Concept Learning with Co-occurrence Network for Image Retrieval

Linan Feng* Bir Bhanu†
Visualization and Intelligent Systems Laboratory
University of California, Riverside
Riverside, California 92521
fengl@cs.ucr.edu* bhanu@vislab.ucr.edu†

Abstract—**In this paper, we address the problem of semantic concept learning in the context of image retrieval. We introduce two types of semantic concepts in our system which are the image individual concept and scene concept. The individual concept is usually explicitly provided in the vocabulary of training concepts, wherein each image could possess multiple individual concepts. Furthermore, we define the scene concept as a potential co-occurrence pattern of individual concepts shown up with high frequency among images. In human learning, it is common to understand simple ideas prior to shape a more sophisticated one. However, the co-occurrence pattern usually has more discriminative power than individual concepts. We present a novel method for deriving scene concepts by investigating the hierarchical community structures and closed subgraph in a weighted concept co-occurrence network (graph). Our study is activated by the continuing requirement of bridging the semantic gap in intelligent image retrieval. We present a semantic image representation called after scene concept signature. The goal is to assign each image a scene concept signature, which tells the probabilities for observing certain hidden patterns in an image and makes it possible to compare and retrieve images based on a higher level of semantic similarity. We conduct extensive experiments on publicly available dataset to demonstrate the effectiveness of our system in semantic concept modeling and concept based image retrieval.**

Index Terms—**concept learning, image retrieval, individual concept, scene concept, co-occurrence network, interactive learning, long-term learning.**

I. Introduction

THE continuing growth of digital image collections shows great demand in efficient concept based searching systems capable of affording satisfying results within an acceptable retrieval period. The word ”concept” is often used in these systems [1] [2] [3] [4] to define the semantic meaning emerged from an image, such as the sets of object names(e.g. food, furniture), events (e.g. commencement, Olympic Games), and implicit knowledge (e.g. drive a car, make dinner). The essence of an intelligent image retrieval task is the learning process through which the desired concept resided in human brain can be grasped accurately and promptly by the system. We borrow the term from machine learning area [5] and define the preceding process as ”concept learning”.

In this paper, we address the problem of retrieving images with concept learning in a combined visual and textual setting. In this task, the retrieval system is given a small proportion of images with labeled regions from pre-segmentation as the training set which is normal in real life, and a large proportion of unlabeled images as the newly added images. We define the words in the training label vocabulary as individual concepts, each database image could possess multiple individual concepts. Due to the compositional property of visual objects [6], the frequency of observing certain patterns of co-occurred individual concepts across the entire image set could be high. We consider the co-occurrence property of individual objects presented in a nature scene, for example, ”sky”, ”sand” and ”sea” often appear together in a ”beach” scene, and define the co-occurrence pattern as scene concept. In this work we propose to take into account the co-occurred dependencies and learn the scene concepts explicitly for the purpose of semantic search. The scene concept functions as a higher level of abstraction of the expressed meaning of an image, which is more concise and discriminative than individual concept. Therefore it has great importance to recognize the discriminative scene concepts for image similarity comparison.

Given a vocabulary with N unique individual concepts, the number of scene concept candidates is significant (e.g. 2^N, consider each individual concept could either be included or excluded from a scene concept). We propose to construct a concept co-occurrence network which captures all the co-occurred relations of the appeared labels. The nodes in the network represent the individual concepts and the edges represent the co-occurrences between concepts. The frequency of every pair-wise co-occurrence is accumulated across the entire dataset and recorded as the edge weight. A common property which was found in many real life networks is called the property of community structure [7] [8]. A community structure is qualitatively defined as a group of nodes within which the connections between nodes are denser compared to the connection of itself to the outside nodes. Furthermore, the community structure can be presented hierarchically in the network, by which means there could be subsets of nodes inside a community which are closely connected in themselves while loosely connected to each other. The detection of scene concept in the network transfers to the problem of discovering community structures and detecting the complete subgraph (or under loose condition, the closed subgraph) with large summation of edge weights inside the communities. We introduce a novel algorithm which is called ”cut-and-merge” to solve the problem in acceptable time duration. The algorithm is generally functioned as a mapping procedure from the network

