Scalable Image-based Search-and-Discovery

Eunji Chong  
Georgia Institute of Technology  
Atlanta, GA 30332, USA  
eunjichong@gatech.edu

Matthew K. Hong  
Georgia Institute of Technology  
Atlanta, GA 30332, USA  
mhong@gatech.edu

Jaehoon Lee  
Georgia Institute of Technology  
Atlanta, GA 30332, USA  
jlee3185@gatech.edu

James M. Rehg  
Georgia Institute of Technology  
Atlanta, GA 30332, USA  
rehg@gatech.edu

Abstract
People use online search-and-discovery services, such as Yelp, by first finding a specific item with keywords and then examining the images linked to the item. Images could constitute an important part of users’ decision-making process but users reach them indirectly. Although recently researchers proposed several image-based search interfaces, how they can effectively arrange huge number of images in a scalable manner is still not clear. To address this, we introduce PicNav, an image-driven navigation system that automatically arranges photos according to their semantic similarity. PicNav is built on deep neural networks learned from the Yelp food dataset and enables effective zoom-in/out features. We conducted interviews with ten users to qualitatively assess the system’s usability. The users identified a number of advantages of PicNav, providing insights into the general use of imagery in search-and-discovery services.

Author Keywords
Image-based search; exploratory search; scalable system; deep learning.

ACM Classification Keywords
H.3.3. [Information Storage and Retrieval]: Information Search and Retrieval; H.5.2. [Information Interfaces and Presentation (e.g., HCI)]: User Interfaces
Introduction

Search-and-discovery services are very popular nowadays and widely used [11]. Despite broad use of such tools, the user interface is standardized and limited [2]; for example in restaurant search, when a user types in a keyword (e.g. Italian, or Steakhouse), a list of restaurants is retrieved, and the user clicks further to see its photos. There is a general consensus among researchers about the need for flexibility in data exploration/retrieval systems to accommodate different user's search goals and uncertainty levels. To put it another way, a user's information seeking process is restricted by the characteristics inherent to the search system [7]. The development of interfaces supporting exploratory search [3, 9] is a popular approach to providing a sufficiently flexible search capability. Moreover, recent advances in data collection and machine learning have enabled the creation of tools [1] to help users explore massive data. In such tools, when the data itself is visually meaningful (e.g., food, fashion items) imagery could be an efficient medium to use. While various interfaces suited for exploration have been proposed [10], how to process and organize “big image data” for such interfaces still remains a challenge.

In this work, we introduce PicNav, a scalable implementation of image-driven search-and-discovery system where images are automatically arranged and selected for users.
PicNav is implemented using deep neural networks that learned visual-semantic features from a large number of images of food. These features are used to construct the food graph, a network over food images, from which some representative samples are selected in the front page to serve as an anchor to navigate the network. In the user interface, newly selected representative images are shown upon each refresh for overview, and similar images are retrieved when clicked, which makes PicNav effective for exploratory navigation as well as query-based targeted search. We chose the Yelp food dataset [12] to build PicNav because of the large size and variety of its food images. However, dataset from other domains where visual information is important could also be used in the same regime.

Dataset
The Yelp dataset [12] was used to build PicNav. Originally, 200K pictures from 86K businesses in 10 cities were available. We processed this data, using image class information provided by Yelp, to filter out non-food pictures, which resulted in 16K images of food. Each image was associated with a unique restaurant ID that is linked to its attributes such as category, hours, and location.

We conducted interviews with ten study participants to gather their qualitative evaluations following the use of PicNav. The users identified several aspects of PicNav that were particularly useful compared to Yelp, which suggests that our implementation can make viable interface for large-scale image-driven search-and-discovery system. In the last section, we discuss which implementation details makes PicNav scalable and effective, and we also discuss its limitation and future directions.

Related Work
The concept of using images in search system has been investigated by a number of researchers. Especially, recent advances in machine learning enabled large-scale image encoding such that semantically similar images would be automatically organized nearby [1]. Some user interfaces adopted this image embedding approach for image search [4, 5], but the focus was heavily on the retrieval-based search, where closely related images are fetched given an image query. This is essentially the same as keyword search in which a text keyword is replaced with an image item. A few interfaces were introduced that use imagery to facilitate exploration [10] but they mainly describe different visualization methods using tags or metadata (e.g., date, category) associated with image and does not explain how the interface should support navigating through a visually-linked image manifold. In order to enable image-driven exploratory search, it is critical to address how an interface can allow users to navigate a large collection of images in a sensible way, beyond mere retrieval, beyond the limit of metadata.

System Design
We developed a web application, PicNav, for our image-based search-and-discovery system. Users can easily search for the food they are interested in and identify potential restaurants by navigating the images. In the following, we explain the details of the front-end (user interface) and the back-end (data access and processing) implementation.

Front-end
The user interface of PicNav consists of two main parts: the food search component and the restaurant information component. The food search component corresponds to the orange area in Figure 2, and the green area is the restaurant information component.

Food Search Component
Food search component provides three independent features to get food images: (1) On refresh, new representative set of images are shown (Figure 1: blue arrow). Alternatively, users can input text in the top-left box. (2) Another method to get images is by similar image searching. When users are particularly interested in a certain food image, they can click it and similar images will be retrieved with the support of our back-end algorithm. (Figure 1: orange arrow) (3) Our application also provides methods to filter images.
It offers three filters at the bottom, based on restaurant location (city name), categories (cuisine) and rating (5 to 1). When a user selects the filters, only images that match the conditions are displayed.

