Context Search and Recommendation for Large-Scale Community-Sharing Photos

(大規模社群媒體中的影像搜尋與推薦)

Carson Liao, Liang-Chi Hsieh, Winston Hsu
Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan

■ Background

The explosive growth of digital videos/photos, the prevalence of capture devices, and the phenomenal success in WWW search have helped attract increasing interest in investigating new solutions in image/video search and navigation. This interest is gradually extended to the novel domain of community-contributed multimedia as the growing practice of online public media sharing. Billions of images shared on websites such as Flickr[1] bring profound social impact to the human society, and pose a new challenge for the design of efficient indexing, searching and visualizing methods for manipulating them.

Current image or video search approaches are mostly restricted to text-based solutions which process keyword queries against text tokens associated with the media, such as speech transcripts, captions, file names, etc. For shared consumer photos, text clues come from tags or descriptions that are added by users via some light-weight annotation tools. The associated tags may contain abundant information, yet their qualities are not uniformly guaranteed. In most photo-sharing websites, tags and other forms of text are freely entered and are not associated with any type of ontology or categorization. Tags are therefore often inaccurate, wrong or ambiguous [1]. In particular, due to the complex motivations behind tag usage [2], tags do not necessarily describe the content of the image [3]. Therefore, one primary task of a search system is to retrieve accurate images from the noisy tags.

The other attribute of consumer photos is their diversity in content, context, and aesthetic aspects. To paraphrase Susan Sontag [32], “everything exists to end up in a photograph.” Even the photos of the same object can be visually very different as they capture different aspects of the object. Therefore, to represent this diversity and enhance the search experience, a retrieval system should recommend additional information such as query-related tags [25] or canonical images [16].

In this project, we propose and evaluate a novel system for searching over large-scale shared consumer photos. The proposed methods are to utilize rich context cues of consumer photos and the ability to automatically improve (by reranking) search result and recommend additional information (i.e., query-related tags and canonical images) at query time. By investigating variant ranking algorithms (e.g., ListNet or RankSVM) for the reranking framework, we believe the ordinal information can be better

[1] http://flickr.com

exploited and result in a more efficient and effective reranking method. We also look into some novel context cue selection methods (e.g., wc-tf-idf) to improve the performance of reranking. In addition, we will devise efficient canonical image selection algorithms to generate multiple representative views related to the query without time-consuming clustering procedures. We will conduct the evaluation over large-scale benchmarks such as TECVID [13] and collected photos from social media (i.e., Flickr).

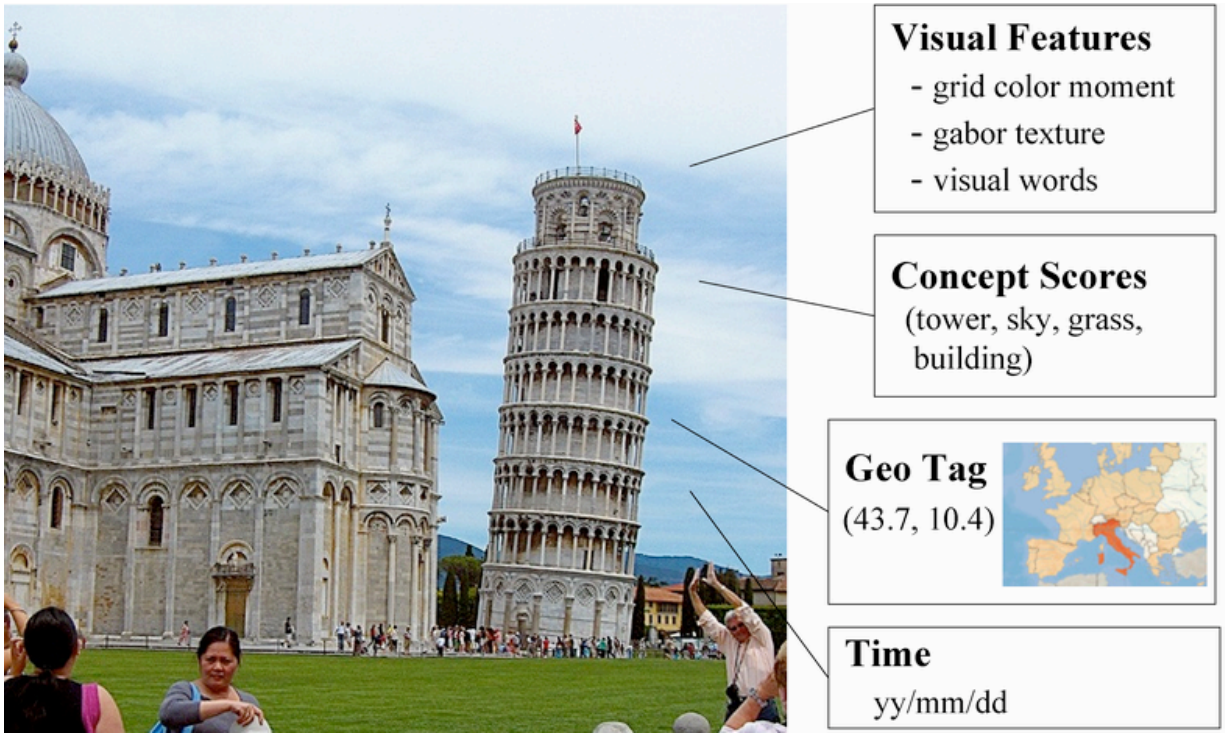


Figure 1: Context cues in consumer photos. Visual features and (high-level) concept scores can be obtained by feature extractors [22] and concept detectors [5]; location and time metadata will be exceedingly available from location-aware camera-phones and digital cameras [6].

