Semantic Middleware: Multi-Layer Abstract Semantics Inference for Object Categorization

Peng Liu, Zhipeng Ye, Wei Zhao, and Xianglong Tang
School of Computer Science and Technology, Harbin Institute of Technology, China

Abstract: *In this paper, we present a hierarchical model, named as Multi-layer Abstract Semantics Inference (MASI), based on Bag-of-Visual-Words (BoVW) to solve the problem of universal image categorization, including typical and zero-shot image categorization. An abstract hierarchical semantics learning method is proposed in the training step by extracting and selecting abstract visual words in a bottom-up way to train abstract semantic classifiers. For a testing image, its category is estimated layer-by-layer from top to bottom according to its corresponding hierarchical categories. Experimental results on popular image datasets have shown that the proposed method achieves better performance compared with traditional learning methods.*

Keywords: *Image categorization, zero-shot learning, semantic abstraction, BoVW.*

Received November 11, 2014; accepted December 21, 2015

1. Introduction

Object categorization, including object and scene classification and annotation, is a developing field in computer vision, which is also the precondition of scene interaction in artificial intelligence. This task involves many sub-tasks such as depth estimation, scene categorization, saliency detection, object detection, event categorization, etc. Nowadays, Bag-of-Visual-Words (BoVW) [4] is one of the most commonly used approaches in image retrieval and scenario classification, which simplicity and surprisingly effectiveness has been tested during these years. In BoVW, visual words are firstly obtained by k-means clustering local features. Then the image is represented by Bag-of-Features (BoF) to train classifier. However, it has three major drawbacks: First, quality of visual vocabulary is sensitive to dataset size [27]. Second, spatial relationships among image patches are ignored [17]. Moreover, the hard-assignment strategy of k-means does not necessarily generate optimized semantic visual words [25]. Several attempts have been made to improve the process of generating visual words, which can be classified as:

1. Eliminating the background which has nothing to do with visual words generation by segmentation based methods [3, 6].
2. Improvement on coding strategies to solve the problem that BoVW ignores spatial information between image patches [2, 15].
3. Introducing ambiguity of visual words that maps continuous image features to discrete visual words by using two or more visual words to describe one image features. Introducing ambiguity will dramatically increase description power of visual words [1, 8, 26].
4. Proposing semantic compression is proposed to improve efficiency and performance of BoVW by narrowing semantic gap [8, 11, 19, 21].

As pointed out in [5], BoVW faces the fact that knowledge generated from a small dataset does not necessarily hold when the content of datasets changes, which often happens in object categorization. To deal with this problem, several studies have been taken, including hierarchical-based and zero-shot learning methods. Hierarchical-based methods tend to build up higher-level semantic vocabularies to narrow semantic gap. Li and Perona [10] proposed a bayesian hierarchical model to represent training set by unsupervised learning technique [10, 13] Avrithis and Kalantidis [1] proposed a hierarchical-based classifier training method by decomposing the problem in several independent tasks. Lampert *et al.* [14] presented a cross-modal approach to extract semantic relationships between concepts using tagged images calculating a representative distribution of latent variables for each concept, which is suitable for concept clustering and image annotation. Zero-shot learning methods investigate the learning problems that each object category has zero training example, which can be treated as knowledge transfer problem. Palacutti *et al.* [21] proposed a knowledge based Semantic Output Code classifier (SOC) to solve the zero-shot learning problem that some categories in testing sets are omitted in training stage. Yu and Aloimonos [30] proposed an attribute-based transfer learning framework to solve both zero-shot and one-shot learning problem. Rohrbach *et al.* [23] evaluated three popular knowledge transfer methods on large-scale

dataset [22, 23] Krapac *et al.* [13] proposed an online incremental zero-shot learning method in which attribute labelling was obtained via online interaction with users. The method is based on indirect-attribute-predictionto learn new and update attributes while retaining as high accuracy as offline method.

