UNIFIED HYPERGRAPH FOR IMAGE RANKING IN A MULTIMODAL CONTEXT

Jiejun Xu[⋆] *Vishwakarma Singh*[⋆] *Ziyu Guan*[⋆] *B.S. Manjunath*[†]

[⋆] Department of Computer Science, University of California, Santa Barbara, CA 93106
[†]Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106

ABSTRACT

Image ranking has long been studied, yet it remains a very challenging problem. Increasingly, online images come with additional metadata such as user annotations and geographic coordinates. They provide rich complementary information. We propose to combine such multimodal information through a unified hypergraph to improve image retrieval performance. Hypergraphs allow for the simultaneously capture of higher order relationships among images using different modalities, e.g. visual content, user tags, and geo-locations. Each image is represented as a vertex in the hypergraph. Each hyperedge is formed by a vertex and it's k-nearest neighbors. Three types of hyperedges exist in our unified hypergraph, which are in correspondence to the three different modalities. Image ranking is then formulated as a ranking problem on a unified hypergraph. The proposed method can easily be extended to incorporate additional modalities as long as a similarity function exists to compare the features. Experimental results on large datasets are promising.

***Index Terms*—** Multimodal, Hypergraph, Image Retrieval

1. INTRODUCTION

Ranking lies at the heart of any image retrieval system. With the number of images growing rapidly on the web, it is essential to develop effective image ranking algorithms to help users finding images from large databases. Often time, ranking based solely on image content yields unsatisfactory results due to the well known semantic gap. Thanks to the online social sharing sites, such as Flickr, many user-uploaded images come with additional metadata such as tags and geo-locations, which provide complementary information describing the semantics of the images. In this work we address the image ranking problem by jointly analyzing image content and its associated metadata.

There are a number of works focused on finding iconic images from a large database. One typical approach involves matching low level features, building clusters of images and then identifying the most representative images based on intra-cluster similarity [1]. Another approach by Hörster et al. [2] directly estimates a density model of images, and they show that images at the peaks of the distribution are the iconic images.

There is also a growing trend to utilize graph-based approach for image ranking [3]. Typically a graph is constructed in which vertices represent images and edges correspond to visual similarity between two images. Once the graph is constructed, standard graph mining techniques such as Pagerank [4] can be used to identify the "authority" vertices (images). Besides image visual features, it has been shown in previous work [5] that integrating other modalities can boost retrieval performance. In the work of [6], graphs have been used to convey multimodal information for search and retrieval. However, all of the graph-based approach above use simple graphs, which can only capture pairwise image relations. In many cases it would be helpful to consider the relationship among three or more vertices collectively. Such relationships can be easily represented through a hypergraph [7]. Recent works by [8, 9] have demonstrated the effectiveness of such representation. Towards this, we propose a hypergraph-based framework to exploit and utilize information from multiple sources for image ranking. In addition to traditional visual features, we take into consideration of other noisy yet important image metadata to capture the high level semantics of an image. Further more, instead of dealing with typical controlled datasets, we focus on diverse uncontrolled online images.

The major contributions of this paper is to integrate multimodal information, including image content, user-generated tags and geo-locations, in a unified hypergraph framework for image ranking. The proposed method seamlessly captures the high-order relationships among local groups of images through different modalities. In addition, we analyze the gains of multimodal ranking from uni-modal ranking, as well as the advantages of hypergraph models in comparison to simple graph models.

The rest of the paper is organized as follows. Section 2 explains the concept of hypergraphs and how to perform ranking on it. Section 3 details formulating the multimodal image ranking problem in the hypergraph-based framework, and subsequently computing the relevance of images through hypergraph ranking. In Section 4, we will give the experimental results. Finally, we conclude our paper in Section 5.

2. HYPERGRAPH MODEL

A hypergraph is an extension of a simple graph, where a set of vertices is defined as a weighted hyperedge [10]. The magnitude of the hyperedge weight indicates the "compactness" of the vertices in a cluster.

