SEMANTIC-VISUAL CONCEPT RELATEDNESS AND CO-OCCURRENCES FOR IMAGE RETRIEVAL

Linan Feng, Bir Bhanu

Center for Research in Intelligent Systems, University of California, Riverside, USA

ABSTRACT

This paper introduces a novel approach that allows the retrieval of complex images by integrating visual and semantic concepts. The basic idea consists of three aspects. First, we measure the relatedness of semantic and visual concepts and select the visually separable semantic concepts as elements in the proposed image signature representation. Second, we demonstrate the existence of concept co-occurrence patterns. We propose to uncover those underlying patterns by detecting the communities in a network structure. Third, we leverage the visual and semantic correspondence and the co-occurrence patterns to improve the accuracy and efficiency for image retrieval. We perform experiments on two popular datasets that confirm the effectiveness of our approach.

Index Terms— Image retrieval, image semantics, concept signature, complex images

1. INTRODUCTION

Semantic image retrieval is the problem of acquiring images that have similar semantic concepts to the supplied target from large image databases. Visual complexity arises when images embrace complex scenes comprising a group of single concepts. Traditional content-based image retrieval paradigm loses its effectiveness when the low-level features not correlate well with the high-level semantics anticipated by users. Similarly, the performance of text-based retrieval systems deteriorate due to the ambiguous meanings of keywords.

Consequently, several approaches that help learn visual and semantic concepts simultaneously have been proposed from different perspectives, e.g., exploring implicit correspondences between visual features and high-level knowledge in a hierarchical structure [1]; utilizing semantic contexts to disambiguate visual word meanings in the bag-of-visual-word model [2]; improving visual concept detector performance by measuring semantic word similarity and co-occurrences [3] for video retrieval. Also, the idea of concept co-occurrences has become popular for recognizing objects by the co-occurred attributes such as material, shape, etc [4].

The problems of existing work are: 1) Most methods [1, 4] take for granted that all the semantic concepts are similarly discriminable by visual contents, without investigating the actual relatedness, for example, the semantic concept "dog" could have stronger links to visual features such as "four-legged" and "has paw" than "animal". 2) Although semantic concept co-occurrence is explored as context of high-level descriptions, there are no explicit co-occurrence patterns that been discovered for scene understanding and image retrieval [3, 8]; 3) Semantic concept similarity is useful for comparing image contents. Current methods measure the word distances defined in WordNet [1, 3] based on meanings, thus, "horse" is closer to "tiger" than to "windmill". We argue that for complex scene images, the pair of "horse" and "tiger" may have less chance to appear in the same image than "horse" and "windmill". Therefore, the measure from concept co-occurrence could be more important.

The main contributions of this paper is to uncover concept co-occurrence patterns for image retrieval. The idea is to detect the communities (graph clusters) from a proposed co-occurrence network, which to our knowledge has not been explored in any previous work. Other contributions include: using concept signature to represent and retrieve images, semantic concept selection by evaluating the semantic-visual relatedness, and the distance metric based on concept signature to compare images. We demonstrate the approaches for image retrieval using the Outdoor Scene Recognition (OSR) dataset and Scene Understanding dataset (SUN09).

2. TECHNICAL APPROACH

The key ingredients of our proposed approach are: (a) Semantic and visual concept relatedness measure for concept selection; (b) Co-occurrence detection from selected concepts; and (c) Image retrieval with concept signatures.

2.1. Semantic and visual relatedness

In this section, we describe the proposed approach for selecting the most visually correlated semantic concepts. We measure the semantic and visual relatedness by evaluating the visual variability within a concept and the visual distances to other concepts. The selected semantic concept draws upon the fact that its within concept visual variability is less significant than the averaged pairwise concept distances.

