Title

Retrieval of Images with Objects of Specific Size, Location and Spatial Configuration

ISBN

9781479966820

Authors

Pourian, Niloufar

Manjunath, BS

Publication Date

2015

DOI

10.1109/wacv.2015.133

Peer reviewed

Retrieval of images with objects of specific size, location and spatial configuration

Niloufar Pourian B.S. Manjunath
Department of Electrical and Computer Engineering
University of California, Santa Barbara, United States
npourian@ece.ucsb.edu manj@ece.ucsb.edu

Abstract

An approach to image retrieval using spatial configurations is presented. The goal is to search the database for images that contain similar objects (image-patches) with a given configuration, size and position. The proposed approach consists of creating localized representations robust to segmentation variations, and a sub-graph matching method to compare the query with the database items. Localized object representations are created using a community detection method that groups visually similar segments. Extensive experimental results on three challenging datasets are provided to demonstrate the feasibility of the approach.

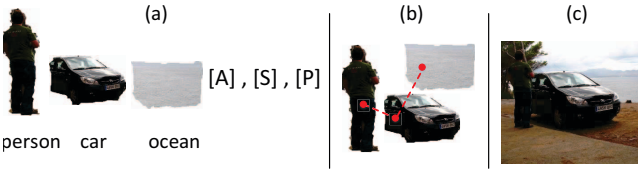

Figure 1: (a) Example of images/images patches provided by the user. Matrix $[A]$ is the associated adjacency matrix that denotes the desired spatial relationship between the provided patches, matrix $[S]$ represents the desired size of each image-patch, and matrix $[P]$ represents the desired position of each image-patch. (b) The system automatically creates a graphical query representation for the image-patches based on matrices $[A]$, $[S]$, and $[P]$. (c) An ideal retrieved image corresponding to the specified configuration, size, and position is depicted.

1. Introduction

Searching for images with a specific visual content has been a topic of intense research in recent years [3]. However, majority of the prior work focuses on searching for images using global (whole image) attributes, and lacks discrimination based on localized objects or their relative spatial positioning in the images. We consider here a more generalized problem in which the objects of interests are provided by a query set that includes multiple image-patches or images along with the desired spatial configuration, size, and location of such patches in the target image.

Our goal is to develop localized representations that would enable queries similar to the one shown in Figure 1. Here the user provides objects/image-patches, together with specifications on their spatial configuration, size and position in the image. From such a specification, an image query is generated and matched against the database. To achieve this goal, localized representations are needed. We propose a robust graph-based representation that is learned from image-part groupings, and encodes size, location, and spatial configurations of objects/patches. Sub-graph matching is used to search and retrieve similar configurations.

Conventional methods usually represent a query (for instance a "dog") through a single image or a set of images, possibly along with some textual description of the query [1, 34]. Many approaches focus on the global image representations [1, 16, 23], while some encode the spatial information of image features to improve the discriminative power of the feature representations [13, 11, 19]. In [19, 15, 33], a large number of key-point based descriptors are computed and their relative spatial relationships are encoded. Also, [34] calculates the location offset of two matched features. [14] utilizes the spatial co-occurrence information of visual words mined from database images to boost the retrieval performance. The work in [11] incorporates spatial layout by introducing a Gaussian location model per visual word and encoding only the absolute spatial information. Utilizing localized grids into the feature representation is also a common approach to integrate spatial information [2, 10, 13, 24]. These methods often result in high-dimensional representation and rely on a predefined partitioning of the image which is independent of its content. Moreover, they are not generally concerned with retrieving an exact spatial configuration that exists between objects of interest.

Alternatively one can compute localized features using segmented image regions. [4] investigates object segmen-

978-1-4799-6683-7/15 $31.00 © 2015 IEEE
DOI 10.1109/WACV.2015.133

tation of database images for image retrieval and [32, 30, 31, 17] focus on semantic segmentation and propose models to recover the pixel labels of the training images. However, the aforementioned methods require every image in the training set to be comprehensively labeled which makes it impractical in most scenarios. In contrast, the proposed approach does not require detailed image annotations. Instead, it automatically groups related image parts across the training set using spectral clustering. Furthermore, most existing methods do not focus on matching the spatial configuration of the query with database images. [9] proposes an approach based on soft-matching tree-walks for classification, however it requires that every image be segmented into equal number of regions.

[25] and [29] are approaches based on fast approximate spatial verification. However, due to the high computational cost, these methods are only applied to the top ranked images. In contrast, our approach is able to apply re-ranking to all images in the database by introducing new graphical representations that significantly reduce the graph matching cost.