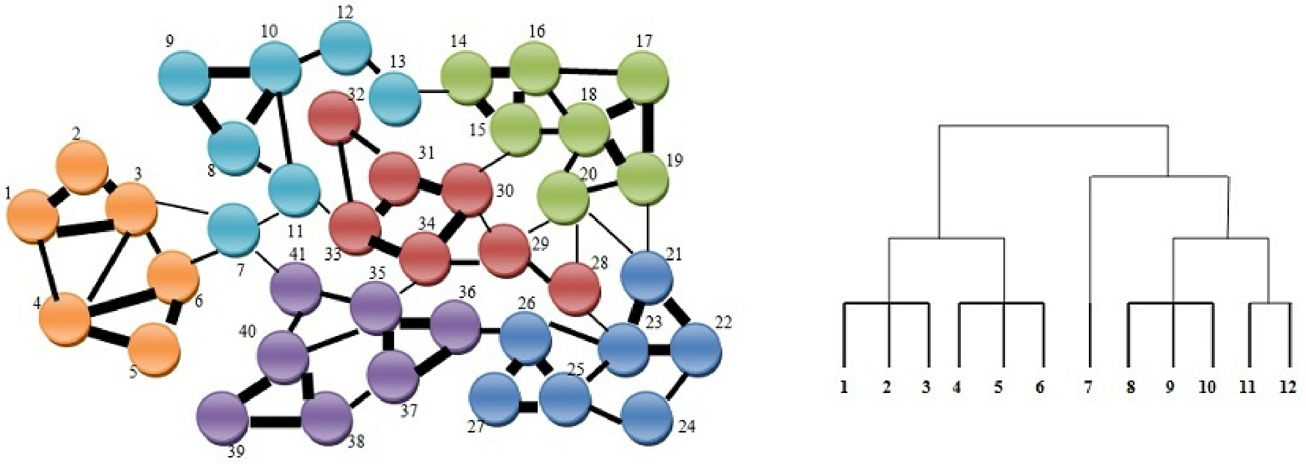


Fig. 1. An example of the structures derived for scene concept detection. The number in the co-occurrence network (left side) represents the individual concept, the color of the nodes shows the community property of the network, the weight of the edge shows the degree of co-occurrence between individual concepts. Part of the generated dendrogram is shown on the right side, the leaf in the structure represents the nodes in the network, the dendrogram shows the hierarchical structure of the community property, the weight of the connecting line shows the degree of co-occurrence between individual concepts.

into a structured model (called a dendrogram in social science) by which the scene concepts can be concluded in different granularities. To give an idea of the co-occurrence network and the deduced structured model, we illustrate the correspondence in Figure 1.

The final goal of our work is to retrieve the most relevant images to a query ranked on their semantic similarity. The most common approach for comparing similarity of two images is to calculate the distance in various dimensions of the image feature representation. We combine the individual concepts and scene concepts possessed by one image as signatures which are the semantic representation for computing similarity. The element in the signature is in numeric form indicating the occurring probability of a concept in an image. Before acquiring the scene concept signature, we rely on an intermediate process to obtain the individual concept signature by learning the correlations between image features and training individual concepts, and further predicting the occurring probabilities to the new images. This process shares common features with the tasks such as image classification, image auto-annotation, and object recognition, where the image visual features are extracted and analysed, and the aim arrives eventually at predicting the relevant class names, annotations, or object categories. To that end, we modified the generative model in [9], where a relatively better solve of the individual concept learning was suggested. However, our approach is unique in two facets: first, we circumvent the class/label prediction which is theoretically appealing to the image retrieval task, and second, we do not impose any controls on the size of the concepts (synonymously, class names/labels in the context of their work) an image could possess, consequently, additional useful details about the image could be kept, exploited, and even be adapted later on. According to the independence property of individual concepts, the occurring probability of a scene concept can be directly calculated by the product of occurring probabilities of individual concepts under detected scene configuration. The comparison of conventional approach and the proposed approach is illustrated in Figure 2.

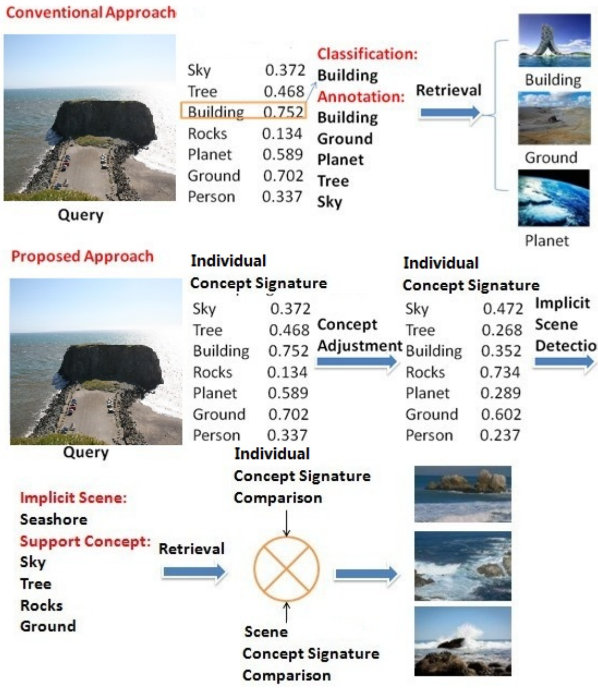


Fig. 2. Conventional approaches only keep the concepts with significant occurring probabilities. The proposed approach considers building the individual concept and scene concept signatures as semantic representations of the image. The learned concept signature is further used in the semantic image search.

In accordance with the above setting, we also set up the distance metrics for measuring the similarity between concept signatures. Since the individual concept signatures could have differences in the number and name of the remarkable elements, the Earth Mover's distance is introduced as an effective measure allowing for a partial match between two individual concept signatures. To compare two scene concept signatures, we use the χ^2 distance for the purpose of measuring the similarity between two modalities. We test our proposed retrieval model using a public benchmark dataset: the Scene Understanding dataset [21] (SUN'09), our results are comparable or better than the state-of-art reported on this dataset.

The rest of this paper is organized as follows. Relevant previous work on image retrieval, concept learning, co-occurrence patterns is discussed in Section II, the components of our approach (concept co-occurrence network, scene concept detection, etc) are described in Section III. Section IV addresses the problem of the semantic based image retrieval. Section V.A describes our experimental setup and the databases included in the current study. Section V.B evaluates the effect of the concept signatures on retrieval performance. Finally, Section VI concludes the paper with a summary of our findings and future extensions.

II. Related Work

Most existing systems address the problem of image retrieval either in a fully visual way (e.g. Content-based image retrieval systems (CBIR) as FIDS [10], CIRES [11]), or in a fully textual way (e.g. web image search engines as Google, Yahoo). The defects of either system are obvious. The CBIR systems rely on the visual contents which can be derived from the image itself, however, the extracted content in the form of color, shape, and texture features are very loosely connected to the actual meaning of the image and may override the expected concept under search. In such systems, the potential intent can only be captured by techniques such as relevance feedback [12].