2.1. Preliminaries

Let V represent a finite set of vertices, and E a family of subsets e of V such that $\cup_{e \in E} = V$. $G = (V, E, w)$ is called a hypergraph with the vertex set V and the hyperedge set E, and each hyperedge e is assigned a positive weight $w(e)$. A hypergraph can be represented by an incidence matrix $H \in R^{|V| \times |E|}$ whose entry $h(v, e)$ is 1 if $v \in e$ and 0 otherwise. The degree $\delta(e)$ of a hyperedge e and the degree $d(v)$ of a vertex v is defined as $\delta(e) = \sum_{v \in V} h(v, e)$ and $d(v) = \sum_{e \in E} w(e)h(v, e)$ respectively [10]. Essentially $\delta(e)$ is the sum of a column in the incidence matrix; while $d(v)$ is the weighted sum of a row in the incidence matrix. We will use D_e and D_v to denote the diagonal matrices consisting of hyperedge and vertex degrees respectively. W is a $|E| \times |E|$ diagonal matrix containing hyperedge weights. The computation of $W(e_i)$ is application dependent, and will be shown in later sections.

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

A unified hypergraph is a hypergraph which has multiple types of vertices or hyperedges. In a multimodal image database, images can be viewed as vertices, and different type of relations among images in individual modality can be viewed as different type of hyperedges. In a ranking scenario, a query vertex will be given, and all other vertices in the hypergraph will be ranked based on their relevance to the query. We use $y \in R^{|V|\times 1}$ to denote the query vector containing the the initial scores. Typically only the entry corresponding to the query vertex is set to be 1 and all others are set to 0. Similarly we use $f \in R^{|V|\times 1}$ to denote the final ranking scores.

2.2. Ranking on Unified Hypergraph

Given a unified hypergraph and a query vertex, we can perform ranking for all the vertices on the graph using similar idea of [11]. That is to minimize the cost function defined as follows:

$$\begin{aligned}\Omega(f) = \frac{1}{2}\sum_{e\in E}\sum_{u,v\in V}&\frac{w(e)h(u,e)h(v,e)}{\delta(e)}\\ &\left(\frac{f(u)}{\sqrt{d(u)}} - \frac{f(v)}{\sqrt{d(v)}}\right)^2 + \mu\sum_{u\in V}(f(u)-y(u))^2.\end{aligned} \tag{1}$$

The first term in the right hand side of Eq.1 imposes the constraint that vertices sharing many incident hyperedges in common should have similar ranking scores. For instance, if two image are similar to many common images, they will probably have a similar ranking score. The second term enforces that the initial score of the query vector should be changed as little as possible. μ is a parameter that controls the tradeoff between the two competing terms.

The first term of the cost function contains what is commonly known as the normalized hypergraph Laplacian, and it can be derived as the following:

$$\begin{aligned}&\sum_{e\in E}\sum_{u,v\in V}\frac{w(e)h(u,e)h(v,e)}{\delta(e)}\left(\frac{f^2(u)}{d(u)} - \frac{f(u)f(v)}{\sqrt{d(u)d(v)}}\right)\\ =&\sum_{u\in V}f^2(u) - \sum_{e\in E}\sum_{u,v\in V}\frac{f(u)w(e)h(u,e)h(v,e)f(v)}{\sqrt{d(u)d(v)}\delta(e)}\\ =&f^Tf - f^TD_v^{-1/2}HWD_e^{-1}H^TD_v^{-1/2}f.\end{aligned} \tag{2}$$

By defining a matrix $\Theta = D_v^{-1/2}HWD_e^{-1}H^TD_v^{-1/2}$, we can rewrite the whole cost function Eq.1 in a matrix form:

$$\Omega(f) = f^T(I-\Theta)f + \mu(f-y)^T(f-y), \tag{3}$$

Differentiating $\Omega(f)$ with respect to f, followed by a few algebraic steps, the final ranking score can be computed as:

$$f = (\frac{\mu}{1+\mu})(I - \frac{1}{1+\mu}\Theta)^{-1}y, \tag{4}$$

Since $\frac{\mu}{1+\mu}$ is a constant coefficient and does not change the relative ranking results, we can rewrite the above equation simply as $f = (I - \frac{1}{1+\mu}\Theta)^{-1}y$.

3. HYPERGRAPH FORMULATION OF IMAGE RANKING

As mentioned previously, a unified hypergraph can have multiple type of vertices or edges. In our scenario, images are the only type of vertices. There are three types of hyperedges corresponding to three types of modalities we are integrating, e.g. image visual content, user tags, and geo-locations. Assuming a similarity function exists in each modality, we take each image as a "centroid" vertex and form a hyperedge by itself and its k-nearest neighbors in each modality. For example, given a database with N images, we can construct a hypergraph $H \in R^{|N|\times|3N|}$.

3.1. Feature Extraction and Similarity Measure

A hyperedge is formed by an image and its nearest neighbors. In the following, we will explain the distance measures used between images in each modality.