The authors of [12] and [27] investigate image retrieval with structured object queries by encoding object-names and certain relations among objects with textual phrases like "car on road". While the query in [12] is restricted to queries with word descriptions (such as $Q = \{\text{car on road}\}$), our approach can be applied to cases where the queries can not necessarily be represented by textual descriptions. An example of that is when a graphic designer searches for illustrations of a specific design that is a combination of two or more designs and those designs can not be easily described by words. The authors in [35] develop a method that matches the objects present in the image. However, they assume that the ground-truth bounding boxes are available in the training images which would not be feasible if the number of images and the number of object classes increase. In addition, during testing, in order to reduce the number of possible configurations (i.e. locations and scales) for each object category, they run an object detector on all locations/scales in a standard sliding window manner. This results in bounding objects with a rigid box which may not be adaptive to all object shapes. Consequently, not only it might not be able to get the regional representation that is solely corresponding to each object, but also it might not be able to provide a good measure for the size and position of each object (although matching the size and the position of objects were not investigated by authors of [35]). In contrast, our approach is based on segmentation followed by learning image-parts enabling one to highlight the region associated with each object and therefore providing the ability to measure the object's size and position accurately.

In this paper, we consider a retrieval problem in which a query is defined by a set of images/image-patches along with their desired spatial configuration, size, and/or location in an image (Figure 1). We use an attributed graph for each of the training images based on segmented regions to capture the relative spatial information and adopt an algorithm to collectively learn image parts across all training images. This is done by discovering different groups (communities) of related image parts based on spatial and visual characteristics using a spectral clustering approach. This provides a way to compensate for variations in segmentation. Each segmented region in an image is represented by a community with the highest strength of association. Based on these communities, a robust graph representation is derived for sub-graph matching between the query configuration and training images. The highest matching scores would correspond to images that are most similar to the constructed query through a formulation that will be discussed in Section 2.3.

In summary, our contributions is twofold: First, a new graphical image representation based on segmentation is proposed. Second, an approach to a query retrieval problem using image-patches and spatial configuration is presented.

The remainder of this paper is organized as follows. In Section 2, we describe the overall framework of the proposed retrieval system. The applicability of the proposed approach is illustrated in Section 3 through a query retrieval problem on three challenging datasets. Finally, we conclude the paper with some final remarks and directions for future research in Section 4.

2. Proposed approach

In this section, we describe the details of the our model as illustrated in Figure 2.

2.1. Attributed graph structure

Suppose I is an image in a set containing all training images $\{1, 2, ..., D\}$, and $G^{(I)} = (V^{(I)}, E^{(I)})$ is a graph with $V^{(I)}$ and $E^{(I)}$ representing the nodes and edges of this graph, respectively. Each node corresponds to an image part indicated by a segmented region. Two nodes are connected by an edge if the corresponding regions are adjacent, i.e. $E_{ij}^{(I)}$ is 1 iff i is adjacent to j and 0 otherwise.

To retrieve images with similar spatial configurations, a query is also defined using a graph-based representation denoted by $G^{(Q)} = (V^{(Q)}, E^{(Q)})$, where the nodes correspond to the provided image-patches and edges represent adjacent regions defined by the specified configuration.

The segmentation is computed using the method of [7]. To represent regions (nodes), we extract densely sampled SIFT features [18] from each image, and map each 128 dimensional feature vector to a segment that they belong to. Each node is then represented by vector $h^{(i)}$ using the Bag of Words (BOW) [28] model. In what follows, the appearance of node i is denoted by $h^{(i)}$.

