Image Search by Concept Map *

Hao Xu† Jingdong Wang‡ Xian-Sheng Hua‡ Shipeng Li‡
†MOE-MS KeyLab of MCC, University of Science and Technology of China, Hefei, 230026, P. R. China
‡Microsoft Research Asia, Beijing 100190, P. R. China
xuhao657@ustc.edu, {jingdw, xshua, spli}@microsoft.com

ABSTRACT

In this paper, we present a novel image search system, *image search by concept map*. This system enables users to indicate not only what semantic concepts are expected to appear but also how these concepts are spatially distributed in the desired images. To this end, we propose a new image search interface to enable users to formulate a query, called *concept map*, by intuitively typing textual queries in a blank canvas to indicate the desired spatial positions of the concepts. In the ranking process, by interpreting each textual concept as a set of representative visual instances, the concept map query is translated into a *visual instance map*, which is then used to evaluate the relevance of the image in the database. Experimental results demonstrate the effectiveness of the proposed system.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.5 [**Information Storage and Retrieval**]: Online Information Services

General Terms

Algorithms, Experimentation, Performance

Keywords

image search, concept map

1. INTRODUCTION

Digital image is nowadays the second most prevalent media in the Web only after text. Image search engines play an important role in enabling people to easily access to the desired images. A variety of search interfaces have been employed to let users submit the query in various forms, e.g., textual input, image input, and painting based input, to indicate the search goal. To facilitate image search, query formulation is required not only to be convenient and effective for users to indicate the search goal clearly, but also to be easily interpreted by image search engines. Therefore, recently more and more research attention has been paid on search interface design [26, 4, 15, 18, 23] in developing image search engines. In this paper, we focus on designing a new image search system to address a specific kind of image search scenario: the user concerns not only the presence but also the spatial distribution of the specified semantic concepts. An example is illustrated in Fig. 1, where the user wants to "find the images containing a butterfly on the top-left of a flower". Such image search intention concerning the spatial distribution of concepts, which is called *concept layout* hereafter, is common when people try to visually convey ideas, such as making illustration for a slide show.

Existing commercial image search engines, e.g., Google image search [10] and Microsoft Bing image search [16], provide a textbox for users to type one or more keywords to indicate the search goal. Then the text-based search technique is adopted to match the textual query with the textual metadata associated with images. This type of search interface is easy to use. However, besides the limitation that the associated texts may not reveal the image content, it is not easy to perform image search with the requirement on the concept layout. The search results by "butterfly flower", as shown in Fig. 1, do not come up to expectation (only two relevant images in top ten images). Adding the spatial description, "on the top-left of", to the query will not improve the search results, since such spatial description rarely appears in the metadata of Web images.

Conventional content-based image retrieval techniques [12] usually require users to input a visual query, e.g., an example image or a painted sketch, to indicate the desired visual content, such as shape or color distribution [20, 7, 22]. Then visual features extracted from the visual query are used to match with the images in the database. Such system is often inconvenient to use because the users have to submit an image that may not be available at hand, or paint a sketch that may not be easy to precisely indicate the search goal. Besides, such techniques suffer from at least two drawbacks: visually similar images may have different semantic contents; visual query has limited capability to represent the semantic search goal, since same semantic content may have different visual appearances. Therefore, content-based image search techniques are difficult to handle the image search task in Fig. 1. This is in line with the unsatisfactory search results by visual queries in Fig. 1.

Recently, some efforts have been made to take advantage of both textual queries and visual queries. The "show similar images" feature of Microsoft Bing image search enables users to pick an image as the example image from the search results of a textual query, and then promote the visually similar images. Another investigation was conducted by painting color strokes to indicate the desired spatial distribution of colors and promoting the images satisfying the color distribution from the search results of a textual query [21].

*This work was performed at Microsoft Research Asia.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

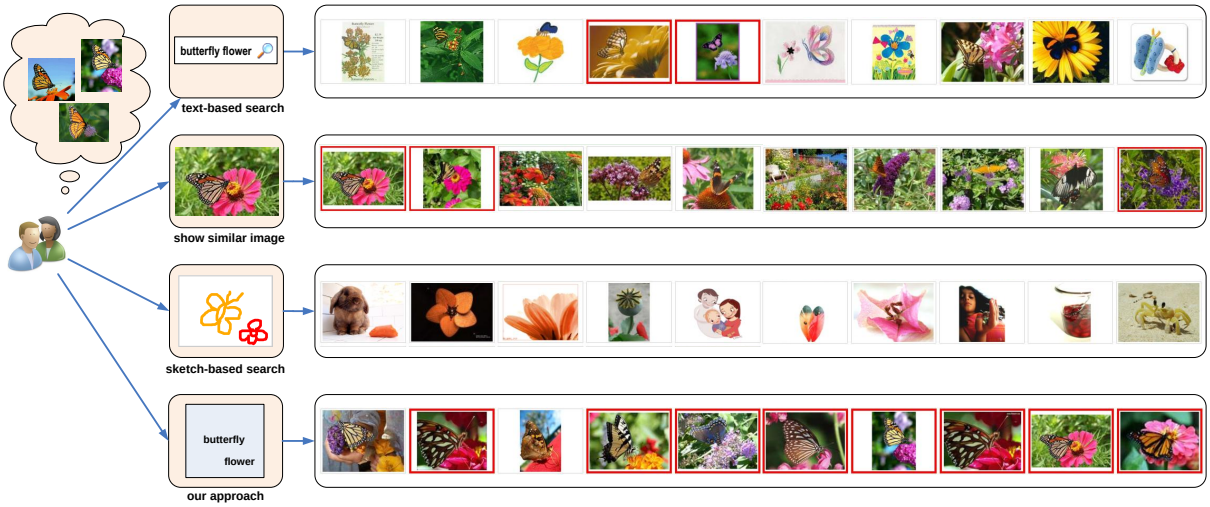


Figure 1: Illustration of four types of image search schemes: text-based image search, "show similar images" of Microsoft Bing image search, sketch-based image search, and our approach. The user's search intention is "finding the images containing a butterfly on the top-left of a flower". The images accord with the search intention are highlighted in red boxes.