■ Objectives and Observations

To address these issues, we propose to utilize rich *context cues*[2] of shared consumer photos to improve the search result and to recommend relevant tags and canonical images. Generally, the context cues of consumer photos contain the low-level visual features, (high-level) concept scores, and time and location metadata[3], as illustrated in **Figure 1**.

Though text-based search model (query by keywords over surrounding text tokens) is not perfect, certain *context cues* do exist in the retrieved images. **Figure 2**gives an illustrative example of text-based search result for the query "eiffel tower." While **Figure 2**(a) correctly contains the query object in sight, the remaining ones fail to for noisy text descriptions (e.g., **Figure 2**(b) is actually a picture of Tokyo Tower, by which Eiffel Tower is usually compared with.) or noisy tags (e.g., **Figure 2**(c)–(e) are the images taken from, underneath and nearby Eiffel Tower; naturally they are tagged with the keyword "eiffel tower" by the photo owners.). However, by investigating the visual content, a search system might be able to learn that the information needs (or *target semantics*) of the user is related to tower-like objects and exclude Figs. 2(c)–2(e) from the search result. Likewise, **Figure 2**(b) can be detected as false positive by examining the location metadata (i.e., geo tags). Time metadata can also be important for activity-related

[2] The meanings of *context* are usually application-dependent [4]. Here, we refer to context as those attributes describing *who*, *where*, *when*, *what*, etc., shared by documents forming the recurrent patterns.

[3] Visual features and concept scores can be obtained by feature extractors or pre-trained concept detectors [5]. Location metadata, or geo tags, will be exceedingly available, primarily from location-aware camera-phones and digital cameras, and initially from user input [6]. Time stamps are readily associated with the photos in digital capture devices.

Figure 2: Example results of a text-based search model for the query "eiffel tower." Because the associated tags or text descriptions are noisy, (b)–(e) are retrieved inaccurately. The incorporation of context cues (e.g., visual words, geo-locations, time, etc.) can mitigate this problem.

queries, such as "christmas eve" or "oktoberfest[4]." In other words, via mining the co-occurrence of context cues, or *contextual patterns*, we can uncover rich information that the user is looking for.

[4] See explanations in http://en.wikipedia.org/wiki/Oktoberfest. Generally, "Oktoberfest" represents (beers) festival in October and originates from Germany.

■ Prior Work and Comparisons

The use of context cues such as visual content and concept scores has been studied and shown to improve upon text-based video search systems, but such multi-modal approaches mostly require either extra training data [19] or multiple query example images [22], which could be difficult for users to prepare. To utilize context cues in an unsupervised fashion and maintain the text-based search paradigm preferred by most users [12], the reranking framework is recently proposed [7]-[11]. Approximating the initial search result of a baseline model as the *pseudo* target semantics, reranking mines the contextual patterns directly from the initial result. The learnt contextual patterns are then leveraged to refine (by reordering) the search result such that relevant images are ranked higher, i.e. assigned with higher relevance scores. This is corresponding to the intuition that users usually put more emphasis on the relevance of top-ranked images.

Reranking for keyword-based photo and video search has been explored in prior work for use in searching for broadcast news videos [7]–[11] and shown to offer 10%–30% relative performance gain to text-based model. Therefore, it is promising to apply reranking to the consumer photo domain, whose content are more unorganized and diverse, but closer to our ambient life. Additionally, previous works typically rely on mining contextual cues from visual features (e.g., color and texture) for reranking [7], [8], which may enhance the visual coherence of the top-ranked images, yet at the same time loses the diversity. To mitigate this problem and achieve semantic coherence, the exploitation of other context cues such as time and location metadata will also be studied in this work.

Besides effectiveness, time efficiency is also critical for the success of an on-line retrieval system. To achieve satisfactory query-time performance (i.e., after issuing the query, the result is obtained instantly such that the user is unaware of the additional computation), we propose a novel reranking method, ordinal reranking, to formulate reranking as a ranking problem and employ linear neural network model [21] to solve it. We also propose to investigate effective selection method for rich context cues in order to automatically select informative context cues to improve the performance of reranking.

On the other hand, to select canonical images for photo/video search recommendation, most existing methods involve the computation of pairwise similarity or the employment of clustering algorithms, which would be overwhelmingly time-consuming for online applications [24]. We will take a fundamentally different approach and select the images that contain the most informative context cues (i.e., visual words [29]) measured by tf-idf-like methods (to be investigated) in a greedy fashion, which is expected to be finished in seconds into order to deliver prompt responses. Relevant tag recommendation can also be conducted efficiently as they are mined directly from the associated tags of retrieved images using tf-idf-like methods at query time.

■ Proposed Methods

In this proposal, we will investigate the usage of context cues to enhance search quality and recommend query-related tags or canonical images. As depicted in **Figure 3**, the flow diagram of the proposed framework follows the text-based search paradigm and only requires a user to key in (arbitrary) text queries. The context cues of the image database are extracted in advance, and utilized in both the reranking and recommending methods. The proposed framework is totally unsupervised and requires no extra training data or off-line learning processes.