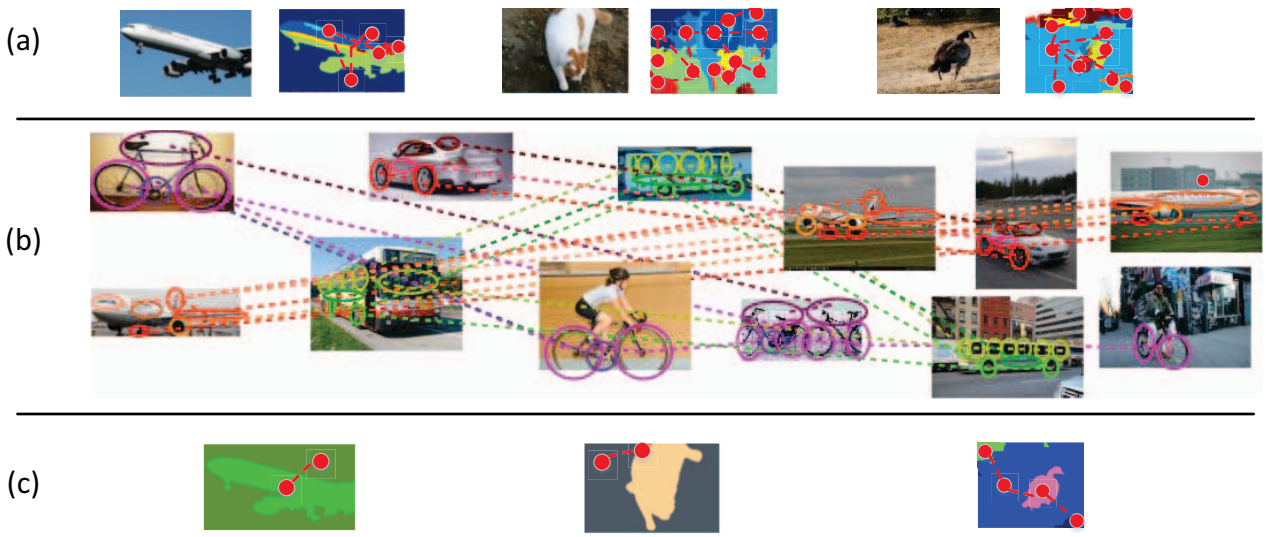

Figure 2: (a) Sample images, their corresponding segmented images and their initial graphical image representations. (b) An example of network of segmented images. Dotted lines represent connections between segmented regions based on spatial adjacency or visual similarity. Each color of an ellipse represents a different community. (c) Resulting images after mapping each of the segmented regions to the detected communities and its corresponding updated graphical image representation. This graph is used for computing the graph matching score. Figure is best viewed in color.

2.2. Learning image parts

The proposed model utilizes a spatially localized feature representation that captures attribute similarity with the relative spatial information without a strong dependence on segmentation. In this section we follow the localized feature representation introduced by [20] and provide a brief summary. We define a network of segmented image regions to integrate the visual similarity between segmented regions across all training images with the localized spatial information. In this network, two nodes i and j are considered related if they are spatially adjacent or if node i/j belongs to the set of T most similar nodes to node j/i based on their visual characteristics denoted by $h^{(i)}$ and $h^{(j)}$. We use the Hellinger metric [21] to compute the distance between $h^{(i)}$ and $h^{(j)}$. For $\mathcal{L}_1$ normalized $h^{(i)}$ and $h^{(j)}$, distance $d(h^{(i)}, h^{(j)})$ is computed by:

$$d(h^{(i)}, h^{(j)}) = \left(\sum_{k=1}^{K} \left(\sqrt{h_k^{(i)}} - \sqrt{h_k^{(j)}} \right)^2 \right)^{1/2} \quad (1)$$

with K denoting the size of the codebook for BOW (number of clusters found by approximate kmeans).

In constructing this network, first, two nodes i and j are connected by a weighted edge equal to their attribute similarity (defined in equation 2) if node i/j belongs to the set of T most similar nodes to node j/i. Second, two spatially adjacent nodes are connected with a weighted edge equal to the average of the weights of all edges connected to the corresponding nodes.

The attribute similarity between two nodes i and j is given by the following:

$$\omega(i,j) = \underbrace{e^{-d(h^{(i)},h^{(j)})}}_{regional\ similarity} \underbrace{\gamma^{\mathcal{I}\{L(i)==L(j)\}}}_{label\ similarity} \quad (2)$$

where $d(h^{(i)}, h^{(j)})$ represents the distance between appearances of two nodes i and j, $L(i)$ denotes the label associated with the image that node i belongs to, and γ is a constant larger than 1. We set $\gamma > 1$ to give a higher weight to the visual similarity of two nodes that belong to images with the same label. The function $\mathcal{I}\{x\}$ represents the indicator function and is equal to 1 if x holds true and zero otherwise.

Next, a spectral clustering technique is applied to this network to aggregate related regions. For graph partitioning, we use the normalized cut method [26]. Each partition is referred to as a community. One can think of each community as a bag that contains all parts of an object.

Let $\mathcal{H}_i$ denote the set of all nodes in the spatial neighborhood of node i, ϕ_c be a community with $c \in \{1, \ldots, C\}$, and $\mathcal{T'}_i$ denote the set of all nodes that are in the top T' nearest neighbors of node i. The strength of association of a node i to a community ϕ_c is measured by two factors: first by the attribute similarity between node i and community ϕ_c, second by considering the attribute similarity between neighbors of node i and different communities in the network along with the relation between community ϕ_c and each of the communities in the network.

