail, the mean AP of our method is 60.82%, and has an improvement with 16.1% than SML method [29], which is also a supervised learning approach. Further, the precision with 60% in the piratical scenario is a satisfied precision for the large scale annotation task, and after all there are usually less than 0.01% web images with useful label.

6.4. Image annotation on Corel5K

To evaluate the transferring performance of our method, semantic dictionary is evaluated on another standard benchmark (Corel5K). This database contains 5000 images with 260 concepts, where we find that the 217 concepts from ImageNet25K cover some of the Corel5K database. In this paragraph, we complement the annotation task with the semantic dictionary learnt from ImageNet25K.

Comparisons methods: (1) Binary SVM (BSVM) [27]. Similar to the above procedure, we train binary SVM classifiers separately. In the training stage, the Corel5K is decomposed into three image subsets: the training set of 4000 images, the validation set of 500 images, and the test set of 500 images. (2) Supervised Multi-class Labeling (SML) [29], the parameter setting is the same with that in Section 6.3. (3) Semantic dictionary (Sdict), the parameters of the learning are the same with that in Section 6.2. (4) Semantic dictionary + semantic expansion model (Sdict + Sexp) (the details of this approach are provided in Section 4).

Evaluation rule: average precision (AP) is used as the measurement metric. Different from that in Section 6.3, we take a more strict precision evaluation, where every method provides a top-3 annotation result for the query image, and this annotation is valid if one of the top-3 results is the same with the benchmark. 10 common categories between the ImageNet267K database and Corel5K database are chosen as the test images.

Fig. 5 shows the average precision comparisons on the Corel5K database, the performance of our two methods unexpectedly exceeds the classical BSVM and SML methods. Particularly, the coalescent method from semantic dictionary and semantic expansion achieves a 9.17% and 8.13% improvement than BSVM and SML separately, and in fact our method was trained not on the Corel5K database but on the ImageNet database. This means that our method has the potential extendibility to be directly used in the other image sets, but it does not need the training process. The inherent reason results from two sides: one is the robust semantic description from the original semantic dictionary, and the other is the semantic expansion technology, which transfers the implicit relevancy between concepts. When the concept set of our semantic dictionary covers more concepts in our daily life, the robust transferring ability of our model can bring in some significant applications, such as the real-time object annotation on the mobile devices. We also notice that our method does not work well on the "swim" category because there are the great visual difference between the instances in the Corel5k dataset and those in the ImageNet. The feasible solutions for such problem are as follows:

- to increase the number of training data and choose the diverse training images for each class;
- to enlarge the concept set in the semantic dictionary learning procedure;
- to learn the concept transferring matrix on the basis of the more accurate semantic network, and improve the generalization ability of the semantic expansion.

As our method can be used directly without relying on any outside information. Thus we could have a coarse annotation for large/web scale image datasets, such as Flickr and Google Image, when the concept set of our semantic dictionary coverers most concepts in our daily life.

7. Conclusion

In this paper, we proposed the semantic dictionary to solve the paradox of visual polysemia and concept polymorphism in the large scale image understanding, which characterizes the probability distribution between visual appearances and semantic concepts. The learning of semantic dictionary is formulated into a minimization optimization problem and mixed-norm regularization learning is introduced to solve the above optimization problem, where the common visual patterns shared by the related concepts are learnt. To improve the generalization ability of the semantic description, we propose the semantic expansion technology, where a concept transferring matrix is learnt to quantize the implicit relevancy among the concepts. Finally, to speeding up, we propose a distributed framework for applying the semantic dictionary into the large scale image understanding. In the future, on one hand, we will focus on the scalability of semantic dictionary for web scale image database, which covers usually ten thousand concepts. On the other hand, restricted in the high complexity of web-scale model learning, we plan to find highly parallel framework for large scale image understanding on many-core platform.

Acknowledgements

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400 and 2015CB351802, in part by National Natural Science Foundation of China: 61332016, 61025011, 61402431, 61303154, 11301517 and 61472203, in part by Project Funded by China Postdoctoral Science Foundation.