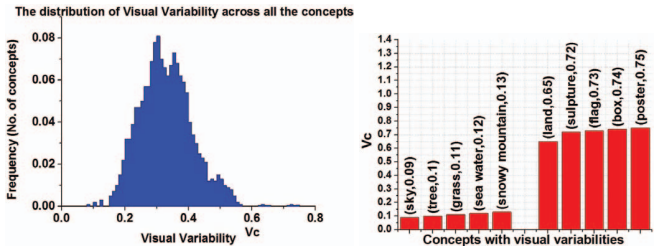
2.1.1. Intra-Concept Visual Variability

The within concept visual variability measures how much visual variation exists among instances of a concept. For a given

concept $\mathcal{C}$, we collect all the regions from entire dataset containing $\mathcal{C}$. The within $\mathcal{C}$ visual variability $V_{\mathcal{C}}$ is measured by the sum of distances between the mean feature descriptor f_{mean} of $\mathcal{C}$ and all the region feature vectors $f_i, i \in \mathcal{C}$ in $\mathcal{C}$,

$$V_{\mathcal{C}} = \frac{1}{\mid \mathcal{C} \mid} \sum_{i \in \mathcal{C}} D_{\chi^2}(f_i, f_{mean}) \tag{1}$$

where D_{χ^2} is the the χ^2 distance between two feature vectors. The distribution of $V_{\mathcal{C}}$ across all the concepts is shown in Figure 1(a) as a rough Gaussian distribution with two peaks at variability 0.30 and 0.35. The top 5 examples of concepts with smallest and largest intra concept visual variations are exhibited in Figure 1(b). As expected, we observe semantic concepts have large variability when they are diversified in visual properties indicating a weak semantic-visual relatedness.



(a) (b)

Fig. 1. (a) The frequency of visual variabilities accumulated through all the concepts. (b) shows the concepts with extreme variability values.

2.1.2. Inter-Concept Distance

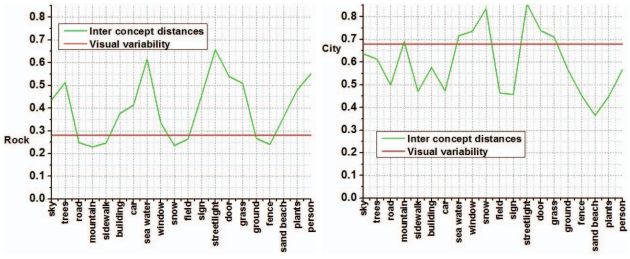
We measure the visual distance between one concept $\mathcal{C}$ and the rest of the concepts as the inter-concept variation. The distance $D(\mathcal{C}, \mathcal{C}^{'})$ between $\mathcal{C}$ and another concept $\mathcal{C}^{'}$ is defined by the averaged χ^2 distance between the mean feature vector of $\mathcal{C}$ and all the region feature vectors in $\mathcal{C}^{'}$

$$D(\mathcal{C}, \mathcal{C}^{'}) = \frac{1}{\mid \mathcal{C}^{'} \mid} \sum_{i \in \mathcal{C}^{'}} D_{\chi^2}(f_i, f^c_{mean}) \tag{2}$$

We measure the distance between each pair of $\mathcal{C}$ and $\mathcal{C}^{'} \in \overline{\mathcal{C}}$, where $\overline{\mathcal{C}}$ is the complementary set. Finally we determine semantic concept $\mathcal{C}$ is visually discriminative if its visual variability and averaged distance to other concepts satisfies the following inequality

$$V_C - \frac{1}{N-1} \sum_{C^{'} \in \overline{C}} D(C, C^{'}) < 0 \tag{3}$$

The intra-concept variability for "rock" is shown in Figure 2(a) and for "city" is shown in Figure 2(b) as the red flat line. The distances to other concepts are shown in the figures as the green curve. The inequality (2) can be estimated by the difference between the areas formed by the two lines under the straight line and above the straight line. As illustrated by the figure, "rock" is more visually discriminative than "city", thus, it is more likely to be selected in the concept signature.



(a) (b)

Fig. 2. (a) The difference between the visual variability of concept "rock" and the distances to other concepts is illustrated as the areas between the two curves and (b) similarly for concept "city".

2.2. Co-occurrence detection of concepts

One way to analyze the relationships between concepts is to represent them in a network structure where the nodes corresponds to individual concepts and the edges indicate co-occurrences. A very common property of complex networks is called community structure, i.e., groups of nodes have tight internal connections and loose external connections to each other. We consider that if a group of semantic concepts always occur together, they reflect a co-occurrence pattern. Co-occurrence detection is close to Graph Clustering (GC) or Graph Partition (GP). The modularity optimization [6] analysis has been adopted as an efficient way to solve it.