Let $g(i \in \phi_c)$ denote the attribute similarity between node i and community ϕ_c. The function $g(i \in \phi_c)$ is defined by the fraction of top T' nearest neighbors to node i that belong to community ϕ_c:

$$g(i \in \phi_c) = \frac{\sum\limits_{j \in \mathcal{T}_i'} \mathcal{I}\{j \in \phi_c\}}{T'}. \tag{3}$$

Moreover, $f(\phi_{c'}, \phi_c)$ is defined to measure the relation between two communities $\phi_{c'}$ and ϕ_c:

$$f(\phi_{c'}, \phi_c) = \frac{\sum\limits_{i \in \phi_{c'}} \sum\limits_{j \in \phi_c} \mathcal{I}\{A_{i,j} > 0\}}{\sum\limits_{i \in \phi_{c'}} \sum\limits_{j=1}^{N} \mathcal{I}\{A_{i,j} > 0\}} \tag{4}$$

where $N = |V|$ denotes the total number of nodes in the network. In particular, $f(\phi_{c'}, \phi_c)$ measures the number of links between the two communities $\phi_{c'}$ and ϕ_c divided by the total number of links between community $\phi_{c'}$ and all other communities. Thus, the strength of association of a node i to a community ϕ_c can be determined by $\mathcal{P}_c^{(i)}$:

$$\mathcal{P}_c^{(i)} = \frac{\sum\limits_{j \in \mathcal{H}_i} \left[\sum\limits_{c'=1}^{C} f(\phi_c, \phi_{c'}) g(j \in \phi_c') \right] g(i \in \phi_c)}{\sum\limits_{c''=1}^{C} \sum\limits_{j \in \mathcal{H}_i} \left[\sum\limits_{c'=1}^{C} f(\phi_{c''}, \phi_{c'}) g(j \in \phi_c') \right] g(i \in \phi_{c''})}. \tag{5}$$

where C denotes the total number of detected communities. Now one can use a maximum likelihood classifier to classify each node by the community with the strongest association:

$$h_u^{(i)} = \arg\max_c \quad \mathcal{P}_c^{(i)} \tag{6}$$

where $h_u^{(i)}$ represents the updated representation of each node i. Learning these image part groupings enables one to illustrate the image using much smaller number of pieces. Choosing an appropriate number of detected communities allows us to set the number of pieces equal to the number of objects that are present in an image.

To find images with the same configuration as the query, one can adopt a sub-graph matching approach and retrieve images with the highest matching score between their corresponding graph representations. However, such a sub-graph matching do not generally perform well due to variations in segmentation. This effect can be reduced by an updated graph representation (Figure 3) that enables one to perform a more robust sub-graph matching for retrieval. Two nodes i and j in graph $G^{(I)}$ are merged if their updated representations are the same and also they are adjacent:

$$j = \begin{cases} i & \text{“merging”} \quad \text{if} \quad h_u^{(i)} = h_u^{(j)} \\ j & \text{“not merging”} \quad \text{if} \quad h_u^{(i)} \neq h_u^{(j)} \end{cases}. \tag{7}$$

In the remainder of this paper, we represent the updated graph of image I by $G_u^{(I)}$. One can compute the sub-graph matching score between the updated graph representations for the query and the database images to find images that best match the query of interest while preserving the spatial configuration, as well as objects' sizes and positions.

Figure 3: Updating graph structures by mapping nodes to detected communities and applying the merging rule. (Left) represents the inital graph representtaion as described in Section 2.1. (Middle) Each color represents an updated representtaion for each node using the concept of communities. (Right) The final updated graph representation with adjacent nodes of same color merged. Figure is best viewed in color.

2.3. Graph matching

Let Q be the query image. We are interested in retrieving images that match the exact spatial configuration, size, and position of the segmented regions of Q. Let $G_u^{(Q)} = (V_u^{(Q)}, E_u^{(Q)})$ and $G_u^{(I)} = (V_u^{(I)}, E_u^{(I)})$ be the updated attributed graph representations for a test image Q and a training image I, respectively. Let $n_Q = |V_u^{(Q)}|$ and $n_I = |V_u^{(I)}|$ represent the number of nodes in each updated attributed graph. In this section, a mapping is found between $V_u^{(Q)}$ and $V_u^{(I)}$ that best preserves the attribute between the two graphs. We seek a set of matches $M = \{i_Q i_I\}$ to maximize the graph matching score.

To get a matching score between $G_u^{(Q)}$ and $G_u^{(I)}$, a modified version of balanced graph matching algorithm [5] is used. Let $x \in \{0,1\}^{n_Q n_I}$ be a binary vector, such that $x_{i_Q i_I} = 1$ iff $i_Q i_I \in M$. We require to have a one-to-one mapping constraint, this is $\sum_{i_Q} x_{i_Q i_I} = 1$ and $\sum_{i_I} x_{i_Q i_I} = 1$. The matching score between graphs of Q and I is defined by solving the following optimization problem which takes the form of an Integer Quadratic Program:

$$\hat{s}(Q, I) = \max_x \quad \frac{x^T W x}{x^T x} \qquad s.t. \quad Bx = b \tag{8}$$

where $\hat{s}(Q, I)$ represents the graph matching score between Q and I, and W is a $n_Q n_I \times n_Q n_I$ comparability matrix indicating the similarity between nodes and between edges. An example of matrix W is illustrated in Figure 4. It is worth noting that $Bx = b$ represents the mapping constraint. For one-to-one matching, we let $b = 1$ in (8). This optimization problem is solved using spectral matching and by computing the leading eigenvector x of W.

