Multimed Tools Appl (2018) 77:22185–22198
https://doi.org/10.1007/s11042-018-5796-9

# Semantic binary coding for visual recognition via joint concept-attribute modelling

**Xing Xu[1,2] · Haiping Wu[2] · Yang Yang[2] · Fumin Shen[2] · Ning Xie[2] · Yanli Ji[2]**

Received: 29 August 2017 / Revised: 1 January 2018 / Accepted: 13 February 2018 /
Published online: 28 February 2018

**Abstract** Recent years have witnessed the unprecedented efforts of visual representation for enabling various efficient and effective multimedia applications. In this paper, we propose a novel visual representation learning framework, which generates efficient semantic hash codes for visual samples by substantially exploring concepts, semantic attributes as well as their inter-correlations. Specifically, we construct a conceptual space, where the semantic knowledge of concepts and attributes is embedded. Then, we develop an effective on-line feature coding scheme for visual objects by leveraging the inter-concept relationships through the intermediate representative power of attributes. The code process is formulated as an overlapping group lasso problem, which can be efficiently solved. Finally, we may binarize the visual representation to generate efficient hash codes. Extensive experiments have been conducted to illustrate the superiority of our proposed framework on visual retrieval task as compared to state-of-the-art methods.

✉ Xing Xu
xing.xu@uestc.edu.cn

Haiping Wu
haipingwoo@gmail.com

Yang Yang
dlyyang@gmail.com

Fumin Shen
fumin.shen@gmail.com

Ning Xie
seanxiening@gmail.com

Yanli Ji
yanliji@uestc.edu.cn

[1] Guizhou Provincial Key Laboratory of Public Big Data, Guizhou University, Guiyang, China

[2] Center for Future Media & School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China

**Keywords** Visual computing · Binary representation · Semantics · Attributes

## 1 Introduction

Recently years, visual representation has become a vital part of computer vision applications (e.g. image recognition [20, 23, 39] and image retrieval [19, 28, 34]). Methods vary from raw feature extraction, to high-level feature statistics such Bag-of-Words, and today more complex feature extraction framework like deep neural network. However, as compared to our humans, who can easily recognize more than tens of thousands of objects, the above artificial frameworks are obvious inferior [36]. Further, it has been widely recognized that a good visual representation should integrate both low-level visual features addressing the more detailed perceptual aspects and high-level semantic features underlying the more general conceptual aspects of visual data [24, 37, 38, 42]. In other words, there remains semantic gap between the low-level features extracted using current methods and high-level semantics. Although many efforts [8, 9, 31, 43] have been devoted to reduce this semantic gap by combining these two types of visual representation methods, the gap is still a challenge for us. In this work, we aim to reduce this semantic gap through finding the correlation between low-level features and high-level concepts to obtain better visual representation.

In order to capture high-level semantics and cross category properties, attributes [4] has been proposed, which offers an important intermediate representation. As human recognize objects by finding their main attributes in high-level concepts, attributes description is more explainable and could be used to describe unknown category objects. Thus, by leveraging high-level attribute information with low-level features, we could obtain more compact and discriminative features across categories and further reduce the semantic gap. Specifically, the more semantic attributes two concepts share, the similar the two concepts are, such as “cat” and “dog” share many attributes together like “fur”, “has head”, and they are conceptually similar. Then we could use dictionary learning scheme in the coding process to efficiently solve this correlation embedding and feature selection problem.

In this paper, we propose a novel binary visual representation learning framework which exploits the semantic attributes to reduce semantic gap and generate binary codes for visual representation. As illustrated in Fig. 1, the proposed framework comprises two stages, the

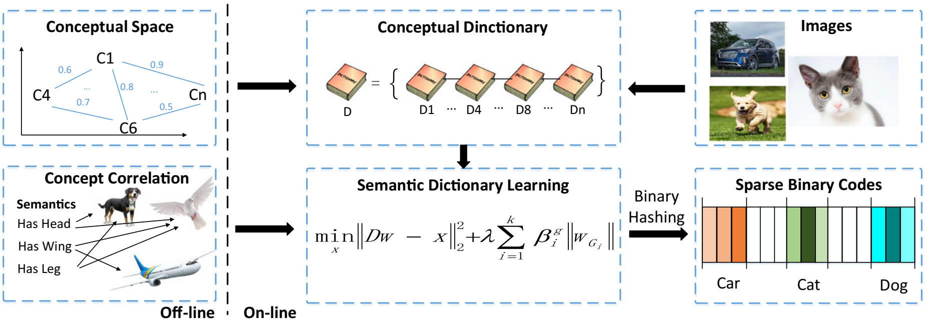

**Fig. 1** The proposed visual representation framework. In the off-line stage, we learn dictionaries in the conceptual space where each concept correspond to one dictionary base. Then inter-concept correlation are exploited by leveraging semantic attributes. In the on-line stage, overlapping groups are formed using the correlation information and then group sparsity are imposed to generate features using dictionaries obtained in the off-line stage. At last, hashing methods are used to generate compact binary representation

off-line stage and the on-line stage. In the off-line stage, we utilize semantics to find the correlation relationships among concepts. To this end, we project the original objects into the conceptual space and further divide dictionary bases of conceptual space into groups using semantic attributes. In the on-line stage, we devise a novel semantic coding approach by exploiting semantics, including concept-rich dictionaries and semantic-based concept correlation information obtained in the off-line stage. In order to achieve this goal, we incorporate group-based sparsity in the dictionary learning process, which efficiently selects salient concepts in the data. Then, we adapt hashing schemes to generate compact and efficient binary codes. The generated sparse codes are considered as the final visual representation and further used in specific tasks.

