Photo Recall: Using the Internet to Label Your Photos

Neeraj Kumar
University of Washington
neeraj@cs.washington.edu

Steven M. Seitz
University of Washington
seitz@cs.washington.edu

ABSTRACT

We describe a system for searching your personal photos using an extremely wide range of text queries, including dates and holidays (*Halloween*), named and categorical places (*Empire State Building* or *park*), events and occasions (*Radiohead concert* or *wedding*), activities (*skiing*), object categories (*whales*), attributes (*outdoors*), and object instances (*Mona Lisa*), and any combination of these – all with **no** manual labeling required. We accomplish this by correlating information in your photos – the timestamps, GPS locations, and image pixels – to information mined from the Internet. This includes matching dates to holidays listed on Wikipedia, GPS coordinates to places listed on Wikimapia, places and dates to find named events using Google, visual categories using classifiers either pre-trained on ImageNet or trained on-the-fly using results from Google Image Search, and object instances using interest point-based matching, again using results from Google Images. We tie all of these disparate sources of information together in a unified way, allowing for fast and accurate searches using whatever information you remember about a photo.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Keywords

photo organization; image search; web mining; content-based image retrieval; gps; on-the-fly visual classification; events

1. INTRODUCTION

Searching through one's personal photo collection is currently much tougher than searching for images online, due to the lack of labels and surrounding context in personal photo collections (few people label their photos). However, a surprisingly broad range of personal photo search queries are enabled by **correlating information in your photos to information mined from the Internet**. Figure 1 describes the many kinds of queries our system supports:

- **dates and holidays**: *August 2012*, *Thanksgiving*
- **named places**: *Grand Canyon*, *Sea World*, *FAO Schwartz*
- **categorical places**: *zoo*, *hotel*, *beach*
- **activities**: *skiing*, *cricket*, *paintball*
- **named events**: *Radiohead concert*, *Knicks game*, *Olympics*
- **events by type**: *wedding*, *birthday*, *graduation*
- **things**: *whales*, *green dress*, *Santa Claus*
- **attributes**: *portrait*, *black-and-white*, *blurry*
- **instances**: *Mona Lisa*, *Eiffel Tower*, *Mickey Mouse*

Furthermore, these types of queries can be combined, *e.g.*, *wedding in New York*, to provide even richer queries and more specific results. Fundamentally, the use of Internet data enables an enormous shift in user experience, where *the user chooses the search terms* rather than being limited to a predefined set of options, or requiring manual labeling. Fig. 2 shows additional search results, and many more are displayed in our supplementary video: `http://youtu.be/Se3bemzhAiY`

We represent all information in our system as a hierarchical knowledge graph, with layers corresponding to language, semantics, sensors, image, and grouping constructs. The graph provides a unified representation of all data and lets us perform inference operations via propagations through the graph, including search, auto-complete, and query-dependent description of matched images.

2. DATA SOURCES

Our personal photo search system takes text queries as input and returns a ranked list of matching images; this requires associating text labels with images. We extract all "sensor readings" taken in modern cameras – timestamps, GPS coordinates, and image pixels – and use them to lookup data from existing online public data sources.

We first associate timestamps with named holidays (like *Christmas*, Fig. 1a) using the Wikipedia article, "Public holidays in the United States"[1] to associate any photos occurring within ±2 days of a listed holiday with its name. Next, we take the GPS coordinates of each photo and look up all places within 50 meters on Wikimapia[2], an online crowdsourced database which focuses on geographic information. We store place names as well as the provided list of type categories and activities, enabling queries like *Grand Canyon*, (Fig. 1b), *skiing* (Fig. 1c), *etc.*

[1] `http://en.wikipedia.org/wiki/List_of_US_holidays`
[2] `http://wikimapia.org`

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577360 .

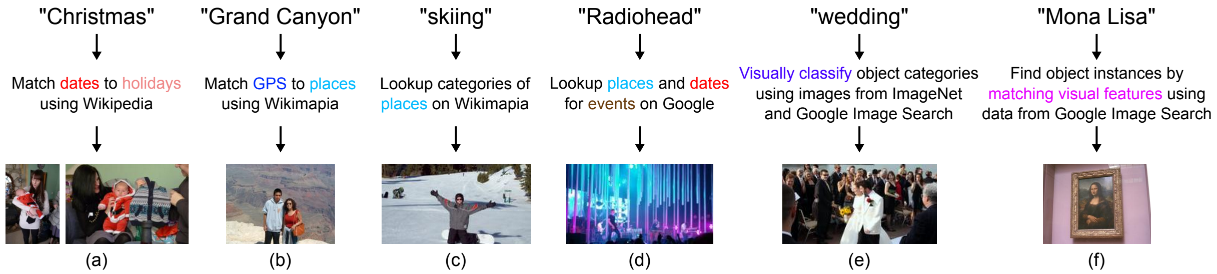
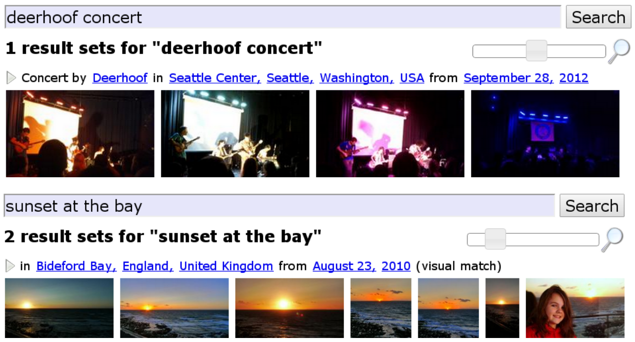

