Augmented Image Retrieval using Multi-Order Object Layout with Attributes

Xiaochun Cao[1,2], Xingxing Wei[1,*], Xiaojie Guo[2], Yahong Han[1], Jinhui Tang[3]
[1]School of Computer Science and Technology, Tianjin University.
[2]State Key Laboratory of Information Security, IIE, CAS.
[3]School of Computer Science and Engineering, Nanjing University of Science and Technology.
{caoxiaochun, guoxiaojie}@iie.ac.cn, {xwei, yahong}@tju.edu.cn, jinhuitang@mail.njust.edu.cn

ABSTRACT

In image retrieval, users' search intention is usually specif ed by textual queries, exemplar images, concept maps, and even sketches, which can only express the search intention partially. These query strategies lack the abilities to indicate the Regions Of Interests (ROIs) and represent the spatial or semantic correlations among the ROIs, which results in the so-called semantic gap between users' search intention and images' low-level visual content. In this paper, we propose a novel image search method, which allows the users to indicate any number of Regions Of Interest (ROIs) within the query as well as utilize various semantic concepts and spatial relations to search images. Specif cally, we f rstly propose a structured descriptor to jointly represent the categories, attributes, and spatial relations among objects. Then, based on the def ned descriptor, our method ranks the images in the database according to the matching scores *w.r.t.* the category, attribute, and spatial relations. We conduct the experiments on the aPascal and aYahoo datasets, and experimental results show the advantage of the proposed method compared to the state of the arts.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing; Information Search and Retrieval.

Keywords

Image retrieval; Attribute; Region of interest; Object layout

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 conf ict, pictures in the stricken area, etc. In these situations, users need to search more such images. 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 ref ned and specif ed by our proposed method. In particular, after obtaining the ROIs, the system will infer their categories, attributes, and spatial relationships. With the inference, we consequently constitute a high-order semantic structure to represent the query and then retrieve the similar images from the database. The f owchart is illustrated in Figure 2.

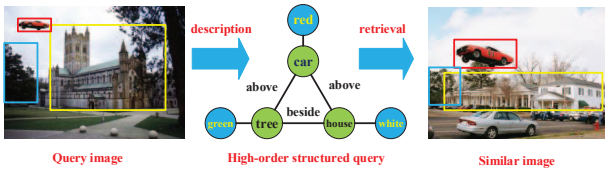


Figure 1: Given the query image, users may be interested in the object layout highlighted with the rectangle boxes. The object layout can be depicted by using categories of objects (car, tree, house), their attributes (red, green, white), and spatial interactions between them (above, beside). This paper aims at retrieving the images having the same layout, including the same categories, attributes, and spatial interactions.

*Corresponding author

MM'14, November 03 - 07 2014, Orlando, FL, USA

http://dx.doi.org/10.1145/2647868.2654972.

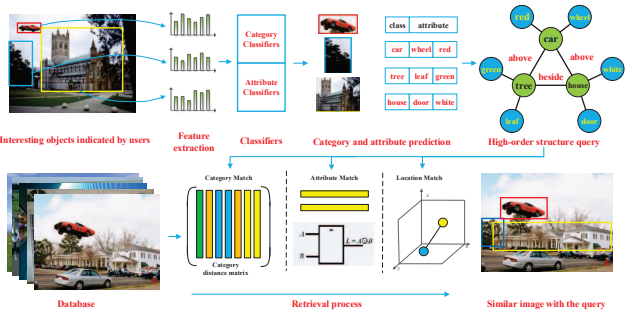


Figure 2: Illustration of our framework. Note that our method is an unsupervised framework, which is different from the work in [7]. In their work, the CRF is used to construct a similar triple with ours. Moreover, ours aims at image retrieval, while [7] aims to generate sentences to describe images.

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.

2. PROPOSED FRAMEWORK

2.1 Structured Description for Images

To represent the images, we first train the classifiers in terms of category and attribute using SVM. When users indicate their ROIs, the low-level features of the corresponding bounding boxes are extracted. These features are input to the SVM classifiers to infer the category and attribute labels of the ROIs.