We summarize our contribution as follows:

- **Semantic Binary Learning Framework.** The proposed framework adapts hashing schemes to generate binary codes for visual representation. Extensive experiments for visual retrieval tasks have been conducted compared to several state-of-art hashing methods and salient performance boost could be observed.
- **Dictionary Learning in Concept Space.** The proposed framework learns dictionaries in conceptual space, which holds objects that if are conceptually similar, then they are close in the conceptual space.
- **Exploit Concept Correlation** The proposed framework utilizes the semantic attribute information to exploit the inter-concept correlation, which combines low-level features and high-level semantics in efficient way and thus reduce the semantic gap.

This paper makes further improvement based on our previous work in [30]. Specifically, the procedures of dictionary learning in concept space and concept correlation exploration are integrated to a unified framework for joint optimization in the proposed model, whereas in [30], they are two separate steps. In addition, we also developed an efficient iterative algorithm to solve the complex optimization problem in the proposed model. We conducted extensive experiments with various settings to evaluate the effectiveness of the proposed method on visual recognition task, and made comprehensive comparison with several latest state-of-the-art approaches on various aspects.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 details the proposed semantic binary visual representation learning framework, including the formulation, algorithms. Experimental results and analysis are reported in Section 4, followed by conclusions in Section 5.

## 2 Related work

### 2.1 Dictionary learning and visual representation

The state-of-the-art visual representation methods comprise two stages: First, using local descriptors (e.g., SIFT [18] and HOG [2]) to extract feature, then, an over-complete representation is encoded using these features. Recently, convolution neural network (CNN) has shown its great power in learning visual representations on various tasks like image classification [25]. CNN differs from the previous one that it learns features automatically while the previous one need manually set the extraction scheme. In this work, we combine these two methods to achieve better performance.

In [10], an efficient sparse coding pipeline was proposed where the dictionary is the feature matrix of the whole training images. Zhang et al. [41] used group sparsity to select different image pairs in visual retrieval tasks. Yang et al. [33] proposed the spatial group sparse coding to considered regions of an image as a group to achieve region-level image annotation. Gao et al. [6] proposed a multi-layer group sparse coding framework for concurrent image classification and annotation. Wu et al. [29] used trace lasso-based sparsity and form a weekly supervised dictionary learning framework. Similar to those works, our framework also adapt group sparsity to select features of the same group, which indicates concept in our work. However, our work further exploits the group correlation and impose overlapping group sparsity to better select variables and thus obtain a better representation.

### 2.2 Semantic attribute

Semantic attribute has been proposed to better describe visual objects [4, 14, 23]. Ouyang et al. [21] shows that semantic attributes are, if used in a proper way, helpful in learning feature representations that improve large-scale object detection. Semantic attributes express high-level meanings that could help represent the visual objects. By combining low-level features and high-level semantic attributes information, semantic gap could be reduced and the combined features are more meaningful and explainable.

Many approaches used predictions on attributes as middle-level features for recognizing new object categories [4, 13]. Chiang et al. [1] proposed a new distance matrix by jointly incorporating data and attribute similarities. Farhadi et al. [5] used the functionality, superordinate categories, viewpoint, and pose of segments as attributes to improve detection accuracy. Wu et al. [29] learned a dictionaries by exploiting visual attribute correlations rather than label priors. Unlike these methods, we utilize the attribute information in a different stage of the coding process, where its main role is to group the dictionaries and thus find the correlation relationships of concepts. Also different from [29] where they cluster visual objects for each concept and then impose group sparsity penalty, we choose to make the concepts overlapped and thus make the eventual space could hold concept similarity. Our dictionaries and concepts correlation are learned in the off-line stage and can be used in the on-line stage to generate visual representation which shows the generalization ability of our framework. Besides, the proposed method leverages explicit concept correlation, that is, first concepts are divided into groups, and then binary codes are generated using the group information and hashing methods. Various grouping mechanism such as [32, 35] could be adopted.

## 3 The proposed framework

In this section, we elaborate our proposed framework in terms of the off-line and on-line stages.

### 3.1 OFF-LINE: dictionary learning in the conceptual space

In off-line stage, we obtain the dictionary bases in the conceptual space and exploit the concepts correlation by utilizing attribute correlation information. Then the dictionary bases and concepts correlation information could be used in the on-line stage to produce new binary visual representation.

