Image Retrieval using Categories, Attributes, and Locations

Xingxing Wei[1], Xiaojie Guo[2], Yahong Han[1]
[1]School of Computer Science and Technology, Tianjin University.
[2]State Key Laboratory of Information Security, IIE, CAS.
{xwei, yahong}@tju.edu.cn, guoxiaojie@iie.ac.cn

1 Introduction

Image retrieval plays an important role in enabling people to easily access to the desired images. A variety of retrieval methods have been developed. The input is of various forms such as text [11], image [5], and sketch [1, 2], to represent the query. However, these query strategies can only express the users' search intention partially. For instance, given the query image in Figure 1, users may be interested in the regions of "car", "tree", and "building" (highlighted with the rectangular boxes). These regions attract users' attention because the "car" is above the "building" and "tree" in the vertical direction, which is not coincident with the usual scenario. This abnormalism may appear in many images in the internet, such as photos of the violent conflict, pictures in the stricken area, etc. In some situations, users need to search more images with such abnormal object layout. Furthermore, in real world, each object has its own attribute informations (such as "red car", "green tree", and "white building" etc). Thus, in order to express the exact search intentions, it would be better to also specify the semantic attributes of each object, such as the color attributes, size attributes, and material attributes etc. However, all these mentioned semantic cues cannot be appropriately incorporated into the textual queries, exemplar images, color or concept maps [12, 8], and even the sketches. To bridge the "semantic gap" between users' search intention and the low-level visual features [10], we should augment the image search strategies to enable users to indicate their Regions Of Interest (ROIs) within the query image, and can handle these high-level semantic concepts as well as the spatial relations. What's more, because the number of ROIs within an image may be more than two, the framework also should deal with the high-order case. i.e., has the ability to retrieval multiple ROIs within an image. In this paper, we propose a novel strategy to improve users' search experience. The interface we provide to users only requires users to indicate the ROIs in a selected exemplar query image. The users' search intention is automatically refined and specified by our proposed method. The flowchart is illustrated in Figure 1.

In summary, the contribution of this paper is two-fold: 1) We present a novel image search strategy that not only allows users to indicate their ROIs, but also can properly handle various high-level semantic queries and spatial relations. 2) We propose a structured descriptor to jointly represent the categories, attributes, and spatial relationships among objects, and design a ranking method to accomplish the image retrieval. The rest of this paper is organized as follows. In Section 2, we detail the proposed framework. Section 3 reports the experimental results, and Section 4 gives the conclusion.

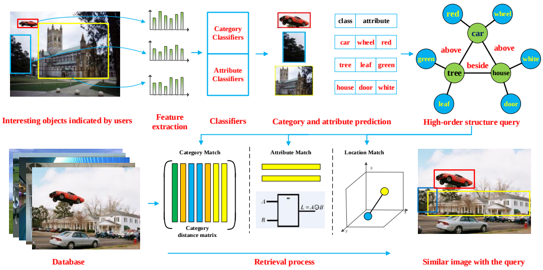


Fig. 1. Illustration of our framework. Different from [7], ours is an unsupervised framework. Moreover, ours aims at image retrieval, while [7] aims to generate sentences to describe images.

2 Proposed Framework

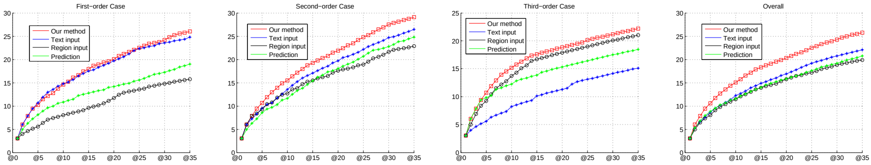
To represent the images, we first use [4] to detect the bounding boxes for objects. Then, the category and attribute labels of these boxes are jointly predicted by our another work [6]. In this way, the categories, attributes and spatial relations of objects in large scale database are easily obtained. Suppose there are F instances within an image labeled by R different category labels ($R \leq F$), each instance is assigned a category label ranging from 1 to Q and some attribute labels ranging from 1 to M. To encode the spatial interactions between two instances, we compute the location prior between each category pair via statistically analyzing the training set. For example, the prior may be “sky” should be above “building”. Based on this prior, we compare the location of two instances, and design a $F \times F$ matrix. If they satisfy the prior, the corresponding location in the matrix is set 1, otherwise, the value is set to -1. If two instances have the same category labels, the value is 0. A matrix is corresponding to a fixed location relationship. If users want to add a new location relationship (for example, the horizontal relationship), they need to compute the prior in horizontal direction, and then use another matrix to represent it.