In the contrary, the other predominant line of work assigns the descriptive metadata in the text form (e.g. labels, annotations, keywords, captions) to the digital images either manually or automatically. The metadata is the most straightforward way to deliver image concepts in most of the cases. In this setting image retrieval seems trivial at first sight since the image contents are mapped into keywords and the retrieval could be simply implemented by query-based on keyword (QBK). As manually labeling images is laborious and expensive due to the large volume of image collections, most of the works are done by voluntary in cooperation, e.g. image manual annotation site such as LabelMe [13]. Recently, much research emphasis have been put on automatic image labeling/annotation (e.g. ALIPR [14]). Typically, machine learning techniques are adopted to analyze the relations between extracted image pictorial features and the training concept words defined in a vocabulary and to automatically predict the relevant concepts to new images without any user interaction. The automated annotation engines greatly accelerate the process of concept learning. Nevertheless, it also suffer imperfections for image retrieval task. First, in these systems the primary goal is to acquire accurate image labels rather than accurate retrieval results. In the unsupervised setting, the assigned concepts of images have no opportunity to be corrected or refined with the aid of human determination. At the same time, since the retrieval is solely based on keywords, it is impossible to perform ranking of the results according to similarity measure. Therefore, the automatic labeling technique is useful for image classification, organization and indexing, while it has not occupied the superiority in the image retrieval task.

To address the difficulty of semantic search, there has been a research trend on semantic similarity. For example, in the transfer learning of new concepts [15], the knowledge are transferred from known concepts which are determined by hidden semantic links. The linguistic knowledge bases used in the research include WordNet, Wikipedia, or the World Wide Web. Further, the relations between visual and semantic category similarity were investigate in [16] in accordance with the appearance of semantically organized image database, such as ImageNet [17]. Similarly, the sematic hierarchy was utilized for similar image retrieval in [18], although the semantic co-occurrence was also considered for similarity comparison, the structure of the proposed hierarchy was not clearly defined in the paper. Many papers were presented on modeling the relations of concepts in tree structured models [2] [19] [20] where the parent-child relation was presented with inherent include, inheritance property. In their structured models, the "jaguar" concept is closer to the "Africa tiger" rather than "forest" as they overlook the co-occurrence relation intentionally. Recently, focuses have been put on the contextual models [21] [22] which take into account the spatial arrangement of concepts, it can efficiently exclude unlikely combinations of concepts with their location information on the image, and produce a semantically coherent interpretation of a scene, such that the concept of "water" on the bottom could suppress the occurring of a "car" on the top while favor the "ship". However, the consideration of spatial relationships between pairs of concepts make it computational expensive.

The attribute-based concept learning has gained popularity in the object recognition and classification literature. Since the attributes are shared in different object categories and one object category could have multiple attributes, the relation between attributes and objects is quite similar to our individual concept and scene concept relation, In [23] the Gaussian Mixture Model was integrated with several newly developed feature descriptors to learn the concepts and they show the performance of attribute-based image retrieval gives comparable results to the state-of-art. An method provided in [24] builds a vocabulary of discriminative attributes which are understandable to human. And also human annotators are invented in a interactive loop where they are asked to provide the attribute name for detected hidden concepts. The co-occurrence property of local attributes of concepts have been explicitly considered in [25], the discriminative patterns are discovered by machine learning techniques such as boosting and are used to distinguish different concept categories.

The hidden affiliation patterns of nodes were found in real life networks with many applications such as social networks, acquaintance networks, collaboration networks, and the biological networks including the metabolic networks, epidemiology and ecological webs, the phenomenon was called the property of community structure to indicate the co-occurrence relation of nodes. Many algorithms were proposed to detect the community structures, e.g. in [26] they solve the problem by using hierarchical clustering, since the edge weights representing how closely connected of two nodes are not clearly defined, they appeal to the number of non-independent paths between nodes which could be computed by using polynomial-time "max-flow" algorithms. And also in [27] they consider detecting the least central edges which are the edges most between communities by using Freeman's edge betweenness centrality. The challenge of these systems is that the network edges do not possess the weights intrinsically, thus additional computing was required in the initial.

III. Concept co-occurrence network and scene concept detection

We now describe our concept detection models, starting with individual concept learning in Section III.A, and extending to the concept co-occurrence network with scene concept detection method in Section III.B.

A. *Individual concept learning*

We use a multiple Bernoulli relevance model to learn the correlations between training individual concepts and the pre-segmented image regions. The model assumes a training set of images in which the specific correspondence between labels and regions is provided in advance, which is usually publicly available (e.g. the benchmark datasets as SUN'09, ImageCLEF'10, Animal with Attributes), one image can have multiple labels, all the training labels contribute to the vocabulary each of which represents an individual concept. We extract the local features of the training image segments which contains GIST (the orientation histogram of the object boundary in the segmented region), PHOG (Pyramid of histograms of oriented gradients), PHOG with oriented edges (which considers the direction of the salient Canny edges), and Pyramid of self similarity descriptor (a log-polar histogram of correlations between central and surrounding pixels). The features are aggregated into a vector as the pictorial descriptor of the image.

We propose to model the joint distribution of associating concepts with the extracted features by a Bernoulli process. Since we are focusing on the presence or absence instead of the prominence of each individual concept, and the occurring probability of each concept is independent of the other ones, the Bernoulli model is the best choice in this scenario. The individual concept signature can be viewed as a continuous valued occurrence vector where each element of the vector could vary in the range [0,1]. all the elements in the vector are independent and identically distributed in Bernoulli process.

