Learning Query-Specific Distance Functions for Large-Scale Web Image Search

Yushi Jing, Michele Covell, David Tsai, and James M. Rehg, *Member, IEEE*

***Abstract*—Current Google image search adopt a *hybrid* search approach in which a text-based query (e.g., "Paris landmarks") is used to retrieve a set of relevant images, which are then refined by the user (e.g., by re-ranking the retrieved images based on similarity to a selected example). We conjecture that given such hybrid image search engines, learning *per-query* distance functions over image features can improve the estimation of image similarity. We propose scalable solutions to learning query-specific distance functions by 1) adopting a simple large-margin learning framework, 2) using the query-logs of *text-based* image search engine to train distance functions used in content-based systems. We evaluate the feasibility and efficacy of our proposed system through comprehensive human evaluation, and compare the results with the state-of-the-art image distance function used by Google image search.**

***Index Terms*—Image search, image processing, content based retrieval, search engine, distance learning.**

I. Introduction

ESTIMATING image distances is central to all content-based image retrieval systems. Commonly used distance functions for image retrieval include Euclidean distance and Earth Mover distance [27]. In some cases, distance functions are learned from data [39], [37]. Such methods have been generally adopted to learn a single distance for all images in the training data.

This work studies the problem of learning distance functions to be used in a hybrid image retrieval systems such as the one used by Google image search. Shown in Fig. 1, such systems adopt a *hybrid* search approach in which a text-based query (e.g., "Paris landmarks") is used to retrieve a set of relevant images, which are then refined by the user (e.g., by re-ranking the retrieved images based on similarity to a selected example). Unlike standard content-based image retrieval system, the goal is not to retrieve an similar image from a common set of images from the Web, but rather to re-rank the *query-specific* search results produced by a text query. For this reason, the same query image can produce different similar images depending on the context of the text query (concrete example retrieval results can be seen using the following two link-shortener: goo.gl/gSRVN and goo.gl/xtZbN). Therefore, we are interested in learning *query-specific* distance functions. Similar observation has made for *exemplar-specific* distance functions [11].

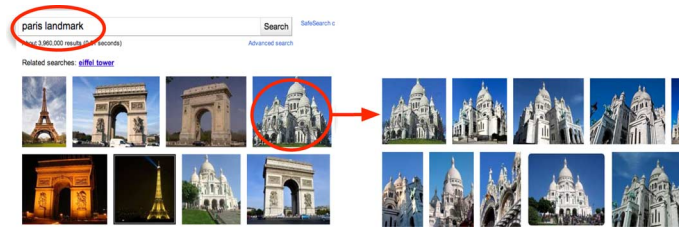


Fig. 1. Google or Bing images adopt a *hybrid* search approach in which a text-based query (e.g., "Paris Landmarks") is used to retrieve a set of relevant images, which are then refined by the user (e.g., by re-ranking the retrieved images based on similarity to a selected example).

The motivation for learning query-specific distance functions stems from our hypothesis that the appropriate choice of feature depends upon the query. For example, consider the problem of identifying a photo of Eiffel Tower. If the query is "Paris landmarks" as shown in Fig. 2(a), then shape feature will be valuable as it differentiates Eiffel tower more clearly from other architectural structures. On the other hand, if the query is "Eiffel Tower" as shown in Fig. 2(b), then color feature would be relatively more useful than shape. Since the context (e.g., the query "Eiffel Tower") is expected to already restrict the images to the correct landmark, the measure of similarity should instead group the images on a less constrained dimension, such as time-of-day, as the color distribution corresponding to time of day.

This work proposes a simple query-specific distance learning framework: 1) the system groups training images and meta-data by their associated text labels, 2) the system learns the optimal distance function from each group of training data using large-margin formulation introduced by [30]. This learning framework can be easily implemented using parallel computational MapReduce [9] framework. Once learned, the distance functions are indexed together with the images annotated with the specific text-query.

Our work is partially inspired by Frome *et al.* [11], who proposed to learn local distance function for each image exemplar. Such exemplar-specific distance functions, however, are computationally expensive for large-scale Web image retrieval. Therefore, based on the observation that search query frequency follows power law distribution [29], we believe query-specific distance functions can be applied to only the most popular search queries and still service large portion of the overall search engine traffic.

Manuscript received July 01, 2012; revised November 29, 2012 and January 23, 2013; accepted April 10, 2013. Date of publication August 23, 2013; date of current version November 13, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Francesco G. B. De Natale.

Y. Jing and M. Covell are with Google Research, Mountain View, CA 94043 USA.

D. Tsai and J. M. Rehg are with the Georgia Institute of Technology, Atlanta, GA 30332 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2013.2279663

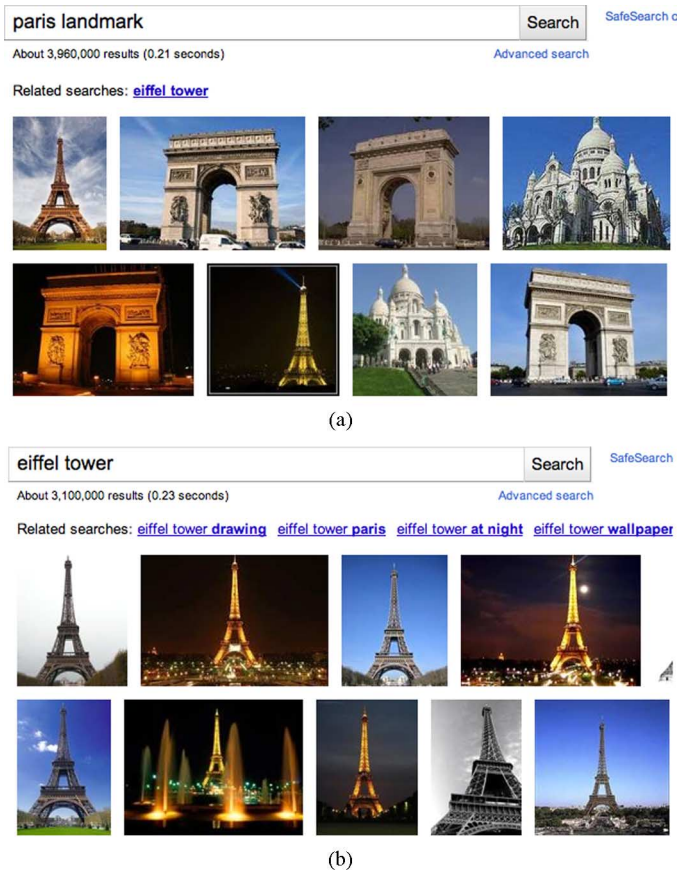


Fig. 2. Top search results with the query *Paris landmarks* and *Eiffel Tower*. (a) Paris landmarks. (b) Eiffel Tower.

This work focuses on addressing two questions related to learning query-specific distance functions for large-scale Web image search.

1) **How can we collect sufficient amount of *labelled* training data to learn distance function for each query?** Standard methods to collect training labels, such as manual labelling and relevance feedback, are costly as they require active human participation, and the resulting annotations are usually insufficiently descriptive of the image contents [23].
2) **Does learned query-specific distance functions significantly improve the estimation of image similarity from the perspective of human raters?** Note that due to the added degrees of freedom, using query-specific function will, in many situations (especially when training and testing data are drawn from the same underlying distribution), produce more accurate image comparisons than query-independent distances. Therefore, any analysis needs to be conducted in a way such that evaluation method are not directly tied to how training data is produced.

This work makes the two contributions. First, we demonstrate that user click-patterns made on text-based search engine (e.g., Google image search) can be used to train content-based image distance functions. As text-based search engines attract large user traffic, such "transfer" of labelled data to train content-based retrieval system makes it possible to apply supervised metric learning techniques such as [30], [11] to large number of Web queries and images. Although learning from click-data has been explored previously [18], [30], [16], there is very little work on learning image distances from an open and unconstrained text-based system, on which users do not have a predefined guideline on tasks to accomplish. For example, the images clicked by a casual surfer may not exhibit any semantic or visual relations at all. This work conducted a detailed study in order to verify that, by aggregating the query sessions made by large number of Web users, the averaged click-patterns largely reflect the results of perceptual similarity test.