### 3.1.1 Conceptual space dictionary

Dictionary learning [26] plays a vital role in multimedia applications. In our work, we aim to construct a semantic-enriched dictionary by mapping the original images to the conceptual space. Close objects in conceptual space should be conceptual similar. Suppose we have $n$ samples $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n}] \in \mathbb{R}^{m \times n}$. We map the samples from original space to conceptual space by the projection:

$$\|\mathbf{X} - \mathbf{D_c S}\|, \tag{1}$$

where $\mathbf{D_c} \in \mathbb{R}^{m \times k}$ is the bases of the conceptual space, and $\mathbf{S} \in \mathbb{R}^{k \times n}$ is the new representation in the conceptual space.

### 3.1.2 Learning bases of concepts

Category is a nature choice for conceptual space, where each concept is implicitly corresponding to one category. In order to effectively exploit the images and category label information to construct conceptual space, we choose to obtain the bases' representations of each category using the classification hyperplanes learned from logistic regression, which split categories from each other:

$$\min_{\mathbf{d},c} \frac{1}{2}\mathbf{d^T d} + C \sum_{i=1}^{n} \log(exp(-\mathbf{y^i}(\mathbf{X_i^T d} + c)) + 1), \tag{2}$$

where $y^i$ is the category label for each sample $\mathbf{X_i}$, $C$ is the reverse of regularization strength, and smaller $C$ specify stronger regularization, $\mathbf{d}$ is the classification hyperplane we need ($\mathbf{D_c} = [\mathbf{d_1}, \mathbf{d_2}, \cdots, \mathbf{d_k}]$). For each category, we learn the hyperplane that could tell samples from whether belong to the category. Then, bases of conceptual space are made up of the hyperplanes. In this way, our dictionary has the natural power of generating discriminative features that could be easily used in the classification task.

### 3.1.3 Modeling concepts correlation with semantic attributes

Intuitively, the co-occurrence statistics include meaningful information about correlation relationship. The more semantic attributes two concepts share, the similar the two concepts are. Thus, conceptual correlation relationship could be found through attributes. To utilize the attribute correlation relationship, we propose to group the conceptual dictionary bases by leveraging attribute information. Any reasonable grouping schemes could be adapted in our framework. In a more specific case which we use in the experiment, a group is formed by selecting which bases share the same attribute, that is, if there are $m$ attributes, there should be $m$ groups where each group all share the same attribute. The formed groups are defined as:

$$\mathbf{G_i} = [j \text{ for } concept_j \text{ has } attribute_i]. \tag{3}$$

If two conceptual dictionary bases are divided into one group, it means that they are somehow conceptually connected through attributes. Notice that, groups could be overlapping due to classes often share more than one attributes, which means one concept is similar to another concept in multiple attribute aspects.

## 3.2 ON-LINE: binary visual representation generation

As the dictionary bases of conceptual space and the relationships of concepts have been obtained in the off-line stage, in this stage, we adapt dictionary learning and hashing methods to produce our semantic binary representation.

### *3.2.1 Semantic visual coding*

New representation in the conceptual space could be obtained by using the projection illustrated by the (1). $S$ is the reconstruction in the conceptual space. However, the conceptual space bases contain the whole classes in the dataset, out of which many bases would not be used to reconstruct the feature for one specific image. Thus, variable selection methods could be used to impose sparsity on the reconstruction weight of bases which generates more explainable features. We use group-based lasso [11] to solve the variable selection problem. The forming groups could be non-overlapping or overlapping, we obtain our new feature representation by separately handle these two situation. The lasso [27] estimate is defined by

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left( \mathbf{y_i} - \beta_0 - \sum_{j=1}^{p} \mathbf{x_{ij}} \beta j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{4}$$

Here, λ is a tuning parameter that controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage. The $l_1$-norm penalty induces sparsity in the solution. In our work, each conceptual dictionary base may be divided into one or more than one groups, thus we select our variables in two cases, non-overlapping and overlapping.

### *3.2.2 Visual coding with non-overlapping concept correlation*

If the concept correlation is non-overlapped which means the formed groups are non-overlapped, we could simply impose group sparsity in the dictionary learning process. And the problem could be described as the follow:

Given a set of groups $G$ which form a partition of $[1, p]$, the group lasso penalty [40] is a norm over $\mathbb{R}^p$ defined as:

$$\forall w \in \mathbb{R}^p, \Omega^G_{group} = \sum_{g \in G} \|w_g\|. \tag{5}$$

When the groups in $G$ form a partition of the set of covariates, then $\Omega^G_{group}(w)$ is a norm whose balls have singularities when some $w_g$ are equal to zero.