These techniques may, to some extent, help users to find the images that satisfy some specific semantic and visual requirements, but still suffer from the drawback that the search intention concerning the concept layout is difficult to indicate. The search results of "show similar image" are shown in Fig. 1. Selecting a good example image that exactly accords with the search intention does not improve the search results significantly.

In this paper, we propose a novel image search system, which presents a novel interface to enable users to intuitively indicate the search goal by formulating the query in a visual manner, i.e., explicitly indicating where the concepts should appear. This new interface is very intuitive and only requires users to type textual queries at the arbitrary positions of a blank canvas to formulate the query, which is called *concept map* in this paper. Compared with the traditional textual query, the concept map query allows users to indicate not only what concepts should appear in the desired images, but also where these concepts should appear. Compared with the visual query, the concept map query allows users to semantically indicate the search goal rather than submitting an image or a sketch that cannot convey the clear semantic intention. We have previously demonstrated an early version of our system [22]. In this paper, we report on extensions, technical details and experimental analysis of the system. Fig. 1 shows a concept map and the corresponding search results of the proposed system, in which most retrieved images accord with the search intention.

Technically, it becomes possible to compute the relevance of an image according to the user's spatial intention, since desired spatial distribution of concepts is visually expressed by a concept map query in an explicit manner. We introduce an effective scheme to translate a *concept map* to a *visual instance map*, which represents each concept as a set of visual instances exploiting the techniques of text-based image search and representative image selection. The visual instance map is then used to evaluate the relevance of the image in the database. Besides, we present an interactive tool to allow users to select a few visual examples to assist to visually specify what a concept looks like. Experiments demonstrate the effectiveness of the proposed image search system, including the new query formulation interface and the relevance evaluation scheme.

In summary, the key contributions of this paper are as follows: 1) We present a novel image search system to enable users to search images with the requirement on the spatial distribution of semantic concepts. 2) We propose a novel image search interface to enable users to intuitively input a so-called concept map by typing textual queries in a blank canvas to indicate what concepts should be included and where they should appear in the desired images. 3) We introduce an effective technique to translate the concept map into visual instance map, and a robust scheme to evaluate the relevance of an image with the visual instance map.

2. RELATED WORK

The image search intention concerning the spatial layout has been investigated in the previous works [20, 7]. The VisualSEEk system presented in [20] uses joint color/spatial queries to perform image search. The users are allowed to freely draw color strokes in a canvas to indicate the desired color and spatial position of visual content in the desired images. A similar idea is explored in [7], only that, instead of color, the desired shape and spatial position of the visual content is considered. These works concern for the search intention of the spatial layout of visual content from the perspective of the low level signal similarity. In contrast, the proposed system concerns for the search intention of the layout of semantic concepts, that is, what expected to appear in a specific position of the desired images is defined by a semantic keyword but not particular color or sketch.

Zavesky and Chang propose an image search scheme called CuZero in [25]. It adopts a concept navigation map to assist users to view the image search results. Note that the concept navigation map in this work is distinct from our concept map, because concept navigation map is a tool for users to straightforwardly select the weights of the concepts to generate a weighted textual query, thus it has nothing to do with the spatial distribution of the concepts in the desired images.

Even more recently, Chen et al. propose a novel system called Sketch2Photo in [3] to synthesize an image by drawing some sketches in a blank canvas and defining each sketch with a keyword. Assisted by the keyword, it is able to find the visual object matched with the input sketch more accurately from the Web, and thus synthesize a more satisfactory image. Our proposed system differs from this work at least in two aspects. First, our goal is to search for the existing images but not synthesize a new one. Second, the Sketch2Photo system requires users to input some sketches and

Figure 2: User interface of the proposed system.

generate the image by matching these sketches, while our system does not require users to draw sketch to describe the shape of the concept but focuses on the spatial relation of the concepts.

3. OVERVIEW

The snapshot of the user interface of the proposed system is shown in Fig. 2. To perform image search, the user first formulates a concept map query, by typing one or more keywords (concepts) in *query formulation canvas*. As the example in Fig. 2, the concept map consists of three concepts, "sky", "house" and "lawn", which are expected to appear from top to bottom. After submitting the query, the system returns a list of images according to their relevances with the query, shown in *search result panel*. We also find that in some cases it may be insufficient to merely use a textual keyword to indicate a desired concept, and hence provide an advanced tool, called *visual assistor*, to allow the user to select a few visual examples to assist to visually specify or narrow down what the desired concept looks like. The user may click *advanced function* button to popup the *visual assistor* window. For each concept, a set of visual examples are presented, as shown in Fig. 2. The user may select the examples that are visually close to what are desired.