Let $\widetilde{U}$ be the entire concept vocabulary and T be the labeled training set, we denote each labeled image in T as I, thus I can be represented as a set of visual regions $R_I = \{r_1, r_2, \ldots, r_n\}$ and the corresponding subset of labeled individual concepts $IC_I = \{c_1, c_2, \ldots, c_m\}$, the number of the regions are the number of the concepts are not necessary to be identical, since one concept can be shared with several regions, while n has to be greater than m, since each region can only possess one concept in the SUN'09 dataset. For each component in IC_I, we assume it is generated from certain multiple-Bernoulli distribution $Pr_{Bernoulli}(IC_i|r_i)$. And for each r_i we have its pictorial feature vector v_{r_i}, we assume the feature vector satisfies certain underlying multi-variate distribution $Pr_{multi-variate}(v_{r_i}|r_i)$. Now we are going to model the joint probability of any n feature vectors of n regions in arbitrary image I_a denoted as $\widetilde{V}_{I_{a_n}} = \{v_1, v_2, \ldots, v_n\}$, and any subset of the vocabulary with m concepts denoted as $IC_{I_{a_m}}$. We assume that the probability of observing $\widetilde{V}_{I_{a_n}}$ and $IC_{I_{a_m}}$ denoted as $Pr(\widetilde{V}_{I_{a_n}}, IC_{I_{a_m}})$ should be generated in the same process as some subset of regions and concepts in the training set. We estimate the joint probability by calculating the expectation over all the regions in the training set. It is in the form as below:

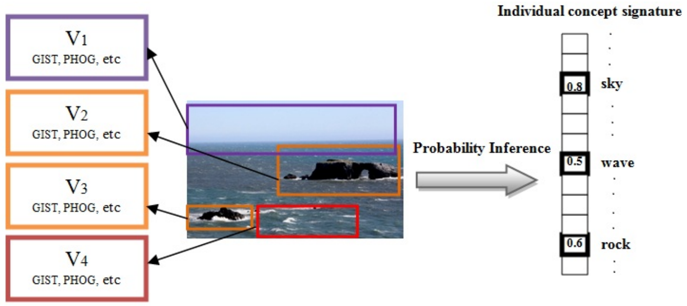


Fig. 3. The new image with detected salient regions, the pictorial features are extracted and used for the probability inference of individual concepts occurrence, the probabilities are built into the individual concept signature.

$$Pr(\widetilde{V}_{I_{a_n}}, IC_{I_{a_m}}) = \sum_{r_i \in R_T} \{Pr_T(r_i) \prod_{a=1}^{n} Pr_{multi-variate}(v_a|r_i) \times \prod_{c \in IC_{I_{a_m}}} Pr_{Bernoulli}(c|r_i) \prod_{c \notin IC_{I_{a_m}}} (1 - Pr_{Bernoulli}(c|r_i)) \quad (1)$$

The reason we model the joint probability is that if we have a set of pictorial features in image I, the occurring probability of each concept can be solved by:

$$Pr(c) = \frac{Pr(\widetilde{V}_I, IC_I)}{Pr(\widetilde{V}_I)} \quad (2)$$

The probability $Pr_T(r_i)$ in Eqs.(1) is the probability of picking region r_i from training set T, which can be simply solved by $Pr_T(r_i) = \frac{1}{R_T}$, where R_T is the number of regions in T. The other two probabilities are estimated as follows: first, the distribution $Pr_{multi-variate}$ is estimated by non-parametric kernel-based density estimation, we use the Gaussian kernel in the process. Second, the distribution $Pr_{Bernoulli}$ is estimated by Bayes estimation with a beta prior (the reason is beta prior is known as the conjugate to the Bernoulli).

For each newly added image, we compute their region features and map to occurring probabilities of the individual concepts. The probabilities are then combined into the individual concept signature of the image. The process is further illustrated in Figure 3.

B. *Concept co-occurrence network*

The original data in SUN'09 dataset is in XML format. We export all the individual concepts in the training set, and we rank them based on the frequency of occurrence. Due to the large amount data, the number of the individual concepts have appeared is significant (5,847 in total), we threshold and only keep the concepts with high occurring frequency. The group of concepts possessed by each training image is called a candidate tuple which shows the co-occurrence relationship among those concepts. We proposed the following algorithm to accumulate all the candidate tuples into a co-occurrence network.

Algorithm 1: Co-occurrence network generation

```
1. Initial a $N \times N$ adjacency matrix $\mathcal{A}$ with empty cells, where $N$ is the number of the individual concepts.
2. Initial a individual concept array $\mathcal{D}$ which is indexes in the decreasing order of the occurring frequency. Set all the values into 0.
3. For each candidate tuple $t$, get each of the concept name and index, if index is equal to 0, associate a new node with the concept in the network, set the value of $\mathcal{D}[index]$ into 1, for each pair of concepts $C_i, C_j$ in the tuple $t$, plus the value of $\mathcal{A}[C_i][C_j]$ by 1.
4. Traverse all the elements in $\mathcal{A}$, connect two nodes $i, j$ with edge weight in the network in accordance with the value of $\mathcal{A}[C_i][C_j]$.
```

The generated concept co-occurrence network is shown in Figure 4.

C. Scene concept detection

In this section we deal with the problem of scene concept detection based on a given co-occurrence network, the task is performed by investigating the topological property of the network, namely the community structure. The co-occurrence network can be mapped into a tree structured model which is called a dendrogram to represent the co-occurrence relation hierarchically. And the low level in the hierarchy with small granularity of the community is considered as potential scene concept. To develop an efficient procedure in an algorithm for the identification of the hierarchical community structure is a matter of great concern for us. While the process is non-trivial.