The second contribution of this work is a comprehensive study on the feasibility and effectiveness of query-specific distance functions for Web image search. This work applies distance learning technique introduced in [30] to learn query-specific distance function from click-data derived from text-based search engines. Our method is more scalable than previously proposed exemplar-specific distance functions [11] when applied to large-scale Web search. Our work also demonstrated that learning query-specific image distances produces more accurate measurement of image similarity than the state-of-the-art Google similar image search system. Specifically, we conducted two types of experiments: 1) perceptual experiments [6] that compare estimated image similarity with those derived from judgement of human raters, 2) target-search experiments [24] that measure users' efficiency in finding the relevant images using various distance functions. Our results demonstrate that query-specific distance functions outperform the L2 distance function used Google image search. The difference in accuracy is especially significant given polysemy queries.[1]

The rest of the paper is divided into six sections. Section II introduces related works in learning distance functions for Web image retrieval. Section III proposes methods to derive measurement of image similarity from text-based image-search query logs, and then use this information as training data. Section IV introduces methods to learn query-specific distance functions. Sections V and VI present evaluation results with perceptual similarity and target-search experiments respectively. Section VII presents the conclusions and future work.

II. Related Work

Our work is mainly related to two areas of research. The first area is related to distance learning research in machine learning, and the second is related to exploring log data as relevance feedback in Web search. This work briefly reviews some representative works in both areas.

A. Learning Distances

In spite of the observation [26] that human perception sometimes cannot satisfy triangular equality, distance metrics, such as Euclidean distance, have been used extensively in large-scale Web image retrieval systems for its simplicity and efficiency [32], [8]. Methods to improve the accuracy of Euclidean distance has been proposed previously, including unsupervised learning techniques such as Metric Multidimensional Scaling,

[1]Polysemy queries are keywords with multiple semantic interpretations such as *apple*.

Locally-Linear Embedding, Isomap, Pyramid Match Kernel [12], and supervised distance learning techniques [38], [30], [36], [11] that learn a *weighted* Euclidean distance function.

Previous work in learning weighted Euclidean distance functions differ in how training samples are collected and how they formulate the optimization. For example, Xing *et al.* [38] learns a weight vector that minimizes the number of violated constraints in the training data, structured in the form of pairwise comparisons ("A and B are similar"). Schultz [30] adopted an optimizing approach that is analogous to a soft-margin SVM trained with relative comparison ("A is more similar to B than A to C"). In these approaches, a single distance function is learned and used to compare all images in the database. In contrast, our work studies metric learning in the context of hybrid image retrieval system, and proposes to adopt the learning approach in [30] to learn query-specific distance functions.

B. Use of Search Engine Logs

The use of logged search data as a form of relevance feedback [40], [28] has been explored previously by Web information retrieval and content-based image retrieval communities [18], [30], [16], [22], [14], [2]. Joachims [17] conjectured that click-through statistics often convey judgment of document relevance with respect to the query, and confirmed this hypothesis with an eye-tracking study [18]. Uchihashi *et al.* [35] proposed a *content-free* image retrieval system entirely based on modeling the click statistics of the image retrieval systems through collaborative filtering [14] techniques. Radlinski *et al.* [22] demonstrated that click-through data is not reliable for deriving absolute relevance judgment as it is affected by the retrieval quality of the underlying system, but relative comparisons ("A is more relevant to the query than B") are reasonably accurate. Schultz *et al.* [30] proposed to learn a ranking function from co-click statistics for Web document retrieval, and Jain *et al.* [16] applied similar techniques to the retrieval of Web images.

Our work is also related to the work by Hoi *et al.* [15], who demonstrated that the click patterns made by users of content-based image retrieval system can be used as relevance feedback signals to refine image distance function. This work has several key differences compared with [15]. First, the training data used in this work is in the form of relative comparison, as opposed to "relevant" or "irrelevant" labels. Second, this work proposes to learn a unique distance function for each query, as opposed to a global distance for all images. Third, this work derives training data from *text-based* image retrieval system. This approach allows us to leverage large quantities of feedback data from popular incumbent Web retrieval system, and to improve the performance of new system that may not have sufficient relevance feedback data of its own.

III. Measure Image Similarity With Co-Click Statistics

A search session [19] starts when the user initiates an image search task (perhaps by typing the URL of a commercial search engine), and ends when the user leaves the search engine, or no longer actively searches on the site. During this time, users usually have viewed a large set of images, and may have *clicked* on one or more images that satisfy his or her search criteria. Such browsing behaviors are recorded as a part of the image search engine query logs. In a single image search session, if image x_i and image x_j are both clicked by the user, we say they are co-clicked.

Fig. 3. Image x_j is more similar to query image x_i than image x_k is to x_i.

This work studies whether two images that are co-clicked more often are more similar to each other than to a third image co-clicked less often. The intuition is that when conducting an image retrieval task, many users have a pre-determined mental image of what they are looking for. Therefore, during the process of browsing through the search results, users may conduct an implicit comparison between the images retrieved with the target image. Only images similar to the target image are selected while others are seen but ignored. Therefore if we aggregate the co-click statistics over all search sessions conducted within a sufficient period of time, then images that are clicked more often are more similar to each other. Our goal is to derive reliable measurement of image similarity from such aggregated co-click statistics, and use it to train query-specific distances.

One can imagine several situations when such hypothesis is not true. For example, a person may not have concrete search criteria (e.g., casual browsing) or the search criteria may change over time. In this case, the images clicked may not exhibit any semantic or visual relationships at all. The hope is that by aggregating the query sessions made by billions of Web users, the distinctive click patterns may emerge to capture how majority of the people perceive image similarities.

A. Image Comparison With Co-Click Statistics

In this work, we propose to derive relative comparisons ("image A is more similar to image B than A is to C") from co-click statistics, and use such information to learn query-specific distance functions. In contrast with with pairwise comparisons (e.g., "image A is similar/dissimilar to B"), relative comparisons is context-dependent, and contain richer set of information that can be used to derive the relative ordering of the images. For example, given the three images shown Fig. 3, although both image x_j and x_k are related to the image x_i (e.g., all Paris landmarks), most would agree that x_j is more similar to x_i than x_k is to x_i.

One can use co-click statistics as absolute and quantitative measurements of pairwise similarity to compare or rank images. For example, given a query image x_i and two candidate images x_j and x_k, one can determine which candidate image is more similar to the query image with the following equation:

$$\delta(x_i, x_j, x_k) = \left\{ \begin{array}{ll} x_j, & \text{if } C(x_i, x_j) > C(x_i, x_k) \\ x_k, & \text{otherwise} \end{array} \right\} \quad (1)$$

where $C(x_i, x_j)$ is the number of search sessions where image x_i and x_j are co-clicked. Adopting (1) for image comparison assumes that we have accurate measurement of pairwise distances (or similarity). However, as this work proposes to derive