In our case, we aim to project the original feature into the conceptual space, derived from (4) and (5), the whole objective function could be written as:

$$\min_{\mathbf{v}} \|\mathbf{D_c v} - \mathbf{x}\|_2^2 + \lambda \sum_{i=1}^{k} \beta_i^g \|\mathbf{v_{G_i}}\|, \tag{6}$$

where $\mathbf{D_c} \in \mathbb{R}^{m \times k}$ is the bases of the conceptual space. $\mathbf{x} \in \mathbb{R}^{m \times 1}$ is the original feature, $\mathbf{v} \in \mathbb{R}^{k \times 1}$ is the new feature. The $\mathbf{v}$ is divided into k non-overlapping groups $\mathbf{v_{G_1}}, \mathbf{v_{G_2}}, \cdots, \mathbf{v_{G_k}}$, and $\beta_i^g$ denotes the weight for the $i$-th group.

### 3.2.3 *Visual coding with overlapping concept correlation*

In the case of overlapping concept correlation where formed groups are overlapped, we impose overlapping group sparsity penalty when learning the visual representation. The variables selection and visual representation learning process could be formalized as the follow:

$$\min_{\mathbf{v}} \|\mathbf{D_c v} - \mathbf{x}\|_2^2 + \lambda_1 \|\mathbf{v}\|_1 + \lambda_2 \sum_{i=1}^{k} \beta_i^g \|\mathbf{v_{G_i}}\|, \tag{7}$$

where $\mathbf{D_c} \in \mathbb{R}^{m \times k}$ is the bases of the conceptual space. $\mathbf{x} \in \mathbb{R}^{m \times 1}$ is the original feature, $\mathbf{v} \in \mathbb{R}^{k \times 1}$ is the new feature. The groups $\mathbf{G_i}$ may overlap, and $\beta_i^g$ is the weight for the $i$-th group. The obtained new visual representation $\mathbf{v}$ is not only sparse, but also contains salient conceptual and attribute information. Equation (7) is a standard sparse dictionary learning problem, and in practice it can be efficiently solved by the the SLEP software [16].

### 3.2.4 *Binary representation generation*

In order to generate compact binary codes, we utilize existing hashing methods to generate binary hashing codes. Unsupervised learning methods such as ITQ [7], LSH [22] and SKLSH [22], supervised hashing methods such as COSDISH [12], KSH [17] and FastHash [15] are adapted to evaluate our generated features. Take ITQ for example, we first do PCA projection converting $\mathbf{V} = [\mathbf{v_1}, \mathbf{v}_2, \cdots, \mathbf{v_n}]$ obtained from (6) or (7) to $\mathbf{V}' \in \mathbb{R}^{l \times n}$, $l$ is the code length. Then the objective function is described as:

$$\|\mathbf{B} - \mathbf{V'^T R}\|_F^2, \tag{8}$$

where $\mathbf{B} \in \{-1, 1\}^{l \times n}$ is the final binary codes, $\mathbf{R} \in \mathbb{R}^{l \times l}$ is the rotation matrix, and $\| \cdot \|_F$ denotes the Frobenuis norm.

# 4 Experiments

## 4.1 Datasets

**Animals with attributes (AWA)** [13] The AWA dataset consists of 30475 images. Images are categorized into 50 animals classes and there are 85 numeric attribute values for each class. Each image is represented by a 4096-dimension vector, fc7 layer of very deep 19-layer CNN, pretained on ILSVRC2014 [25]. We split the dataset into two parts, which are separately used in off-line and on-line stages. More specifically, we randomly select two-thirds of each concept (class) as the off-line stage training dataset, and the left one-third as the on-line stage testing dataset, which gives us 15226 images for the off-line stage and 10172 images for the on-line stage.

**aPascal & aYahoo dataset** [4] aPascal dataset contains 12695 visual objects, 20 categories and 64 semantic attributes to describe objects. aYahoo dataset contains 2644 visual objects, 12 categories. In our experiment, we use the whole aPascal dataset in the off-line stage to obtain concept dictionaries and exploit concept correlation. And we use the whole aYahoo dataset in the on-line stage to evaluate our framework. This also helps evaluate the generalization abilities of our framework.

## 4.2 Evaluation settings

For evaluation metrics, we chose commonly used Average Precision (AP) and mean Average Precision (mAP). For AWA dataset, we randomly choose 1000 samples in the testing phase, and for aPascal dataset, we randomly choose 100 samples in the testing phase.

For AWA dataset, We compare our proposed binary visual representation framework with the state-of-art deep neural networks methods (VGG-19 [25]. For aPascal dataset, we compare our proposed binary visual representation framework with the method provided by [4] which is carefully designed for this dataset. We use several hashing methods to in our framework, including unsupervised hashing methods, LSH [3], SKLSH [22], ITQ [7] and supervised hashing methods, COSDISH [12], KSH [17] and FastHash [15]. All the hash methods are implemented by the source code provided by the corresponding authors. For KSH, following the setting in [17], we randomly sample 2000 points as training set in the off-line stage and the number of support vectors is 300. For FastHash, boosted decision trees are used for out-of-sample extension.

## 4.3 Results and discussions

In this part, first we show how our proposed semantic binary learning framework achieve better performance by comparing it with other popular methods using several popular hashing methods including both unsupervised ones and supervised ones. Then we explore the sensitivity of hashing methods and code length. Finally we show our proposed attribute-based group scheme boost performance by comparing with representation learned by not leveraging attribute information.

