ICIC Express Letters
Part B: Applications
Volume **8**, Number **6**, June **2017**

ICIC International ©2017 ISSN 2185-2766
pp. 989–997

IMAGE RETRIEVAL WITH EXTENDED ATTRIBUTES BASED ON WEB SEARCH AMOUNT

Fen Zhang[1], Xiangwei Kong[1], Ze Jia[2] and Fei Ning[3]

[1]School of Information and Communication Engineering
Dalian University of Technology
No. 2, Linggong Road, Dalian 116024, P. R. China
dlut_zhangfen@mail.dlut.edu.cn; kongxw@dlut.edu.cn

[2]Unit 91439 of PLA
Dalian 116041, P. R. China
11346282@qq.com

[3]China Ship Development and Design Center
Shanghai 221108, P. R. China
701sh@701sh.com

Received September 2016; accepted December 2016

Abstract. *Existing attribute-based image retrieval approaches restrict the users to utilize only the pre-labeled attributes for searching the desired targets. To compensate for this, we focus on extended attributes learning in this paper. Firstly, we introduce both Wiktionary and WordNet as external lexical semantic resources to learn corresponding extended attributes of the pre-defined attributes. After that, we consider the web search amount obtained from both Baidu Index and Google Trends during a specified period as user preference, to get rid of some not commonly used words. As a result, one can use not only the pre-labeled attributes, but also the extended attributes to retrieve the intended images. Experiments on several attribute benchmarks demonstrate significant performance improvements over several state-of-the-art methods.*
Keywords: Extended attribute, Image retrieval, Semantic relation measure, User preference

1. **Introduction.** In the recent years, the idea of "attribute" has drawn much attention in the computer vision community. Attributes are considered as an expressive middle layer which plays an important role in bridging the gap between low level features and high level semantic concepts. So far, a large number of related work referring to attributes have been done ranging from attribute learning [1] to various applications such as object recognition [2], object classification [3], and image retrieval [4-7].

In this paper, we focus on attribute based image retrieval. Currently, most attribute based image retrieval systems limit the users to select query attributes simply from the pre-defined attribute set, which is extremely inconvenient when the attribute changes. Farhadi et al. [2] introduced a novel feature selection method for learning attributes and built independent classifiers for each attribute. The end result was the summation of all individual classifiers, which ignored the correlations among the query attributes. Siddiquie et al. [4] analyzed the dependencies between different query attributes and leveraged such multi-attribute interdependence to allay the noises generated from the classifiers. However, the approach proposed in [4] relied only on the pre-labeled query attributes to build the dependency model, which is insufficient in forming an expressive feature space. An alternative method was proposed in [5], which introduced a middle layer named weak attribute for large-scale image retrieval. Although the dimensionality of weak attributes is much higher than that of the pre-labeled query attributes, which

ensures the weak attribute space to be expressive enough, it makes no contribution to expanding the set of the actually used query attributes.

Liu et al. [6] proposed to learn extended attributes for image retrieval, but it has two defects: (i) it seemed simplex to learn extended attributes merely based on WordNet; (ii) it considered the polysemy count of candidate attributes in the WordNet as user familiarity, which may be biased. For example, *shoes* is an extended attribute of the pre-labeled attribute *foot*, and its polysemy count is low, but we are, to be honest, very familiar with this word. Accordingly, we first propose to exploit another external lexical semantic resource of Wiktionary to enrich the attribute representations, and combine the result with that of WordNet [6]. Wiktionary is a large collaboratively-constructed online dictionary, which provides a larger and more up-to-date vocabulary than WordNet. So far, Wiktionary has played an important role in semantic similarity measurement [8, 9]. After that, the web search amount obtained from both Baidu Index and Google Trends is utilized as user preference, to reject some not commonly used words. Baidu index is a data-sharing platform based on massive user behavior, through which one can study the search trends of keywords, see the interests and needs of users, etc. [10]. Google Trends is a public web facility based on Google Search that allows the users to compare the volume of searches between two or more terms [11]. Hence, the web search amount is more comprehensive and practical than the polysemy count.