There are two categories of community detection methods in the literature: the agglomerative method and the divisive method distinguished by the order they construct the dendrogram. The former generates a community by iteratively adding edges between pairs of nodes in order of decreasing weight. In this way the nodes are first grouped into small communities and then agglomerate into larger ones. The dendrogram is built up bottom-up respectively. The latter category of method deal with the problem in different direction. Given the entire network with all the connecting edges, the network is divided into small disconnected parts progressively by cutting the edge between them, and the disconnected subnetwork is identified as community. The challenge part is the selection of the edges to be cut. A concept of "edge betweenness" is introduced in [27] to measure the degree of closeness between pairs of nodes. The betweenness of an edge is defined as the number of shortest path between any two nodes in the network running through that edge. Clearly, if an edge is connecting two communities, the betweenness of it should be high, since the nodes in one community have to cross that edge to reach the nodes in the other community. However, the edge betweenness is computationally expensive, evaluating the scores for all the edges in one iteration needs a time in the order of $E \times N$, where E is the number of edges in the network and N is the number of nodes. In our situation, N is moderately large (N varies from 100 to 500), and E is in quadratic to N, thus in the worst case if all the edges have to be dropped and each node itself represents a community, the scale of the computation is O(E^2N) which equals to O(N^3).

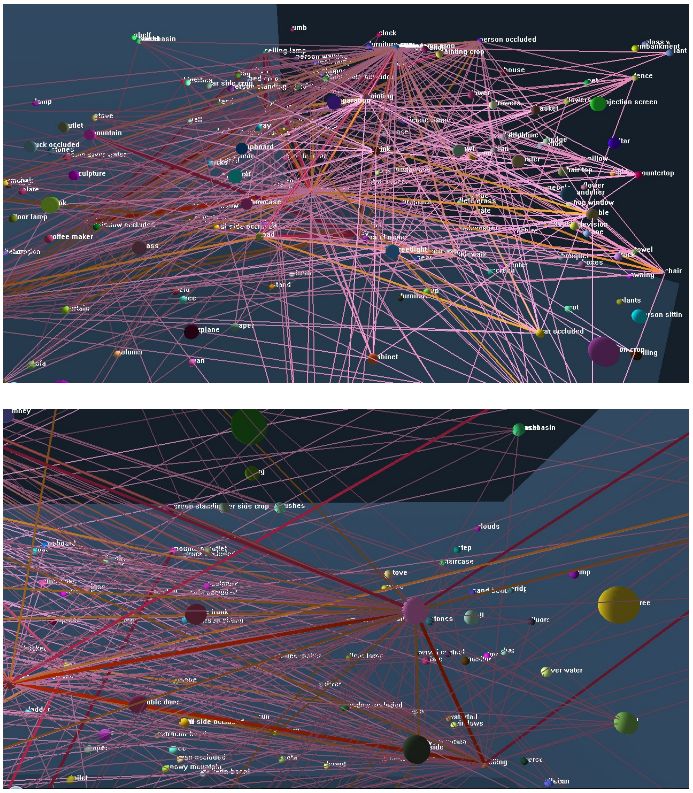


Fig. 4. Top: A macroscopic view of the generated co-occurrence network in a three-dimensional space, The nodes are labeled with the corresponding concepts, the edges are colored and weighted based on the frequencies of co-occurrence. Down: A detailed view inside the network, it can be observed clearly that a hidden pattern of co-occurrence is marked with heavy red lines. We propose to detect all the co-occurrence patterns in the network.

To deal with the problem, we present the "cut-and-merge" algorithm which first cut the network into large communities with acceptable time and then detect the small communities inside a large one as scene concept patterns by merging the nodes with large co-occurrence frequencies. The effectiveness of the detected scene concept is measured by quantitative definitions. Here we first give the definition of an effective community and an effective scene concept. Suppose we obtained a subnetwork of community by cutting from the co-occurrence network and node a is inside the subnetwork. In the $N \times N$ adjacency matrix $\mathcal{A}$, the connected nodes to a can be categorized into two parts which are within the subnetwork and outside the subnetwork. We define the sum of weights in the adjacency matrix $\mathcal{A}$ of each category as inward edge connectivity D_a^{in} and outward edge connectivity

D_a^{out}. A subgraph S is an effective community if for all the nodes in S, we have

$$\sum_{i \in S} D_i^{in} > \sum_{i \in S} D_i^{out} \tag{3}$$

And, likewise, a co-occurrence pattern P in S is an effective scene concept if for all the nodes P, we have

$$\sum_{j \in P} D_j^{in} > \sum_{j \in S} D_j^{out} \tag{4}$$

The algorithm for detecting the scene concept is given as follows.

Algorithm 2: Scene concept detection

Define the maximum granularity of an effective community as ϱ_{max}, define the minimum number of individual concept in an effective scene concept as ρ_{min}.

1. Remove the edges in the co-occurrence network whose weights do not surpass the thresholding value $\tau = (0.1\%) \cdot |T|$, where $|T|$is the size of the training set.

2. For all the remaining edges in the co-occurrence network, calculate the betweenness and remove the edges with large betweenness values.

3. If the removing does not generate any subnetwork S, continue doing step 2.

4. If the removing generates some subnetworks, verify their effectiveness by the above quantitative definition. If any effective subnetwork exists, and its granularity is smaller than ϱ_{max}, draw the corresponding part in the dendrogram, put the effective subnetwork S_e into the community set CS, otherwise, repeat step 2.

5. If any ineffective subnetwork exists, repeat step 2.

6. For each the candidate community S_e in the set CS with granularity less than ϱ_{max}, merge the nodes in the order of increasing edge weights, if a complete subgraph P (or under loose condition, a closed subgraph) is formed with size larger than ρ_{min}, verify its effectiveness based on the above quantitative definition, if P is effective, draw the corresponding part in the dendrogram, and put it into the scene concept configuration set $\widetilde{P}$, otherwise, repeat step 6.