### *4.3.1 Comparison with other methods*

We evaluate our proposed framework by comparing with deep neural network representation [25] on AWA dataset, and method designed by [4] on aPascal dataset. As illustrated, Tables 1 and 2 shows the results of mAP on AWA dataset using unsupervised and supervised

**Table 1** Comparisons in terms of mAP on AWA Dataset with unsupervised hashing methods

| Method | | 8 bits | 16 bits | 32 bits | 48 bits |
|---|---|---|---|---|---|
| ITQ | Fc-7 | 0.6067 | 0.6871 | 0.7529 | 0.7926 |
| | Ours | 0.5839 | 0.6867 | **0.7660** | **0.8064** |
| LSH | Fc-7 | 0.1934 | 0.2557 | 0.3319 | 0.4143 |
| | Ours | **0.3060** | **0.4528** | **0.5609** | **0.6538** |
| SKLSH | Fc-7 | 0.1705 | 0.1794 | 0.1928 | 0.2677 |
| | Ours | **0.1851** | **0.2341** | **0.2805** | **0.3123** |
| COSDISH | Fc-7 | 0.4226 | 0.7590 | 0.8317 | 0.8622 |
| | Ours | **0.6970** | **0.8924** | **0.9152** | **0.8907** |
| KSH | Fc-7 | 0.4566 | 0.6254 | 0.7018 | 0.7385 |
| | Ours | **0.5317** | **0.6418** | **0.7407** | **0.7619** |
| FastHash | Fc-7 | 0.7093 | 0.8314 | 0.8644 | 0.8835 |
| | Ours | **0.8374** | **0.8912** | **0.9016** | **0.9084** |

Better scores are marked with bold font in the proposed method

**Table 2** Comparison in terms of mAP on AWA Dataset with Supervised hashing methods

| Method | | 8 bits | 16 bits | 32 bits | 48 bits |
|---|---|---|---|---|---|
| COSDISH | Fc-7 | 0.4226 | 0.7590 | 0.8317 | 0.8622 |
| | Ours | **0.6970** | **0.8924** | **0.9152** | **0.8907** |
| KSH | Fc-7 | 0.4566 | 0.6254 | 0.7018 | 0.7385 |
| | Ours | **0.5317** | **0.6418** | **0.7407** | **0.7619** |
| FastHash | Fc-7 | 0.7093 | 0.8314 | 0.8644 | 0.8835 |
| | Ours | **0.8374** | **0.8912** | **0.9016** | **0.9084** |

Better scores are marked with bold font in the proposed method

hashing methods, respectively. We can observe from the results that our proposed framework can achieve obvious improvements in terms of retrieval performance over other methods on AWA dataset. Our proposed framework not only boosts the performance of unsupervised hashing methods but also supervised learning methods. Take 48 bits codes for example, our proposed framework outperforms CNN representation by around 1.74, 57.80, 20.39% for ITQ, LSH and SKLSH hashing methods, and around 3.31, 3.17, 2.82% for COSDISH, KSH, and FashHash hashing methods. The underlying reasons are two folds. First of all, our framework works in the conceptual space where similar concepts could be close in this space. Second, our framework efficiently exploit the concepts correlation by using semantic attribute information, thus reducing semantic gap.

Table 3 reports the mAP performance for our proposed framework and the method designed by [4] on aYahoo dataset. Figure 2 shows the detailed performance (AP) of individual concepts on aYahoo dataset. We can see from the results that our proposed framework achieve better performance of all individual concepts for all the hashing methods we use. This consistently validates the advantages of our framework. Notice that, the dictionary bases and concepts correlation relationships for aYahoo dataset are not derived from the dataset itself but from aPascal dataset. This further shows the salient performance and generalization abilities of our framework.

### *4.3.2 The sensitivity of hashing methods and code length*

We can observe from Tables 1 and 3 that the improvement of our proposed framework highly depends on the hashing methods we adapt. For example, we obtain 1.74% improvement using ITQ, compared to 57.80% using LSH for 48 bits on AWA dataset. We infer that this is because some hashing methods utilize the data better than others, thus our framework boosts

**Table 3** Comparison in terms of mAP on aYahoo Dataset

| Method | | 4 bits | 8 bits | 12 bits | 16 bits | 20 bits |
|---|---|---|---|---|---|---|
| ITQ | [4] | 0.7117 | 0.7621 | 0.7872 | 0.8003 | 0.8124 |
| | Ours | **0.8532** | **0.8721** | **0.9064** | **0.9207** | **0.9282** |
| LSH | [4] | 0.5261 | 0.5242 | 0.5384 | 0.5448 | 0.5704 |
| | Ours | **0.7213** | **0.7201** | **0.8090** | **0.8181** | **0.7990** |
| COSDISH | [4] | 0.1541 | 0.2219 | 0.3082 | 0.3720 | 0.3191 |
| | Ours | **0.3893** | **0.4701** | **0.5146** | **0.5665** | **0.5539** |