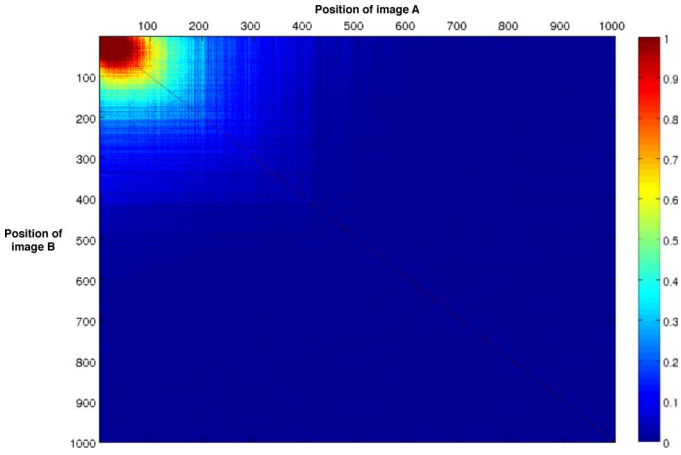
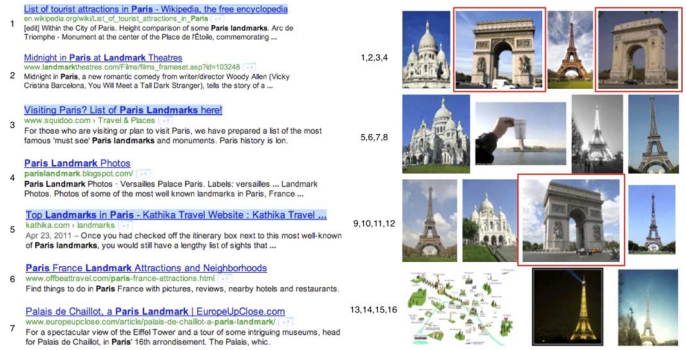


Fig. 4. The correlation between co-clicks between two images and their respective position in the search results. The point (x, y) on the two dimensional plot (x, y) represents the average amount of co-clicks received by images with x and y as their respective position in the search results. On average, the likelihood of a user click on an image tends to decreases as the rank increases.



(a) (b)

Fig. 5. An example of Web/Image search results. Web documents or images that are clicked on by the user during a search session are highlighted. We can reasonably expect that those images (or documents) ranked ahead of the clicked images (or documents) are observed but not clicked. (a) Web search. (b) Image search.

image similarity from user click-patterns, the order in which images are presented to the user can significantly affect whether the likelihood of images being clicked by the users.

For example, Fig. 4 illustrates the position bias by showing the correlation between co-clicks between two images and their respective position in the search results, generated by averaging the co-click statistics generated from the top 1000 search results produced by 40 popular queries on Google image search. The point (x, y) on the two dimensional plot (x, y) represents the average amount of co-clicks received by images with x and y as their respective position in the search results given a set of popular queries. On average, the likelihood of a user click on an image tends to decreases as the rank increases. Such position bias is due to the well document tendency [17], [7] for search engine users to exist search when the first relevant image is found, regardless of whether there is a more or equally relevant images positioned further down in the list of search results.

To address this problem, we proposes to incorporate the *average position* of the images into the comparison function. This is based on the observation that, when a ranked list of Web documents are presented to the Web search users, documents that are clicked on are more semantically relevant to the query than those that are observed but not clicked on [17]. In the absence of information on what documents users have observed, a commonly used assumption is that user examine search results sequentially and therefore all the documents ranked ahead of the last clicked document is considered observed. For example, in Fig. 5(a), the documents that are clicked are highlighted. One can reasonably expect that document 2 and 4 are observed but not clicked.

We extend this intuition to the domain of image search: images that are clicked more frequently are more similar to each other than those ranked higher but clicked less frequently. The process of labeling image triplets contains the following two steps: first, we count the number of search sessions where a pair of images is co-clicked, denoted as $C(x_i, x_j)$ for image x_i and x_j. Note that $C(x_i, x_j)$ is aggregated over all possible queries. Next, we computed the average position of each image relative to other images in the same query during the time the data is collected. In this work, we refer the average position for the image x_i given query q as $P_q(x_i)$, where $P_q(x_i) < P_q(x_j)$ when x_i is ranked ahead of x_j. Note that the position of the images are query-dependent.

The resulting relative comparison function is shown below,

$$\delta(x_i, x_j, x_k) = \begin{cases} x_j, & \text{if } C(x_i, x_j) > C(x_i, x_k), \\ & P_q(x_k) < P_q(x_j) \\ x_k, & \text{if } C(x_i, x_j) < C(x_i, x_k), \\ & P_q(x_j) < P_q(x_k) \\ \emptyset, & \text{otherwise} \end{cases} \quad (2)$$

If the position-constraints is not satisfied, then the function will output $\emptyset$, indicating that we do not have sufficient information to determine which of the candidate images is similar to the query image. We apply (2) to all permutations of image triplets x_i, x_j, x_k sampled from the images produced by query q. We remove the triplets when image comparison cannot be reliably estimated from the co-click statistics (labeled with $\emptyset$).

One can also combine co-click statistics with other types of distances, such as Euclidean distance (L2) derived from image features, using the following equation:

$$\delta(x_i, x_j, x_k) = \begin{cases} x_j, & \text{if } C(x_i, x_j) + d_2(x_i, x_k) \\ & \quad > C(x_i, x_k) + d_2(x_i, x_j), \\ & \quad P_q(x_k) < P_q(x_j) \\ & \text{or } d_2(x_i, x_k) > d_2(x_i, x_j), \\ & \quad P_q(x_k) > P_q(x_j) \\ x_k, & \text{otherwise} \end{cases} \quad (3)$$

where $d_2(x_i, x_k)$ is L2 distance computed over image features. Comparing with (2), (3) combines co-click statistics with the distances produced using L2 distance over image features when the rank constraint is satisfied, otherwise only L2 distance is used.

As Web search engines typically do not have control over how a user may interpret the search results and interact with the retrieval system, it is possible that the images clicked may not exhibit any semantic or visual relationship at all in some search sessions. Our hope is that despite the subjectivity in human perception of image similarity, one can still find distinctive click patterns for subsets of queries and images that capture

how majority of the people perceive image similarity. For this reason, instead of considering image triplets generated from each search session as a separate measurement of image similarity, we use the aggregated statistics over 1-year worth of image search query logs.

IV. Learning Query-Specific Distance for Web Image Search

Our goal is to learn a weighted Euclidean distance d_{w_q} for each query q. Each query is associated with a set of images X_q and $x_i \in X_q$ is represented by a M dimensional feature vector $\vec{x_i} = \{\vec{x_i}^1, \ldots, \vec{x_i}^M\}$. We define the query-specific Euclidean distance between image x_i and x_j

$$d_{W_q}(\vec{x_i}, \vec{x_j}) = \sqrt{\sum_{m=1}^{M} \vec{W_q}^m (\vec{x_i}^m - \vec{x_j}^m)^2} \tag{4}$$

where W_q is a M dimensional weight vector over the features.

Given a training set T_{train} of n relative comparisons, our goal is to learn the weight vector W_q over the features such that the training error (i.e., the number of violated constraints) is minimized. Using the training image triplet shown in Fig. 3 as an example: it is clear that image x_j should be considered more similar to image x_i than x_k is to x_i. The learning goal is to find distances between images such that relationships of this type holds, for example, that the distance $d_{W_q}(\vec{x_i}, \vec{x_j}) < d_{W_q}(\vec{x_i}, \vec{x_k})$. Of course, if all our served results images are from the training set, then we don't need the distance functions at all; we can simply rank images based on the comparison based on co-click statistics. However, such supervised information typically is not available for all image triplets in the database (due to the position bias), and it is certainly not available for new images.