In this paper, we present a hierarchical knowledge transfer model, named As Multi-layer Abstract Semantics Inference (MASI), to improve the performance of traditional BoVW for zero-shot learning problem. One of the related works is attribute-based learning method proposed by Li *et al.* [16], in which both Indirect Attribute Prediction (IAP) and Direct Attribute Prediction (DAP) classification strategies are introduced in detail. Our model is different from previous work [16] in two aspects: The first is that in [16], IAP and DAP are evaluated individually, while MASI contains two steps including training and testing. At the training stage, MASI uses IAP to generate semantic attributes. For the testing stage, MASI uses a structure combining DAP and flat classification. The second difference lies in the fact that the selection strategy of attributes. In [16], different parts of images are selected separately while MASI uses semantic visual words from whole image as one semantic feature to train classifiers. The proposed model contains three hierarchical structure, including layers of Upper-Abstract Semantics (UAS) and Middle-Abstract Semantics (MAS), and the concrete layer. Layers of UAS and MAS are introduced as knowledge between real-world and training image dataset to narrow semantic gap. Concrete layer is constructed by categories of image datasets. We explore on existing popular datasets, making it appropriate for zero-shot learning problem by proposing hierarchical structures. The rest of the paper is organized as follows: Section 2 introduces the MASI method. Section 3 shows the experimental results of MASI under various datasets. Finally, we present our conclusions in section 4.

2. Multi-Layer Abstract Semantics Inference Method

Traditional methods for image categorization assume that categories of objects in testing images exist in the training set, i.e., the classifier has "seen" the object before [23]. However, this assumption is very weak in practice since the categorization system is not able to imagine what kinds of input it may receive. This is the main reason that traditional methods are not appropriate for such tasks. In this paper, we extend BoVW by introducing the category inference process with techniques of abstraction for the purpose of knowledge transfer. We start with the description of architecture of MASI model in section 2.1. Then, we propose a semantics learning method in section 2.2.

2.1. MASI Model

Human beings try to distinguish objects by learning visual knowledge. For example, we will first categorize a hairy, long-ear and four-leg object as an animal, then refine the result as a cat. If we see an object with similar features next time, we will follow the same steps to infer its category. Although, there are many kinds of objects in the real world, they share a limited number of common attributes, which can be generated by techniques of abstraction [20]. In this paper, original categories of each image datasets are named as concrete categories. An abstract category is formed by some concrete categories that share common features provided by concept abstraction, which is manually done offline as prior information.

The structure of MASI is shown in Figure 1. Compared with original BoVW that directly training the classifier by concrete categories [16], we extend it by adding upper and middle abstract categories, which is constructed by semantic visual words extracted from concrete categories. From the figure we can see that MASI is a superset of BoVW, and if the abstract layers of MASI were omitted, the whole model would degrade into BoVW. According to the principle of abstraction, abstract level increases from bottom to top. Consequently, the descriptiveability increases while the differences between concrete categories are dimmed and the common attributes are preserved. In other words, the range of knowledge is extended, while the impact of knowledge from each concrete category decreases. For example, concrete category "chicken" is different from "sparrow", however, when the abstract level raises, we get abstract category "bird" that is able to describe both categories by aggregating their common features such as feather and wings, but the abstract category is mixed with 'diluted' knowledge from different concrete categories.

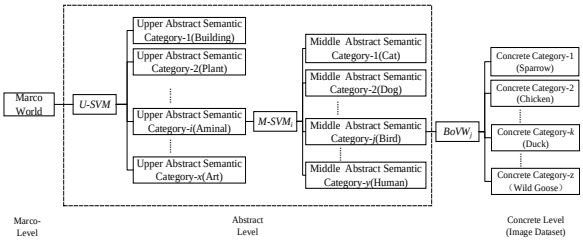


Figure 1. Architecture of MASI model.

The abstract layer works just like a middleware of knowledge: It connects between the real world and image datasets, which is independent from both parts to transfer knowledge. Information can be transmitted through the abstract layer through certain directions: Semantic visual words are transmitted from bottom to top, while categories of images are transmitted in the

reverse direction. Concrete classes under the same abstract class share common features, while concrete classes under different classes are much different. This leads to the fact that every concrete classes under the same abstract classes are similar yet distinguishable, the inner-class distances are small while the intra-class distancesare large, which are beneficial for classification.

The categorizing process is described below where F_i^u, F_j^m and F_k are the visual features of upper and middle abstract layers, *d* is the measurement function.

$$u = arg\ min\, d(F_t, F_i^u)\ \text{®}$$
$$m = arg\ min\, d(F_t, F_j^m)\ \text{®}$$
$$c = arg\ min\, d(F_t, F_k)$$

Figure 2 shows the comparative results between BoVW and MASI on zero-shot learning for better illustration. We can see that for a zero-shot learning problem, traditional BoVW outputs a completely non-relative result while MASI returns a result of the same upper abstract category.