3.1.1. Visual Distance

To compute the visual distance between images, we employed two common descriptors, GIST [12] and SIFT Signature [13]:

GIST: the GIST descriptor describes the spatial layout of an image using global features derived from the spatial envelope of an image. It encodes the texture information of horizontal or vertical lines in an image to help matching scenes with similar layouts. The Gist feature is computed on a gray scale image by convolving it with a Gabor [14] filter at different orientations and scales. This way the high and low frequency repetitive gradient directions of an image can be measured. The pixel responding scores from the filter convolutions are stored in an array, which is the GIST feature descriptor for that image. In our work, We compute the Gist descriptors using a Gabor filter at 8 orientations and 4 different scales. The results are then averaged on a 4-by-4 grid. This gives us the final descriptor a dimension of 512. The GIST descriptor is particularly useful for scene recognition.

SIFT Signature: the SIFT Signature is a variant of the widely used bag-of-word models. First, 5000 random keypoints are sampled from an image, then the 128 dimension SIFT descriptor is extracted from each of the keypoints. A quantization step is followed to pushed each descriptor into a pre-trained vocabulary tree with 4 levels and a branching factor of 10. As a result, each image is represented by a feature vector of dimension 11111.

Once feature vectors are extracted for each image, traditional $L1$ norm is used to compute the pairwise distance.

3.1.2. Tag Distance

All the tags in the dataset are first converted to lower case, then a dictionary of unique words are generated. In addition, all the tags whose frequency is 1 is removed from the dictionary. A histogram of tag occurrence is computed for each image. Finally Jaccard's coefficient is used to compute the distance between two tag vectors.

3.1.3. Geo Distance

Pair-wise geodesic distances are computed using the Vincenty formula [15] based on the image latitude and longitude. In addition, we assume that any images taken more than 50 miles away are not geo-correlated, and prevent them from forming a hyperedge.

3.2. Computation of Image Rank

Based on the distance measure in each modality, we derive the affinity/similarity for image i and image j as follow:

$$A_k(i,j) = \begin{cases} exp\left(-\frac{D_k(i,j)}{\bar{D}_k}\right), & \text{if } i \neq j\\ 0 & \text{else}\end{cases} \tag{5}$$

where D_k is the distance matrix computed on the k^{th} modality, and $\bar{D_k}$ is the median value of the entries in the D_k matrix

Subsequently we can derive the hyperedge weight for the k^{th} modality as:

$$w_k(e_i) = \sum_{v_j \in e_i} A_k(i,j). \tag{6}$$

The intuition is that a higher weight should be assigned to the hyperedge if the images within a hyperedge are close to each other or have a higher inner group similarity.

The proposed algorithm for multimodal image ranking on Hypergraph is summarized in **Algorithm** 1:

Algorithm 1 Image Ranking on Unified Hypergraph

```
Compute the image distance matrix $D_k$ in each modality.
Compute the affinity matrix $A_k$ from $D_k$.
for Each modality do
    for Each vertex do
        Collect its k-nearest neighbors based on $A_k$, and form a hyperedge.
    end for
    Compute the incidence matrix $H_k$.
    Compute the weight matrix $W_k$ based on (6).
end for
Generate the unified incidence matrix $H$ by column concatenating $H_k$, and similarly for $W$.
Compute the matrix $\Theta$ as shown in Eq.(2).
Given a query image (vertex in the hypergraph), compute the rank scores of other vertices by $f = (I - \frac{1}{1+\mu}\Theta)^{-1}y$.
```

4. EXPERIMENTS

4.1. Dataset

To evaluate the performance of the proposed method, we first collected a "base" set of images from Flickr by querying a list of fifteen well known landmarks, such as *Sagrada Familia*, *Parthenon*, *Great Wall* etc. We downloaded 100 images for each landmark using its name as query. This results in a total of 1500 images. Note that we did not do any filtering or post processing on these images. Thus there is no guarantee that each image will contain the actual landmark. After that we downloaded a "distractor" set of images again from Flickr, which comprises of 10k geotagged images sampled in the nearby locations of the landmarks. We then combine the "base" set and "distractor" set into one large database to conduct our ranking experiments.

As no ground truth is available for these online images, a group of 6 volunteers participated to rate our algorithm. In our first experiment, we would like to compare the performance of unified hypergraph ranking with other single modality ranking. We selected 15 images from our "base" set, one for each landmark, and use them as the queries. In our hypergraph framework, they are the query vertices. These images are shown in Figure 1.