Following [30], finding a solution with minimal training error is equivalent to finding a W_q that fulfills the following constraint.

$$\forall (i,j,k) \in T_{train} : d_{W_q}(\vec{x_i}, \vec{x_k}) - d_{W_q}(\vec{x_i}, \vec{x_j}) > 0. \tag{5}$$

As the solution is typically not unique, learning methods have been proposed to select W, such that the learned distance remains as close to an un-weighted Euclidean distance as possible. Following [30], we adopt the max-margin framework that minimizes the norm of weight vector $\vec{w_q}$. This leads to the following optimization problem:

$$\begin{aligned} &\min \frac{1}{2}\|\vec{w_q}\|^2 \\ s.t. \quad &\vec{w_q} \cdot (\vec{\Delta}^{x_i,x_k} - \vec{\Delta}^{x_i,x_j}) > 0 \\ &\forall (i,j,k) \in T_{train} \end{aligned} \tag{6}$$

$$\vec{w_q}^m \geq 0 \quad \forall m \in \{1, \ldots, M\} \tag{7}$$

where $\vec{\Delta}^{x_i,x_k} = (\vec{x_i} - \vec{x_k})^T(\vec{x_i} - \vec{x_k})$. Compared with standard quadratic programming such as SVM, this optimization has an additional constrain on $\vec{w_q}$, which needs to be positive such that it meets triangle inequality of distance. We add slack variables ξ_{ijk} to each triplet to account for constrains that cannot be satisfied, we then get the following optimization problem:

$$\begin{aligned} &\min \frac{1}{2}\|\vec{w_q}\|^2 + C\sum_{i,j,k} \xi_{ijk} \\ s.t. \quad &\vec{w_q} \cdot (\vec{\Delta}^{x_i,x_k} - \vec{\Delta}^{x_i,x_j}) > 1 - \xi_{ijk} \\ &\forall (i,j,k) \in P_{train} \\ &\xi_{ijk} \geq 0, \quad \vec{w_q}^m \geq 0 \quad \forall m \in \{1, \ldots, M\} \end{aligned} \tag{8}$$

where the scalar C is the trade-off parameter between the empirical loss term and the regularization term. The form in (8) is similar to the soft-margin SVM. We solve this optimization problem using sub-gradient method based on [33]. This method does not directly depend on the number of training samples and is very fast in practice.

V. Experiment I

A. Perceptual Similarity Test

This work adopts an evaluation approach that first asks human raters to compare sets of images, and apply the resulting human ratings as labels to evaluate a particular distance function. Such perceptual similarity experiments have been proposed previously [5], [31] to evaluate content-based image retrieval systems.

Typically human raters are asked to compare sets of images and assign either an quantitative similarity score to a pair of images (e.g., image A and B are very similar) or qualitative and relative comparisons (e.g., image A is more similar to image B than image A to C). Compared with alternative approaches, such as using class labels or image annotations, this perceptual comparison test is closely related to the task at hand (e.g., comparison and ranking of images).

This work uses a variation of relative comparison test [5] as shown in Fig. 6. The query image is displayed at the top and two candidate images are displayed at the bottom of the screen. Human raters are instructed to indicate which of the two candidate images is more similar to the query image. As candidate images can all be similar or dissimilar to the query image, the experiment also allows users to select *cannot decide*. Also, since the perception of image similarity can be subjective with respect to the experiences of the raters, we assign each sets of images to multiple raters, and only consider images with consistent label from all raters. Therefore, we propose to measure the accuracy of a distance function by comparing its output with rater selection on the testing images.

Note that due to the added degrees of freedom, using query-specific distance functions will, in many situations (especially when training and testing data are drawn from the same underlying distribution), produce more accurate image comparisons than query-independent distances. Therefore, this experiment is conducted in a way such that evaluation method (e.g., perceptual comparison test) are not directly tied to how training data is produced (e.g., co-click statistics in Web image search).

B. Methodology

1) Sampling Queries From Image Search Logs: We selected queries belonging to four categories of visual concepts, *person*, *product*, *animals* and *places*, as these categories contain many of the most frequent terms people use to query commercial image search engines. Also, as such categories usually have distinctive visual appearances, they are commonly used as a part of benchmark database for evaluation of recognition systems [13], [10]. We also included a fifth category referred as *polysemy*, which

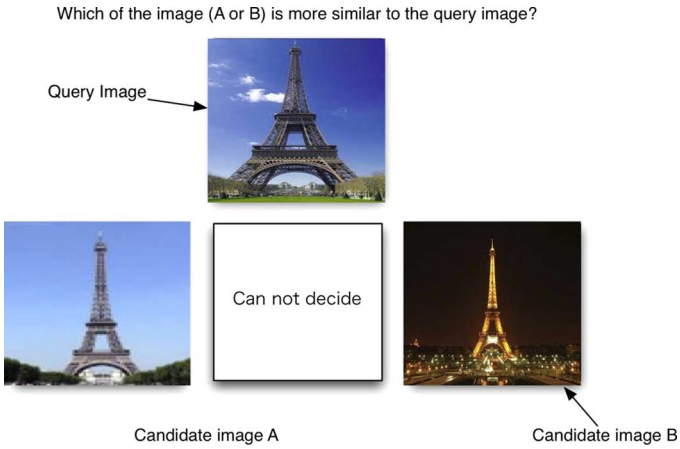


Fig. 6. The triplet rating interface. Users are asked to select the image (A or B) more similar to the query image.

TABLE I
A LIST OF 50 (44 UNIQUE) QUERIES WERE SAMPLED FROM A SET OF 10,000 MOST POPULAR QUERIES ON GOOGLE IMAGE SEARCH

People	lady gaga, steve jobs, bill gates, barrack obama, brad pitt, taylor swift, kobe bryant, david beckham, paris hilton, allen iverson
Product	iphone, bmw z8, ipod, coca cola, nokia phone, electric guitar, dell computer, zune, paper clips, alarm clock
Animal	tiger, fish, cat, dog, pig, beetle, zebra, chicken, bird, jaguar
Places	golden gate, eiffel tower, stanford university, beach, fuji, great wall, washington, store, lincoln memorial, notre dame
Polysemy	apple, tiger, jaguar, washington, notre dame, cup, fuji, beetle, darwin, crane

are queries with multiple semantic and visual concepts. For example, the query *apple* produces images related to both the company and the fruit. We conjecture that ranking based on image similarity is mostly helpful on the search results retrieved with such ambiguous queries.

We selected 10 queries for each category with the following selection methodology: first, we collected 10,000 of the most frequently searched for queries on Google images during the month of July 2010; second, we uniformly sample queries from this list, and manually assign each query to one of the five categories illustrated above. A query is removed from consideration if it does not fall into any of the five category, or if the retrieved images contain pornography or other inappropriate content. This process is repeated until each category contains 10 queries. The complete list of queries are shown in Table I.

2) Sampling Image Triplets From Search Results: Given each query, we extracted the top 100 search results from Google image search, with the strict safe search filter. The top 100 images are used instead of the 1000 images due the following two considerations: first, we observed that the relevance between the retrieved images and text query degrades significantly beyond the top 100 results retrieved from Google images; second, as users usually follow the order in which search results are presented to them, they are more likely to select a query image from the top search results to conduct hybrid image retrieval.

We randomly sampled 25 testing images from the top 100 search results. Since 2300 unique combinations image triplets (25-choose-3) can be sampled from the testing images, and each of image in a triplet can be the query image, there are 6900 possible testing image triplets. We randomly sampled 1000 (14.5% of 6900) image triplets and have them labeled by the human raters.

3) Experiment User Interface and Procedure: The interface is shown in Fig. 6. For each query, we partition the 1000 triplets into 20 triplet groups, and each group is presented to three different human raters for labeling. 7 raters participated in this experiment, each rater spent 4 hours a day (50 minutes rating-time with 10 minutes rest-time) for a total of 10 business days. We only consider testing images that received consistent labels from the three raters.

C. Experiment Results

This section presents a set of experiments designed to evaluate the quality of the relative comparisons generated from the query logs and the accuracies of distance learned from such information. We used the click-data recorded for the entire year of 2008 from Google image search. Note that this is before Google introduced the "hybrid search" functionality with visual similarity. For this reason, the click-data used in our experiment is derived from open and unconstrained text-based image system. Since the image search results do not change significantly within a short period of time, we sampled the query-based ranking of the images at the beginning of each month, and then compute the average ranking of image x_i within the text query q according to the following formula: $P_q(x_i) = \sum_{t=1}^{12} P_q^t(x_i)/12$, where $P_q^t(x_i)$ is the position of image x_i in the results list produced by query q at the beginning of the month t.