"Christmas"
Match dates to holidays using Wikipedia
(a)
"Grand Canyon"
Match GPS to places using Wikimapia
(b)
"skiing"
Lookup categories of places on Wikimapia
(c)
"Radiohead"
Lookup places and dates for events on Google
(d)
"wedding"
Visually classify object categories using images from ImageNet and Google Image Search
(e)
"Mona Lisa"
Find object instances by matching visual features using data from Google Image Search
(f)

Figure 1: Our personal photo search system associates images with labels by matching different types of image data to various online sources automatically, thus requiring no manual tagging by the user. Users search using natural language queries, allowing for flexible and accurate search results based on whatever information they remember about a photo.

Putting time and place together yields *events* – the natural way in which people recall their memories. Most public events such as concerts, sports, *etc.*, are now documented online, and so by issuing queries on Google for a venue name and date, we can find labels related to the event in question. For example, a search for *"Madison Square Garden" "December 8, 2013"* brings up a page of results, many of which contain the terms *Boston Celtics* and *New York Knicks*, the two basketball teams playing that night. We thus perform queries for all pairs of place names and dates from the user's photo collection and store the most frequently occurring n-grams from the Google results page, allowing for a wide range of event queries, such as *Radiohead* (Fig. 1d) or *Deerhoof concert* (Fig. 2a).

Many other photos you'd like to be able to find – from your sister's wedding, to the exotic flowers you saw in Brazil – correspond to visually distinctive categories, which are amenable to classification by modern computer vision recognition techniques. Unlike most vision systems, however, we do not limit users to a pre-trained list of classifiers. Instead, we take advantage of the fact that online image search engines can return relevant images for almost *any* query by associating images with surrounding text on webpages. When a user performs a search on our system, we issue the same query on Google Image Search, immediately, download their top results, and run the classifier training and evaluation pipeline as described next, returning results within 10 seconds to the user. Our low-level features are histograms of color, gradient magnitude, and gradient orientation. For our classifier, we use linear Support Vector Machines (SVMs) trained via stochastic gradient descent. For speed, we only use thumbnails of the first 64 images from Google as positive examples, and a random set of $30,000$ images as negatives, and interleave downloads and feature extraction across several concurrent threads. Consequently, the entire process takes under 10 seconds on a single machine in our prototype system. This allows for search results like *wedding* (Fig. 1e) and *sunset at the bay* (Fig. 2b).



deerhoof concert

1 result sets for "deerhoof concert"

Concert by Deerhoof in Seattle Center, Seattle, Washington, USA from September 28, 2012

sunset at the bay

2 result sets for "sunset at the bay"

in Bideford Bay, England, United Kingdom from August 23, 2010 (visual match)

Figure 2: Additional search results.

Finally, if the user is looking for specific instances of objects (*e.g.*, *Mona Lisa*, Fig. 1f) rather than broad visual categories, we run an interest point-based matching pipeline [2]. During indexing, we downsample user images to thumbnail size, extract SIFT [1] features, quantize them into a $10,000$-word vocabulary using k-means and store these in an inverted index. At query time, we issue the user query on Google Images, download thumbnails of the top 10 results, extract SIFT features, project them using the learned vocabulary into a list of visual words, and use the inverted index to accumulate scores for each user image. This process also takes about 10 seconds, and runs in parallel with the on-the-fly visual category training.

3. QUANTITATIVE EVALUATION

We evaluated place coverage by manually labeling geotagged images from flickr and comparing these ground truth annotations with search results generated using our system. We found that 73.0% of all places were successfully found by our system when searching by name and 28.9% when searching by category. We evaluated event coverage similarly, and found that 30.2% of all labeled images were found using our system. For both places and events, the biggest problem was generally that the place was not present on Wikimapia; as it continues to expand, recall rates will increase for both places and events. For evaluating visual classifiers, we downloaded the photo collections of 5 users from Google's Picasa Web Albums, and labeled the visual categories found in them. We then queried our search engine with these tags, and got a recall@10 of 49.6%, *i.e.*, half of all queries returned at least one relevant result in the top 10.

These results, coupled with qualitative examinations, show that our system results in a completely transformed search experience: unlike all existing work, the user now has complete flexibility in deciding how to search through their images, and all without any manual labeling.

4. ACKNOWLEDGEMENTS

This work was supported by funding from National Science Foundation grant IIS-1250793, Google, Adobe, Microsoft, Pixar, and the UW Animation Research Labs.

5. REFERENCES

[1] D. Lowe. Distinctive image features from scale-invariant keypoints. *Intl. Jnl. of Computer Vision (IJCV)*, 2004.

[2] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching. In *ICCV*, 2003.