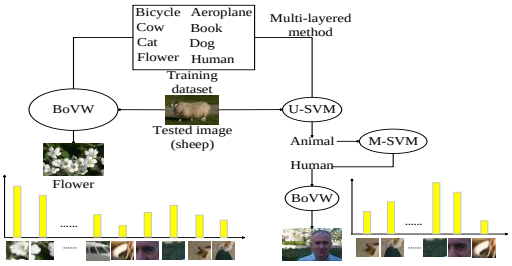


Figure 2. Example of zero-shot categorization.

2.2. Semantics Learning and Categorization

Visual Semantic Attribute (VSA) composed by semantic visual words is extracted to train classifiers of abstract layer. The inputs of concrete and abstract layers are concrete images collected from visual datasets. Semantic visual words are generated by Semantic Preserving BoW (SPBoW) [29] from concrete layer. The algorithm is given below. Classifiers are trained withone-vs-all strategy. CC_k is short for the k^{th} concrete category, MAC_j is short for the j^{th} middle abstract category, UAC_i is short for i^{th} upper abstract category stands for Semantic Visual Vocabulary Set (SVVS).

Algorithm 1: Generating SPBoW.

Preparation stage:

1. *For each* CC_k *under* MAC_j, *generate its SVVS* $Inh_k = \{(v_q, s_q)\}_{q=1}^{c}$, *where* v_q *and* s_q *are visual words and corresponding semantic information; c is the size of codebook.*
2. *Calculate SVVS of* MAC_j *under* UAC_i: $M\_A_i = \bigcup_{j=1}^{y} M_j$, *where* $M_j = \bigcup_{k=1}^{z} Inh_k^j$.
3. *For each* UAC_i, *randomly select SVVS with equal probability from each* Inh_k^j *to get the training set of the upper abstract* $category_{U\_A} = \bigcup_{i=1}^{x} U\_ABS_i$, $where_{U\_ABS_i} = \bigcup_{j=1}^{y}\bigcup_{k=1}^{z} Inh_k^j$.

Training stage:

4. *For every* UAC_j $TRAIN\ (BoVW_j, Inh_k^j)$.
5. *For every* UAC_i, $TRAIN\ (M\_SVM_i, M\_A_i)$.
6. $TRAIN\ (U\_SVM, U\_A)$.

In this paper, classifier is trained by images from concrete categories set *T* to categorize an image *I* with ground truth label C_i. If $C_i \in T$, it is considered as typical categorization, i.e., the testing and training samples are under same category distribution. Otherwise, it is considered as zero-shot categorization. We will evaluate their performance respectively with corresponding strategies.

For typical situation, the correct rate is evaluated by:

$$correct_{rate} = \frac{\sum_j \delta(C_j - C_i)}{N} \times 100\% \quad (1)$$

Where

$$\delta(C_j - C_i) = \begin{cases} 1, C_j = C_i \\ 0, otherwise \end{cases}$$

For zero-shot learning problem, according to our MASI model

$$correct_{rate} = \frac{\sum_j \delta(U_j - U_i)}{N} \times 100\% \quad (2)$$

Where $\delta(U_j - U_i) = \begin{cases} 1, U_j = U_i \\ 0, otherwise \end{cases}$.

The categorization process is achieved in a top-down way. First, for each *I*, the candidate values of upper and middle abstract hierarchies $p^{(u)}$ and $p^{(m)}$ are predicted by corresponding classifiers. Middle abstract category for next categorization is decided by the following criterion:

$$C_{middle} = \sum_{i=1}^{U}\sum_{j=1}^{M} \arg\max(P_i^{(u)} + P_j^{(m)}) \quad (3)$$

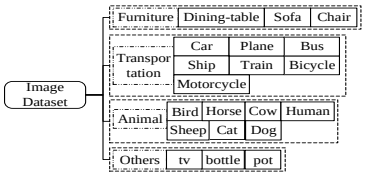
At last, *I* is passed through the BoVW classifier of concrete layer with *n* outputs p_1, p_2, ..., p_n, where *n* depends on the number of categories under each classifier. The concrete category of *I* is decided by the classifier that outputs the largest value:

$$C = arg_t \max_{t=1}^{n} p_t \quad (4)$$

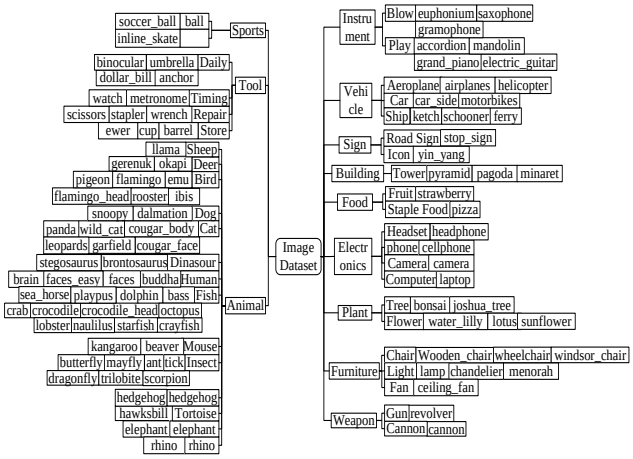
3. Experiments and Analysis

To evaluate the performance of the proposed model on typical categorization, two popular computer vision datasets are used, including PASCAL VOC 2007 [7] and caltech-101 [18]. PASCAL VOC 2007 has full object class annotation file for each image. It contains 20 categories with 9963 images. This dataset is more challenging due to the object's size and position are not always located in the middle of image. Caltech-101 contains 101 categories with 9197 images. The size of

each image is roughly 300×200. Outlines of each object are carefully annotated. Most images contain only one centred object, reducing the difficulty of object recognition. Most images contain only one major object with simple and uniform background. Inspired by previous researches [1, 19] hierarchical structures of PASCAL VOC 2007 and caltech-101 in this paper are shown in Figure 3. To our knowledge, the structure of Caltech-101 is firstly organized and proposed in detail. We utilize Library for Support Vector Machine (LIBSVM) in our experiments. We compare our method with other two methods: BoVW [4] and Locality-Constrained Linear (LLC) [27]. Mean Average Precision (MAP) is used to evaluate experimental results. The experimental results are given in Figure 4.

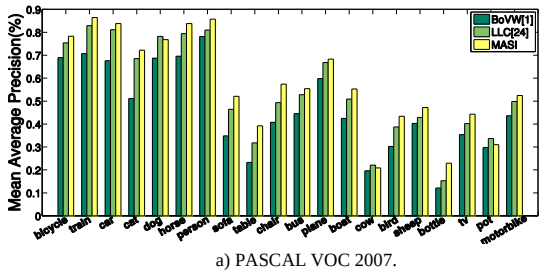


a) PASCAL VOC2007.

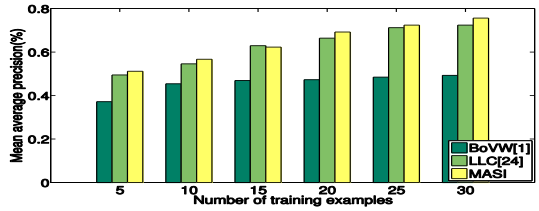


b) Caltech-101.

Figure 3. Hierarchical structures of datasets.



a) PASCAL VOC 2007.

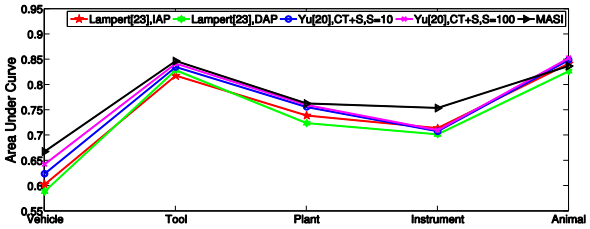


b) Caltech-101.

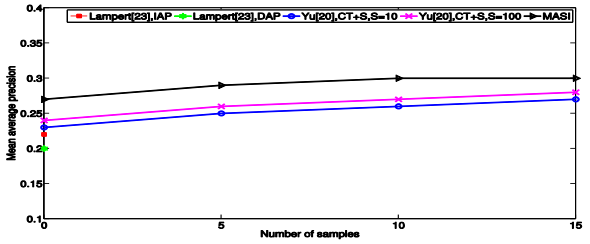
Figure 4. Typical categorization results.

In typical categorization tests, performances of MASI achieves substantial improvement over traditional non-hierarchical categorization methods on most tests, which agrees with the purpose of abstraction technique. It is due to the reason that, when building abstract semantic visual vocabularies, MASI does not exclude ambiguity of visual words [26]. Mean while, MASI randomly selects lower semantic visual words with equal probability constructing upper semantics to make sure each semantics could have a chance to be selected training a much more balanced classifier that enhances the generalized ability of each abstract category. Last but equally important, MASI introduces middle abstract category to further narrow semantic gap between visual words and image patches.