The testing images are collected with the procedure listed in Section V-B. We used **Google-L2** distance function as benchmark for comparison. Google-L2 distance is a highly optimized distance function over the image features used by Google Similar Images.

1) Analysis 1: Examples of Results: Fig. 7 contains a sample of testing triplets. In particular, it contains triplets such that the comparison decision derived from co-click statistics disagrees with those derived by applying Euclidean distance over the image features. Each numbered row represents a testing triplet. For each triplet, the first image is the query image and the second and third image are candidate images. The candidate images are arranged such that the second image is more similar to the query image based on the **Google-L2**, and the third image is more similar based on **co-click** statistics using (2). In order to not "double-count" polysemy queries, we removed polysemy queries from People, Product, Animal and Places when computing the averages for each category.

We observe that **Google-L2** distances are sufficiently accurate when two images contain the same objects (row 1, 2, 3, 5, 20) or share dominant visual cues (4, 6, 7). Co-click similarities are less accurate in such cases—indicating that images clicked during a search session are likely to be semantically and visually similar only up to a point. Images that are duplicate or near-duplicate of each other are not necessarily the most frequently clicked pair during a search session. On the other hand, when a particular visual concept (such as apple logo) has high intra-class variance with respect to the image features, co-click statistics tends to be more accurately capture the semantic similarity among the images (8 – 15, 16, 17, 18, 19, 21, 22, 23). It is our hope that by learning feature weights from the co-click statistics, those most discriminative features in this query, such

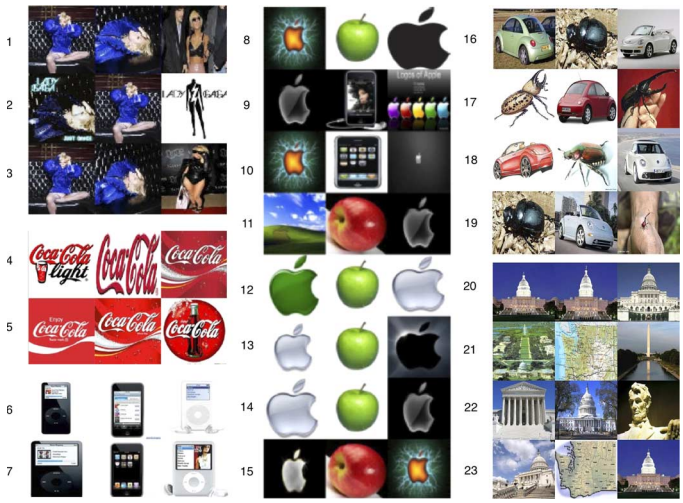


Fig. 7. A sample of testing triplets where comparison results derived from co-click statistics disagree with those based on applying Euclidean distance over the image features.

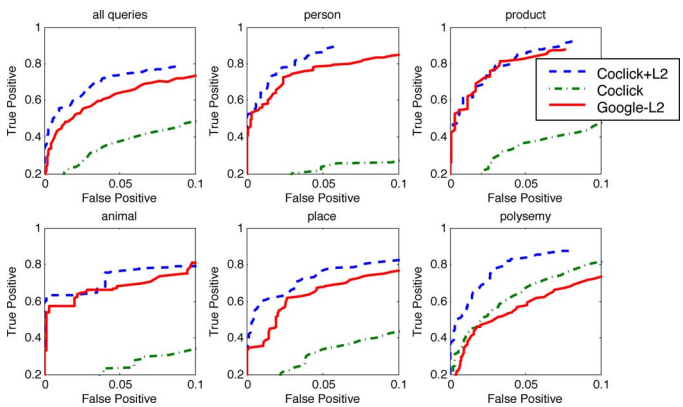


Fig. 8. The accuracy of co-click statistics (Coclick) v.s. image distances (L2), and combined co-click and image distances (Coclick+L2).

as shape of the logo, are likely to have more weights over other features (color, etc.).

2) Analysis 2: Accuracy of Co-Click Statistics: Fig. 8 compares the accuracy of co-click statistics with other measurement of image similarity such as Euclidean distance over image features. In order to compute the ROC curves, we a threshold ϵ to (2) and (3), such as $C(x_i, x_j) > C(x_i, x_k) + \epsilon$ in (1). By adjusting ϵ, we can get various operating points. **Google-L2** represents the highly optimized distance function used by Google Similar Images. **Coclick** represents comparison with (2); and **Coclick+L2** compares images using a combination of co-click statistics and *Google-L2* distance derived from image features using (3), where the distances are mean/variance normalized. We scaled **Google-L2** and **Coclick** such that they have the same variance with approach similar to [11]. To ensure that one can fairly compare **Coclick** with three outcomes (e.g., x_j, x_k, $\emptyset$) with the other two methods with two outcomes (e.g., x_j, x_k), we removed all the testing data where the output of co-clicks resulted in $\emptyset$ from evaluation (2).

Fig. 8 shows that when used separately, for all categories of queries other than polysemy, **Google-L2** distance is more accurate than co-click statistics. There are three likely reasons for such results: first, **Google-L2** distance, used by Google Similar images, is highly optimized over the image features; Second, the perceptual similarity test we used to obtain the labels ("which of the image is more similar to the query image") is conducted without giving raters a specific search task in mind. For this reason, raters are more likely to base their judgement on what they perceive as the most dominant visual properties of the images (e.g., dominant color of the background, etc.).

Third, as shown in column 1 of Fig. 7, images that are near-duplicate of each other are frequently shown in the top search results (especially in queries associated with *product* categories) and therefore in the testing image triplets. In such cases, when one of the near-duplicate image is used as the query image, raters often choose the other one as the more-similar image. Such near-duplicate images can be easily identified using global features. However, we observed that near-duplicate images do not usually receive more clicks than images that are visually less similar but semantically related. We conjecture that this is due to the fact that when searching for photo from the image search results, search engine users are unlikely to exam near-duplicate images during the same search sessions. For this reason, such task-dependent click patterns have different properties than those derived from perceptual similarity test.

Fig. 8 also shows that for queries in *polysemy* categories. **Co-click** statistics outperforms **Google-L2** distances. This is due to the observation that, as shown in column 2 and 3 of Fig. 7, a particular visual concept (such as apple logo) has high intra-class variance with respect to the image features. Images can be similar in multiple feature dimensions such as color (green apple, green logo) or shape. Query-independent distance functions, such as **Google-L2**, have limited capacity to select features that are important to disambiguate images produced by the text-query. For this reason, **Google-L2** is less accurate than distance generated from co-click statistics. By learning feature weights from the co-click statistics, those most discriminative features in this query, such as shape of the logo, are likely to have more weights over other features (color, etc.).

Fig. 8 also shows that combining both co-click statistics and Euclidean distance over the image features, **Coclick+L2**, produces more accurate estimation of image distances than when either is used separately. This result is not surprising as combining two largely independent sources of information usually produce more accurate results than when either one is used separately.

Fig. 9 presents the accuracy of co-click statistics derived from the query logs. **TP/FP** represents true positive/false positive rates—the percentage of testing triplets where the label agrees/disagrees with the output of (2). **0** (Difficult-to-decide) represents the percentage of testing triplets where co-click statistics are not sufficient to estimate the more similar image (output 0 in (2)). **Other** represents the percentage of testing triplets where the selections made by the three raters differ from one another (or they all choose to skip the triplet by choosing *cannot decide*.).