Better scores are marked with bold font in the proposed method

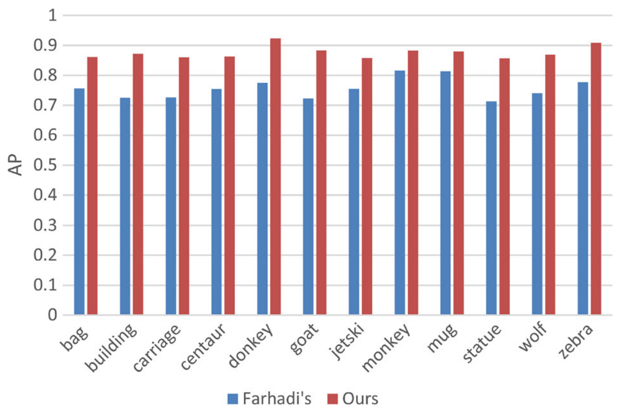

**Fig. 2** Average precision of individual concepts for visual retrieval on aYahoo dataset

less using these methods compared to others. We could also observe that the improvement varies with code length. For example, mAP improvement on AWA dataset using SKLSH varies from 8.56, 30.49, 45.49, 16.66% with code length from 8 bits to 48 bits, the max improvement occurs using 32 bits. We assume the reason that there is a best code length suitable for the hashing methods to interpret the dataset, thus influencing the performance boost of our framework.

### *4.3.3 Effectiveness of semantics*

Exploiting concept correlation using semantics is a crucial part of our proposed framework (Semantic). In order to evaluate the effectiveness of our semantics scheme, we design another framework which does not consider any semantic attribute information (Non-Semantic). For Non-Semantic method, we assume each concept is a group and then follow the process the same as our proposed Semantic framework. Tables 4 and 5 shows how semantics leading concept correlation help improve the performance on AWA and aYahoo datasets, respectively. As we can see, the Semantic method which utilizes semantics to find concept correlation achieves better performance than the Non-Semantic method the two datasets and all the hashing methods we use. For instance, the Semantic method outperform

**Table 4** Comparison on representations with semantics and without semantics on AWA dataset

| | | AWA Dataset | | | |
|---|---|---|---|---|---|
| Method | | 8 bits | 16 bits | 32 bits | 48 bits |
| ITQ | Non-semantic | 0.4580 | 0.5794 | 0.6537 | 0.6902 |
| | Semantic | **0.5839** | **0.6867** | **0.7660** | **0.8064** |
| LSH | Non-semantic | 0.2801 | 0.4204 | 0.5017 | 0.5480 |
| | Semantic | **0.3060** | **0.4528** | **0.5609** | **0.6538** |
| COSDISH | Non-semantic | 0.6263 | 0.8147 | 0.8707 | 0.8873 |
| | Semantic | **0.6970** | **0.8924** | **0.9152** | **0.8907** |

Better scores are marked with bold font in the proposed method

**Table 5** Comparison on representations with semantics and without semantics on aYahoo dataset

| | | aYahoo Dataset | | | | |
|---|---|---|---|---|---|---|
| Method | | 4 bits | 8 bits | 12 bits | 16 bits | 20 bits |
| ITQ | Non-semantic | 0.8072 | 0.8371 | 0.8657 | 0.8871 | 0.8987 |
| | Semantic | **0.8532** | **0.8721** | **0.9064** | **0.9207** | **0.9282** |
| LSH | Non-semantic | 0.6082 | 0.7155 | 0.7387 | 0.7719 | 0.8348 |
| | Semantic | **0.7213** | **0.7201** | **0.8090** | **0.8181** | 0.7990 |
| COSDISH | Non-semantic | 0.3231 | 0.4629 | 0.4895 | 0.5466 | 0.5062 |
| | Semantic | **0.3893** | **0.4701** | **0.5146** | **0.5665** | **0.5539** |

Better scores are marked with bold font in the proposed method

Non-Semantic method by around 17.18, 11.80 and 5.11% for ITQ, LSH and COSDISH hashing methods over AWA dataset for 32 bits codes. The reason behind this is because the concepts are related, exploiting concept correlation relationship using semantics reduces the semantic gap. Thus, this experiment shows the effectiveness of our proposed semantic-based concept correlation scheme.

### 4.3.4 *Parameter sensitivity*

In this part, we empirically test the parameter sensitivity in our approach. Specifically, we evaluate the effects of λ on visual retrieval over AWA dataset. We tune λ in the range of [0.005, 0.01, 0.05, 0.1, 0.2, 0.5]. As we can see from Fig. 3, the sensitivity of the parameter depends on the hashing method our proposed framework adapts. For hashing methods