The similarity between nodes and between edges is de-

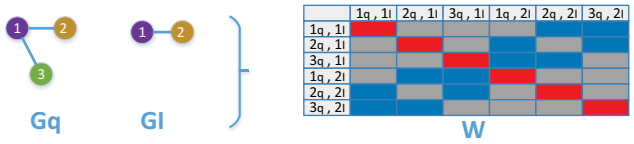

Figure 4: Encoding the edge and node similarities using matrix W. Red represents the node similarities, blue represents the edge similarities, and gray corresponds to the comparison of the similarity of nodes and edges and is set to zero. Figure is best viewed in color.

fined by the following:

$$W_{i_Q i_I, j_Q j_I} = \exp\left(-\frac{\Delta\left(A^{(Q)}_{i_Q j_Q}, A^{(I)}_{i_I j_I}\right)}{\alpha}\right) \tag{9}$$

where Δ is a function that represents the distance between two nodes or between two edges and will be defined shortly. Here, $A^{(I)}$ denotes an adjacency matrix associated with the updated graph of image I. Diagonal elements of the adjacency matrix $A^{(I)}$ are vectors corresponding to the updated node representations of $G^{(I)}_u$, and their corresponding normalized size and position. The off-diagonal entries contain scalar binary values representing the edges between the nodes of $G^{(I)}_u$. For different values of $\alpha \in [0, 1]$, one can emphasize more on the importance of the node similarity versus edge similarity. When matching nodes characteristics, smaller values of α correspond to a less emphasis on node similarities compared to edge similarities.

The Δ function in (9) is defined by the following:

$$\begin{cases} \text{if} \quad (i_Q = j_Q \text{ and } i_I = j_I) \rightarrow \text{“comparing nodes”}: \\ \Delta\left(A^{(Q)}_{i_Q j_Q}, A^{(I)}_{i_I j_I}\right) = \beta_1\left[1 - \delta\left(h^{(i_Q)}_u, h^{(i_I)}_u\right)\right] + \\ \beta_2 |S^{(Q)}_{i_Q} - S^{(I)}_{i_I}| + \beta_3 |P^{(Q)}_{i_Q} - P^{(I)}_{i_I}| \\ \\ \text{if} \quad (i_Q \neq j_Q \text{ and } i_I \neq j_I) \rightarrow \text{“comparing edges”}: \\ \Delta\left(A^{(Q)}_{i_Q j_Q}, A^{(I)}_{i_I j_I}\right) = 1 - \delta\left(A^{(Q)}_{i_Q j_Q}, A^{(I)}_{i_I j_I}\right) \\ \\ \text{Otherwise} \rightarrow \text{“comparing nodes and edges”}: \\ \Delta\left(A^{(Q)}_{i_Q j_Q}, A^{(I)}_{i_I j_I}\right) = 0 \end{cases} \tag{10}$$

where $S^{(I)}_{i_I}$ and $P^{(I)}_{i_{(I)}}$ indicate the normalized size and position of node i in image I. The parameters β_1, β_2, and β_3 determine how much one emphasizes on the objects' appearance similarities, their associated sizes and positions in an image. In addition, $\delta(m, n)$ is equal to 1 if $m = n$ and 0 otherwise. One should note that the Δ function is set to 0 when comparing a node with an edge since they are incomparable.

Algorithm 1 Update the graph representation $G^{(I)}$

Input: $G^{(I)} = (V^{(I)}, E^{(I)}) \quad \forall\, I \in \{1, \ldots, D\}$, network of segmented regions, detected communties

Output: $G^{(I)}_u = (V^{(I)}_u, E^{(I)}_u) \quad \forall\, I \in \{1, \ldots, D\}$

$V = \cup_{I=1}^{D} V^{(I)}$
$N = |V|$

Comment: update node representations:
for $i = 1 \rightarrow N$ **do**
  for $c = 1 \rightarrow C$ **do**
    compute $\mathcal{P}^{(i)}_c$: The likelihood of node i belonging to community c
  end for
  $h^{(i)}_u = \arg\max_c \quad \mathcal{P}^{(i)}_c$
end for

Comment: apply merge rule:
for $I = 1 \rightarrow D$ **do**
  for $i = 1 \rightarrow |V^{(I)}|$ **do**
    for $j = 1 \rightarrow |V^{(I)}|$ **do**
      if i and j are adjacent **then**
        if $h^{(i)}_u == h^{(j)}_u$ **then**
          merge nodes i and j of image I
        end if
      end if
    end for
  end for
end for