We select arbitrary three instances with different category labels to construct a triangle. In the triangle, each vertex point is associated with an instance $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_q$, and the corresponding value indicates its category label (i, j, q, and $i \neq j \neq q$). As a result, it will produce totally $\binom{R}{3}$ different triangles (when $R \geq 3$). If $R < 3$, we add 0 to the corresponding vertex, and still use a triangle to represent it. Note because a category label may be assigned to multiple instances, the number of triangles within an image will be above $\binom{R}{3}$. In addition, we assign a M dimensional binary vector to each vertex to represent its corresponding attributes, where 1 denotes the instance has this attribute, and 0 denotes not. If the value of the vertex is 0, we assign M zeros to it.

Now we introduce how to rank the images based on the constructed triangles. We define a category distance matrix to represent the semantic correlation between categories. Specifically, we use Eq.1 to compute the correlation between two categories:

$$\theta_{ij} = \log \frac{P_{00}P_{11}}{P_{10}P_{01}}, \tag{1}$$

where P_{ij} denotes the probability when i=$\{0,1\}$, and j=$\{0,1\}$. By using Eq. 1, we obtain a $Q \times Q$ matrix. If two categories are relevant, the corresponding value in the



(a) (b) (c) (d)

Fig. 2. Performance curves. (a), (b), (c) list the comparisons between different methods under the first-order, second-order, and third-order cases. The average score under all cases is illustrated in (d). “Prediction” denotes the output using inferred category and attribute labels by our method.

matrix will be large (In the diagonal, the value is largest, because the category is the most relevant with itself). Based on this matrix, we could search the most relevant category according to a given category. In this way, the objects between two triangles will be assigned. We use the category matched score S_t to represent the matched degree between the query triangle and the *t*-th triangle in dataset. S_t is defined by the number of objects that are exactly correctly matched with the query triangle.

Next, we compute the attribute matched score between the assigned objects. Suppose $\mathbf{y}_i$, $\mathbf{y}_j$ and $\mathbf{y}_q$ are the attribute vectors of the query objects, and $\mathbf{y}'_i$, $\mathbf{y}'_j$ and $\mathbf{y}'_q$ are the attribute vectors for the assigned objects.

$$R_t = \sum([\mathbf{y}_i, \mathbf{y}_j, \mathbf{y}_q]^T \odot [\mathbf{y}'_i, \mathbf{y}'_j, \mathbf{y}'_q]^T), \qquad (2)$$

where $[\mathbf{y}_i, \mathbf{y}_j, \mathbf{y}_q]$ denotes concentrating the $\mathbf{y}_i, \mathbf{y}_j, \mathbf{y}_q$ into a long vector. $\odot$ denotes the xnor operator. Eq.2 computes the similarity of the attributes between query triangle and triangles in dataset, and R_t denotes the attribute matched score.

Consequently, we compute the spatial matched score. Suppose the spatial vector of query triangle is $\mathbf{z}_p$, and the spatial vector of the *t*-th triangle is $\mathbf{z}_t$. $\mathbf{z}_p$ and $\mathbf{z}_t$ both compose of 1 and -1. We use $Q_t = \|\mathbf{z}_i - \mathbf{z}_j\|$ to compute the spatial matched score. The final matched score of the *t*-th triangle is obtained by:

$$F_t = \sum_{j=1}^{n} \theta_j \psi_j(I, q), \qquad (3)$$

where ψ_j is a metric between the image in the database and the query image. Here, we only explore S_t, R_t, and Q_t introduced above. θ_i are the weights. Using F_t, we rank the triangles in the dataset, and search the most similar images with the query image.

3 Experiments

Two datasets: aPascal [3] and aYahoo datasets [3] are used here. The aPascal dataset contains 20 categories and 64 attributes, covering 4340 images. The aYahoo dataset has 2237 images, including 12 categories and 64 attributes. We use the same low-level features as in Farhadi *et al* [3], which results in a 9751 dimensional feature vector. Some volunteers are recruited to label the ground truth. We assign a three-level relevance score

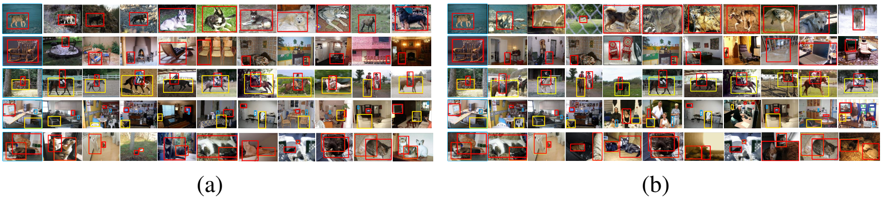


Fig. 3. Five examples output by our method. (a) and (b) are the results using ground-truth and inferred category and attribute labels, respectively.