■ Reranking

To maintain the text-based search paradigm while improving the search result, the reranking framework is proposed to automatically rerank the initial text search results based on the auxiliary information[5], thereby contextual cues, from the retrieved objects in the initial search results [7]. Approximating the initial result as the *pseudo ground truth* of the target semantic, reranking algorithms mine the contextual patterns directly from the initial search result and further rerank it.

Rooted in pseudo-relevance feedback for text search [19], several reranking algorithms have been proposed in the literature. The IB reranking method [7] finds the optimal clustering of images that preserves the maximal mutual information between the initial relevance scores and extracted features based on the information bottleneck principle. The context reranking method [8] mines the contextual cues by a biased random walk along the context graph, where video stories are nodes and the edges between them are weighted by multimodal similarities of the extracted features. The classification based reranking [9] takes the higher-ranked and lower-ranked images of a baseline system as pseudo-positive and pseudo-negative examples to train a discriminative model (i.e., support vector machines, SVM), and regards the normalized classification score for each image as its reranked score. Despite the promising relative performance gains that can be obtained, these approaches neglect the fact that reranking is also a kind of ranking problems and make no use of the underlying ordinal information of the initial list. In addition, the classification-based reranking method suffers from the ad-hoc nature of the mechanism for determining the threshold for noisy binary labels.

More recently, we propose a novel ordinal reranking method [10] to incorporate and compare the learning to rank algorithms such as RankSVM [20] and ListNet [21] to the reranking framework. Since the

[5] Google's PageRank [33] is one of the representative works to utilize auxiliary information (i.e., hyperlinks) to gauge the quality of text-based retrieved web pages; the authors in [8] also shortly correlated the video reranking method with PageRank.

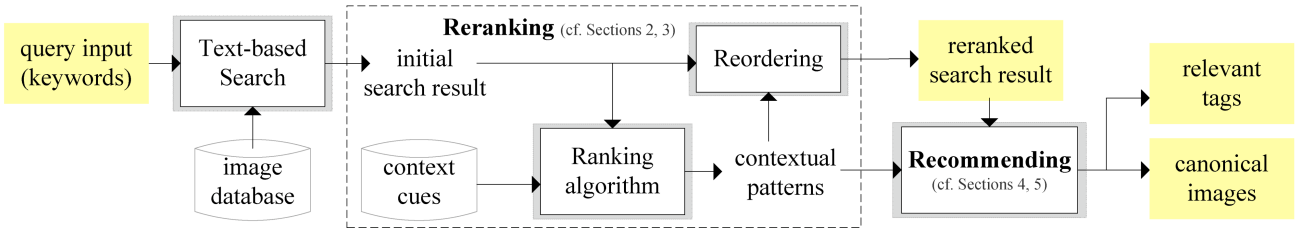


Figure 3: Flow diagram of the proposed framework. The contextual patterns are mined from the initial search result of a text-based model and then utilized to refine (by reranking) the initial result. Search-related tags and multiple canonical images are further recommended to the users.

objective of learning to rank algorithms is to minimize errors in object ranking, ordinal reranking is by nature more effective for mining ordering information and free of the ad-hoc thresholding problem. When evaluated on TRECVID 2005 video search benchmark, ordinal reranking outperforms existing methods in a great margin [10]. Moreover, thanks to the linear kernel of ListNet, ordinal reranking is remarkably efficient; it takes less than one second to rerank the result of a query. Below we first introduce the learning to rank task and the ListNet algorithm, and then describe the ordinal reranking algorithm.

Learning to Rank and ListNet

Any system that presents results to a user, ordered by a utility function that the user cares about, is performing a ranking, in contrast to classification problem which aims to determine class labels. A common example is the ranking of search results from the search engine (e.g., Google). A ranking algorithm assigns a relevance score to each object, and ranks the object by that. The ranking order represents the relevance of objects with respect to the query. In the literature, the task of training a ranking model which can precisely predict the relevance scores of test data is commonly referred to as *learning to rank*.

For learning to rank, a query is associated with a list of training data $D = (d_1, d_2, \ldots, d_N)$, where N denotes the number of training data and d_j denotes the j-th object, and a list of manually annotated relevance scores $Y = (y_1, y_2, \ldots, y_N)$, where $y_j \in [0, 1]$ denotes the relevance score of d_j with respect to the query. Furthermore, for each object d_j a feature vector $X_j = (X_{j1}, X_{j2}, \ldots, X_{jM})$ is extracted, where M is the dimension of the feature space. The purpose of learning to rank is to train a ranking algorithm f that can accurately predict the relevance score of test data by leveraging the co-occurrence patterns among $\mathbf{X}$ and Y. For the training set D we obtain a list of predicted relevance score $Z = (z_1, z_2, \ldots, z_N) = (f(X_1), f(X_2), \ldots, f(X_N))$. The objective of learning is formalized as minimizing the total losses $L(Y, Z)$ with respect to the training data, where L is a loss function for ranking. Conventional ranking algorithms such as RankSVM [20]formulate the learning task as classification of object pairs into two categories (correctly ranked and incorrectly ranked) and define L as the number of wrongly classified objects. These approaches are thus time-consuming as they take every possible pair in the data and runs at a complexity of $O(N^2)$.