3. Evaluation

Database: We have performed experiments using three challenging datasets: PASCAL VOC2007 [8], ImageNet ILSVRC2010 [6] and TREC. PASCAL VOC2007 is a publicly available dataset containing $9,963$ images and 20 object classes. A subset of ImageNet ILSVRC2010 [6] which contains roughly about 1 million images is added to the VOC2007 dataset as distractors to test the scalability of our system with respect to the size of the dataset. The resulting combined dataset is referred to as "VOC+ImageNet". To evaluate the performance of different methods in identifying the object of interest when it occupies only a small portion of the image in a cluttered background, a set of images is collected by extracting frames from TRECVID 2012 instant search (INS) dataset [22] and it is referred to as TREC dataset. Since the groundtruth is only published for a subset of the data, only classes that have sufficient numbers of true positives are considered. TREC dataset contains $10,289$ images and 10 object classes.

Qualitative analysis: To show the effectiveness of our approach, we consider a query consisting of multiple

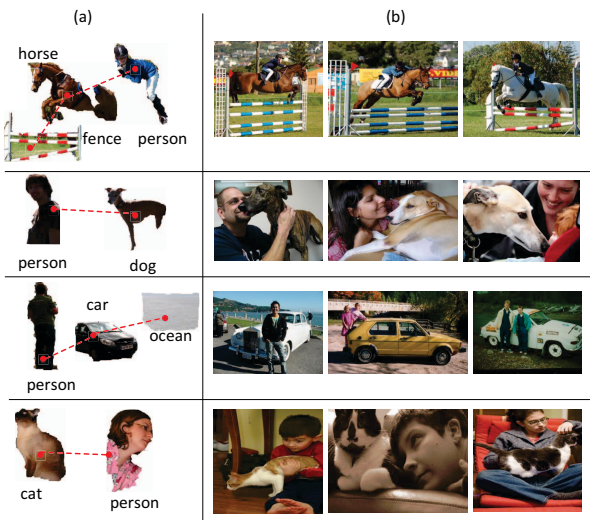

Figure 5: (a) Sample queries and (b) their corresponding top 3 retrieved images using the proposed retrieval system on "VOC07" dataset.

image-patches (objects). Our goal is to retrieve images from the dataset that not only include such objects but also satisfy a set of requirements such as size, position, and spatial configuration provided by the user. Figure 5 (a) illustrates the image-patches of interest that should appear in the target image with the desired spatial configuration, size, and position. Using the proposed approach the top 3 retrieved images from dataset "VOC07" are depicted in Figure 5 (b). As one can see, our method returns images that satisfy the requirements provided in Figure 5 (a).

Baseline methods: Since the state of the art image retrieval systems can not be applied to queries as presented in Figure 1, for fair comparison, we consider the case in which the query is a single image with the goal of retrieving an exact spatial configuration of its segmented regions. This can be viewed as a sub-class of problems that can be handled by our approach.

The performance of the proposed system is compared with the retrieval system based on the Spatial Pyramid Matching (SPM) by encoding the global positions of features in the image [13].

We further compare our method with the case in which each image in the dataset is segmented and represented by a graph of its segmented regions. Here, each segmented region is represented solely by a BOW representation [28]. This will be referred to as "Basic Graph Matching".

Performance: Each database is divided into a training set and a testing set. For each class, a model is learned using image features, their corresponding class types, and a SVM classifier. These models are used to estimate the class type of each image as wells as their associated confidence score. The estimated class types for each image are then sorted based on their confidence scores. Each image is only associated with a set of class types that have the top Y confidence scores.

Given a query image, the above model is used to associate the query image with a set of class types with top Y confidence scores. In what follows, a "positive set" refers to a set containing test images that share at least one class type with the query. The rest of the database images are referred to as a "negative set". Images included in the positive set are ranked higher than the ones in the negative set. In addition, images in each set are re-ranked based on their similarity score to the query image. The similarity is defined by $\omega = e^{-d(R^{(I)},R^{(Q)})}$ where $d(R^{(I)}, R^{(Q)})$ denotes the Hellinger distance between two image representations indicated by $R^{(I)}$ and $R^{(Q)}$.

In "Basic Graph Matching" and the proposed approach, the retrieved images for each query in the positive and negative sets are re-ranked separately. This ranking is based on combined graph matching scores and visual similarity scores. This process is summarized in Figure 6.