In such a scenario, users can utilize not only the original pre-labeled attributes, but also the learned extended attributes to retrieve all the related images that best match the multi-attribute queries. As shown in Figure 1, attributes such as *jet engine* and *fly* can be used to search images of an airplane, while *fly* is an extended attribute corresponding to the pre-defined attribute *wing*. The light-colored arrows represent the process related to the extended attribute *fly*, while the deep-colored arrows represent that related to the pre-labeled attribute *jet engine*.

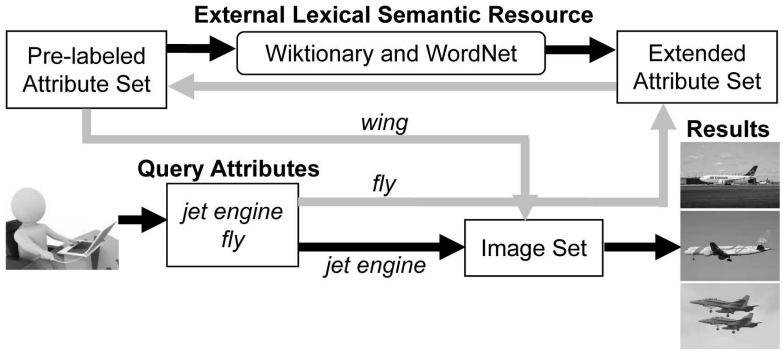


FIGURE 1. Retrieval process based on extended attributes

The organization of this paper is as follows: Section 2 introduces the retrieval model based on extended attributes; Section 3 describes the proposed method for semantic relation (SR) measure. Section 4 demonstrates the experimental results and comparisons over several state-of-the-art approaches. Section 5 presents the conclusion and future work.

2. **Retrieval Model Based on Extended Attributes.**

2.1. **Retrieval.** Let $X_q \in \mathbb{R}^m$ be the pre-labeled query attribute set and $X_e \in \mathbb{R}^n$ the extended attribute set, and the complete attribute set X is then defined as their concatenation, i.e., $X = \{X_q, X_e\}$, where $X \in \mathbb{R}^{m+n}$. Given a multi-attribute query Q, where $Q \subset X$, and the set of images Y, our goal is to select a subset of images $y^* \subset Y$ that are

most relevant to Q. Therefore, the prediction function $f_\omega : Q \to y$ returns the set y^* as the structured response to the given Q:

$$y^* = \arg\max_{y \subset Y} \omega^T \psi(Q, y) \tag{1}$$

We define $\omega^T \psi(Q, y)$ as:

$$\omega^T \psi(Q, y) = \sum_{x_i \in Q} \sum_{x_j \in X} \omega_{ij} \phi(x_j, y) \tag{2}$$

where

$$\phi(x_j, y) = \begin{cases} \sum\limits_{y_k \in y} \varphi(x_j, y_k), & \text{where } x_j \in X_q \\ \sigma_{jq} \sum\limits_{y_k \in y} \varphi(x_q, y_k), & \text{where } x_j \in X_e \text{ and } x_q \Rightarrow x_j \end{cases} \tag{3}$$

Here, $\varphi(x_j, y_k)$ is the feature vector that indicates the presence of attribute x_j in image y_k. We set $\varphi(x_j, y_k)$ to be the output of an independently trained attribute detector. However, we are interested in learning the interdependency model ω on the entire attribute set X, not just within the pre-defined query attribute set itself. For the pre-labeled attributes $x_j \in X_q$, Equation (2) represents the weighted contribution to each query attribute $x_i \in Q$. While for the extended attributes $x_j \in X_e$, we adopt a simpler method that defines the output as a weighted score of corresponding pre-labeled attributes, as shown in the second part of Equation (3). $x_q \Rightarrow x_j$ denotes the pre-labeled attribute working in concert with x_j, and σ_{jq} is a parameter that weighs the semantic relation between x_q and x_j. Accordingly, both the pre-labeled attributes and their semantically related extended attributes are used in interdependency model learning and further for image retrieval.