Fig. 9 shows that the number of testing triplets that are correctly predicted by using co-click statistics (TP) is significantly larger than those with incorrect predictions (FP) and this difference is consistent across all queries. This result suggests co-click statistics, even when used independent of other sources

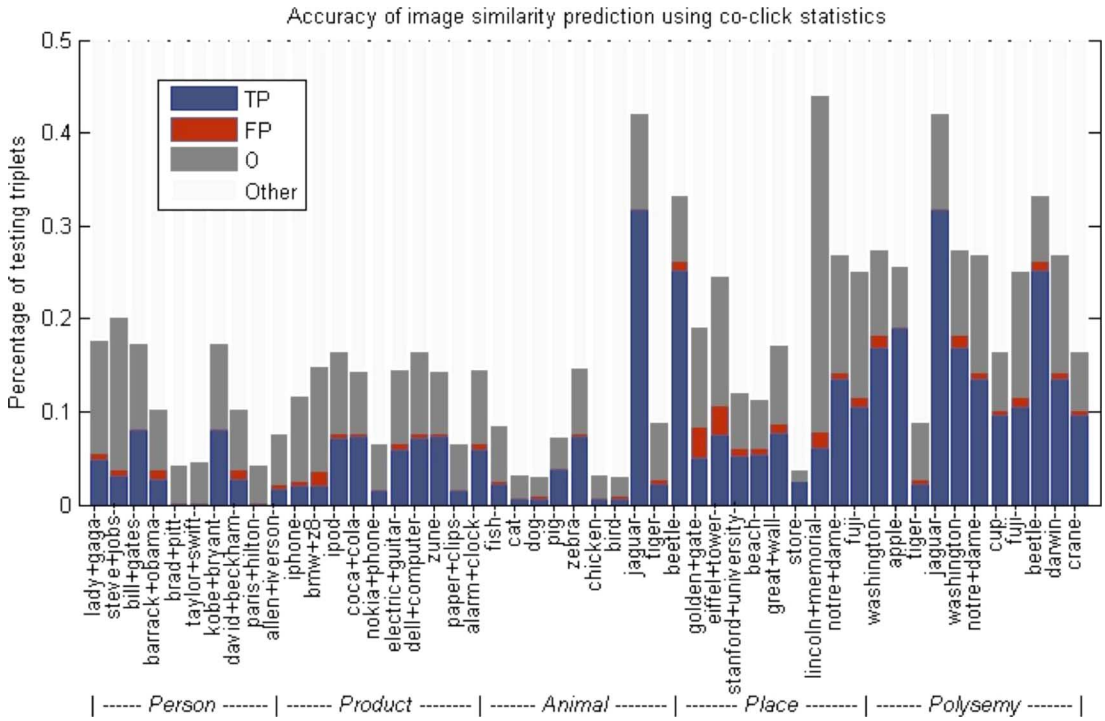


Fig. 9. The accuracy of co-click statistics. TP/FP represents true positive/false positive rates—the percentage of testing triplets where the label agrees/disagrees with the output of (2). 0 represents the percentage of testing triplets where co-click statistics are not sufficient to estimate the more similar image (output 0 in (2)). Other represents the percentage of testing triplets where the selections made by the three raters differ from one another (or they all choose to skip the triplet by choosing *cannot decide*.).

of signal, is a reliable measurement of image similarity. Note that we removed the testing triplets labeled as **other** from consideration, as the difference in perceived similarity relative to the query image is either too small or too subjective. Fig. 9 also shows that for all queries, majority (more than 50%) of the image triplets received the rating of **other**, more so in categories such as person, product and animal than polysemy.

3) Analysis 3: Accuracy of Query-Specific Distance: This section describes a set of experiments designed to evaluate the accuracy of learned query-specific distance functions, and compare them against the global (query-independent) [30] distance functions. We sample training triplets from the combined distance using (3), and use them to train query-specific distance functions. All methods use the same type of features—each image is represented as a fixed dimension feature vector derived from first concatenating and quantizing various types of image features such as color, texton and wavelets and then use kPCA (with histogram intersection kernel) to reduce the dimensionality of the feature space. We used the most significant 59 dimensions. The difference between Google-L2, $L2_W q$ and $L2_w$ are the learned feature weights.

Fig. 10 shows the accuracy of these query-specific distances. We compute the average true positive and false positive rates for each categories of queries. $\mathbf{L2}_{w_q}$ represents query-dependent distance; $\mathbf{L2}_w$ represents query-independent distance learned from the same training data. Fig. 10 shows that $\mathbf{L2}_{w_q}$ outperforms both $\mathbf{L2}_w$ and **Google-L2**. The improvement is most significant in category *person, place* and *polysemy*. Note that in most cases, learning a single query-independent distance function resulted in worse performance than query-specific distance functions.

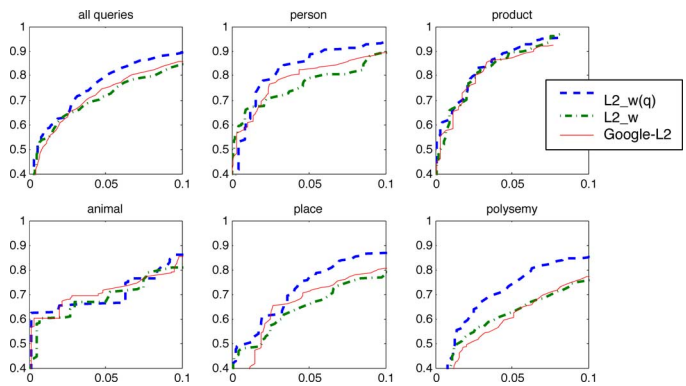


Fig. 10. The accuracy of query-specific distances ($L2_w(q)$) against state-of-the-art distance function used by Google (Google-L2), and query-independent distances ($L2_w$).

Fig. 11 presents a set of ranking results where the learning query-specific distance is particularly beneficial. Each row presents the top 10 nearest neighbor images retrieved with the first image as the query image. The odd number of rows (1, 3, 5, …) are ranking based on **Google-L2**, while the even number of rows (2, 4, 6, …) are those base on query-dependent distance ($\mathbf{L2}_{w_q}$).

D. Analysis 4: Size of Training Data on Accuracy

As it is computationally expensive to train with all available training triplets,[2] this section studies the effect of number of training data on the accuracy of the learned distance. Fig. 12

[2]In our experiment, we obtained an average of 2 million training triplets for each query.

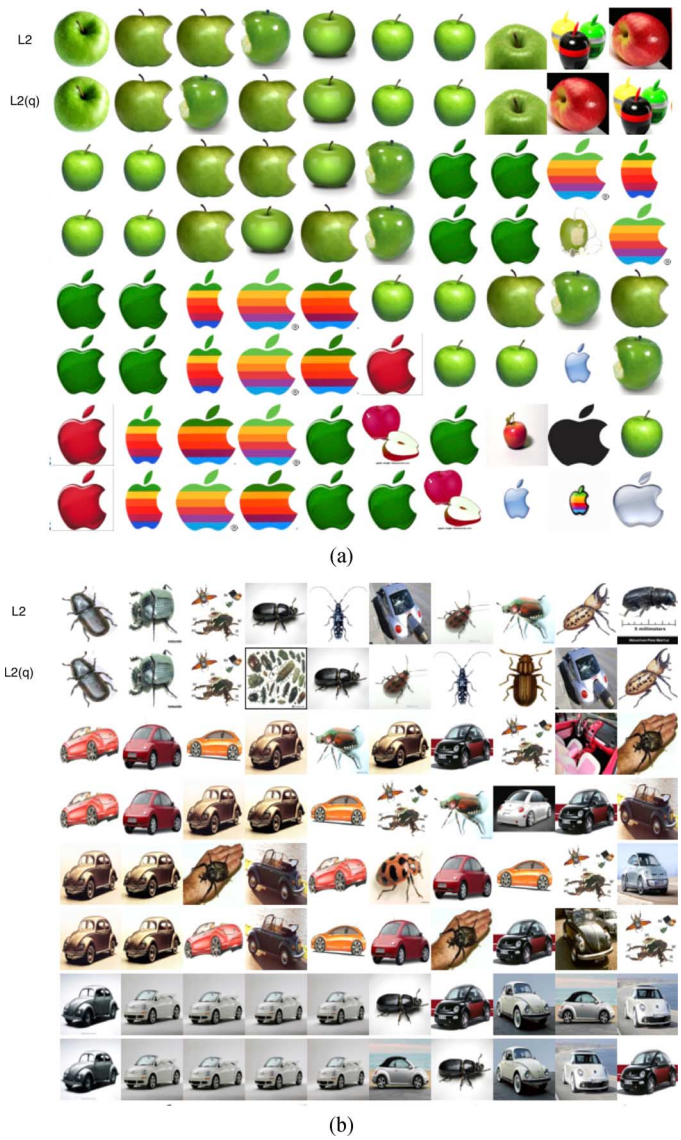


Fig. 11. Examples of image ranking results. Each row presents the top 10 nearest neighbor images retrieved given the first image as the query image. The odd number of rows (1, 3, 5, …) are ranking based on Google-L2, while the even number of rows (2, 4, 6, …) are those base on query-dependent distance $\mathbf{L2}_{w_q}$. (a) apple. (b) beetle.