References

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

[2] K. Weinberger, L. Saul, Fast solvers and efficient implementations for distance metric learning, in: Proceedings of ICML, 2008, pp. 1160–1167.
[3] G. Qi, C. Aggarwal, Y. Rui, Q. Tian, S. Chang, T. Huang, Towards cross-category knowledge propagation for learning visual concepts, in: Proceedings of IEEE CVPR, 2011, pp. 897–904.
[4] L. Li, S. Jiang, Q. Huang, Learning hierarchical semantic description via mixed-norm regularization for image understanding, IEEE Trans. Multimed. 14 (2012) 1401–1413.
[5] S. Bucak, R. Jin, A. Jain, Multi-label multiple kernel learning by stochastic approximation: Application to visual object recognition, in: Proceedings of NIPS, 2010, pp. 1145–1154.
[6] L. Li, S. Jiang, Z.-J. Zha, Z. Wu, Q. Huang, Partial-duplicate image retrieval via saliency-guided visually matching, IEEE Multimed. 20 (2013) 13–23.
[7] R. Rockafellar, Convex Analysis, Princeton University Press, Princeton, NJ, 1970.
[8] D. Bertsekas, Nonlinear programming, Athena Scientific, 1999.
[9] L. Grippo, M. Sciandrone, On the convergence of the block nonlinear gauss-seidel method under convex constraints, Oper. Res. Lett. 26 (3) (2000) 127–136.
[10] L. Li, C. Yan, X. Chen, S. Jiang, S. Rho, J. Yin, B. Jiang, Q. Huang, Large scale image understanding with non-convex multi-task learning, in: Proceedings of International ICST Conference on Game Theory for Networks, 2014.
[11] R.W. White, Beliefs and biases in web search, in: Proceedings of SIGIR, 2013.
[12] S. Hingmire, S. Chakraborti, Topic labeled text classification: a weakly supervised approach, in: Proceedings of SIGIR, 2014.
[13] Y. Zhang, M. Zhang, Y. Zhang, Explicit factor models for explainable recommendation based on phrase-level sentiment analysis, in: Proceedings of SIGIR, 2014.
[14] C. Yan, Y. Zhang, J. Xu, F. Dai, J. Zhang, Q. Dai, F. Wu, Efficient parallel framework for hevc motion estimation on many-core processors, IEEE Trans. Circuits Syst. Video Technol. 24 (2014) 2077–2089.
[15] C. Yan, Y. Zhang, J. Xu, F. Dai, L. Li, Q. Dai, F. Wu, A highly parallel framework for hevc coding unit partitioning tree decision on many-core processors, IEEE Signal Process. Lett. 21 (2014) 573–576.
[16] C. Yan, Y. Zhang, F. Dai, X. Wang, L. Li, Q. Dai, Parallel deblocking filter for hevc on many-core processor, Electron. Lett. 50 (2014) 367–368.
[17] J. Sivic, A. Zisserman, Video google: a text retrieval approach to object matching in videos, in: Proceedings of IEEE ICCV, 2003, pp. 1470–1477.
[18] D.G. Lowe, Distinctive image features from scale invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[19] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: Proceedings of IEEE CVPR, 2009, pp. 248–255.
[20] W. Wu, H. Li, H. Wang, K. Q. Zh, Probase: a probabilistic taxonomy for text understanding, in: Proceedings of SIGMOD, 2012, pp. 481–492.
[21] M.A. Hearst, Automatic acquisition of hyponyms from large text corpora, in: Proceedings of COLING, 1992, pp. 539–545.
[22] Y. Jing, S. Baluja, Visualrank: applying page-rank to large-scale image search, IEEE Trans. PAMI 30 (11) (2008) 1877–1890.
[23] P. Duygulu, K. Barnard, J. Freitas, D. Forsyth, Object recognition as machine translation: learning a lexicon for a fixed image vocabulary, in: Proceedings of ECCV, 2002, pp. 97–122.
[24] D. Nister, H. Stewenius, Scalable recognition with a vocabulary tree, in: Proceedings of IEEE CVPR, 2006, pp. 2161–2168.
[25] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Lost in quantization: improving particular object retrieval in large scale image databases, in: Proceedings of IEEE CVPR, 2008, pp. 1–8.
[26] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation, in: Proceedings of IEEE CVPR, 2010, pp. 3304–3311.
[27] C. Chang, C. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27:1–27:27, software available at ⟨http://www.csie.ntu.edu.tw/cjlin/libsvm⟩.
[28] A. Torralba, R. Fergus, W.T. Freeman, 80 million tiny images: a large dataset for nonparametric object and scene recognition, IEEE Trans. PAMI 30 (11) (2008) 1958–1970.
[29] G. Carneiro, A.B. Chan, P.J. Moreno, N. Vasconcelos, Supervised learning of semantic classes for image annotation and retrieval, IEEE Trans. PAMI 29 (3) (2007) 394–410.

Liang Li received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2013. He received the B.S. degree in computer science from Xi'an Jiaotong University, Shaanxi, China, in 2008. He is currently a post-doc with University of 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.

Chenggang Clarence Yan received his Ph.D. from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2013. He is a Post-Doctoral Research Fellow with the Department of Automation, Tsinghua University, Beijing, China. His research interests include parallel computing, image processing, computational biology, computer vision, and pattern recognition.

Xing Chen received the Ph.D. degree from Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China, in 2012. He received the B.S. degree in Information and Computing Science from Shandong University, Weihai, China, in 2007. He is currently an Assistant research fellow with Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China. His research interests include computational biology, bioinformatics, systems biology, data mining, pattern recognition, and image semantic understanding.

Chunjie Zhang received his Ph.D. degree in Pattern Recognition and Intelligent Systems from Institute of Automation, Chinese Academy of Sciences, China, in 2011. He received his B.E. degree from Nanjing University of Posts and Telecommunications, Jiangsu, China, in 2006. He worked as an engineer in the Henan Electric Power Research Institute during 2011–2012. He is a Postdoc at School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China from July 2012 to January 2015. He is now an assistant professor at School of Computer and Control Engineering, University of Chinese Academy of Sciences. Dr. Zhang's current research interests include image processing, machine learning, cross media content analysis, pattern recognition and computer vision.

Jian Yin received the B.S. degree in the machine electricity from Harbin Institute of Technology, China, in 1998. He is currently an associate professor with Department of Computer, Shandong University, Weihai, China. His research interests include parallel computing, image processing, computational biology.

Baochen Jiang received his M.S. degree in Department of Electronic Engineering from Tsinghua University, Beijing, China, in 1990. He received the B.S. degree in Department of Electronics from Shandong University, in 1983. He is currently a professor with School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai, China. His research interests include image processing, signal processing, and electrical system.

Qingming Huang received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1994. He was a Postdoctoral Fellow with the National University of Singapore from 1995 to 1996 and was with the Institute for Infocomm Research, Singapore, as a Member of Research Staff from 1996 to 2002. He joined the Chinese Academy of Sciences, Beijing, China, under Science 100 Talent Plan in2003, and is currently a Professor with the University of Chinese Academy of Sciences. His current research areas are image and video analysis, video coding, pattern recognition, and computer vision.