2.2. **Training.** Given a set of multi-label training images Y_{tr}, as well as their ground truth pre-labeled query attribute set X_q and corresponding extended attribute set X_e, we are dedicated to learning the dependency model ω, in order that the retrieved image set y^* has the highest score of all $y \subset Y_{tr}$ for each query $Q \subset X$, where $X = \{X_q, X_e\}$. This can be performed through a standard max-margin training formulation:

$$\begin{aligned} &\arg\min_{\omega, \xi} \qquad \omega^T \omega + C \sum_t \xi_t \\ &\forall\, t \quad \omega^T \psi(Q_t, y_t^*) - \omega^T \psi(Q_t, y_t) \geq \Delta(y_t^*, y_t) - \xi_t \end{aligned} \tag{4}$$

where C is a parameter controlling the trade-off between the training error and regularization, ξ_t is the slack variable corresponding to the training query Q_t, and $\Delta(y_t^*, y_t)$ is the loss function, which can be defined on different performance metrics. In this paper, hamming loss is used as the representation of $\Delta(y_t^*, y_t)$:

$$\Delta(y_t^*, y_t) = 1 - \frac{|y_t \cap y_t^*| + |\bar{y}_t \cap \bar{y}_t^*|}{|Y_{tr}|} \tag{5}$$

In order to solve the quadratic optimization problem of Equation (4), we resort to the cutting plane method [12], which consists of starting with no constraints and iteratively adding the most violated constraint for the current solution of the optimization problem. The most violated constraint at each iteration is given by:

$$\xi_t \geq \max_{y_t \subset Y_{tr}} \left[\Delta\left(y_t^*, y_t\right) - \left(\omega^T \psi\left(Q_t, y_t^*\right) - \omega^T \psi(Q_t, y_t) \right) \right] \tag{6}$$

which can be solved in $O\left(|Y_{tr}|\right)$ with hamming loss.

3. **Extended Attributes Learning Based on Web Search Amount.** Given a set of pre-labeled attributes $X_q \in \mathbb{R}^m$, we aim at learning the most related extended attribute set $X_e \in \mathbb{R}^n$ from both Wiktionary and WordNet. In this work, we adapt the semantic relation (SR) measure of [6] for WordNet and [8] for Wiktionary respectively, to learn candidate extended attributes. Then relative average retrieval amount between predefined attributes and their corresponding candidate extended attributes obtained from both Baidu Index [10] and Google Trends [11] during a specified period, is exploited to measure user preference.

Given a pre-labeled attribute $x \in X_q$, [6] proposed to use synonyms S_x, coordinate terms C_x, explanations E_x and derivations D_x in the WordNet to form the candidate set F_x:

$$F_x = \{S_x, C_x, E_x, D_x\} \tag{7}$$

Hence, our goal is to find a subset of words $W \subset F_x$ which have the most close relations with x.

Given a word $w_i \in W$ and its corresponding set of senses Sen_{w_i}, as well as the senses set Sen_x of x, the similarity score between w_i and x is then defined as follows:

$$Sim^1(x, w_i) = MS(x, w_i) + MS(w_i, x) \tag{8}$$

where

$$MS\,(x, w_i) = \frac{\sum\limits_{s_m \in Sen_x} \max\limits_{s_n \in Sen_{w_i}} S(s_m, s_n)}{|Sen_x| + |Sen_{w_i}|} \tag{9}$$

$$MS\,(w_i, x) = \frac{\sum\limits_{s_n \in Sen_{w_i}} \max\limits_{s_m \in Sen_x} S(s_n, s_m)}{|Sen_x| + |Sen_{w_i}|} \tag{10}$$