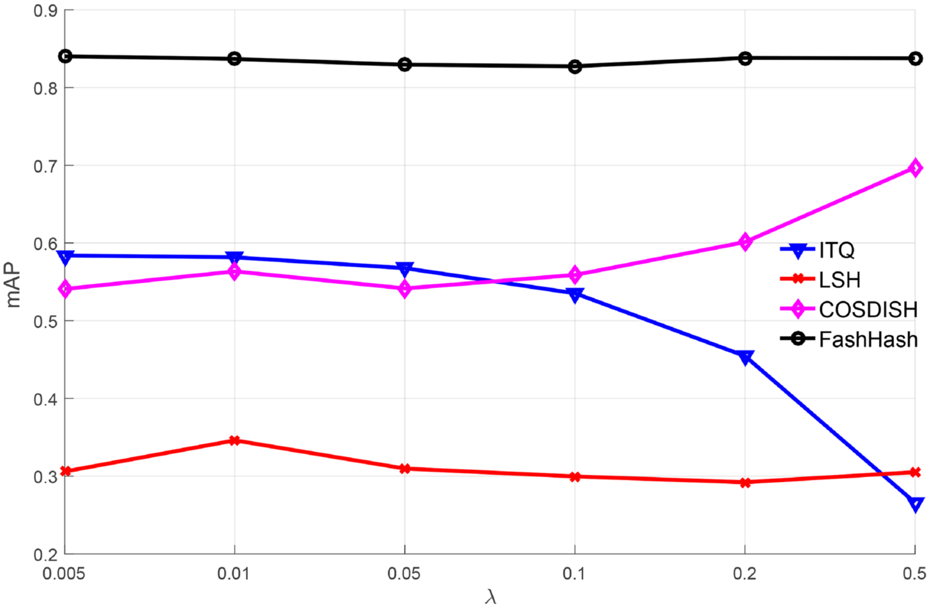

**Fig. 3** Evaluation of effects of λ on visual retrieval (mAP) over AWA dataset

FastHash and LSH, the parameter λ is nearly no influence on them. For COSDISH, the performance improves as λ rises from 0.1 to 0.5, while for ITQ, the performance decreases as λ rises from 0.1 to 0.5. This is because some hashing methods is sensible to the sparsity of the data while others are not.

## 5 Conclusion

In this paper, we introduced a novel binary visual representation learning framework to effectively exploit inter-concept correlation. Through the designed framework, we generated a concept-enriched dictionary and exploit concepts correlation by utilizing semantic attribute information. We proposed to impose overlapping group sparsity on the conceptual dictionaries, which achieves a better variable selection process. Also, we show the generalization ability of our framework by using different datasets in the off-line and on-line stages. In the future, we will further investigate an automatic way for group the dictionaries as well as exploit the inter-concept correlation.

**Acknowledgements** This work was supported in part by the National Science Foundation of China under Project 61572108, Project 61602089, Project 61502081, Project 61632007, and the Fundamental Research Funds for the Central Universities under Project ZYGX2014Z007, Project ZYGX2015J055 and the 111 Project No. B17008.

## References

1. Chiang C-K, Su T-F, Yen C, Lai S-H (2013) Multi-attributed dictionary learning for sparse coding. In: CVPR, pp 1137–1144
2. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: CVPR, vol 1, pp 886–893
3. Datar M, Immorlica N, Indyk P, Mirrokni VS (2004) Locality-sensitive hashing scheme based on p-stable distributions. In: SCG. ACM, pp 253–262
4. Farhadi A, Endres I, Hoiem D, Forsyth D (2009) Describing objects by their attributes. In: CVPR, pp 1778–1785
5. Farhadi A, Endres I, Hoiem D (2010) Attribute-centric recognition for cross-category generalization. In: CVPR, pp 2352–2359
6. Gao S, Chia L-T, Tsang IW-H (2011) Multi-layer group sparse coding—for concurrent image classification and annotation. In: CVPR, pp 2809–2816
7. Gong Y, Lazebnik S (2011) Iterative quantization: a procrustean approach to learning binary codes. In: CVPR, pp 817–824
8. Hu M, Yang Y, Shen F, Zhang L, Shen HT, Xuelong L (2017) Robust web image annotation via exploring multi-facet and structural knowledge. IEEE Trans Image Process 26(10):4871–4884
9. Hu M, Yang Y, Shen F, Xie N, Shen HT (2018) Hashing with angular reconstructive embeddings. IEEE Trans Image Process 27(2):545–555
10. Huang J, Liu H, Shen J, Yan S (2013) Towards efficient sparse coding for scalable image annotation. In: MM. ACM, pp 947–956
11. Jacob L, Obozinski G, Vert J-P (2009) Group lasso with overlap and graph lasso. In: ICML, pp 433–440
12. Kang W-C, Li W-J, Zhou Z-H (2016) Column sampling based discrete supervised hashing. In: AAAI, pp 1230–1236
13. Lampert CH, Nickisch H, Harmeling S (2009) Learning to detect unseen object classes by between-class attribute transfer. In: CVPR, pp 951–958
14. Li C, Feng Z, Han Y (2016) Image attribute learning with ontology guided fused lasso. Multimedia Tools Appl 75(12):7029–7043