To evaluate the relevance of an image in the database with the concept map, we propose to transform the concept map to the visual instance map, which replaces each keyword with a set of visual instances and generates a probabilistic map, by estimating the spatial intention, to tolerate the roughness of the input spatial position in the concept map. A layout-sensitive relevance evaluation scheme is then used to compare the visual instance map with the images in the database.

4. APPROACH

4.1 Query Formulation

The procedure of formulating a concept map query is not only convenient, as it only requires the user to simply select a position and type textual keywords, but also intuitive, since the user can express the spatial intention in a visually straightforward manner. To make the interface even more user-friendly, a series of intuitive manipulations are supported to let the user formulate and edit the query more easily, i.e., the user may delete or edit an existing keyword, or modify the position of a concept by dragging the corresponding keyword.

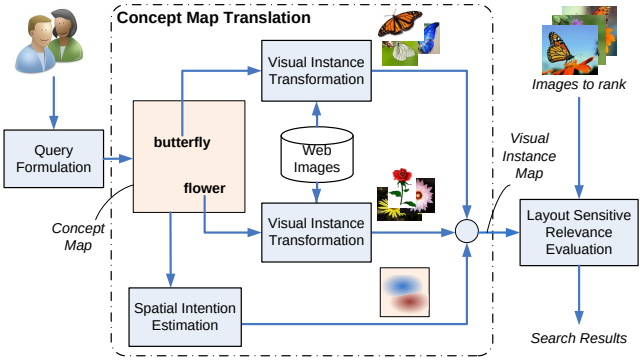


Figure 3: The flowchart of concept map translation. A concept map with two concepts "butterfly" and "flower" is used as the example input.

In practice, besides the positions of the concepts, the user may also care about the sizes of the concepts, i.e., the sizes of the regions occupied by the concepts. Therefore, a rectangle is associated with each keyword to indicate the influence scope of the keyword, i.e., the scope of the spatial distribution of the corresponding concept. Once a keyword has been input, a rectangle with the default size, one-ninth of the size of the query formulation canvas, is assigned to it. Our interface allows the user to explicitly specify the influence scope of a keyword by stretching the corresponding rectangle. The rectangle is visible around a keyword for editing when the user moves the mouse over the keyword.

4.2 Concept Map Translation

Relevance evaluation will be straightforward if the images in the database are also represented in the form of concept map, i.e., the information of what concepts appear in the image and where they appear is available. Recently, a lot of efforts, including human tagging and automatic annotation, have been taken to extract such information. Flickr [8] provides a feature to let users add notes in local regions of the images. The LabelMe project [19] also presents a tool to users to help manually assign tags to local regions of the images. The precision of manual annotation may be well guaranteed, but it has some difficulties in the practical applications since we are facing Web-scale images and Web-scale concepts. The computer vision community has tried to investigate automatic techniques [5, 2, 14], which need to collect training images and train a discriminative classifier for each concept, and hence also suffer from the Web-scale concepts.

Instead, we follow the instance-based classification methodology to translate a concept map into a visual instance map, which encodes both the visual appearance and desired spatial distribution of each concept in the concept map, and use the visual instance map to evaluate the relevance of the image in the database. The flowchart of concept map translation is illustrated in Fig. 3.

4.2.1 Visual Instance Transformation

Existing text-based image search engines are successful if the textual query is relatively simple, e.g., a single concept without spatial intention. This means that the search results in some degree can visually represent the concept. On the other hand, the research in the pattern recognition community has shown that the instance-based classifier is promising and can obtain competitive performance with discriminative classifier. These motivate us to exploit the text-based image search technique to transform the textual concept to a few visual instances for the subsequent relevance evaluation.

We first collect a set of images by querying a text-based image search engine with the keyword of the concept and then adopt the affinity propagation (AP) algorithm [9] to find the visual instances, so as to obtain the representative appearances of the concept. AP has been shown to be effective and efficient to find a set of exemplars from a set of images [6]. To reduce the computation cost, we run AP only for the top ranked images returned by the text-based image search engine as we found that the main visual appearances of a concept are almost covered by the top ranked images. In our implementation, the visual similarities are evaluated on the regions containing salient objects instead of the entire images. This helps alleviating the influence of the background. The salient object is detected offline using the learning based technique [13]. Considering that small image groups are more likely to be formed of irrelevant images, we sort the obtained exemplars in a descending order of the sizes of their associated image groups, and take the detected salient objects of the top V exemplars as the visual instances of the concept. Experimentally, we found our algorithm is to some extent robust to the quality of the mined visual instances, and feeding the top 50 images of the text-based image search results to AP is adequate to guarantee the satisfactory performance.

4.2.2 Spatial Intention Estimation

A concept map includes raw information related to the spatial intention, the position and the influence scope, of each concept. Our goal here is to estimate the spatial intention and represent it by a group of spatial distributions, each of which corresponds to a concept. The spatial intention estimation follows two principles: 1) a concept has larger probability to appear near the specified position; 2) a concept is not expected to appear in the position where another concept should appear.