We build concept co-occurrence network of the selected concepts from the training images. We model the concepts in the candidate vocabulary obtained from previous section as nodes in the network. We model the co-occurrence relationship as connecting edges between nodes. The co-occurring frequency between two concepts across the entire training set is further assigned to the corresponding edge as weight.

Modularity is introduced as a measure of the quality of a particular partition of the network. Given a adjacency matrix A, we define the modularity of the partition between two communities C_i and C_j as

$$Q = \frac{1}{2d} \sum_{i,j} [A_{ij} - \frac{k_i k_j}{2d}] \delta(C_i, C_j) \tag{4}$$

where d equals half of the summation of all the edge weights in the adjacency matrix, A_{ij} represents the edge weight between node i and j, k_i (k_j) equals the summation of the weights of the edges attached to node i (j), C_i and C_j are the community IDs, $\delta(C_i, C_j) = 1$ if $C_i = C_j$, otherwise $= 0$. Experiments show the value of Q equals to or is greater than 0.3 indicating a good community. The modularity is calculated over all the pairs of nodes in the network.

We consider iteratively merging the nodes into a hierarchical community structure with different levels of resolution by maximizing the modularity gain in each iteration. The modularity gain of moving an outside node i into a community C is evaluated by

$$\begin{aligned} \Delta Q = [\frac{\Sigma_{in} + k_{i,C}}{2d} - (\frac{\Sigma_{out} + k_i}{2d})^2] - \\ [\frac{\Sigma_{in}}{2d} - (\frac{\Sigma_{out}}{2d})^2 - (\frac{k_i}{2d})^2] \end{aligned} \tag{5}$$

where Σ_{in} represents the sum of edge weights inside C, $k_{i,C}$ equals the sum of weights of edges that link i to C, d is the same as defined in equation (7), Σ_{out} is the sum of weights of edges that link outside nodes to nodes in C, k_i is the sum of weights of the edges incident to i.

We divide the algorithm for uncovering co-occurrence patterns into *separating* and *merging* phases.

Algorithm 1: Uncovering concept co-occurrence patterns

Separating phase:

1. Suppose N nodes in the network, assign each node a different community tag $C_i, i = 1, ..., N$.

2. For each node V_i in the network, attempt to remove it from its own community C_i and add it into each of its neighboring nodes V_j's community C_j, $j = 1, ...n$.

2.1 If placing V_i from C_i to C_j produces a positive maximum modularity gain evaluated by equation (5), examine the value of Q_{C_i} and Q_{C_j} with V_i assigned to each community by equation (4).

2.1.1 If both Q_{C_i} and Q_{C_j} are ≥ 0.3 which implies a potential share of individual concept between scenes, split node V_i into V_i and $V_i^{'}$ and put into C_i and C_j separately, the edges incident to other nodes are copied between the them.

2.1.2 Else place node V_i into C_j.

2.2 Otherwise, all the nodes stay put.

3. The first phase stops when every node is traversed and no further improvement can be achieved.

Merging phase:

1. Replace each of the uncovered communities by a single node and replace the edges by a single edge with weight being equal to the sum of the weights of the edges it represents.

2. Represent the edges in the same community as a self-looped edge with weight equaling to the sum of the weights of the inside edges.

Iteration:

Repeat above two phases until no modularity gain given by eq.(5) can be achieved.

The algorithm iteratively generates a hierarchical structure of communities, in other words, the communities of concepts, and the communities of communities.