To encode the spatial interactions between two instances, we propose a matrix-based representation method. 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 the attribute label ranging from 1 to M. Thus, a $F \times F$ matrix is established. We then 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 next compare the location of two instances. 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. The 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 corresponding prior in horizontal direction, and then use another matrix to represent it.

We select arbitrary three instances that have 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 to vary with the values in vertex ($R \geq 3$). 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 the attributes of the corresponding instance, where 1 denotes the instance has this attribute, and 0 denotes the instance does not have this attribute. If the value of vertex is 0, we assign M zeros to it.

2.2 Image Ranking Based on Triangles

We use the above triangle to represent each image in the dataset. In this way, the image retrieval problem is converted into a triangle matching problem. In order to solve this problem, we first define a category distance matrix to represent the semantic correlation between each two categories. Specifically, we use Eq.1 to compute the correlation strength 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 the above formula, we obtain a $Q \times Q$ matrix. If two categories are relevant, the corresponding value in the matrix will be large (In the diagonal, the value is largest, because the category is the most relevant with itself). Based on the proposed category distance 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.

After the above steps, we assign the objects of query triangle with the objects of the triangles in dataset. Next, we compute the attribute matched score R_t between the corresponding 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.

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

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 the dataset.

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 the following formula to compute the spatial matched score: $Q_t = \|\mathbf{z}_i - \mathbf{z}_j\|$. In order to incorporate S_t, R_t, and Q_t to represent the match degree of images, we normalize S_t, R_t, and Q_t. The final matched score of the *t*-th triangle is:

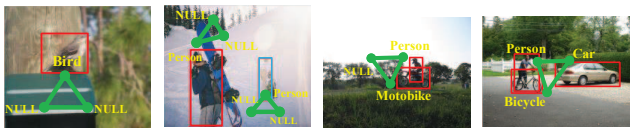
$$F_t = \sum_{j=1}^{n} \theta_j \psi_j(I, q), \tag{3}$$

where ψ_j is one metric between the image in the database and the query image. In this work, we only explore the S_t, R_t, and Q_t introduced above. $\theta_i, i = 1, ..., n$ are weight parameters. According to F_t, we can rank the triangles in the dataset, and search the most similar images with the query image.

3. EXPERIMENTS

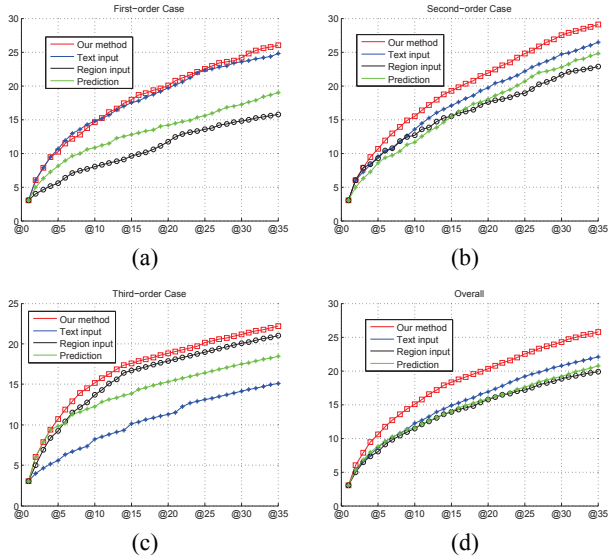
3.1 Dataset

Two benchmark datasets: aPascal dataset [3] and aYahoo dataset [3] are used in our experiments. aPascal is built based on the Pascal VOC 2008 dataset by assigning the *attribute* labels to each image within it. There are totally 4340 images in the dataset, which are split around 50% train/val and 50% test. In our experiment, we use the 4340 images as our search database, and random select some



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

Figure 3: Four example triangles are illustrated. In each triangle, if the number of different categories is less than 3, we use "NULL" to replace the corresponding nodes (as shown in (a) and (c). If there are multiple identical instances, we use multiple triangles to represent them (as shown in (b)).