ListNet [21] conquers these shortcomings by using score lists directly as learning instances and

minimizing the listwise loss between the initial list and the reranked list. In this way, the optimization is conducted directly on the list, and the computational cost can be reduced to $O(N)$, making online reranking applications possible. Our previous studies [10] show ListNet is surprisingly efficient and even outperforms the conventional pairwise approaches such as RankSVM.

To define the listwise loss function, ListNet transforms both the (pseudo-) ground truth scores and the predicted scores into probability distributions by sum-to-one normalization, and uses cross-entropy to measure the distance (listwise loss function) between these two probability distributions. Let $P(y_j)$ denotes the normalized score of d_j, then the loss function is defined as:

$$L(Y,Z) = -\sum_{j=1}^{N} P(y_j)\log(P(z_j)). \qquad (1)$$

To minimize Eq. (1), ListNet employs a linear neural network model to assign a weight to each feature and forms the prediction of ranking score by linear weighted sum as:

$$z_j = f(X_j) = < W, X_j >, \qquad (2)$$

where $<\cdot,\cdot>$ denotes an inner product operation and $W = (w_1, w_2, \ldots, w_M)$ is a weighting vector for feature dimensions. We can then derive the gradient of Eq. (2) with respect to each w as:

$$\Delta w_m = \frac{\partial L(Y,Z)}{\partial w_m} = \sum_{j=1}^{N} \Big(P(z_j) - P(y_j)\Big) X_{jm}, \qquad (3)$$

where w_m denotes the weight for the m-th feature. The above formula is then used in gradient descend. Initially each w is set to zero, and then updated by

$$w_m = w_m - \eta \times \Delta w_m \qquad (4)$$

at a learning rate η. The learning process terminates when the change in W is less than a convergent threshold δ. The values of η and δ are determined empirically and a parameter sensitivity test over them will be investigated in this work.

Learning to Rank vs. Reranking

Reranking and learning to rank differs in a number of aspects. First, while learning to rank requires a great amount of supervision, reranking takes an unsupervised fashion and approximates the initial results as the pseudo ground truth Y. Second, for learning to rank the ranking algorithm f is trained in advance to predict the relevance scores for arbitrary queries, while for reranking f is particularly trained at runtime to compute the reranked relevance scores for each query itself. However, since both the targets are to minimize errors in object ranking, it is reasonable to incorporate learning to rank algorithms to the task of reranking as long as the pseudo ground truth can be used to train a ranking algorithms. The proposed ordinal reranking achieves this by employing the cross-validation technique [10], as described below.

Ordinal Reranking

The inputs to ordinal reranking is a list of objects D and the corresponding relevance scores Y assigned by a baseline model. We assume that the features (context cues) for each image are computed in advance. For these N objects in D, the corresponding M-dimensional features can form an $N \times M$ feature matrix X. The major steps of ordinal reranking are as follows:

1. *Context cue selection*: Select informative context cues via some feature selection method to reduce the feature dimension to M'.
2. *Employment of ranking algorithms*: Randomly partition the data set into F folds $D = \{D^{(1)}, D^{(2)}, \ldots, D^{(F)}\}$. Hold out one fold $D^{(i)}$ as test set and train the ranking algorithm $f^{(i)}$ using the remaining data. Predict the relevance scores $Z^{(i)}$ of the test set $D^{(i)}$. Repeat until each fold is held out for testing once. The predicted scores of different folds are then combined to form the new list of scores $Z = \{Z^{(1)}, Z^{(2)}, \ldots, Z^{(F)}\}$.
3. *Rank aggregation*: After normalization, the initial score y_j and new score z_j (for object d_j) are fused to produce a merged score s_j by taking the weighted averaged as follows:

$$s_j = (1-\alpha) y_j + \alpha z_j, \qquad (5)$$

   where $\alpha \in [0, 1]$ denotes a fusion weight on the initial and reranked scores; α=1 means totally reranked.
4. *Rerank*: Sort the fused scores to output a new ranked list for the target semantics.

To accommodate the supervised ranking algorithms to the unsupervised environment of reranking, we employ the F-fold cross-validation technique as [9]to partition the dataset, and train F ranking algorithms with different folds of data held out as the test set. Though the new scores Z are not assigned by a unified ranking algorithm, the nature of cross-validation ensures that $F-2$ folds of data are commonly used in the training of two different ranking algorithms. Experimental result reported in [9] also shows the effectiveness of applying cross-validation.

Eq. (5) is used for rank aggregation of the initial relevance scores and newly predicted scores. Such a linear fusion model, though simple, has been shown adequate to fuse visual and text modalities in video search and concept detection [22]. The fusion weight α controls the degree of reranking and may be influential to the reranked result. Therefore, we also test variant fusion weights experimentally in this work.

■ Context Cue Selection

As reported in [9] and [15], the performance of a search model can degenerate significantly as the feature dimension increases arbitrarily. To automatically select informative context cues for reranking in an unsupervised fashion, we are to develop a novel measurement, wc-tf-idf, by improving the concept tf-idf[6]

[6] "tf-idf" stands for term-frequency inverse-document-frequency.