$|Sen_x|$ and $|Sen_{w_i}|$ are the number of senses for x and w_i, respectively. $S(s_m, s_n)$ is the similarity between the two senses s_m and s_n, which is represented as:

$$S(s_m, s_n) = S(s_n, s_m) = \frac{1}{R_{s_m} R_{s_n}} \frac{Q(S) + Q(C) + Q(E) + Q(D)}{Q(F_{s_m}) + Q(F_{s_n})} \tag{11}$$

where

$$Q(S) = \sum_{q_i \in S_{s_m} \cap S_{s_n}} \lambda_S \; idf(q_i)^2 \tag{12}$$

$$Q(F_{s_m}) = \sum_{q_i \in F_{s_m}} \lambda_1 \; idf(q_i)^2 \tag{13}$$

R_{s_m} is the rank of sense s_m in word x, while R_{s_n} is the rank of sense s_n in word w_i. $Q(S)$ explicitly measures the similarity between s_m and s_n on their synonyms, with $q_i \in S_{s_m} \cap S_{s_n}$ meaning the presence in the synonyms of both s_m and s_n. λ_S is a parameter controlling the weight of the synonym set S_x, and $idf(q_i)$ is the inverse document frequency of q_i acquired from WordNet. $Q(C)$, $Q(E)$ and $Q(D)$ are similar functions corresponding to C_x, E_x and D_x respectively. $Q(F_{s_m})$ is the summation of the entire candidate set of s_m, where $F_{s_m} = \{S_{s_m}, C_{s_m}, E_{s_m}, D_{s_m}\}$ and $\lambda_1 \in \{\lambda_S, \lambda_C, \lambda_E, \lambda_D\}$.

For Wiktionary, it is more appropriate to employ the concept vector (CV) based measure proposed in [8] to compute SR. Concept vector based approaches represent a word in a document vector space. Given a pre-labeled attribute x, the meaning of x is represented as a high dimensional concept vector $\vec{v}(x) = (v_1, \ldots, v_n)$, where n is the number of documents. The value of v_j depends on the occurrence of the word x in the document d_j. If the word x can be found in the document, the word's tf.idf score in the document d_j is assigned to the CV element v_j. Otherwise, v_j is 0. As a result, the vector $\vec{v}(x)$ represents the word x in a concept space. Similarly, we can acquire the concept vector $\vec{v}(w_i)$ of the

candidate word w_i. The SR of x and w_i can then be computed as the cosine of their concept vectors [9].

$$Sim^2(x, w_i) = \frac{\vec{v}(x) \cdot \vec{v}(w_i)}{\|\vec{v}(x)\| \, \|\vec{v}(w_i)\|} \tag{14}$$

Thus, the total similarity score between x and w_i can be represented as follows:

$$Sim(x, w_i) = Sim^1(x, w_i) + Sim^2(x, w_i) \tag{15}$$

If the candidate attribute w_i is absent in WordNet but present in Wikitionary, then $Sim^1(x, w_i) = 0$, vice versa.

As there are some words not commonly used by users (e.g., users tend to use common words such as *head* or *brain* to retrieve their desired images, rather than the not commonly used words *caput*), we have to take user preference into account. To this end, we propose to exploit the relative average retrieval amount obtained from both Baidu Index and Google Trends from September 2015 to August 2016, to measure user preference. Observations suggest that the effects of individual data floating on the statistical properties could be negligible, as the overall trend of average retrieval amount of the attributes is basically consistent in both Baidu Index and Google Trends. For Baidu Index, letting N_x be the average retrieval amount of the pre-labeled attribute x, and N_{w_i} be that of the corresponding candidate attribute w_i, the relative average retrieval amount between x and w_i is then defined as:

$$RAR_1(x, w_i) = \frac{N_{w_i}}{N_x} \tag{16}$$

While for Google Trends, the relative average retrieval between x and w_i can be represented as:

$$RAR_2(x, w_i) = \frac{M_{w_i}}{M_x} \tag{17}$$

where M_x and M_{w_i} are the average retrieval amount of x and w_i respectively. As a result, we can remove those not frequently used words. The final score of w_i corresponding to pre-labeled attribute x is then denoted as:

$$Score(w_i) = Sim(x, w_i) + \beta \left(RAR_1(x, w_i) + RAR_2(x, w_i)\right) \tag{18}$$

where β is a trade-off parameter controlling the semantic similarity and user preference.