7. Output the effective scene concept set $\widetilde{P}$.

The above algorithm avoids the extensive calculation of edge betweenness, while still maintains the effectiveness of the detected community and scene concept. We have measured the speed and effectiveness of our algorithm on both the practical network and randomly generated network by computer program compared to the traditional agglomerative and divisive methods.

D. Scene concept signature

After we have the individual concept signature of each database image, and the co-occurrence patterns indicated in the scene concept configuration set $\widetilde{P}$, we can calculate the scene concept signature easily by multiplying the occurring probabilities in the individual concept signature based on each scene concept configuration. The reason we can do it simply in this way is that the occurring of individual concept is independent to each other, thus,

$$\underset{P_e in \widetilde{P}}{Pr(P_e)} = \prod_{i=1 \cdots n} Pr(c_i) \tag{5}$$

where $P_e = \{c_1, c_2, \cdots, c_n\}$. We combine all the scores of all scene concept occurrences into a vector for each database image defined as the scene concept signature of it.

IV. Semantic based image retrieval

We present the proposed distance functions for similarity measure of semantic image search and retrieval in this section. First, let the individual scene concept formed in each image be represented as $IC_s signature = (c_1^I, e_1^I), \cdots, (c_m^I, e_m^I)$, where m is the number of individual concepts, c_i^I denotes the individual concept, and e_i^I denotes the supporting score of occurrence. Earth Mover's Distance (EMD) is evaluated as suitable measure of two image signatures given the predefined ground distance between each pair of signature elements. In our setting, the ground distance can be easily obtained by reversing the edge weights between concepts in the co-occurrence network, we use $d(c_i^I, c_j^I)$ to denote the ground distance of two individual concepts i and j. For two images A and B, the Earth Mover's Distance between their signatures $IC_A = (c_{A_1}^I, e_{A_1}^I), \cdots, (c_{A_m}^I, e_{A_m}^I)$ and $IC_A = (c_{B_1}^I, e_{B_1}^I), \cdots, (c_{B_m}^I, e_{B_m}^I)$ is defined as:

$$D_{EMD}(IC_A, IC_B) = \frac{\sum_{i=1}^m \sum_{j=1}^n f_{ij} d(c_{A_i}^I, c_{B_j}^I)}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}} \tag{6}$$

where f_{ij} is called a flow that is transferred from one signature to the other. The EMD is computed by first solving all the f_{ij} by linear programming. The problem is further defined as:

$$f_{ij} = \underset{min}{arg} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij} \tag{7}$$

subject to the following constraints:

$$\begin{gathered} f_{ij} \geq 01 \leq i \leq m, 1 \leq j \leq n \\ \sum_{j=1}^{n} f_{ij} \leq e_{A_i}, 1 \leq i \leq m \\ \sum_{i=1}^{m} f_{ij} \leq e_{B_j}, 1 \leq j \leq n \\ \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = min(\sum_{i=1}^{m} e_{A_i}, \sum_{j=1}^{n} e_{B_j}) \end{gathered} \tag{8}$$

The EMD can be viewed as a measure of the least amount of work needed to transfer one signature into the other, a unit of work in the process is evaluated by the ground distance.

Two scene concept signatures $SC_A = c_{A_1}^S, \cdots, c_{A_m}^S$ and $SC_B = c_{B_1}^S, \cdots, c_{B_m}^S$ are compared based on χ^2 distance measure,

$$D_{\chi^2}(SC_A, SC_B) = \frac{1}{2} \sum_{i=1}^{m} [\frac{(c_{A_i}^S - c_{B_i}^S)^2}{c_{A_i}^S + c_{B_i}^S}] \tag{9}$$

where only the pair-wise distances on the same scene concepts are considered, the reason is that there could not be any co-occurrence or other relations between different scene concept pairs.

Further, the distance between query image Q and database image D considering both the individual and scene concept signature, is defined in a weighted manner:

$$Dist(Q, D) = \omega D_{EMD}(Q, D) + (1 - \omega) D_{\chi^2}(Q, D) \quad (10)$$

The weight ω can be adjusted through the retrieval iterations based on the retrieval precision determined from relevance feedback. The updating of the weight is given by:

$$\omega_{new} = \omega_{old}(1 + Precision_{pre} + Precision_{current}) \quad (11)$$

Therefore, if the precision is increasing, which means the scene concept contributes more in the retrieval, we put more weight on the χ^2 distance.

V. Experimental evaluation

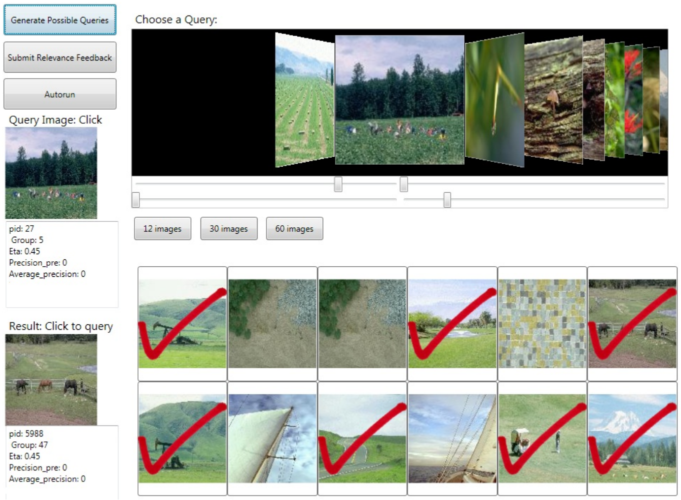
A. Experimental setup and dataset

For our experimental evaluation, we use the SUN'09 dataset which is suitable for scene concept learning for the reason that each image could contain multiple co-occurred concepts compared to other popular datasets, such as PASCAL 07. The original dataset includes 12,000 labeled images spreading over 200 object categories and covering a large scope of individual concepts (more than 6,000). The images are pre-segmented into salient regions each of which contains a certain concept. The labels are given by human annotator in the on-line annotation system LabelMe, and the consistency of labeling has been verified to be acceptable. The scene concepts are learned with randomly dividing the dataset into training set with a third of the images and the newly added image set taking the rest. We conduct the dividing for 10 times, and the retrieval performance is averaged over different divisions of the dataset.