(c-tf-idf) proposed in [14]. Though c-tf-idf is originally proposed for selecting concepts, it is of high generality and can be applied to other real-valued context cues. As its name implies, c-tf-idf is adapted from the best-known term-informativeness measurement tf-idf [15] by viewing images as documents and (real-valued) context cues as term frequencies. The authors in [14] construct the document-term occurrence table from the initial search list, and compute the c-tf-idf of context cue c in a query q as:

$$\text{c-tf-idf}(c,q) = freq(c,q) \log(\frac{T}{freq(c)}),\ c \in C \qquad (6)$$

where $freq(c,q) = \sum_{j=1}^{N} X_{jc}$, $freq(c) = \sum_{j=1}^{T} X_{jc}$ is the occurrence frequency of c in the initial list and the whole corpus, respectively, X_{jc} is the context cue of c in d_j, T is the size of corpus (typically $T \gg N$), and C is the set of adopted context cues. The intuition is the relevance of a context cue increases proportionally to the frequency it appears in the return list of a query, but is offset by the frequency of the context cue in the entire corpus to filter out common context cues. In this way, c-tf-idf offers a good combination between popularity (idf) and specificity (tf) [15].

c-tf-idf is unsupervised and efficient, which meets the requirement of reranking quite well. However, one major drawback of c-tf-idf is that it regards each object as equally relevant to the target semantics, and thus totally neglects the underlying ordinal information among the objects. Intuitively, the context cues of lower-ranked objects should be less important than those of higher-ranked ones. To address this issue, we improve c-tf-idf by weighting the context cues of each object according to the associated relevance scores, as

$$\text{wc-tf-idf}(c,q) = \sum_{j=1}^{N} y_j X_{jc} \log(\frac{T}{\sum_{j=1}^{T} X_{jc}}). \qquad (7)$$

In this way, we preserve the merit of c-tf-idf, but put more emphasis on the higher ranked objects. We will conduct intensive evaluations over large-scale benchmarks such as TRECVID (for broadcast news videos) or self-collected Flickr photos to see the effectiveness of wc-tf-idf vs. c-tf-idf.

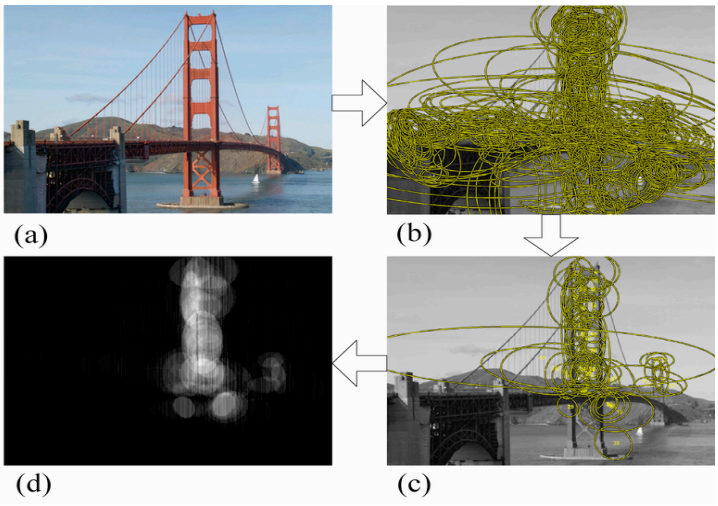


Figure 4: Preliminary experiments for computing the coverage scores of informative visual words for a text query (i.e., "golden gate bridge"): (a) an image from the returned result, (b) its extracted visual words [29], (c) top 50 selected visual words by wc-tf-idf, and (d) the covered areas of informative visual words relevant to the query. Apparently, the informative visual word selection can cover certain salient objects relevant to the query. We will utilize the method for multiple canonical image selection.

■ Challenges and Opportunities

Variant Ranking Methods

Ranking methods is a promising direction for improving search quality either in text information retrieval or photo and video search. In this work, we adopt ranking methods into *ordinal reranking*, in which ranking algorithms such as RankSVM [20] and ListNet [21] are employed to learn the co-occurrence patterns between target semantics and extracted features from the initial result, and then further rerank it. Since the objective of ranking functions is to minimize errors in ranking objects, ordinal reranking is expected to be more effective and efficient in mining ordering information, while being ease of the ad-hoc thresholding problem.

Improving Recall for Visual Reranking Methods

In this proposal, we will also address the "low recall" issue, which has not been addressed in keyword-based reranking methods, however, essential for improving search quality over large-scale photo/video data. We will investigate theoretical and practical methods for selecting informative concepts and context cues (e.g., geo-tags, visual words, etc.) over consumer photos. We will further propose efficient methods to retrieve the large-scale consumer photos based on the selected informative concepts or context cues. Especially, we will devise inverted–document-like indexing methods for real-valued multi-dimensional content and context cues.

Effectiveness of Visual Features

Besides the ranking methods, variants of visual feature representations will affect the search quality and recommendation results. In this work, we will investigate variant visual features such as Visual feature extractors and concept detectors are then applied to generate context cues. The low-level visual features adopted by our system include grid color moments (gcm), Gabor textures (gbr), edge detection (edh) [7] and visual words (vw) [29]. We will mainly focus on visual words for its effectiveness of capturing salient object characteristics in images and its property of being invariant to changes in image/video capturing conditions such as variation in scale, viewing angle or lighting [29]. The construction of VW involves image feature point detection, description and quantization. Feature points are detected by either Difference-of-Gaussian or Hessian-Affine detector [30] and then described by Scale-invariant feature transform (SIFT) [31]. Typically an image can contain hundreds to thousands interest points. The descriptors from all images are quantized into 3500 clusters using K-means, and the resulting centroid of a cluster is then defined as a VW. A feature point is assigned to as a VW by comparing its descriptor to the cluster centroids. Therefore, every image can be viewed as a collection of (discrete) VWs and empirically represented as a *frequency histogram* (FH) by counting the number of each VW it contains (empirically FH is normalized to ensure the largest value being one).