4. **Experiments and Results.**

4.1. **Experiment setups.**

We perform experiments on two well known datasets: a-Pascal and a-Yahoo [2]. a-Pascal dataset contains 12695 images (6340 images for training and 6355 images for testing) collected from the PASCAL VOC 2008 challenge[1]. Images in a-Pascal dataset are divided into 20 categories, such as aeroplane, and bicycle. And each image has been labeled with a set of 64 pre-labeled attributes, for example, *jet engine*, and *window*. a-Yahoo dataset includes 2644 images for 12 object classes, collected from the Yahoo image search[2] and considered as a supplement of a-Pascal dataset. Images in a-Yahoo are described by the same set of 64 pre-labeled attributes, but with different category labels. As same setting as that in [2], we use the pre-defined training images of a-Pascal for training and the remaining images (including the pre-defined testing images of a-Pascal and all the images of a-Yahoo) for testing. Each image is represented as a 9751 dimensional feature vector referring to color, texture, visual words, and edges.

The structural SVM [13] used in this paper is based on its matlab wrapper[3], under the 1 slack formulation. Our implementation of extended attributes mining in Wikitionary

[1]http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008/

[2]http://vision.cs.uiuc.edu/attributes/

[3]http://www.vlfeat.org/vedaldi/code/svm-struct-matlab.html

is based on a Java-based API called JWKTL[4], and that in WordNet is based on the Java Word Net Library (JWNL1.4)[5]. We learn extended attributes for each pre-labeled attribute in the training phase. The weights used for measuring semantic similarity in Equation (11) are $\lambda_S = 1.5$, $\lambda_C = 1$, $\lambda_E = 0.5$, $\lambda_D = 1$. Moreover, according to massive experiments and personal experience, we discover that $\beta = 0.7$ in Equation (18) gives the best result. We evaluate our performance by the standard mean AUC, which is frequently used to measure performance in binary classification cases.

4.2. **Acquisition of extended attributes.** We learn a set of extended attributes for each of the 64 pre-labeled attributes in the training set. Integrating the result of Wiktionary with that of WordNet results in 528 candidate extended attributes for all the 64 pre-labeled attributes. By rejecting the repeated items, we can gain 281 extended attributes in the end. Moreover, we note that there are certain extended attributes corresponding to several pre-labeled attributes, such as *bicycle⇒pedal* and *handlebars*, *horse⇒rein* and *saddle*. In such *one-vs-many* scenarios, we learn separate SVM detectors for certain extended attributes using training images related to all of their corresponding pre-labeled attributes.

Five randomly selected pre-labeled attributes and their corresponding extended attributes are shown in Table 1. The results indicate that the extended attributes are closely correlated to the pre-labeled attributes in semantic space, e.g., *fly⇒wing* and *sea⇒sail*. Moreover, the semantic relation between the pre-labeled attributes has been reinforced by sharing the same extended attributes, such as *wing* and *jet engine* linking to the same extended attribute *airplane*.

Table 1. Extended attributes of five randomly selected pre-labeled attributes

Pre-labeled Attributes	**Extended Attributes**
Face	facial, visage, appearance, cheek
Window	windowed, glass, windowpane, windowglass
Wing	fly, flying, airfoil, airplane, plane, aeroplane
Hair	hairy, haired, head
Sail	boat, sailing, vessel, canvas, sea, ship