presents the accuracy of query-specific distance given the number of available training triplets. We randomly sampled **2%**, **16%**, and **64%**, of the triplets from the all available training data. The result shows that the testing accuracy improves quickly as we increase the number of training data from 2% to 16%. The accuracy distance functions trained from 15% of the available data is comparable with those trained from all the data.

E. Analysis 5: Visualizing the Learned Feature Weights

In order to give more intuitions behind the learned weights, we conducted a separate experiment: instead of concatenating various features and then use kPCA for dimension reduction (a process that makes the resulting feature weights difficult to interpret), we treat each type of feature (e.g., color, wavelet, texton) separately, and learn the relative importance of these features using the same learning approach. The results are shown in Fig. 13.

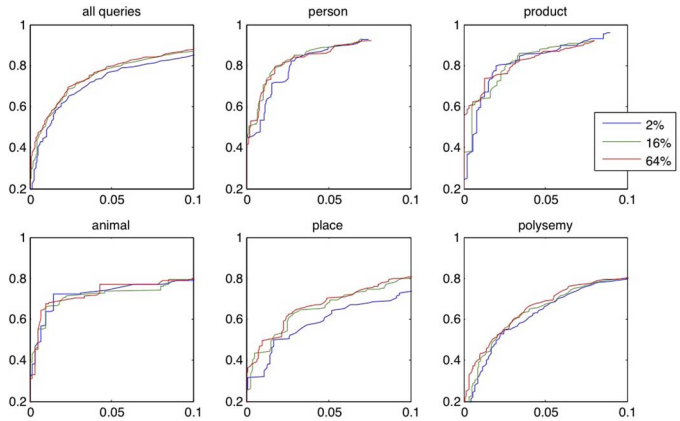


Fig. 12. The accuracy of query-specific distance given the number of available training triplets.

Fig. 13 shows that color (LAB) is most dominant in distance function for queries such as *steve jobs*, *brad pitt*, *jaguar*, *taylor swift*, *allen iverson*, *zune* and *chicken*, and least dominant in *electric guitar*, *fuji*, *iphone* and *cup*. We were initially puzzled by the fact that color is also important for queries related to people (e.g., brad pitt). Further analysis reveals that, when examining the image triplets related to celebrities, if all three images contain the person of interest (which is most likely given the triplets are selected from the top 100 search results), human raters find it hard to choose the more similar image. Very often they choose *cannot decide* given triplets related to such queries, except in cases where one of the candidate image is near-duplicate of the other, or when there is a strong similarity to the color of the dress or background stage. We also found that features sensitive to the shape or texture of the images, such as texton, are dominant in queries including *pig*, *great wall*, *cat*, *cup* and *electric+guitar*.

VI. Experiment II

A. Target Search Evaluation

This section evaluates distance functions in the actual task of image search. Specifically, we adopted the *target-search* experiment methodology [24], designed to simulate the actual retrieval task of locating a specific "target" image using image retrieval systems. The experiment consists of two steps: first, a target image selected from the image database is presented to the subject for a short duration of time; next, the subject is instructed to locate the target image based on their "mental image" using a retrieval or browsing system. The experiments are timed so that the speed with which a task is completed is used to quantify the effectiveness of the retrieval system. This is a variation of a simulated work task situation [1], and has been used previously [6], [21], [3] to evaluate information visualization systems.

We adopted the target-search experiment methodology for two reasons. First, searching for a target image with a specific criteria in mind is one of the most common mode of image search tasks [8], [4], [25], so users' efficiency in conducting target-search is a strong indication of the effectiveness of an

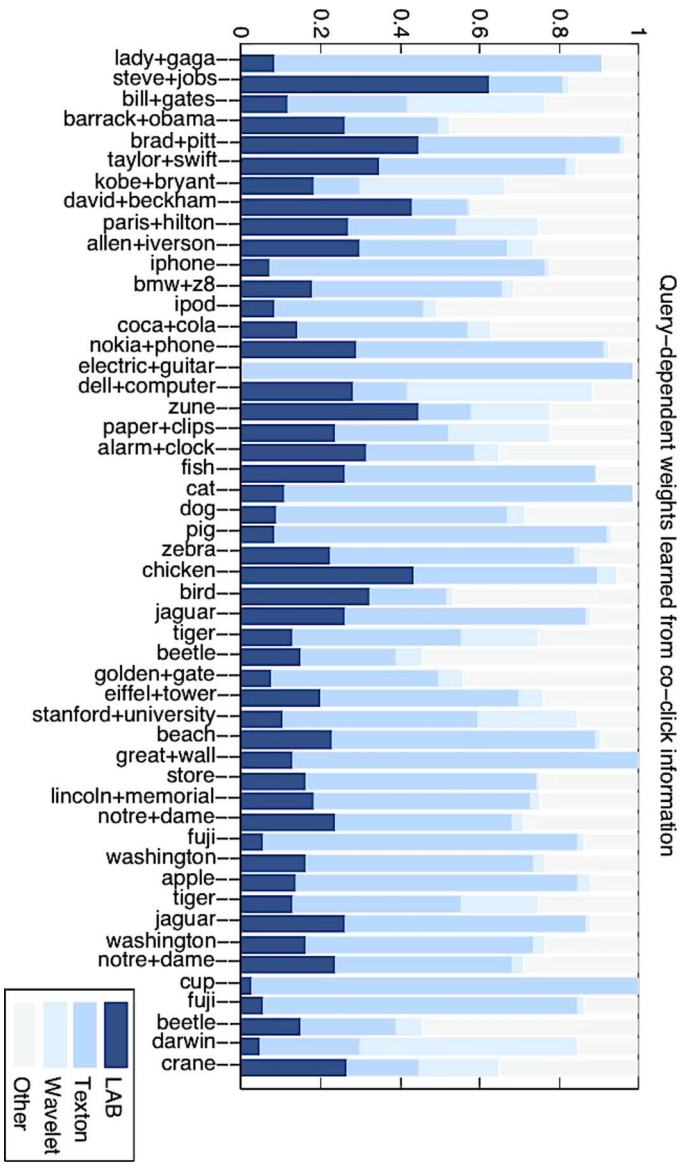


Fig. 13. The relative importance of features with respect to the query.

image retrieval system. Second, an image retrieval system is only meaningful in its service to people, so performance measurement should be anchored in human evaluation, especially when the retrieval system allows users to interact with the search results (e.g., selecting an image exemplar).

We made two choices when designing the study. First, we make the assumption that the search results contain the target image. We conjecture that in practice, if the user cannot find an image in the search results, she or he will formulate another query and repeat this process until the target image is found. Second, in order to better simulate the case where users find images based on a "mental" sketch of the target image, we make the decision to remove the image from the user view after displaying for a short period of time.