Relevant Tag Recommendation

As discussed in [24], log analysis reveals that user queries are often too short (generally two or three words) and imprecise to express their information needs. This is largely due to the fact that it is usually difficult for users to specify words that adequately represent their needs. To paraphrase N. J. Belkin [25], "When people engage in information-seeking behavior, it's usually because they are hoping to resolve some problem, or achieve some goal, for which their current state of knowledge is inadequate." To resolve this difficulty, term suggestion mechanisms [26]-[28] are commonly employed in current search engine design to give users hint on other possible queries and save efforts by providing shortcuts to relevant queries.

We will apply wc-tf-idf to mine the co-occurring tags associated with retrieved images that are ranked high after reranking. Note that this is different from the conventional web search settings, where relevant terms are mined from the web pages or query logs [26]. Tags, though noisy and few, contain abundant information that may be valuable for the user.

Multiple Canonical Image Selection

The problem of selecting views that capture the essence of an object has been studied for over twenty years in the human vision and computer vision communities [23]. Defining precisely what makes a view *canonical* is still a topic of debate. A typical approach is to use image processing methods such as kmeans to cluster the images into visually similar groups, and then utilize a number of heuristic criteria to select images from the clusters [17], [23], [18]. For example, [17] ranks clusters by several criteria such as visual coherence and cluster connectivity, and then selects images from highly ranked clusters. On the other hand, [23] proposes the following three criteria (but only uses the first one in their work): *similarity* to other images in the input set, *coverage* of important features in the scene, and *orthogonality* to

previously selected images. Unlike previous work for mining a representative image from a static collection (especially for landmark images, e.g., [17], [23]) we are to recommend *multiple* search-related canonical images for arbitrary text queries at query time.

However, as noted by [24], clustering or computing the similarity of hundreds of images is not efficient enough for use in practical online processes. Moreover, approaches (e.g., local coherence [17] or point-wise linking [18]) that involve the matching of local interest point descriptors are exceedingly time-consuming. To relieve the computational cost, we take a fundamentally different approach. Instead of computing any similarity measures, we select images according to the last two criteria proposed in [23], i.e., coverage and orthogonality, in a greedy fashion. In this project, we will investigate canonical image selection methods based on visual word (VW) [29] in a greedy manner for efficiency, as motivated by the observation in **Figure 4**.

Flickr Photo Acquisition

Since there are no public benchmarks for the shared consumer photo collections, we will resort to build up the dataset by ourselves. Over 0.5 million consumer photos will be collected from Flickr and a number of queries are annotated for facilitating evaluation.

Data Acquisition: Initially 540,321 public photos are preliminarily downloaded from Flickr through Flickr API[7]. We start with the "European Travel" Flickr group, where 486 members have uploaded their photos before 2007/10/25. From the 486 members, we further trace back to their own public photos and download the images and associated metadata including tags, text description, time stamp and geo tag. Visual feature extractors and concept detectors are then applied to generate context cues. The low-level visual features adopted by our system include grid color moments (gcm), Gabor textures (gbr), edge detection (edh) [7] and visual words (vw) [29]. For detection of high-level semantic concepts, we utilize the SVM-based models donated by Columbia University[8] to detect the concept scores for the LSCOM lexicon (cpt) [22], a set of 374 visual concepts annotated over an 80-hour subset of the TRECVID data. The associated geo tags and time stamps will also be utilized.

Query Design and Annotation: To evaluate the search result quantitatively, we define 21 queries covering several types of information needs and manually annotate them. The queries can be classified to four categories:

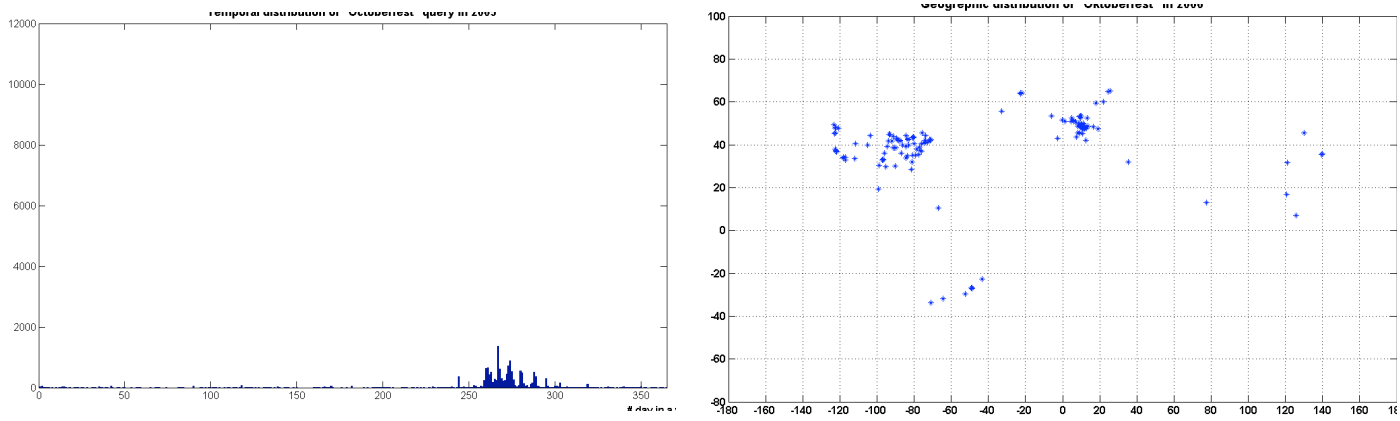
1. *Landmarks*: Colosseum, Eiffel Tower, Golden Gate Bridge, Louvre, Tower Bridge, Torre Pendente Di Pisa, Triomphe
2. *Objects*: beer, *coffee* cup, football, horse, pyramid, ramen, spaghetti, Starbucks, Van Gogh painting
3. *Scenes*: beach, *church*, park, snow
4. *Activity*: *Oktoberfest*