2.3. Retrieval with concept signature

We train concept detectors from the regions of the labeled training images. When applied to a region in the unlabeled testing image, each detector d_c can give a score $s \in \mathbb{R}$ according to how strongly the region contains a particular concept c. Assuming the detector is reliable, which means regions containing the concept can yield higher scores than irrelevant regions. We conduct visual region and semantic concept detection simultaneously within a novel image based on the discovered co-occurrence patterns in the previous section. Suppose we have a pool of co-occurrence patterns $P = P_1, P_2, ..., P_i$ with each P_i denoting a concept pattern $c_1, c_2, ..., c_j$, our goal is to find the best match between regions of a novel image I $R_I = r_1, r_2, ..., r_k$ and the co-occurrence pattern which can generate the maximized score

$$S^* = \underset{P_i \in P, c_j \in P_i, r_k \in R}{\arg\max} \sum_{i,j,k} d_{c_j}(r_k), \text{subject to } d_{c_j}(r_k) \geq 0 \quad (6)$$

The problem is similar to the optimal assignment problem in a complete bipartite graph, where each edge weight w_{ij} denotes the corresponding sub-score obtained from $d_{c_j}(r_k)$. It can be solved by using Kuhn-Munkres algorithm [8].

After finding the optimal pattern and assigning the concepts in the pattern to image regions, we construct the concept signature for each image by combining the scores of corresponding concepts into a vector. We propose to compare and retrieve images based on the signature similarity with Earth Mover's distance metric given the pre-defined ground distance between each pair of concepts as the inverse of the edge weights in the co-occurrence network.

3. EXPERIMENTAL RESULTS

3.1. Datasets and features

OSR and SUN09 are used for evaluation because have segmented regions with hand labeled annotations. OSR has 2,682 images with 520 individual concepts across 8 outdoor scene categories. SUN09 contains 12,000 images and more than 5,800 individual concepts covering a variety of indoor and outdoor scene categories. They offer several advantages as compared to other datasets: 1) Both contain complex scene images. 2) Both datasets have manually labeled concepts associated with corresponding bounding boxes. 3) Both datasets have labels in the same LabelMe format with similar labels. For both datasets, we use the following visual features: 1) Color GIST feature, the orientation histogram of the object boundary. 2) PHOG, the Pyramid of histogram of oriented gradients. 3) PHOG with oriented edges, which considers the direction of the salient Canny edges. 4) Pyramid of self similarity feature, which is a log-polar histogram of correlations between central and surrounding pixels.

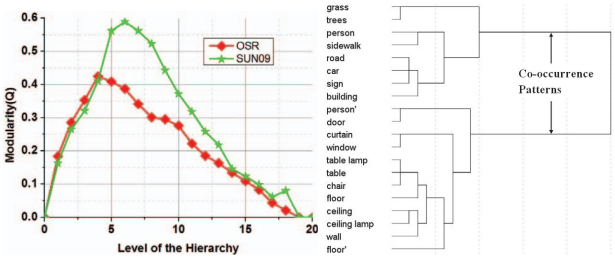
3.2. Performance Measure

The image retrieval performance is evaluated by the number and ranking of the relevant retrieved images to the query. We request three human assessors to launch queries with each image and provide the relevance information. The decision of relevance is made by majority voting by the three human assessors. Further statistic evaluation relies on standard image retrieval measures: 1) Average Precision of top N retrieved images. 2) Precision of top N retrieved images measuring the percentage of relevant images that are able to be encountered by a user within the first N results in the retrieval engine.

3.3. Experimental results

We use 130 selected concepts based on their semantic-visual relatedness from SUN'09 dataset, and 90 concepts from OSR, to uncover the co-occurrence patterns. Figure 3(a) shows the

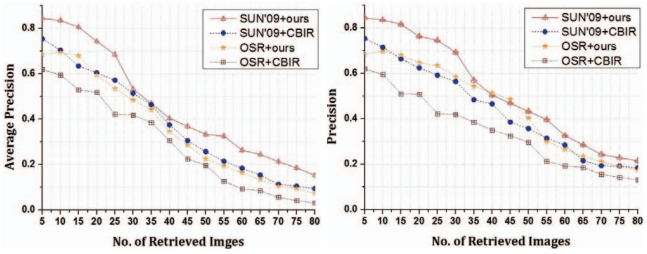
modularity changes over different levels of the hierarchical concept patterns. We observed the maxima of modularity of OSR at level 3 with $Q \approx 4.3$ and the maxima of SUN'09 at level 5, $Q \approx 0.52$. This indicates that individual concepts in SUN'09 have more significant co-occurrence property than OSR, and even at lower level of SUN'09, the co-occurrence is comparable to OSR. Figure 3(b) shows an example of the part of the detected community structure in SUN'09. The hierarchical structure shows two obvious co-occurrence patterns indicated by the longest two lines. We also observed two concepts "person" and "floor" copied and split themselves into two sub-patterns which hints at a overlapping between co-occurrence patterns.