We denote a concept map as $\{(w_k, r_k)\}_{k=1..K}$, with w_k and r_k being the kth keyword and the associated rectangle respectively, and K the number of concepts in the concept map. The desired spatial distribution D_k of concept k is estimated by

$$D_k(x,y) = \begin{cases} G_k(x,y) & G_k(x,y) = \max\limits_{j=1..K} G_j(x,y) \\ 0 & \text{otherwise} \end{cases}, \quad (1)$$

where $G_k(x,y)$ is a 2D Gaussian distribution, with the mean $\mu_k = [x_k, y_k]^T$ and the covariance matrix $\Sigma_k = \mathrm{Diag}((\theta w_k)^2, (\theta h_k)^2)$. Here, h_k, w_k and (x_k, y_k) are height, width and center coordinates of rectangle r_k respectively. The shape of the resulting distribution is determined by the height and the width of the rectangle. θ is set to a constant $\sqrt{(2\log(2))^{-1}}$ to make the distribution degrade to a half near the boundary of the rectangle. The smaller the rectangle is, the more rapidly the distribution degrades to zero from its center, indicating the concept is expected to appear within a smaller area.

4.3 Layout Sensitive Relevance Evaluation

In this section, we present the approach to calculate the relevance of an image with the visual instance map. Two aspects are taken into consideration during the relevance evaluation: 1) whether each concept occurs in the image; 2) whether the occurrence is spatially consistent to user expectation.

4.3.1 Image Representation

We follow the state-of-the-art image representation technique, and extract a Bag-of-Words (BoW) feature. Rather than building a global BoW model, we build a partially spatial BoW model, which will be helpful to compute the spatial occurrences of the concepts. We divide the image uniformly into $n \times n$ cells, and extract visual features for each cell. An image is thus represented as $\{\mathbf{f}_c\}_{c=1..n\times n}$, where $\mathbf{f}_c$ denotes a feature vector extracted from the cell c. In our implementation, each entry f_{cj} in $\mathbf{f}_c$ is the number of the associated visual word (e.g., quantized SIFT feature [17] and color feature) in the corresponding cell. Experimentally we find $n = 9$ works well.

Accordingly, we extract visual features to represent the visual instance map with a feature map, $\{(\mathcal{F}_k, D_k)\}_k$. Here $\mathcal{F}_k = \{\mathbf{f}_v^k\}_{v=1..V}$ is a set of BoW vectors, with each vector corresponding to a visual instance of concept k.

4.3.2 Relevance Evaluation

The relevance evaluation consists of two main steps: calculating a relevance score for each concept and combining these relevance scores to an overall relevance score.

Relevance Evaluation Per Concept

Since a concept is represented as a few visual instances, we reduce the problem of estimating a relevance score for a concept to checking whether there appears a visual instance of the concept at the expected position in the image.

For a visual instance v of concept k, we first compute an occurrence map $O_v^k = [o_{v1}^k, \cdots, o_{vc}^k, \cdots, o_{vn^2}^k]^T$ by calculating the similarity between v and each cell of the image:

$$o_{vc}^k = \mathrm{sim}(\mathbf{f}_v^k, \mathbf{f}_c) = \sum\nolimits_j \min(f_{vj}^k, f_{cj}). \quad (2)$$

This similarity measure is similar to histogram intersection. The slight difference is that the BoW vectors are not normalized because we do not aim to check whether the concept appears in a single cell, but aim to inspect how much part of the concept appears in the cell by counting how many common visual words they share.

Appearance consistency: Given the occurrence map O_v^k, we evaluate the appearance consistency, i.e., the degree that visual instance v appears in the image, as $\alpha_{v1}^k = \sum_c o_{vc}^k$, which can be viewed as the count of common visual words of visual instance v and the image.

Spatial consistency: In order to check whether the occurrence of the concept is spatially consistent to user expectation, we compare the spatial distribution of visual instance v in the image, which is approximated by the normalized occurrence map $\tilde{O}_v^k$, with the desired spatial distribution of the concept k, say D_k. The spatial consistency, denoted as α_{v2}^k, is calculated as:

$$\alpha_{v2}^k = \sum\nolimits_c \mathrm{sign}(d_c^k)\min(|d_c^k|, \tilde{o}_{vc}^k), \quad (3)$$

$$\tilde{o}_{vc}^k = \frac{o_{vc}^k}{\sum_{c'} o_{vc'}^k}, \quad (4)$$

$$d_c^k = \begin{cases} \frac{1}{z_1}(D_k(x_c,y_c) - \omega) & D_k(x_c,y_c) \geq \omega \\ \frac{\rho}{z_2}(D_k(x_c,y_c) - \omega) & \text{otherwise} \end{cases}, \quad (5)$$

where (x_c, y_c) is the center coordinates of cell c. ω and ρ control the degree of penalty to the case that the concept appears at the position not expected. $\omega = \frac{1}{3}\max_{x,y} D_k(x,y)$ and $\rho = 0.5$ in our implementation. z_1 and z_2 are partition factors to make $\sum_{c:d_c^k>0} d_c^k = 1$ and $\sum_{c:d_c^k<0} d_c^k = -\rho$. There are two parts contribute to α_{v2}^k: a positive part by $\{d_c^k | d_c^k > 0\}$, and a negative part by $\{d_c^k | d_c^k < 0\}$. Note that the negative part is introduced to penalize the inconsistent spatial distribution in a harsher manner and the degree of penalty is controlled by parameter ρ.