In zero-shot tests, we compare MASI with other zero-shot learning methods[16, 29], on animals with attributes [16] and hierarchical caltech-101 datasets. "Animals with Attributes" dataset includes 30475 images from 50 animal categories, and 85 attributes to describe these categories. All images are resized such that the longest side has 300 pixels. For each images, we extract SIFT feature to construct semantic visual words as described in [28]. We implement [29]. With parameters of CT+S, S=10 and CT+S, S=100. DAP and IAP are implemented respectively for [16]. We choose five upper abstract categories for the consideration of implementation on knowledge transfer since all of them contains more than one concrete categories. The overall performance of zero-shot testing results on both datasets are shown in Figure 5. For animals with attributes, we evaluate the result with MAP as adopted in [29], while for caltech-101, we evaluate the result by area under curve.



a) Caltech-101.



b) Animal with Attributes.

Figure 5. Zero-shot testing results.

The zero-shot testing results show that on most tests, MASI achieves better performance over existing zero-shot learning methods. This is due to the reason that MASI utilizes semantic vocabulary to store and

transfer knowledge. High-level semantic features are used to narrow semantic gaps between image features and semantic features. Meanwhile, descriptive ability of visual vocabulary for categories are also increased. Another improvement over existing learning methods is the combination of DAP and IAP meta-strategies in the whole model for training and testing, which makes MASI more flexible and robust to training and testing samples.

4. Conclusions

We introduced a highly abstract and easy-to-extend MASI model to deal with typical and zero/one-shot object categorization problem. Abstract semantics are extracted from low level features to learn hierarchical semantic classifiers. The advantages include:

1. Increasing the descriptive ability of model to real world.
2. Better performance in both typical and zero-shot object categorization. We have demonstrated the capabilities of our model on popular computer vision datasets. Future work includes optimizing the structure of MASI model so that there could be a fault tolerant mechanism to further improve categorization results.

Acknowledgments

This research is supported by NSFC program, China (61171184，61201309).

References

[1] Avrithis Y. and Kalantidis Y., "Approximate Gaussian Mixtures for Large Scale Vocabularies," *in Proceeding of European Conference on Computer Vision*, Florence, pp. 15-28, 2012.

[2] Bannour H. and Hudelot C., "Hierarchical Image Annotation Using Semantic Hierarchies", *in Proceeding of the 21st ACM International Conference on Information and Knowledge Management*, Maui, pp. 2431-2434, 2012.

[3] Bolovinou A., Pratikakis I., and Perantonis S., "Bag of Spatio-Visual Words for Context Inference in Scene Classification," *Pattern Recognition*, vol. 46, no. 3, pp. 1039-1053, 2013.

[4] Chai Y., Rahtu E., Lempitsky V., Gool L., and Zisserman A., "Tricos: A Tri-Level Class-Discriminative Co-Segmentation Method for Image Classification," *in Proceeding of European Conference on Computer Vision*, Florence, pp. 794-807, 2012.

[5] Csurka G., Dance C., Fan L., Willamowski J., and Bray C., "Visual Categorization with Bags of Keypoints," *in Proceeding of ECCV Workshop on Statistical Learning in Computer Vision*, Meylan, pp. 1-22, 2004.

[6] Deng J., Berg A., Li K., and Li F., "What does Classifying more than 10,000 Image Categories Tell us?," *in Proceeding of European Conference on Computer Vision*, Crete, pp. 71-84, 2010.

[7] Du R., Wu Q., He X., and Yang J., "Object Categorization Based on a Supervised Mean Shift Algorithm," *in Proceeding of Workshops and Demonstrations ECCV*, Florence, pp. 611-614, 2012.

[8] Everingham M., Van L., Williams C., Winn J., and Zisserman A., "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303-338, 2010.

[9] Fei-Fei L., Fergus R., and Perona P., "Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories," *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59-70, 2007.

[10] Fei-Fei L. and Perona P., "A Bayesian Hierarchical Model for Learning Natural Scene Categories," *in Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 524-531, 2005.

[11] Fernando B., Fromont E., Muselet D., and Sebban M., "Supervised Learning of Gaussian Mixture Models for Visual Vocabulary Generation," *Pattern Recognition*, vol. 45, no. 2, pp. 897-907, 2012.

[12] Gemert J., Veenman C., Smeulders A., and Geusebroek J., "Visual Word Ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1271-1283, 2010.

[13] Ji C., Zhou X., Lin L., and Yang W., "Labeling Images by Integrating Sparse Multiple Distance Learning and Semantic Context Modeling," *in Proceeding of European Conference on Computer Vision*, Florence, pp. 688-701, 2012.