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

Figure 4: Performance curves. (a), (b), (c) list the comparisons between different methods under the first-order, second-order, and third-order cases, respectively. The average score under all cases is illustrated in (d). "Prediction" denotes the output using inferred category and attribute labels by our method.

images as the queries to accomplish our task. The aPascal dataset contains 20 categories and 64 attributes, covering 10,363 annotated objects. As for aYahoo dataset, there are totally 2237 images downloaded from the internet by using Yahoo search engine. 12 categories and 64 attributes are used to label the 2267 objects. The detailed names of corresponding categories and attributes can be found in [3]. To fairly compare the different retrieval methods, in our experiment, we assume the ground truth bounding box of each object have been detected perfectly in both the aPascal and aYahoo datasets, and conduct our experiments on this assumption.

We use the same low-level features as in Farhadi *et al* [3]. The base features in [3] contain four types: color, texture, visual words and edges. They use a bag of words style feature for each feature type, which results in a 9751 dimensional feature vector.

3.2 Evaluation Metric

To perform the quantitative evaluation, some volunteers are recruited to label the ground truth. Given a specified query, the image retrieval system ranks the images in the database. For each image, we assign a relevance score to it according to how well the image accords with the search intention of the task. In our experiment, the relevance score is defined in three levels from level 1 to level 3.

Figure 5: Five examples output by our method. The first row lists the query image, where the indicated ROIs are highlighted with the colorful bounding boxes (one color corresponds to one kind of object). The following ten rows denote the top ten images ranked by our method.

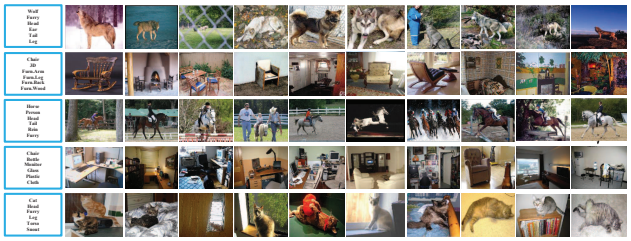


Figure 6: Five examples output by text-input method in terms of the same query used in Figure 5.

Level 3 corresponds to the most relevant (all indicated categories appear, and the object layout is right), and level 1 denotes the least relevant (some categories are missed or the layout of objects is quite different from the query). Level 2 is similar with Level 1, but some attributes are missed. With the ground truth, the Discounted Cumulative Gain (DCG) is used to measure the performance:

$$DCG@p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2 i}, \quad (4)$$

where rel_i is the relevance score of the image ranked at the i-th place. p is the depth of DCG, and DCG@p represents the DCG score when we select the top p ranked images.

3.3 Experimental Results

Some examples of the structured description are given in Figure 3. We compare two state-of-the-art methods: text-input image search engine and Region-Based Image Retrieval (RBIR) engine [9]. For text-input image search engine, we use the ground-truth category and attribute labels of the ROIs as the input keywords to search the images. The input of region-based image search engine is the same as our method, i.e., the ROIs indicated by users. For the methods above, we both use the ground-truth category and attribute labels. To fairly compare the performance of different image retrieval algorithms, we first use the ground-truth category and attribute labels to accomplish the retrieval task (as done in the text-input and region-based methods), and then give the performance using the inferred category and attribute labels. 20 image search tasks are designed to evaluate the proposed system. In these tasks, 10 tasks involve only one object (first-order case), 6 tasks involve two objects (second-order case), and 4 tasks involve three objects (third-order case). The second-order and third-order cases contain the attribute information and the spatial relationship between objects, while the first-order case only contains the attribute information. The quantitative experimental results are given in Figure 4.