The relevance score for concept k is then calculated by combining the appearance consistency and spatial consistency of all related visual instances:

$$\beta_k = \max_v \alpha_{v1}^k \alpha_{v2}^k. \quad (6)$$

Relevance Fusion

Finally, a scalar relevance score of the image is obtained by combining the relevance scores of all the concepts, $\boldsymbol{\beta} = \{\beta_k\}_{k=1..K}$. A fusion function that naively takes the average will overestimates the images having perfect scores for some concepts but very poor scores for the others. To take account of every concept, the final relevance score is calculated with a fusion function sensitive to both the average and the variance of the input vector:

$$rel = E(\boldsymbol{\beta}) - \frac{\gamma}{K}\sum_{k=1}^{K}|\beta_k - E(\boldsymbol{\beta})|, \quad (7)$$

where $E(\boldsymbol{\beta}) = \frac{1}{K}\sum_k \beta_k$ is the average of the vector. γ is a positive parameter controls the penalization degree for the input vector with big variance, which is heuristically set to 0.8 in our implementation. Given two vectors with the same average, it is straightforward that the one with smaller variance will get larger score by Eqn. (7).

4.4 Advanced Functions

In the proposed system, two advanced functions, influence scope selection (ISS) and visual example selection (VES), are provided to assist users to further indicate the search goal.

The user may have a specific spatial intention for a concept, e.g., expecting the concept "lawn" to be filled in a long and narrow region at the bottom of the image. In such a case, the user can indicate the spatial intention in an explicit manner by ISS. Specifically, ISS associates each keyword with a rectangle and enables the user to explicitly control the influence scope of a keyword by stretching the rectangle. The desired spatial distribution for the concept is then estimated taking account of the shape of the rectangle as mentioned in Section 4.2.2.

The user may be interested in a particular appearance of a concept, e.g., expecting only yellow flower for the concept "flower". In such a case, the user can indicate what the concept looks like by VES. Specifically, VES assists the user to express the intention by visual assistor. Candidate visual examples of each concept are listed in a showcase, as shown in Fig. 2, from which the user can select the desired ones for a concept. The visual instances of a concept are then obtained from the selected visual examples.

5. EXPERIMENTS

5.1 Setting

To evaluate the proposed system, we design 42 image search tasks with the requirement on the concept layout. Fig. 4 illustrates some examples of the tasks. Among all the tasks, 22 tasks involve only a single concept, 16 tasks involve two concepts and 4 tasks involve three concepts. There are totally 33 concepts in these tasks, which are categorized as follows: 6 scenes (sky, grass, desert, beach, lawn, fire), 4 landmarks (Pyramid, Sphinx, Great Wall, Colosseum), 5 cartoon characters (hello kitty, garfield, pooh, snoopy, teddy bear) and 18 real world objects (polar bear, tiger, dolphin, seagull, butterfly, panda, bamboo, Christmas tree, tulip, flower, jeep, car, bicycle, house, windmill, keyboard, mouse, us flag). The concepts that do not have stable appearance, such as "rain", are not supported in our system and handling concepts like people's names by introducing face-related features is our future work.

To perform quantitative evaluation, some volunteers are recruited to label the ground truth. Given a task, an image is labeled with a relevance score, according to how well the image accords with the search intention of the task. To differentiate the relevance degrees,

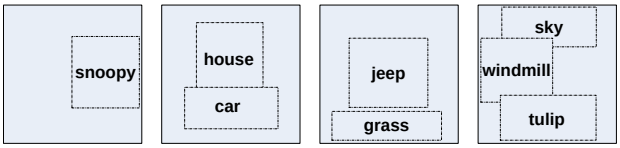


Figure 4: Examples of the image search tasks in the experiment. Each task is represented as a concept map (desired sizes of the concepts are represented by the rectangles) to describe the desired concept layout.

the relevance score is defined in four levels from level 0 to level 3. Level 3 corresponds to the most relevant (all concepts appear as expected), and level 0 the least relevant (some concepts missed or the layout of concepts is quite different from what is expected). With the ground truth, Normalized Discounted Cumulative Gain (NDCG) [11] is adopted to measure the image search performance.

In our implementation, two kinds of visual features are adopted: color feature and SIFT feature. For color feature, each pixel is represented in the HSV color space and quantized into 192 levels. For SIFT feature, the extracted SIFT features are quantized into visual words with a visual vocabulary of size 6K in the way of [1]. Given a concept, we select which feature to use adaptively to measure the visual similarity, exploiting query classification method in [24]. The image database is dependent on the specific task. Specifically for a given task, the image database is obtained by querying a text-based image search engine using the keywords of the task. All possible combinations of keywords are used as the textual query one at a time, and the top 500 images in the search results are merged to form the database.

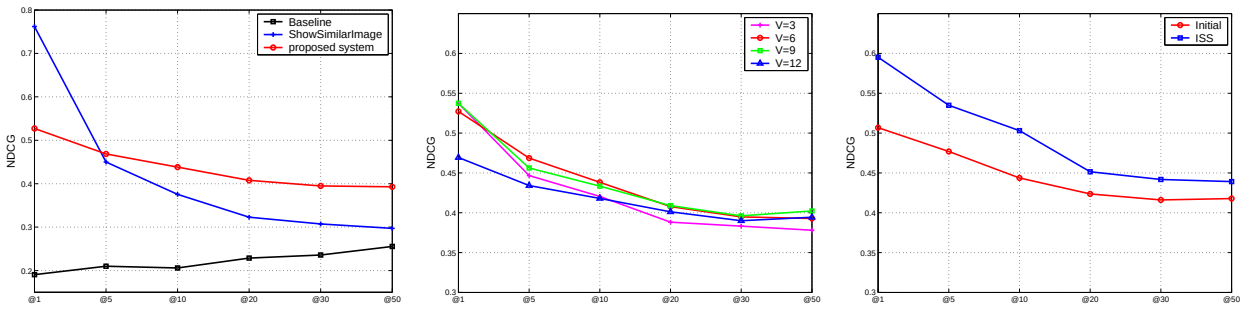
5.2 Quantitative Search Performance

The text-based image search system is taken as the baseline in the quantitative comparison. To accomplish a task with a text-based image search system, the keywords of the task are used to query the system. Taking the task of Fig. 4(c) as an example, four queries, "jeep", "grass", "jeep grass" and "grass jeep", can be formulated to get four different search results. An NDCG score is calculated for each search result and the best one is taken as the baseline performance for this task. Free-text query, such as "jeep at center and grass at bottom", is not adopted in the experiment, since the search results are too noisy.