15. Lin G, Shen C, Shi Q, van den Hengel A, Suter D (2014) Fast supervised hashing with decision trees for high-dimensional data. In: CVPR, pp 1963–1970
16. Liu J, Ji S, Ye J (2009) SLEP: sparse learning with efficient projections. Arizona State University
17. Liu W, Wang J, Ji R, Jiang Y-G, Chang S-F (2012) Supervised hashing with kernels. In: CVPR, pp 2074–2081
18. Lowe DG (1999) Object recognition from local scale-invariant features. In: ICCV, vol 2, pp 1150–1157
19. Luo Y, Yang Y, Shen F, Huang Z, Zhou P, Shen HT (2017) Robust discrete code modeling for supervised hashing. Pattern Recogn 75:128–135
20. Nie L, Yan S, Wang M, Hong R, Chua T-S (2012) Harvesting visual concepts for image search with complex queries. In: Proceedings of the 20th ACM international conference on multimedia, pp 59–68
21. Ouyang W, Li H, Zeng X, Wang X (2015) Learning deep representation with large-scale attributes. In: CVPR, pp 1895–1903
22. Raginsky M, Lazebnik S (2009) Locality-sensitive binary codes from shift-invariant kernels. In: NIPS, pp 1509–1517
23. Ri C, Yao M (2015) Bayesian network based semantic image classification with attributed relational graph. Multimedia Tools Appl 74(13):4965–4986
24. Shih TK (2002) Distributed multimedia databases: techniques and applications. IGI Global, Hershey
25. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
26. Tang J, Shao L, Li X (2014) Efficient dictionary learning for visual categorization. Comput Vis Image Underst 124:91–98
27. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol 73(3):273–282
28. Wang B, Yang Y, Xu X, Hanjalic A, Shen HT (2017) Adversarial cross-modal retrieval. In: ACM multimedia, pp 154–162
29. Wu L, Wang Y, Pan S (2016) Exploiting attribute correlations: a novel trace lasso-based weakly supervised dictionary learning method. IEEE Transactions on Cybernetics 47(12):4497–4508
30. Wu H, Yang Y, Xu X, Shen F, Xie N, Ji Y (2017) Exploiting concept correlation with attributes for semantic binary representation learning. In: ICIMCS
31. Xu X, Shen F, Yang Y, Shen HT, Li X (2017) Learning discriminative binary codes for large-scale cross-modal retrieval. IEEE Trans Image Process 26(5):2494–2507
32. Yan Y, Nie F, Li W, Gao C, Yang Y, Xu D (2016) Image classification by cross-media active learning with privileged information. IEEE Trans Multimedia 18(12):2494–2502
33. Yang Y, Yang Y, Huang Z, Shen HT, Nie F (2011) Tag localization with spatial correlations and joint group sparsity. In: CVPR, pp 881–888
34. Yang Y, Nie F, Xu D, Luo J, Zhuang Y, Pan Y (2012) A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. IEEE Trans Pattern Anal Mach Intell 34(4):723–742
35. Yang Y, Wu F, Nie F, Shen HT, Zhuang Y, Hauptmann AG (2012) Web and personal image annotation by mining label correlation with relaxed visual graph embedding. IEEE Trans Image Process 21(3):1339–1351
36. Yang Y, Zhang H, Zhang M, Shen F, Li X (2015) Visual coding in a semantic hierarchy. In: MM, pp 59–68
37. Yang Y, Zhang H, Zhang M, Shen F, Li X (2015) Visual coding in a semantic hierarchy. In: Proceedings of the 23rd ACM international conference on multimedia, MM '15, pp 59–68
38. Yang Y, Luo Y, Chen W, Shen F, Shao J, Shen HT (2016) Zero-shot hashing via transferring supervised knowledge. In: Proceedings of the 2016 ACM on multimedia conference, pp 1286–1295
39. Yang B, Gu C, Wu K, Zhang T, Guan X (2017) Simultaneous dimensionality reduction and dictionary learning for sparse representation based classification. Multimedia Tools Appl 76(6):8969–8990
40. Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B Stat Methodol 68(1):49–67
41. Zhang S, Huang J, Li H, Metaxas DN (2012) Automatic image annotation and retrieval using group sparsity. IEEE Trans Syst Man Cybern B Cybern 42(3):838–849
42. Zhang H, Zha Z, Yang Y, Yan S, Gao Y, Chua T (2013) Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In: ACM multimedia conference, MM '13, Barcelona, Spain, October 21–25, 2013, pp 33–42
43. Zhang H, Shen F, Liu W, He X, Luan H, Chua T (2016) Discrete collaborative filtering. In: ACM SIGIR, pp 325–334

**Xing Xu** received the B.E. and M.E. degrees from Huazhong University of Science and Technology, China, in 2009 and 2012, respectively, and the Ph.D. degree from Kyushu University, Japan, in 2015. He is currently a lecturer with the School of Computer Science and Engineering, University of Electronic of Science and Technology of China, China. His research interests include multimedia information retrieval and pattern recognition. He has served as a reviewer for IEEE TMM, IEEE TCSVT, PR, Neurocomputing and guest editor for Multimedia Tools and Application.