B. Retrieval performance evaluation

We simulate the human retrieval process by letting the system automatically determine whether the retrieved images is in the same concept as the query. The ground truth labels is used for the determination. It does not override the proposed concept learning approach because the ground truth is only used for evaluating the performance. Thus the proposed method can still work on other unseen databases. To simulate the practice situation of retrieval, each image from the 200 object categories is used as the query, and the performance is evaluated after five iterations of each query. We use the average precision metric for performance measure. And we compare the retrieval performance of our system with the baseline method which only considers the pictorial feature distance (GIST, PHOG, etc), the individual concept based model where the individual concept signature is added. The interface for image retrieval is shown in Figure 5. The results averaged on image categories is provided in Table 1.

There is a clear advantage in incorporating both the individual concept signature and scene concept signature in the similarity comparison in most of the image categories, although the improvement is not huge. We found that the scene concept signature performs outstanding when the concept can not be distinguishedly described by the pictorial features or the concept varies in shape, color or other dimensions of the feature vector in different images, while the concept has close co-occurring relation to other concepts which can be easily told by the visual feature.



Query Image: Click

pid: 27
Group: 5
Eta: 0.45
Precision_pre: 0
Average_precision: 0

Result: Click to query

pid: 5988
Group: 47
Eta: 0.45
Precision_pre: 0
Average_precision: 0

Fig. 5. The proposed image retrieval system based on concept signature similarity. The red mark shows the correctly retrieved relevant images, and the performance is evaluated based on the rank of the relevant images, the initial weight ω is set to 0.45, the retrieval is conducted automatically by the system based on the ground truth labels.

Category	Baseline	Individual	Scene
car	32.71	35.84	36.65
people	29.63	30.11	32.54
wall	19.13	20.78	21.01
tree	54.32	54.17	52.10
building	30.94	32.86	35.77
grass	40.26	41.83	42.96
mountain	28.73	29.68	31.14
sea water	52.66	53.37	59.62
plants	24.45	26.75	28.82
bicycle	11.16	13.32	13.54
book	10.09	11.06	10.59
sofa	15.32	16.55	15.67

TABLE I
THE AVERAGE PRECISION STACKED OVER THE IMAGE CATEGORIES.

C. Scene concept detection evaluation

Figure 6. shows the a sub-tree of the learned dendrogram structure relating to 15 concepts. The sub-tree is correlated to a potential community structure in the entire co-occurrence network and is built up in the "cut" phase. The four scene concept configuration is learned next. Since the setup of the controlling parameters: the maximum granularity ϱ_{max} of an effective community, and ρ_{min} the minimum number of individual concept in an effective scene concept has the great infulence on the generated structure and the detected scene concepts. We also study the retrieval performance with respect to different parameter setups. The retrieval performance on the "people" category is shown in Figure 7. From the figure we observe that when the ϱ_{max} is around the range 25-30 and the

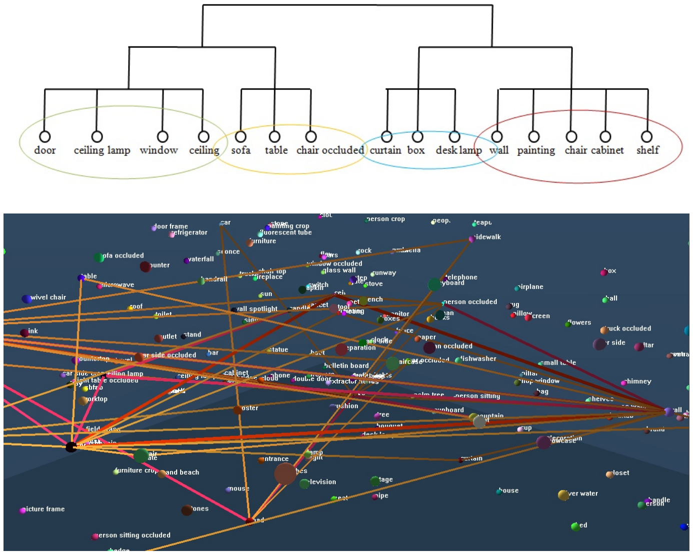


Fig. 6. Top: a sub-part of the generated dendrogram related to 15 concepts. Down: detected community structure and relevant scene concept configurations, the differnt color of edges shows the differnet scene concept.

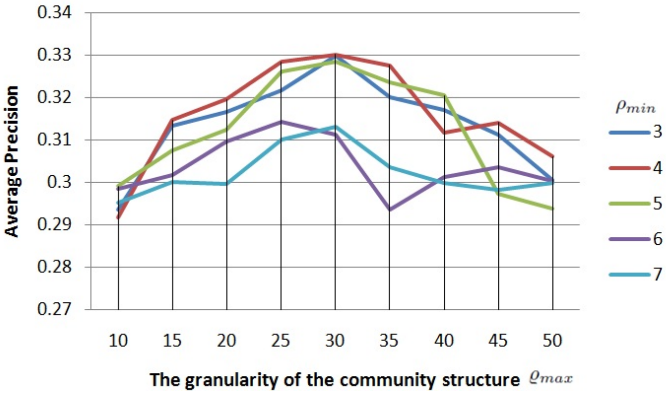


Fig. 7. The relation between the retrieval performance and the controlling parameters ϱ_{max} and ρ_{min}.