For each query, we retrieved the top 15 images using the rankings from each of the three modalities plus the integrated ranking generated from the proposed hypergraph. Each volunteer will rate the retrieved image with 1 (correct), 0.5 (somewhat relevant) or 0 (incorrect). For each query, we compute the mean score per image over all test users. For each modality, we compute the mean scores over all queries. This gives us four rating scores in correspondence to the four methods. To further evaluate our method, we repeated our experiment using three different "distractor" set downloaded from Flickr, each with 10k images.

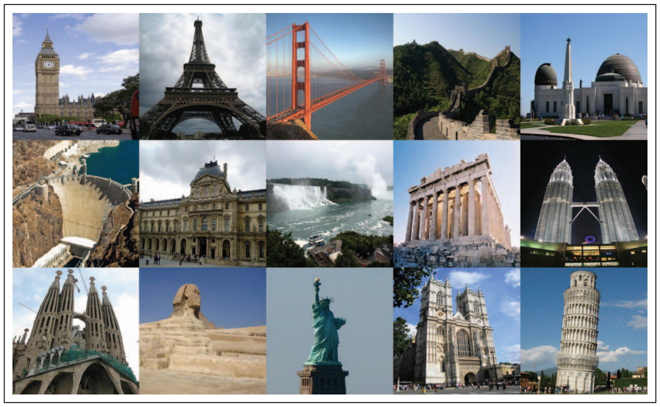

Fig. 1. Query Images

Note that the hypergraph ranking algorithm requires two parameters, the first one is the hyperedge size, and the second one is μ as shown in Eq.(4). The size of the hypergraph is directly controlled by the number of k nearest neighbors. In our experiment, we choose k to be 10, based on empirical observations. We also fix the value of μ such that $\frac{1}{1+\mu}$ equals to 0.1.

Figure 2 shows the performance in each of the three trials. Hypergraph-based multimodal ranking consistently outperforms the other ranking methods which uses only unimodality. Among the three modalities, image feature performs the worst, while tags and geo-locations perform better and give results similar to each other.

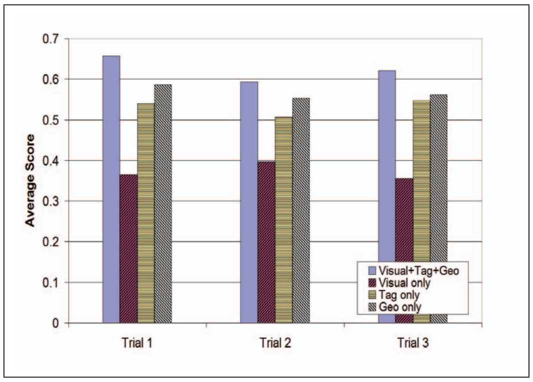

Fig. 2. Performance of hypergraph ranking and other unimodal ranking

In our second experiment, we investigated how would different values of hyperedge size (i.e. k) affect the ranking performance. We choose to vary k among 5, 10, 15, 20 and 25. Due to the constraint on user evaluation, we only performed the analysis using the first dataset from now on. The result is shown as the top curve in Figure 3. The best performance was achieved when k is 10. Note that this number is relatively small. We think that is because our dataset and hyperedge type is quite diverse, and a large hyperedge size introduces unreliable links to the graph.

In [6], a simple graph was used to represent multimodal infor-

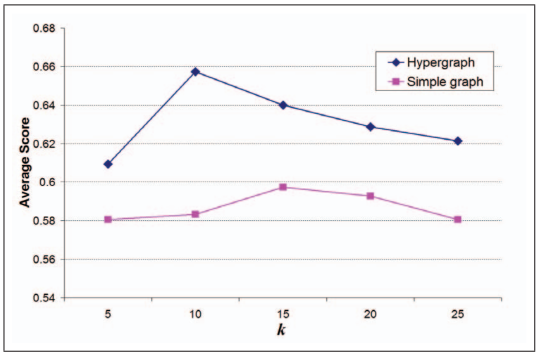

Fig. 3. Sensitivity of k values on hypergraph and simple graph ranking

mation for image ranking. The edges in simple graph are formed by each vertex and its k nearest neighbor, which is the same as the hyperedge size in hypergraph. In our third experiment, we compared the performance of hypergraph ranking with simple graph ranking. We first computed the aggregated affinity matrix by summing up A_k from each modality. Then we constructed the simple graph based on the k nearest neighbors of each vertex. In terms of image ranking, there are many off-the-shelf choices. Previous work by [6, 3] have used Pagerank [4] to compute the scores for each image. However, PageRank gives a static scores for each vertex independent of query. We believe that a better alternative for the "query and ranking" scenario is random walk with restart (RWR), which is similar to Pagerank, except that there is a bias restart probability assigned to the query vertex. We followed the work on [16], and implemented RWR using a power iteration method.