**Restaurant Information Component**
When the user is interested in a food image displayed in the food search component, they can hover over it and the corresponding restaurant information will be shown in the restaurant information component. The upper part displays the basic information about this restaurant, such as name, rating, reviews, etc. At the bottom part we visualize the restaurant’s location in a map, enabling users to include location information in their search.

**Back-end**
Images in the Yelp dataset are only labelled with restaurant categories. (e.g., pasta and a salad can have the same label “Italian”). For this reason, building an image classifier does not make much sense since classifying by category would not be very informative. Instead, we exploit an “embedding” approach where similar images are located nearby in a low dimensional space. With this embedding, we visualize the whole image gallery that preserves collective similarity and a representative set is selected in the front page. (This set is chosen by uniform sampling in 2D, maximizing the inter-sample distances within the embedding space. Figure 1: blue). Also, it can efficiently find related images for an image query by calculating k-nearest-neighbors in the low dimensional space. ($K = 60$. Figure 1: orange). Specifically, we used state-of-the-art image features extracted from Convolutional Neural Networks [6], which gives us 4K dimensional features for an image input, which then are reduced to 30 dimension using PCA and finally projected onto 2D space via tSNE [8]. To effectively arrange all of the images, we trained our network with

![Figure 3: Projection of the embedding onto 2D space, before (top) and after (bottom) training the neural network with the Yelp dataset. It is observed that, after training, same colors (i.e., images belonging to the same cuisine) are more clustered together.](image1)

![Figure 4: Overview of our back-end algorithm; Images of food from Yelp dataset are used to extract visual-semantic features from deep neural networks. With these features, images are arranged in 2 dimensional space. The middle plot shows color-coded coordinates of food images according to their associated restaurant category, and the bottom plot shows the network layout of photos with this embedding.](image2)
restaurant category information from the Yelp dataset so that the model can learn useful food-related low-level image features. Figure 3 shows relevant images are better clustered after training.

User Interview

In the beginning, we briefly explained the basic features of PicNav before the participants were instructed to use it. After their use of PicNav, we asked them how PicNav compares to Yelp and share their experience and thoughts. If they identified differences, we asked what made them think so. Participants were allowed to ask any questions anytime and each session varied between 10 to 30 minutes depending on how long the participant wanted. Below, we share a few of the interview responses.

Participants appreciated the advantage of using PicNav as an exploration tool when they are not certain of which item to search. “When I don’t have specific things to eat in mind, I can quickly explore all of the options in the front page.” - P5 “When in an unfamiliar location trying to find a restaurant, I feel that I can see more dishes quickly. It makes me discover new menu fast without scrolling through 10 pages of restaurants.” - P1 “It is more intuitive. I was able to find many items that I want to try which I didn’t know what they’re called or which cuisine they belonged to.” - P2

Participants also indicated that they found PicNav effective when they have specific items in mind. “When I have food (for example, burgers) that I want to eat, I can click that food to see all of the similar dishes and then use them for my final choice.” - P10 “I want to visit this restaurant because I want to try this particular food. I usually care ratings in Yelp, but in this tool I tend to care less because the rating score is for the restaurant as a whole rather than for this specific dish.” - P2 “I find the optional text box and filtering features useful, it helps me view the items I need to find.” - P4

This initial results demonstrate PicNav’s potential as a navigation tool that can support both exploratory and targeted search.

Conclusion and Discussion

In this work, we proposed a scalable image-driven search-and-discovery implementation, by building a network of images and a user interface that enables flexible navigation. Our system integrates image overview and image-query search to provide more flexibility. Our user interview shows the potential of this approach as an image-driven search-and-discovery system. Main aspects of PicNav that make it particularly suitable for large-scale search-and-discovery system can be summarized as follows.

1. Image feature learning: Each image in the Yelp dataset has a cuisine (superclass of dish) label (e.g., Mexican, or Italian). In our experiment, only with this superclass-level label, it was possible to learn class-level embeddings (Figure 3). We can envision that the proposed model can also be trained similarly for other domains (e.g., “Pants” for leggings, jeans, etc.) and learn class-related embeddings, which then can be used in search interface.

2. Scalability (i.e., applicable to billions of images): The capacity of the chosen deep model is known to handle unbounded number of inputs, and the flexibility of using superclass labels for training makes it more scalable. Moreover, once the model is trained with initial set, it can automatically handle and allocate unseen and new images in the learned embedding space.

3. Search scope control: The learned embeddings can be used either (A) to maximize exploration by zooming out, or.
(B) to maximize similarity-based retrieval by zooming in. (A) corresponds to the “representative images” (Figure 1 blue area), which are given upon startup/refresh of PicNav. These images are chosen by uniform sampling in 2D, maximizing the inter-sample distances within the embedding space. (B) corresponds to “nearby images” (Figure 1 orange area), which are given upon a click on one image. These images are chosen using KNN algorithm in the embedding space ($K = 60$). (A) and (B) are the two extreme cases of how PicNav can be used. Interestingly, the users identified benefits of each case in comparison to Yelp. If one wants to refine the search scope to somewhere between (A) and (B), it can be achieved by adjusting the value $K$ of KNN algorithm, which was not investigated in our study but could be an interesting future work.

Lastly, while our proposed method can be useful if visual information of the data plays a big role (e.g., food, fashion), but probably not as much useful when images are not very informative. (e.g., book, electronics.) In such case, other approaches should be considered to accommodate users’ various search goals.

REFERENCES