ρ_{min} equal to 4, we can have the best retrieval precision due to a good compromise between the discriminating probability and not overfitting the scene configuration.

VI. Discussion

We have introduced a model for individual concept learning and scene concept detection based on concept co-occurrence network. We provide the framework for semantic based image retrieval in accordance with the image semantic signatures. We have tested our retrieval system on practical dataset where the results are comparable or better than the state-of-art reported on this dataset. In the future, we may address the problem of learning the concepts in a long-term interactive scenario where users are requested to provide feedbacks of both visual and textual relevance from the retrieved images. As such, the system has the adaptability to concept transition in a dynamic environment, and also the learned individual and scene concept signatures can be adjusted to further capture the true concepts underlaid in the search.

References

[1] D. Grangier, S. Bengio, *"A Discriminative Kernel-Based Approach to Rank Images from Text Queries,"* IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1371-1384, August, 2008.

[2] T. Mensink, J. Verbeek, G. Csurka, *"Learning Structured Prediction Models for Interactive Image Labeling,"*, In CVPR, 2011.

[3] M. Douze, A. RamisaCombining, *"attributes and Fisher vectors for efficient image retrieval,"*, In CVPR, 2011.

[4] G. Carneiro, A. Chan, P. Moreno, N. Vasconcelos, *"Supervised Learning of Semantic Classes for Image Annotation and Retrieval,"*IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 394-410, March, 2007.

[5] T. Mitchell, *"Machine Learning,"* McGraw-Hill, 1997.

[6] J. Yuan, J. Luo, and Y.Wu, *"Mining compositional features from GPS and visual cues for event recognition in photo collections,"* IEEE Trans. on Multimedia, 12(7):705716, 2010.

[7] K. Ravi, N. Jasmine, T. Andrew, *"Structure and Evolution of Online Social Networks,"* Proceedings of the 11th ACM International Conference on Knowledge Discovery and Data Mining, pp. 611617, 2006.

[8] K. Faust, P. Dupont, J. Callut, J. Helden, *"Pathway discovery in metabolic networks by subgraph extraction,"* Bioinformatics, 2010.

[9] S. Feng, R. Manmatha, V. Lavrenko, *"Multiple Bernoulli Relevance Models for Image and Video Annotation,"* in CVPR 2004.

[10] Q. Iqbal, J. Aggarwal, *"Feature Integration, Multi-image Queries and Relevance Feedback in Image Retrieval,"* 6th ICVISUAL, 2003 in conjunction with 9th ICDMS.

[11] Y. Li and L. G. Shapiro, *"Consistent Line Clusters for Building Recognition in CBIR,"* in ICPR 2002.

[12] X. Zhou, T. Huang, *"Relevance feedback in image retrieval: A comprehensive review,"* International Journal of Multimedia Systems, Volume 8, Number 6, 536-544, 2003.

[13] B. Russell, A. Torralba, K. Murphy and W. Freeman, *"LabelMe: A Database and Web-Based Tool for Image Annotation,"* International Journal of Computer Vision, Volume 77, Numbers 1-3, 157-173, 2008.

[14] J. Li and J. Wang, *"Real-Time Computerized Annotation of Pictures,"* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 6, pp. 985-1002, 2008.

[15] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, B. Schiele, *"What helps where and why? Semantic relatedness for knowledge transfer,"* In CVPR, 2010.

[16] T. Deselaers, V. Ferrari, *"Visual and Semantic Similarity in ImageNe,"* In CVPR 2011.

[17] D. Jia, W. Dong, R. Socher, L.-j. Li, K. Li, and L. Feifei, *"ImageNet: A large-scale hierarchical image database,"* In CVPR, 2009.

[18] D. Jia, A. Berg, L. Fei-Fei, *"Hierarchical Semantic Indexing for Large Scale Image Retrieval,"* In CVPR 2011.

[19] L.-J. Li, R. Socher, and L. Fei-Fei, *"Towards total scene understanding: classification, annotation and segmentation in an automatic framework,"* In CVPR, 2009.

[20] I. Dimitrovski, D. Kocev, S. Loskovska and S. Deroski, *"Detection of Visual Concepts and Annotation of Images Using Ensembles of Trees for Hierarchical Multi-Label Classification'"* In Book: Recognizing Patterns in Signals, Speech, Images and Videos, Springer, Lecture Notes in Computer Science, 2010.

[21] M. Choi, J. Lim, A. Torralba, and A. Willsky, *"Exploiting hierarchical context on a large database of object categories,"* In CVPR, 2010.

[22] C. Galleguillos, A. Rabinovich, and S. Belongie, *"Object categorization using co-occurrence, location and appearance,"* In CVPR, 2008.

[23] M. Douze, A. Ramisa, C. Schmid, *"Combining attributes and Fisher vectors for efficient image retrieval,"* In CVPR, 2011.

[24] D. Parikh, K. Grauman, *"Interactively Building a Discriminative Vocabulary of Nameable Attributes,"* In CVPR, 2011.

[25] J. Yuan, M. Yang, Y. Wu, *"Mining Discriminative Co-occurrence Patterns for Visual Recognition,"* In CVPR, 2011.

[26] D. R White, F. Harary, *"The cohesiveness of blocks in social networks: Node connectivity and conditional density,"* In Sociological Methodology. 31, 305359, 2001.

[27] M. Girvan, M. E. J. Newman, *"Community structure in social and biological networks,"* In Proceedings of the National Academy of Sciences of the USA, vol. 99 no. 12 7821-7826, 2002.