For fair comparison, we run the experiment using the same number of k values. The performance of simple graph ranking is put side-by-side with hypergraph ranking in Figure 3. Both curve peak at a specific k value, and then slowly decrease as k increases. For all the k values, hypergraph ranking consistently outperform simple graph ranking. This indicates that leveraging similarity measure with a local group of images collectively can significantly boost ranking performance.

Another highlight of our framework is that, features from different modalities can be easily included/excluded by keeping/discarding the corresponding types of edges in the unified graph. This could be extremely useful for online user query process, i.e. user has the flexibility to retrieve results based on his/her emphasis on the combinations of modalities.

5. CONCLUSIONS

In this paper, we address the image ranking problem for online community photo database, and focus on combining multimodal information such as image visual features, user tags and geo-locations simultaneously. We model the image ranking problem using a unified hypergraph, which captures the high-order relationships among local groups of images through different modalities. Based on individual similarity measures, the proposed unified hypergraph seeks for the maximum agreement across different modalities to generate good ranking scores.

6. REFERENCES

[1] Y.T. Zheng, M. Zhao, Y. Song, H. Adam, U Buddemeier, A Bissacco, F. Brucher, T.S Chua, and H. Neven, "Tour the world: Building a web-scale landmark recognition engine," in *Computer Vision and Pattern Recognition*, 2009.

[2] Eva Hörster, Malcolm Slaney, Marc'Aurelio Ranzato, and Kilian Weinberger, "Unsupervised image ranking," in *Proceedings on Large-scale multimedia retrieval and mining*, 2009.

[3] Yushi Jing and Shumeet Baluja, "Visualrank: Applying pagerank to large-scale image search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[4] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Seventh International World-Wide Web Conference (WWW 1998)*, 1998.

[5] Rainer Lienhart, Stefan Romberg, and Eva Hörster, "Multilayer plsa for multimodal image retrieval," in *International Conference on Image and Video Retrieval*, 2009.

[6] Fabian Richter, Stefan Romberg, Eva Hörster, and Rainer Lienhart, "Multimodal ranking for image search on community databases," in *Proceedings of the ICMR*, 2010.

[7] Dengyong. Zhou, Jiayuan Huang, and Bernhard Schlkopf, "Beyond pairwise classification and clustering using hypergraphs," Tech. Rep. 143, 08/18/ 2005.

[8] Yuchi Huang, Qingshan Liu, Fengjun Lv, Yihong Gong, and Dimitris N. Metaxas, "Unsupervised image categorization by hypergraph partition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 1266–1273, 2011.

[9] Hung-Khoon Tan, Chong-Wah Ngo, and Xiao Wu, "Modeling video hyperlinks with hypergraph for web video reranking," in *Proceeding of the 16th ACM international conference on Multimedia*, 2008.

[10] S. Agarwal, K. Branson, and S. Belongie, "Higher-order learning with graphs," in *International Conference On Machine Learning (ICML)*, 2006.

[11] Dengyong Zhou, Jiayuan Huang, and Bernhard Scholkopf, "Learning with hypergraphs: Clustering, classification, and embedding," in *Advances in Neural Information Processing Systems (NIPS) 19*, 2006.

[12] Aude Oliva and Antonio Torralba, "Building the gist of a scene: the role of global image features in recognition," in *Progress in Brain Research*, 2006.

[13] David Nister and Henrik Stewenius, "Scalable recognition with a vocabulary tree," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2006.

[14] B.S. Manjunath and W.Y. Ma, "Texture features for browsing and retrieval of image data (derivation of the gabor filter dictionary parameters)," Tech. Rep. 8, Aug 1996.

[15] "Vincenty formula," http://www.movable-type.co.uk/scripts/latlong-vincenty.html.

[16] T. Haveliwala, S Kamvar, and G Jeh, "An analytical comparison of approaches to personalizing pagerank," Technical report, Stanford InfoLab, 2003.