In our experiments, a set of queries are randomly selected from each dataset and the accuracy of the retrieval system is measured using mean-average-precision (MAP). In computing the retrieval accuracy, each image I in the test set is weighted based on how much its characteristics match the query Q. These characteristics are class type $\mathbb{I}$, spatial configuration $\mathbb{A}$, size $\mathbb{S}$ and position $\mathbb{P}$. Given a query Q, we define a label $\mathbb{L}$ for each image I to specify how much its characteristics matches the corresponding ones in the query:

$$\mathbb{L}_{Q,I} := [\mathbb{I}_{Q,I} \quad \mathbb{A}_{Q,I} \quad \mathbb{S}_{Q,I} \quad \mathbb{P}_{Q,I}] \tag{11}$$

where $\mathbb{I}_{Q,I}$ is one iff Q and I both belong to the same class type, otherwise zero. Similarly, $\mathbb{A}_{Q,I}$, $\mathbb{S}_{Q,I}$, or $\mathbb{P}_{Q,I}$ are either one or zero. In particular, they are set to one if Q and I are both from the same class type and their corresponding spatial configuration, size, or position match, respectively.

Finally, the weight that determines how much each image characteristics match the query is given by:

$$\tau(Q, I) = \mathcal{L}_1(\mathbb{L}_{Q,I}). \tag{12}$$

Table 1 illustrates the accuracy of the proposed retrieval system compared with the baseline methods. It is shown that the proposed approach achieves a higher retrieval accuracy than the baseline methods by 11 percent in the VOC dataset and 15 percent in the TREC dataset.

In addition, Figure 7 shows a comparison between the retrieval accuracy of different methods at different depths for the "VOC+ImageNet" dataset. These results emphasize the scalability of our approach.

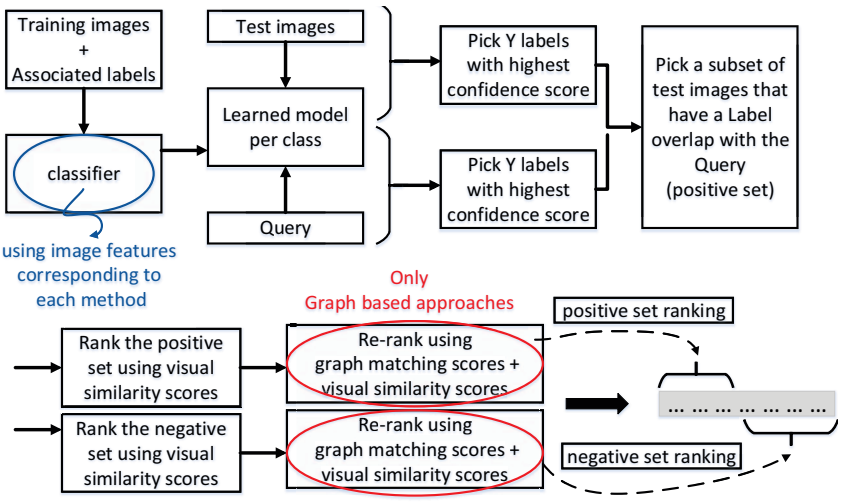

Figure 6: Process of ranking the retrieved images for the methods discussed in Section 3.

Database: VOC07			
Method / K	200	500	1000
Spatial Pyramid ℓ_1	0.39	0.44	0.52
Spatial Pyramid ℓ_2	0.41	0.46	0.54
Basic Graph Matching	0.42	0.48	0.54
Proposed Approach	**0.60**	**0.63**	**0.65**

Database: TREC			
Method / K	200	500	1000
Spatial Pyramid ℓ_1	0.37	0.39	0.42
Spatial Pyramid ℓ_2	0.39	0.41	0.43
Basic Graph Matching	0.40	0.42	0.44
Proposed Approach	**0.52**	**0.56**	**0.59**

Table 1: Comparison of the accuracy of different retrieval systems discussed in Section 3 with different codebook sizes (K). The results are reported for VOC2007 and TREC databases.

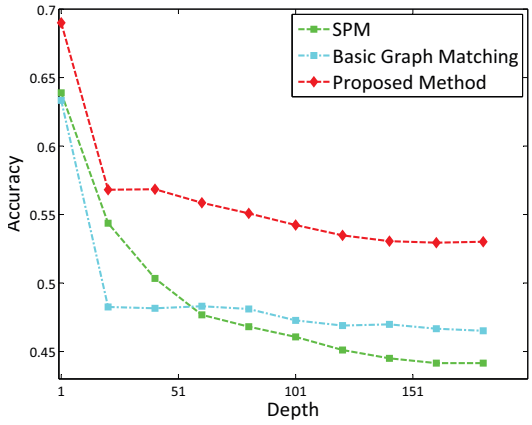

Figure 7: Comparison of the accuracy of retrieval systems on "VOC+ImageNet" dataset at different depths. Results are reported for $K = 1000$.