[14] Kankuekul P., Kawewong A., Tangruamsub S., and Hasegawa O., "Online Incremental Attribute-Based Zero-Shot Learning," *in Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3657-3664, 2012.

[15] Katsurai M., Ogawa T., and Haseyama M., "A Cross-Modal Approach for Extracting Semantic Relationships between Concepts Using Tagged Images," *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 1059-1074, 2014.

[16] Krapac J., Verbeek J., and Jurie F., "Modeling Spatial Layout with Fisher Vectors for Image Categorizationk," *in Proceeding of IEEE International Conference on Computer Vision*, Barcelona, pp. 1487-1494, 2011.

[17] Lampert C., Nickisch H., and Harmeling S., "Attribute-Based Classification for Zero-Shot

Visual Object Categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453-465, 2014.

[18] Lazebnik S., Schmid C., and Ponce J., "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," *in Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, New York, pp. 2169-2178, 2006.

[19] Li L., Wang C., Lim Y., Blei D., and Fei-Fei L., "Building and Using a Semantivisual Image Hierarchy," *in Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3336-3343, 2010.

[20] Liu J., Yang Y., and Shah M., "Learning Semantic Visual Vocabularies Using Diffusion Distance," *in Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 461-468, 2009.

[21] Palatucci M., Pomerleau D., Hinton G., and Mitchell T., "Zero-Shot Learning with Semantic Output Codes," *in Proceeding of the 22nd International Conference on Neural Information Processing Systems*, Vancouver, pp. 1410-1418, 2009.

[22] Penatti O., Silva F., Valle E., Gouet-Brunet V., and Torres R., "Visual Word Spatial Arrangement for Image Retrieval and Classification," *Pattern Recognition*, vol. 47, no. 2, pp. 705-720, 2014.

[23] Rohrbach, M., Stark M., and Schiele B., "Evaluating Knowledge Transfer and Zero-Shot Learning in a Large-Scale Setting," *in Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1641-1648, 2011.

[24] Saitta L. and Zucker J., *Abstraction in Artificial Intelligence and Complex Systems*, Springer, 2013.

[25] Subramanian M. and Sathappan S., "An Efficient Content Based Image Retrieval Using Advanced Filter Approaches," *The International Arab Journal of Information Technology*, vol. 12, no. 3, pp. 229-236, 2015.

[26] Van J., Snoek C., Veenman C., Smeulders A., and Geusebroek J., "Comparing Compact Codebooks for Visual Categorization," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 450-462, 2010.

[27] Varma M. and Zisserman A., "A Statistical Approach to Texture Classification from Single Images," *International Journal of Computer Vision*, vol. 62, no. 1, pp. 61-81, 2005.

[28] Wang J., Yang J., Yu K., Lv F., Huang T., and Gong Y., "Locality-Constrained Linear Coding for Image Classification," *in Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3360-3367, 2010.

[29] Wu L., Hoi S., and Yu N., "Semantics-Preserving Bag-of-Words Models and Applications," *IEEE Transactions on Image Processing*, vol. 19, no. 7, pp. 1908-1920, 2010.

[30] Yu X. and Aloimonos Y., "Attribute-Based Transfer Learning for Object Categorization with zero/one Training Example," *in Proceeding of European Conference on Computer Vision*, Heraklion, pp. 127-140, 2010.

Zhipeng Ye Ph.D. candidate at the School of Computer Science and Technology, Harbin Institute of Technology. He receives Master degree of computer application technology of Harbin Institute of Technology in 2013. His research interest covers image processing and machine learning.)

Peng Liu Associate professor at the School of Computer Science and Technology, Harbin Institute of Technology. He receives Doctoral degree of microelectronics and solid state electronics of Harbin Institute of Technology in 2007. His research interest covers image processing, video processing, pattern recognition and design of very large scale integrated circuit.)

Zhao Wei Associate professor in the Pattern-Recognition Research Center, School of Computer Science and Technology. She receives Doctoral degree of computer application technology of Harbin Institute of Technology in 2006. Her research interest covers image pattern recognition, image processing, character recognition, computer, remote sensing image explain.)

Xianglong Tang Professor at the School of Computer Science and Technology. He receives Doctoral degree of computer application technology of Harbin Institute of Technology in 1995. His research interest covers pattern recognition, aerospace image processing, medical image processing, machine learning. Corresponding author of this paper.