We also involve the "show similar image" search in the quantitative comparison. The basic idea of this method is to pick an example image from the search results of the text-based image search system to rerank the initial search results. To accomplish a given task, we collect the most relevant images of the task from the top 20 images of the search results of the baseline system, and use them one at a time as the example image to perform "show similar image" search. The performance is evaluated by averaging the NDCG scores obtained in all these trials. This setting is to simulate the searching process in a practical case, i.e., from the top 20 images displayed in the screen, the user may select one of the most relevant images she think to conduct "show similar image" search, expecting more relevant images to be returned.

To accomplish a task with the proposed system, we use the concept map of the task as the query, except that we only input the keywords in the specified positions without explicitly indicating the influence scopes of the keywords. The other advanced function, visual example selection, is also disabled.

The performance of each of the three methods is measured by averaging the NDCG scores over all tasks and depicted in Fig. 5(a). As can be anticipated, the text-based image search system works



(a) (b) (c)

Figure 5: (a) Quantitative comparison of the image search performance for the three different methods. (b) Illustration of the influence of the number of visual instances V. (c) Illustration of the effectiveness of the advanced function *influence scope selection*.

poorly, since it is difficult to handle the tasks with the spatial intention. The "show similar image" method gets a good score for NDCG@1, because the image ranked at the first place is supposed to be the example image itself, which must be relevant otherwise the user would not select it as the example image. However, the performance of "show similar image" method drops rapidly along with growth of the NDCG depth. It is not difficult to explain, since visually similar images are not ensured to have the same semantic concepts. The proposed system shows a satisfactory performance. It outperforms the baseline remarkably in all NDCG depth and is superior to "show similar image" method except of NDCG@1. This means the search intention is well interpreted by the proposed system through the concept map query.

In Fig. 5(b), we show the influence of parameter V, which is the number of visual instances selected for each concept. As can be seen, the performance peaks at the medium values, say $V = 6$ and $V = 9$. We explain this as follows. When the number of visual instances is too small, the various appearance of a concept cannot be covered, so that many relevant images in the database are not found. When the number of visual instances is too large, a certain number of noisy visual instances irrelevant to the concept are selected, so that many irrelevant images are mistakenly ranked high. This experiment shows that setting parameter $V = 6$ is a good tradeoff between the search performance and computation cost. The experiments in this section all adopt $V = 6$.

5.3 Visual Results

In this subsection, we present some visual results to visually demonstrate the advance of the proposed system. In Fig. 6, we compare the search results by different methods for the 3 tasks related to the concept "snoopy": "snoopy at left", "snoopy at right" and "snoopy at top". The text-based image search system uses the same query for the 3 tasks (free-text query like "snoopy at left" is not considered in the experiment) and produces the identical search results (Fig. 6(a)). Since it has no sense of the desired layout of the concepts in the task, the position of snoopy is quite inconsistent in the search results. The search results of "show similar image" for the task "snoopy at right" is depicted in Fig. 6(b). The image ranked at the first place is the example image used to perform the search. Observed from the search results, this method ranks the images mainly according to the color similarity, which mistakenly interprets the search intention. In Figs. 6(c), 6(d), 6(e), we show the concept maps adopted for accomplishing the 3 tasks and the corresponding search results of the proposed system. Clear spatial correspondence of the concept "snoopy" between the input concept map and the retrieved images can be observed. We also illustrate the mined visual instances for the concept "snoopy" in Fig. 6(f). Note that not all the visual instances perfectly represent the appearance of "snoopy". The third visual instance is the back of snoopy and the fifth is a dog which has nothing to do with snoopy. This justifies that our algorithm can produce satisfactory search results in the presence of imperfect visual instances, which relaxes the requirement of the accuracy of the algorithm for visual instance transformation.

More visual results for the tasks involving multiple concepts are illustrated in Fig. 7. The demonstrated search results of the text-based image search system is the best one selected from the set of the search results, which is obtained using all the possible keyword combinations as query one at a time. The proposed system produces much better results compared with baseline method. For example, for the task of "finding a jeep at center and grass at bottom" (Fig. 4(c)), only one desired image can be found using the text-based search system, while 8 of top 10 images in the search results accord with the search intention using the proposed system.

5.4 Evaluation of Advanced Functions

In this subsection, we evaluate the effectiveness of the advanced functions provided in the proposed system: influence scope selection (ISS) and visual example selection (VES).

We first perform an experiment to evaluate ISS quantitatively using our image search tasks. To accomplish a task, we use the concept map of the task as the query, indicating explicitly the influence scopes of the keywords this time. The other advanced feature, visual example selection, is still disabled. Fig. 5(c) shows the performance of the proposed system with ISS enabled, together with the initial results with both advanced functions disabled. We can see that the performance is improved considerably by enabling ISS. This shows that the proposed system can well interpret the indication of the influence scopes of the keywords.