Computational cost: A moderate computational complexity is important when considering scaling to thousands of images and hundreds of categories. The computational cost of learning image parts is considerably reduced by performing the search using a space partitioning data structure (k-d tree) with running time of $O(\log N)$. Also, as the size of the database increases, the network can be constructed using a subset of the images from the training data. Therefore, the cost of network construction would not increase. It is also worth noting that the computational cost related to the graph matching algorithm of the proposed approach is relatively small as the number of nodes in the updated graph representations are considerably fewer than the initial graph representations. The overhead cost for the retrieval of each query compared to the global image representations is roughly equal to 0.1 seconds using a quad core computer with 3.0 GHz processor.

4. Conclusion

We presented an approach to image search using image-patches and pre-specified spatial configurations. In general, such queries can not be handled by global image representations. The updated graphical structures are robust to segmentation variations and are suitable for sub-graph matching. Extensive experiments conducted on challenging datasets demonstrate that the proposed approach compares favorable with current state of the art methods. For future work, we plan to explore the applicability of the proposed method for enhanced object tracking.

5. Acknowledgments

This work is supported by ONR grant #N00014-12-1-0503.

References

[1] R. Arandjelovic and A. Zisserman. Multiple queries for large scale specific object retrieval. In *BMVC*, 2012.

[2] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, 2007.

[3] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, 2011.

[4] J.-J. Chen, C.-R. Su, W. E. L. Grimson, J.-L. Liu, and D.-H. Shiue. Object segmentation of database images by dual multiscale morphological reconstructions and retrieval applications. In *IEEE Trans. Image Processing*, 2012.

[5] T. Cour, P. Srinivasan, and J. Shi. Balanced graph matching. In *NIPS*, 2006.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[7] Y. Deng and B. Manjunath. Unsupervised segmentation of color-texture regions in images and video. In *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2001.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2007 results, available at http://www.pascal-network.org/challenges/voc/voc2007/workshop/index.html.

[9] Z. Harchaoui and F. Bach. Image classification with segmentation graph kernels. In *CVPR*, 2007.

[10] A. Hegerath, T. Deselaers, and H. Ney. Patch-based object recognition using discriminatively trained gaussian mixtures. In *BMVC*, 2006.

[11] J. Krapac, J. Verbeek, and F. Jurie. Modeling spatial layout with fisher vectors for image categorization. In *ICCV*, 2011.

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

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[14] Y. Li, B. Geng, Z.-j. Zha, Y. Li, D. Tao, and C. Xu. Query expansion by spatial co-occurrence for image retrieval. In *ACM international conference on Multimedia*, 2011.

[15] D. Liu, G. Hua, P. Viola, and T. Chen. Integrated feature selection and higher-order spatial feature extraction for object categorization. In *CVPR*, 2008.

[16] J. Liu. Image retrieval based on bag-of-words model. In *arXiv preprint arXiv:1304.5168*, 2013.

[17] Y. Liu, J. Liu, Z. Li, J. Tang, and H. Lu. Weakly-supervised dual clustering for image semantic segmentation. In *CVPR*, 2013.

[18] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[19] N. Morioka and S. Satoh. Building compact local pairwise codebook with joint feature space clustering. In *ECCV*, 2010.

[20] N. Pourian, B.S. Manjunath. PixNet: A Localized Feature Representation for Classification and Visual Search. Submitted to Journal Publications.

[21] M. S. NIKULIN. *Hellinger Distance*. Springer, 2001.

[22] P. Over, G. Awad, J. Fiscus, B. Antonishek, M. Michel, A. F. Smeaton, W. Kraaij, G. Quénot, et al. Trecvid 2012-an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID 2012-TREC Video Retrieval Evaluation Online*, 2012.

[23] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *CVPR*, 2010.

[24] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.

[25] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[26] J. Shi and J. Malik. Normalized cuts and image segmentation. In *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2000.

[27] B. Siddiquie, B. White, A. Sharma, and L. S. Davis. Multimodal image retrieval for complex queries using small codes. In *ICMR*, 2014.

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

[29] G. Tolias and Y. Avrithis. Speeded-up, relaxed spatial matching. In *ICCV*, 2011.

[30] J. Verbeek and B. Triggs. Region classification with markov field aspect models. In *CVPR*, 2007.

[31] A. Vezhnevets and J. M. Buhmann. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In *CVPR*, 2010.

[32] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised semantic segmentation with a multi-image model. In *ICCV*, 2011.

[33] T. Weyand, T. Deselaers, and H. Ney. Log-linear mixtures for object class recognition. In *BMVC*, 2009.

[34] Y. Zhang, Z. Jia, and T. Chen. Image retrieval with geometry-preserving visual phrases. In *CVPR*, 2011.

[35] G.-T. Zhou, T. Lan, W. Yang, and G. Mori. Learning class-to-image distance with object matchings. In *CVPR*, 2013.