[7] http://www.flickr.com/services/api/flickr.people.getPublicPhotos.html
[8] http://www.ee.columbia.edu/dvmm/

Since the annotation for 21 queries over 0.5 million images is almost infeasible, we employ the *pooling* strategy to sample a small subset of candidate images for human annotation similar to the tasks in TRECVID benchmark [13]. Several baseline models are employed in the search process, and the top retrieved images by different models are then fused to generate the candidate set. In this work, we build five baseline models using different context cues: M_{text}, M_{gcm}, M_{gbr}, M_{edh} and M_{cpt}. M_{text} is simply the text-based baseline (e.g., by OKAPI-like methods). We employ the MySQL[9] full-text search model and extract text information from associated tags, title, description and comments. We further run example-based search using the top five retrieved images of M_{text} with different features (gcm, gbr, edh and cpt) to generate other baselines. The top 1000 retrieved images of each baseline model are then fused by linear combination to form the candidate set, from which we finally choose the top 3000 images to be annotated by human. See more explanations for pooling method in [13].

Temporal and geographical proximity from social media

Besides conventional contextual features from text and visual modalities, there have been quite discoveries in temporal and geographical cues in social media such as Flick photos. These temporal and geographical tags are along with the user-contributed photos (or videos) and had been shown highly related to certain events [17]. For example, in **Figure 5**, we found that the photos tagged with "Oktoberfest" apparently have strong correlations in temporal and geographical aspects, which can be salient cues to augment the context graph construction. However, we need to further investigate proper contextual mining methods in the temporal and geographical modalities and apply them for photo search and recommendation.



(a) (b)

Figure 5: The temporal (a) and geographical (b) distributions for photos tagged with "Oktoberfest" in Flickr. Apparently, such events are mostly held in October (cf. (a)) and are held in several locations in US and Germany (cf. (b)), where the latter location is the main focus.

■ Prospective Results

- The proposed reranking and canonical image selection algorithms are expected to be efficient and

[9] http://www.mysql.com/

can be finished at query time.

- To enhance the semantic coherence of a text-based model, reranking algorithm is employed to mine the contextual patterns between target semantics and context cues including image content, concept scores, location and time. Novel context cue selection measurements (e.g., wc-tf-idf) will also be investigated and evaluated over large-scale video and photo.
- To further facilitate context-related queries and navigations, we will propose tf-idf-like methods (e.g., wc-tf-idf) to recommend relevant tags from those user-contributed ones, and devise methods for generating multiple canonical images to visualize search results
- Extensive experiments are conducted to evaluate the accuracy and effectiveness over large-scale benchmarks – We will leverage our extensive experience in video retrieval evaluating through NIST TRECVID benchmark [13]. Affiliating with researchers in IBM TJ Watson research center and Columbia University, we had participated in the related tasks since TRECVID 2003 and have achieved one of the top performances respectively. In addition, we will explore the use of scenarios and data that resemble realistic consumer contexts. One possible source is Flickr or YouTube. Users in such community usually are non-professional or semi-professional, sharing behavior patterns of consumers. We plan to use such data sets to simulate the consumer contexts and evaluate the accuracy of the proposed tools for automatic image and video search.

■ Impacts in Academy and Industry

- With the prevalence of digital cameras, video recorders, and ease of media sharing, consumers are embracing the revolution of digital multimedia with great enthusiasm. There arise dire needs for effective and semantic image/video retrieval for large-scale images and videos. Besides, the phenomenal success in WWW search has also helped attract increasing interest in investigating new solutions in video search. There are huge incentives either in industrial and academic fields if entailing such multimedia research capabilities.
- In practice, users are expecting to retrieve relevant image and video contents simply through a few keywords. The proposed work requires no search examples or highly-tuned complex models and utilizes the recurrent patterns commonly observed in large-scale distributed video databases and employs the multimodal similarity between visual documents to improve the initial text search results. It will be the first work attaching social media search in such direction.
- The developed technologies can be extended to search “social media” – multimedia in the social network. For example, Flickr and YouTube boost a large user community of hundreds of thousands members capturing and sharing a large amount of images and videos. It will be very promising if we can entail users to discover those multimedia contents they are interested in.