We demonstrate the effectiveness of VES and ISS with a few visual examples in Fig. 8 and Fig. 9. In Fig. 8, we show two search results for the task of "finding a jeep at the center" together with the selected visual examples for the concept "jeep". It is obvious that the search results are influenced by selecting different visual examples for the concept. The jeeps in front view are retrieved by selecting three front view visual examples, while the jeeps in side view are retrieved by selecting three side view visual examples. In Fig. 9, we show two concept maps with the same keywords but different rectangles, indicating the different spatial distributions of the concept "windmill" are desired. The search results satisfactorily reflect such search intention.

5.5 User Study

A user study with 20 participants is performed to justify our system. The participants taking part in the experiments are college students from nearby universities, and all have the experience of

Figure 6: Visual results for the image search tasks related to the concept "snoopy". (a) shows the top 10 images in the search results of the text-based image search system using textual query "snoopy". (b) shows the top 10 images returned by "show similar image" for the task "snoopy at right". (c,d,e) are the three concept maps (the rectangle reflecting the influence scope of the keyword is with default size and not displayed) and the corresponding search results of the proposed system. (f) shows the mined visual instances of "snoopy" (in greed boxes). The relevant images are highlighted in red boxes.

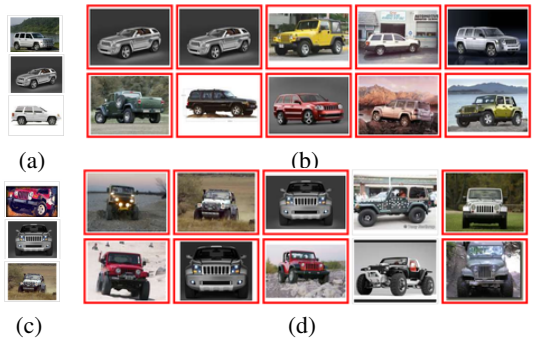


Figure 8: Illustration of the effectiveness of visual example selection. (a) and (c) show two different sets of visual examples selected for the concept "jeep", indicating to find a jeep in side view and a jeep in front view respectively. (b) and (d) are the search results corresponding to (a) and (c) respectively, using the same concept map query.

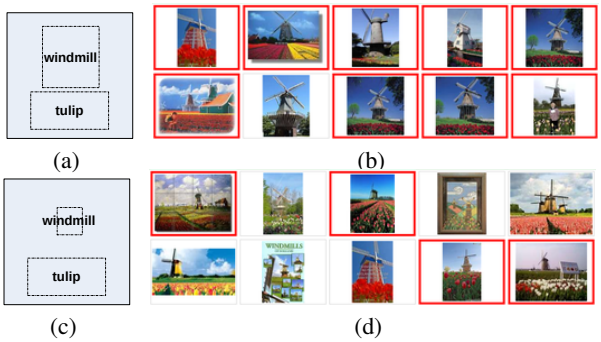


Figure 9: Illustration of the effectiveness of influence scope selection. (a) and (c) show two concept maps with the same concepts but different influence scopes for the concept "windmill", indicating to find a big windmill and a small windmill respectively. (b) and (d) are the search results obtained by (a) and (c) respectively.

using image search engines. After a brief introduction of our system, they are asked to try our system for a few minutes and then fill a questionnaire.

To the question "have you ever had any image search intention concerning the concept layout?", 20% of respondents replied with *"yes"* and 50% of respondents replied with *"no, but probably in the future"*. It shows there are a certain amount of image search intentions concern about the concept layout. Moreover, several respondents gave us the concrete examples that they think the system may help them: *"It is useful for me to search a wallpaper that Michael Jordan appears at the right of the picture." "I was searching for an interesting picture that I saw before to share with my friends. I remembered there is a cute bear lying at the bottom of the picture, but failed to find it by the search engines with the textbox interface. It should be much easier to search it using this interface."* In summary, the user study justifies that the proposed image search system, which handles specifically the image search intentions concerning the concept layout, is meaningful in practical use.

6. CONCLUSION

We have presented a novel image search system to enable users to search images with the particular requirement on the concept layout. A new formulation of image search query called concept map is introduced so that the user can explicitly indicate not only the desired concepts but also the expected layout of the concepts in the desired images. In the implementation, a concept map is first translated into a visual instance map by mining visual instances for each concept from the Web and estimating the spatial intention concerning the concept layout. The relevance of an image is evaluated by comparing the visual instance map with the image. Experiments demonstrate that the proposed system is effective to handle the particular kind of image search tasks.