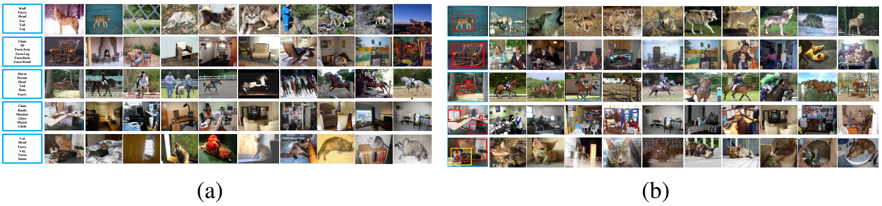


Fig. 4. Five examples output by text and RBIR methods versus the same query used in Figure 3.

to each image. Level 3 corresponds to the most relevant, and level 1 denotes the least relevant. The Discounted Cumulative Gain (DCG) is used to measure the performance.

We compare two state-of-the-art methods: text-input image method and Region-Based Image Retrieval (RBIR) [9]. The quantitative experimental results are given in Figure 2. From the figure, we see our method achieves the similar performance with the text-input method under the first-order case. This is not difficult to explain, since our model degenerates into the text-input model under this case (i.e., searching images only according to the category and attribute labels). In the second-order and third-order case, with adding the spatial relationship, the performance of text-input method drops, and our method gradually outperforms it. This demonstrates the fact that text-input methods can not handle the tasks with spatial relationship. In addition, we see RBIR works poorly under the first-order and second-order case, this is because RBIR ranks the images according to the similarity using the low-level features. Therefore, it can not handle the high-level semantic information (such as attributes), resulting in a poor performance. From (d), we see our method achieves the best average performance. For using inferred labels (green curve), we see the performance falls somewhere between the text-input model and RBIR. Figure 3 and Figure 4 show the qualitative results.

4 Conclusions

In this paper, we have presented the image retrieval problem that specified Regions Of Interest (ROIs). In the ROIs, various high-level semantic concepts like categories of objects, their attributes, and the spatial relationship between them were jointly considered to accomplish the search. Experiments conducted on two benchmark datasets showed that our method achieved the best performance compared with the state of the arts.

References

1. Chen, T., Cheng, M.M., Tan, P., Shamir, A., Hu, S.M.: Sketch2photo: internet image montage. TOG 28(5), 124 (2009)
2. Eitz, M., Hildebrand, K., Boubekeur, T., Alexa, M.: Sketch-based image retrieval: Benchmark and bag-of-features descriptors. TVCG 17(11), 1624–1636 (2011)
3. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR. pp. 1–8 (2009)
4. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. TPAMI 32(9), 1627–1645 (2010)
5. Gordoa, A., Rodríguez-Serrano, J.A., Perronnin, F., Valveny, E.: Leveraging category-level labels for instance-level image retrieval. In: CVPR. pp. 3045–3052 (2012)
6. Han, Y., Wei, X., Cao, X., Yang, Y., Zhou, X.: Augmenting image descriptions using structured prediction output. TMM, doi:10.1109/TMM.2014.2321530
7. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Baby talk: Understanding and generating simple image descriptions. In: CVPR. pp. 1601–1608 (2011)
8. Lan, T., Yang, W., Wang, Y., Mori, G.: Image retrieval with structured object queries using latent ranking svm. In: ECCV, pp. 129–142 (2012)
9. Liu, Y., Zhang, D., Lu, G.: Region-based image retrieval with high-level semantics using decision tree learning. Pattern Recognition 41(8), 2554–2570 (2008)
10. Liu, Y., Zhang, D., Lu, G., Ma, W.Y.: A survey of content-based image retrieval with high-level semantics. Pattern Recognition 40(1), 262–282 (2007)
11. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: ICCV. pp. 1470–1477 (2003)
12. Xu, H., Wang, J., Hua, X.S., Li, S.: Image search by concept map. In: SIGIR. pp. 1–8 (2010)