Table 2. Results of relative average retrieval amount

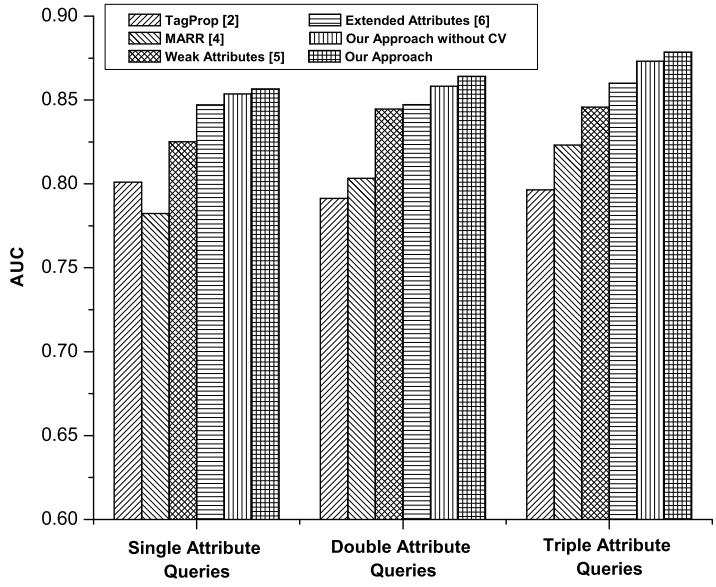
Pre-labeled Attributes	**Extended Attributes**					
Face	facial	visage	appearance	cheek		
	0.413	0.102	0.613	0.260		
Window	windowed	glass	windowpane	windowglass		
	0.501	0.782	0.392	0.682		
Wing	fly	flying	airfoil	airplane	plane	aeroplane
	1.949	0.453	0.068	0.546	1.488	0.554
Hair	hairy	haired	head			
	0.624	0.047	1.546			
Sail	boat	sailing	vessel	canvas	sea	ship
	1.023	0.766	0.782	1.310	1.318	1.521

[4]http://dumps.wikimedia.org/

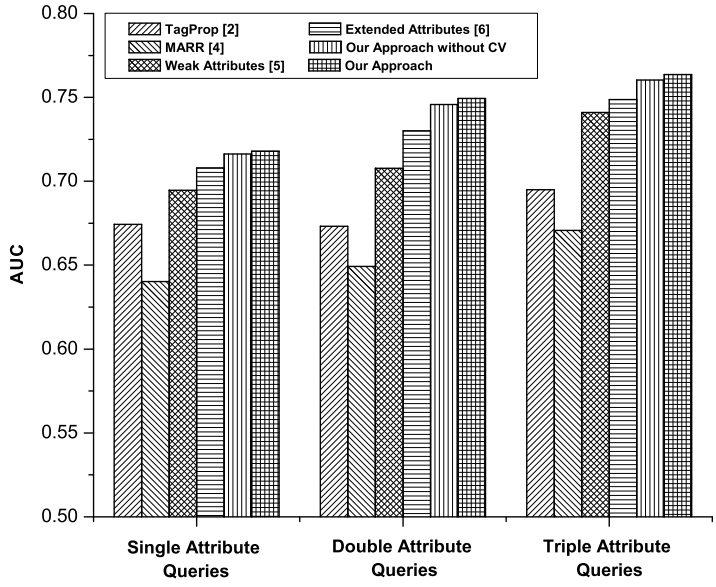
[5]http://cogcomp.cs.illinois.edu/m2repo/net/didion/jwnl/1.4_rc3/

4.3. **Results of relative average retrieval amount.** In this paper, we introduce the relative average retrieval amount between the pre-labeled attribute and its candidate attributes as the measurement of user preference, which are based on the data information of Baidu Index and Google Trends from September 2015 to August 2016. Table 2 shows the results of relative average retrieval amount of five randomly selected pre-labeled attributes and their corresponding extended attributes. We choose the average value of results from Baidu Index and Google Trends as the final results.

4.4. **Performance comparisons and retrieval results.** Figure 2 shows the performance comparisons of our approach to several existing methods, including TagProp [2],



(a) Comparisons on a-Pascal



(b) Comparisons on a-Yahoo

FIGURE 2. Comparisons of retrieval performance on a-Pascal and a-Yahoo datasets respectively

FIGURE 3. Top-10 results of [6] and our approach on a-Pascal and a-Yahoo datasets respectively

MARR [4], Weak Attributes [5] and Extended Attributes [6], on both a-Pascal and a-Yahoo datasets. Results of the top four methods are copied from [6], under the same configurations compared to ours, as well as with the optimal sparsity ($k = 400$). And our approach without CV means that we measure SR in Wiktionary based on the method applied in [6], involving synonyms, coordinate terms, explanations and derivations. The CV based SR measure employed in this paper works better, due to the use of all relation types offered by Wiktionary.