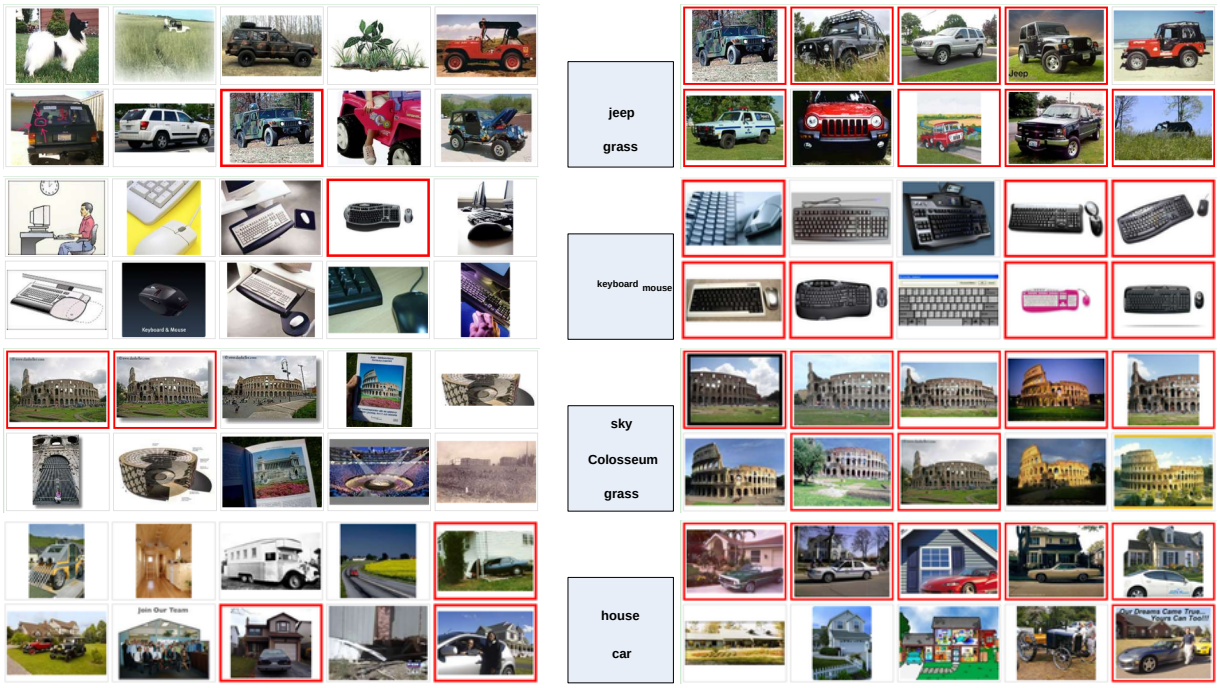


Figure 7: Visual results for the image search tasks involving multiple concepts. The left column shows the search results of the text-based image search system. The middle column shows the concept maps for the tasks (the rectangles reflecting the influence scopes of the keywords are all with default size and not displayed). The right column shows the search results of the proposed system. The relevant images are highlighted in red boxes.

7. REFERENCES

[1] J. Sivic, B. C. Russell, A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *ICCV '05*, 2005.

[2] M. B. Blaschko and C. H. Lampert. Learning to localize objects with structured output regression. In *ECCV '08*, pages 2–15, 2008.

[3] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu. Sketch2photo: internet image montage. In *SIGGRAPH Asia '09*, pages 1–10, 2009.

[4] J. Cui, F. Wen, and X. Tang. Intentsearch: interactive on-line image search re-ranking. In *MM '08*, pages 997–998, 2008.

[5] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV '09*, 2009.

[6] D. Dueck and B. J. Frey. Non-metric affinity propagation for unsupervised image categorization. In *ICCV '07*, 2007.

[7] M. J. Egenhofer. Spatial-query-by-sketch. In *IEEE Symposium on Visual Languages*, pages 60–67, 1996.

[8] Flickr. `http://www.flickr.com/`.

[9] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972, 2007.

[10] Google Image Search. `http://images.google.com/`.

[11] K. Jarvelin and J. Kekalainen. IR evaluation methods for retrieving highly relevant documents. In *SIGIR '00*, pages 41–48, 2000.

[12] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *TOMCCAP*, 2(1):1–19, 2006.

[13] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *IEEE Trans. Pattern Anal. Mach. Intell.*, to appear.

[14] X. Liu, B. Cheng, S. Yan, J. Tang, T. S. Chua, and H. Jin. Label to region by bi-layer sparsity priors. In *MM '09*, pages 115–124, 2009.

[15] Y. Luo, W. Liu, J. Liu, and X. Tang. Mqsearch: image search by multi-class query. In *CHI '08*, pages 49–52, 2008.

[16] Microsoft Bing Image Search. `http://images.bing.com/`.

[17] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004.

[18] G. P. Nguyen and M. Worring. Optimization of interactive visual-similarity-based search. *TOMCCAP*, 4(1), 2008.

[19] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. LabelMe: a database and web-based tool for image annotation. *IJCV*, 77(1):157–173, 2008.

[20] J. R. Smith and S.-F. Chang. Visualseek: A fully automated content-based image query system. In *MM '96*, pages 87–98, 1996.

[21] J. Wang, X. Hua, and Y. Zhao. Color-Structured Image Search. MSR-TR-2009-82. July 7, 2009.

[22] H. Xu, J. Wang, X. Hua, and S. Li. Interactive Image Search by 2D Semantic Map. In *WWW '10*, 2010.

[23] R. Yan, A. Natsev, and M. Campbell. Multi-query interactive image and video retrieval -: theory and practice. In *CIVR '08*, pages 475–484, 2008.

[24] R. Yan, J. Yang, and A. G. Hauptmann. Learning query-class dependent weights in automatic video retrieval. In *MM '04*, pages 548–555, 2004.

[25] E. Zavesky and S.-F. Chang. Cuzero: embracing the frontier of interactive visual search for informed users. In *MIR '08*, pages 237–244, 2008.

[26] Z.-J. Zha, L. Yang, T. Mei, M. Wang, and Z. Wang. Visual query suggestion. In *MM '09*, pages 15–24, 2009.