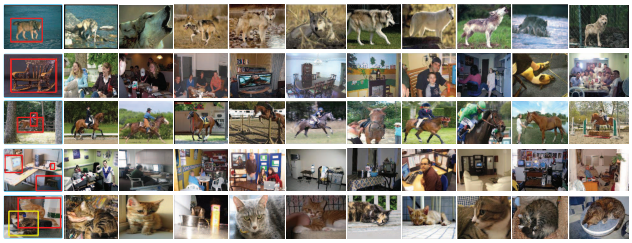


Figure 7: Five examples output by region-based method in terms of the same query used in Figure 5.

Figure 8: Five examples output by our method. The difference from Figure 5 is that the inferred category and attribute labels are used here.

The comparisons between different methods under first-order case, second-order case and third-order case are listed in (a), (b) and (c), respectively. The overall comparison under all cases is given in (d). 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 the region-based method works poorly under the first-order and second-order case, this is because RBIR ranks the images according to the similarity between 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 performance. The performance of our method using the inferred category and attribute labels (both for query and database) is also given in the green curve. We see the performance falls somewhere between the text-input model and region-based method.

For the qualitative comparison, we show the ranked results output by our method, text-input image search engine, and region-baed image search engine in Figure 5, Figure 6, and Figure 7, respectively. In each figure, the first row lists the query image, where the indicated ROIs are highlighted with the colorful bounding boxes (one color corresponds to one kind of object). The following ten rows denote the top ten images ranked by the different methods. The task in the first column is conducted on the aYahoo dataset, and the tasks in the following four columns are conducted on the aPascal dataset. From top to bottom are two first-order cases, one second-order case, one third-order case, and one case that has two same objects within an image. We see our method can retain the spatial relationship between objects. For example, for the task in the third column, the ROIs can be depicted using “person" is above “horse". We see the all the searched images contain the ROIs that “person" is above “horse". For the task in the forth column, the ROIs is depicted using “monitor" is above “bottle", and “bottle" is above “chair". Similarly, the images having the same layout are also searched by our method. In contrast, the text-input and region-based methods can not retain the spatial relationship. We show the qualitative results based on the inferred category and attribute labels in Figure 8. From the figure, we see the categories of some objects are wrong, limiting the performance of our method. Finding out more accurate annotation algorithm will promote our method.

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. Our method can also be used in the large scale image retrieval, where the object bounding boxes can be automatically detected using [4], and the categories and attributes of these boxes can be jointly predicted by our another work [6]. In this way, the categories, attributes and spatial relations of objects in large scale database can be easily obtained.

5. ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (No. 61332012), 100 Talents Programme of The Chinese Academy of Sciences, and the Opening Project of State Key Laboratory of Digital Publishing Technology. X. Guo was supported by Excellent Young Talent of the Institute of Information Engineering, Chinese Academy of Sciences.

6. REFERENCES

[1] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu. Sketch2photo: internet image montage. *TOG*, 28(5):124, 2009.

[2] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *TVCG*, 17(11):1624–1636, 2011.

[3] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, pages 1–8, 2009.

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010.

[5] A. Gordoa, J. A. Rodríguez-Serrano, F. Perronnin, and E. Valveny. Leveraging category-level labels for instance-level image retrieval. In *CVPR*, pages 3045–3052, 2012.

[6] Y. Han, X. Wei, X. Cao, Y. Yang, and X. Zhou. Augmenting image descriptions using structured prediction output. *TMM, doi:10.1109/TMM.2014.2321530*.

[7] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, pages 1601–1608, 2011.

[8] T. Lan, W. Yang, Y. Wang, and G. Mori. Image retrieval with structured object queries using latent ranking svm. In *ECCV*, pages 129–142. 2012.

[9] Y. Liu, D. Zhang, and G. Lu. Region-based image retrieval with high-level semantics using decision tree learning. *Pattern Recognition*, 41(8):2554–2570, 2008.

[10] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, 2007.

[11] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003.

[12] H. Xu, J. Wang, X.-S. Hua, and S. Li. Image search by concept map. In *SIGIR*, pages 1–8, 2010.