(a) (b)

Fig. 3. (a) Modularity changes as a function of the level of the hierarchy of the co-occurrence patterns. (b) a part of the hierarchical structure detected in SUN'09 dataset.

We compare the retrieval performance of our method with conventional content-based image retrieval method introduced in [6] which only considers the visual feature similarity. We provide the same parameter settings as in the previous experiment. Figure 4 summarizes the results. Our model consistently outperforms the other approach on the two datasets. It shows the effect of visually discriminative semantic concept co-occurrence patterns. We note a high precision for OSR, this may be due to the number of concepts presented in the dataset is relatively small, and we find the patterns in more compact form.



(a) (b)

Fig. 4. Comparison of our retrieval approach with the CBIR approach in [6] (a) Curves of the Average Precision on the two datasets as a function of the number of retrieved images. (b) Curves of the Top-N Precision rate on the two datasets as a function of the number of retrieved images.

Finally, we show the effects of our semantic image retrieval approach in Figure 5. Images with indoor and outdoor scenes have been used as queries. We observe that more semantically rather than visually relevant images have been retrieved by our approach. These results underscore the effectiveness of the semantic facilitation by our approach.

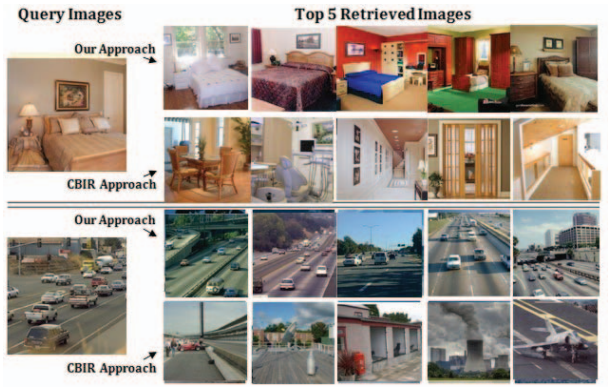


Fig. 5. Image retrieval by considering the concept similarity versus those without exploiting concept similarity. Our approach retrieves more semantically relevant images than conventional CBIR.

4. CONCLUSIONS

In this paper, we presented a novel semantic retrieval approach for complex scene images based on (a) semantic-visual relatedness measure method, (b) semantic concept selection and concept co-occurrence detection method, and (c) concept signature distance metric for image retrieval. By using real world images from OSR and SUN09, the proposed approach shows robustness and efficiency against current content-based image retrieval paradigms. The future research direction may consider including relevance feedback from user to adjust the learned concept signatures.

5. ACKNOWLEDGMENT

This work was supported in part by NSF grants 0641076 and 0727129.

6. REFERENCES

[1] D. Jia, A. Berg and L. Fei-Fei. Hierarchical semantic indexing for large scale image retrieval. In *CVPR* 2011.
[2] S. Yu, J. Frederic. Visual word disambiguation by semantic contexts. In *ICCV* 2011.
[3] Y. Aytar, M. Shah, and J. Luo. Utilizing semantic word similarity measures for video retrieval. In *CVPR* 2008.
[4] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR* 2009.
[5] M. E. J. Newman. Fast algorithm for detecting community structure in networks. In *Physical Review E* 2004.
[6] Y. Peng-Yeng, B. Bhanu, C. Kuang-Cheng and D. Anlei. Long-term cross-session relevance feedback using virtual features. In *IEEE Trans on KDE* 2008.
[7] J. Munkres. Algorithms for the Assignment and Transportation Problems. In *J'SIAM* 1957.
[8] Y. Junsong, Y. Ming and W. Ying. Mining discriminative co-occurrence patterns for visual recognition. In *CVPR* 2011.