As shown in Figure 2(a), our approach outperforms the other methods for all kinds of queries on a-Pascal, especially with a large margin for double and triple queries. This proves the improvement of the accuracy of SR measure by introducing Wiktionary and web search amount, which in turn contributes to a better performance. Figure 2(b) shows the results of our method compared with other approaches on a-Yahoo. It is discovered that our approach can also leverage the cross-dataset information, and still achieves the best performance. However, because of the reduction of correlation, the entire performance on a-Yahoo descends.

Examples of retrieval results on both a-Pascal and a-Yahoo datasets are shown in Figure 3, where the images with a border mean false positive. Experiments conducted on a-Pascal dataset aim to retrieve bicycles over the queries "handle", "pedal" and "wheel", and experiments on a-Yahoo are designed to retrieve carriages over the queries "wooden" and "wheel".

5. **Conclusions.** Note that nearly all of the existing attribute-based image retrieval methods are implemented merely on the pre-labeled attributes, which is inadequate for constantly changing circumstances and inconvenient for large scale databases. Accordingly, we propose a novel method to learn extended attributes, so as to enrich the representations of the pre-labeled attributes. In this paper, we utilize both Wiktionary and WordNet as external lexical semantic resource to mine extended attributes, and adopt web search amount from both Baidu Index and Google Trends as user preference. Extensive experiments have been carried out on a-Pascal and a-Yahoo datasets, demonstrating the superiority of our proposed approach. We tend to study more accurate representations of attributes and explore other linguistic resources (e.g., Wikipedia and WWW) in the future work.

Acknowledgment. This work is supported by the Foundation for Innovative Research Groups of the NSFC (Grant No. 71421001), the NSFC (Grant No. 61172109), the Fundamental Re-search Funds for the Central Universities DUT14QY03 and the Open Projects Program of National Laboratory of Pattern Recognition (No. 201407349).

REFERENCES

[1] S. Huang, M. Elhoseiny, A. Elgammal and D. Yang, Learning hypergraph-regularized attribute predictors, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.409-417, 2015.

[2] A. Farhadi, I. Endres, D. Hoiem and D. Forsyth, Describing objects by their attribute, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.1778-1785, 2009.

[3] J. Zhang, X. Gao and X. Du, Bayesian classifying algorithm of continuous attributes based on weighted error rates, *ICIC Express Letters*, vol.10, no.7, pp.1643-1648, 2016.

[4] B. Siddiquie, R. S. Feris and L. S. Davis, Image ranking and retrieval based on multi-attribute queries, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.801-808, 2011.

[5] F. X. Yu, R. Ji, M.-H. Tsai, G. Ye and S. F. Chang, Weak attributes for large-scale image retrieval, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.2949-2956, 2012.

[6] Y. Liu, X. Kong, H. Fu, X. You and Y. Guo, Model semantic relations with extended attributes, *International Conference on Pattern Recognition*, pp.2549-2554, 2014.

[7] K.-H. Liu, T.-Y. Chen and C.-S. Chen, MVC: A dataset for view-invariant clothing retrieval and attribute prediction, *Proc. of the 2016 ACM on International Conference on Multimedia Retrieval*, pp.313-316, 2016.

[8] T. Zesch, C. Müller and I. Gurevych, Using wiktionary for computing semantic relatedness, *Proc. of the 23rd National Conference on Artificial Intelligence*, vol.2, pp.861-866, 2008.

[9] M. T. Pilehvar and R. Navigli, From senses to texts: An all-in-one graph-based approach for measuring semantic similarity, *Artificial Intelligence*, vol.228, pp.95-128, 2015.

[10] *Baidu Index*, http://index.baidu.com, 2016.

[11] *Google Trends*, http://www.google.com/trends, 2016.

[12] I. Tsochantaridis, T. Joachims, T. Hofmann and Y. Altun, Large margin methods for structured and interdependent output variables, *Journal of Machine Learning Research*, vol.6, pp.1453-1484, 2005.

[13] T. Joachims, T. Finley and C.-N. J. Yu, Cutting-plane training of structural SVMs, *Machine Learning*, vol.77, no.1, pp.27-59, 2009.