■ Impacts for Researchers

Participating students will be guided to go through the experiment design and preparing for high-quality publications and will benefit from:

- Gaining valuable research and development experiences in multimedia search for large-scale

databases

- Understanding and comparing those state-of-the-art techniques for multimedia search and being motivated to devise their novel ideas for these challenging problems.
- Experienced with effective and efficient multimodal feature presentations and their respective capabilities for multimodal fusion.
- Learning rigorous models in information retrieval and machine learning which are essential for advanced retrieval and classification problems in large-scale multimedia databases
- Practicing problem formulation, reasoning, proposing novel solutions, designing experiments, and paper write-ups for high-quality papers.

References:

[1] L. Kennedy et al, "How Flickr helps us make sense of the world: Context and content in community-contributed media collections," ACM Multimedia, pp. 631–640, 2007.

[2] M. Ames et al, "Why we tag: Motivations for annotation in mobile and online media," ACM CHI, pp. 971–980, 2007.

[3] L. Kennedy, S.-F. Chang, and I. Kozintsev, "To search or to label?: Predicting the performance of search-based automatic image classifiers," Proc. ACM Int. workshop on Multimedia information retrieval, pp. 249–258, 2006.

[4] A. K. Dey, "Understanding and using context," Personal and Ubiquitous Computing, vol. 5, no. 1, 2001.

[5] M. Naphade et al, "Large-scale concept ontology for multimedia," IEEE Multimedia Magazine, vol. 13, no. 3, pp. 86–91, 2006.

[6] K. Toyama et al, "Geographic location tags on digital images," ACM Multimedia, pp. 156–166, 2003.

[7] W. Hsu et al, "Video search reranking via information bottleneck principle," ACM Multimedia, pp. 35–44, 2006.

[8] W. Hsu, L. Kennedy, and S.-F. Chang, "Video search reranking through random walk over document-level context graph," Proc. ACM Multimedia, pp. 971–980, 2007.

[9] L. Kennedy and S.-F. Chang, "A reranking approach for context-based concept fusion in video indexing and retrieval," ACM CIVR, pp. 333–340, 2007.

[10] Y.-H. Yang and W.-H. Hsu, "Video search reranking via online ordinal reranking," IEEE ICME, 2008.

[11] A. Natsev et al, "Semantic concept-based query expansion and re-ranking for multimedia retrieval," ACM Multimedia, pp. 991-1000, 2007.

[12] J. Battelle, The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture, 2006.

[13] NIST TREC Video Retrieval Evaluation. http://www-nlpir.nist.gov/projects/trecvid/.

[14] X. Li et al, "Video search in concept subspace: a text like paradigm," ACM CIVR, pp. 603–610, 2007.

[15] A. Aizawa, "An information-theoretic perspective of tf-idf measures," Information Processing and Management, vol. 39, pp. 45–65, 2003.

[16] S. Palmer et al, “Canonical perspective and the perception of objects,” Attention and Performance IX, pp. 135–151, 1981.
[17] L. Kennedy et al, “Generating diverse and representative image search results for landmarks,” WWW, 2008.
[18] Y. Jing et al, “Canonical image selection from the web,” ACM CIVR, pp. 280–287, 2007.
[19] R. Yan, A. Hauptmann, and R. Jin, “Multimedia search with pseudo-relevance feedback,” ACM CIVR, 2003.
[20] R. Herbrich et al, “Support vector learning for ordinal regression,” IEEE ICANN, pp. 97–102, 1999.
[21] Z. Cao et al, “Learning to rank: from pairwise approach to listwise approach,” IEEE ICML, pp. 129–136, 2007.
[22] S.-F. Chang et al, “Columbia University TRECVID-2005 video search and high-level feature extraction,” NIST TRECVID workshop, 2005.
[23] I. Simon et al, “Scene summarization for online image collections,” IEEE ICCV, pp. 1–8, 2007.
[24] S. Wang et al, “IGroup: presenting web image search results in semantic clusters,” ACM CHI, 2007, pp. 377–384.
[25] N. J. Belkin, “Helping people find what they don’t know,” Communication of the ACM, vol. 43, no. 8, pp. 58–61, 2000.
[26] C.-K. Huang et al, “Relevant term suggestion in interactive web search based on contextual information in query session logs,” Journal of the American Society for Information Science and Technology, vol. 54, no. 7, pp. 638–649, 2003.
[27] R. Jones, B. Rey, O. Madani, and W. Greiner, “Generating query substitutions,” ACM WWW, 2006.
[28] J. Xu and W. Croft, “Query expansion using local and global document analysis,” ACM SIGIR, pp. 4–11, 1996.
[29] J. Sivic and A. Zisserman, “Video Google: A text retrieval approach to object matching in videos, ICCV, 2003.
[30] K. Mikolajczyk and C. Schmid, “Scale & affine invariant interest point detectors,” IJCV, vol.60, no.1, pp. 63–86, 2004.
[31] D. Lowe, “Distinctive image features from scale-invariant keypoints,” IJCV, vol. 60, no. 2, pp. 91–110, 2004.
[32] S. Sontag, On Photography, Picador USA, 2001.
[33] L. Page et al., “The PageRank citation ranking: Bringing order to the web,” Stanford University, 1998.