1) Experiment Methodology: We follow the grid layout most commonly used in current Web image retrieval systems. As a typical Web image retrieval system displays 20–30 images per page, our system displays 24 images per page. User can use a scrollbar to see the next page of images. The system caches all the images in the browser memory at any given time, to reduce the display latency. Standard 27-inch monitor is be used, with the browser (chrome) is maximized to occupy the entire screen.

Subjects are instructed to use the scroll-bar or arrow key to browse through the images. At any time, a subject can switch from text query-based ranking to content-based ranking by selecting an image example. If the clicked image is the target image, then the task is completed. We make the assumption that after an image example is selected, the subject needs to browse through the re-ranked images without making further image selections. This is to simplify our analysis of the retrieval system. In practice, a search engine user may choose to go back to the original search results, or issue a new query if they cannot find the target image in the top re-ranked results. We plan to study these more complex interaction processes in future work. We also allow experiment subjects to "skip" any particular task by clicking a button located at the beginning of the search results.

We selected queries using similar methodology to those described in Section V-B. Fifteen human subjects participated in this study. They are recruited by a third party who had no knowledge about the goal of the experiments. Human subjects were asked to complete a set of *randomly* selected tasks within a period of time. The random selection of retrieval tasks ensures that experiments are not biased by the order in which tasks are presented to the subjects—we expect subjects to become increasingly efficient with the hybrid retrieval systems and therefore more adept at finding the target image. We do not inform the subjects about the type of ranking algorithm used. This is to prevent subjects from developing strategies to exploit the artifacts of a particular ranking algorithm. At the beginning of the experiment, we present subjects with an instruction page on how to use hybrid image retrieval systems.

2) Evaluation Criteria: We use three quantitative measurements to evaluate the effectiveness of image retrieval systems, **time-to-completion**, **target-rank** and **abandonment-rate**. **Time-to-completion** is the time it takes for the subject to locate the target image.

Target-rank is the position of the target image in the search results. It is closely related to the number of images that users need to examine before the target image is found. Given a text query-based image retrieval system, target-rank is simply the rank of the target image given text query, as we assume that the subjects observe all the images positioned ahead of the target image. In a *hybrid* image retrieval system, subjects can change the ordering of the search results by selecting of image exemplar, therefore, the value of target-rank depends on a) the position of the target image given the query, b) the position of the selected image example, and c) the position of the target image after re-ranking. Specifically, we define target-rank (TR) for hybrid image retrieval system given the target image t and user selected image exemplar s as:

$$TR(t,s) = P(s) + P(t,s) \tag{9}$$

where $P(s)$ is the position of the similar image s in the search results produced by the text query. $P(t,s)$ is the position of the target image t after reordering based on similarity to the image exemplar s. Note that in the initial search results, if the target image is ranked ahead of the best ranked similar image, then

Fig. 14. In target-search experiment, the user is first briefly shown a target image and then instructed to locate the image from an image database using a specific retrieval system.

target-rank is simply the position of the target image in the initial search results, or $TR(t) = P(t)$.

Compared with time-to-completion which is affected by multiple factors not related to the retrieval system (i.e., concentration of the individual when completing that task), target-rank depends only on the ranking functions used and subjects' selection of the similar image. Experiments presented in Section V-C demonstrated that a collection of query-specific distance functions can outperform a Euclidean distance (used for all queries) in predicting the outcome of image similarity comparison test. Therefore, we conjecture that, on average, the target image will be ranked higher with a query-specific distance function given the selected similar image. This will result in lower target-rank, and if the difference is sufficiently large, also result faster time-to-completion.

Abandonment-rate is the percentage of target-search tasks abandoned by the subjects. A task is abandoned when the subject clicks the "skip" button in the option pane. We believe that abandonment rate offers clear indication on the effectiveness of an image retrieval system. We did not give specific instructions on when the "skip" button should be used. We suspect that subjects are likely to abandon a task when the target photo is difficult to interpret and/or when the subject has experienced frustration in locating the photo in the search results. We aggregated all the abandoned tasks across all subjects and grouped them based on the type of image retrieval system used.

3) Experiment Results: Fig. 16 displays the completion statistics of the each subject using scatter plots. The x-axis represents the average completion statistics using the baseline hybrid image retrieval system (**Google-L2**). The y-axis representing the average completion statistics using the query-specific distance functions ($\mathbf{L2}_w(q)$). Each point on the graph represents the comparison of the performance measurement of a single subject. A point below the diagonal line suggests that a subject is able to find the target image faster using the proposed approach than using baseline.

The results show that the two distance functions are comparable with each other in time-to-completion, but $\mathbf{L2}_w(q)$ consistently produces better target-rank than **Google-L2** (11 out of 15). We conjecture that this result is due to the fact that Google-L2 is already very accurate – a small improvement in the rank (or position) of the target image does not significantly affect the time it takes for user to locate the target image.

Table II shows the percentage of tasks abandoned given each image retrieval systems, averaged across the tasks for all subjects. It shows that query-specific distance functions resulted in lower rate of abandonment. In other words, for every 100 tasks

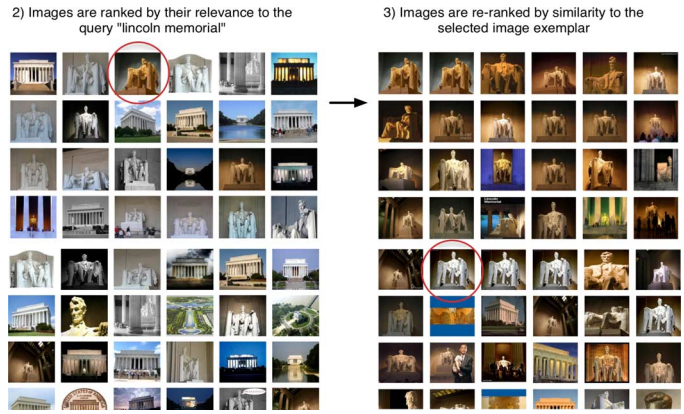


Fig. 15. An example of how an experiment subject locates the target image using hybrid image retrieval system.

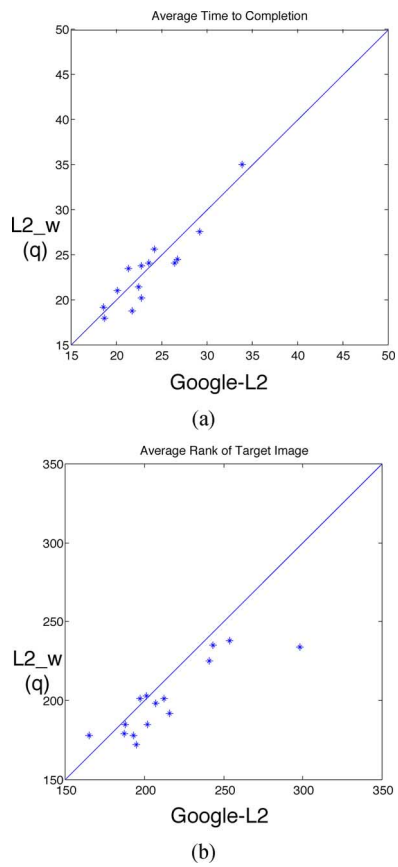


Fig. 16. Google-L2 v.s. Query-specific distances. (a) Time to completion. (b) Target Rank.

conducted using $\mathbf{L2}_w(q)$, subjects abandoned on average 9.4 tasks, 3.4 fewer than **Google-L2**. As abandonment rate is an important criteria to measure retrieval success, this result shows that query-specific distance functions improves the user experience of hybrid image retrieval systems. Fig. 17 shows the correlation between target-rank and the abandonment rate: 1% of the tasks were abandoned when the target image had a rank of less than 200, 12% were abandoned with target-rank with more than 200 but less or equal to 400. The abandonment rate becomes significantly higher when the rank increased to above 400. 78%

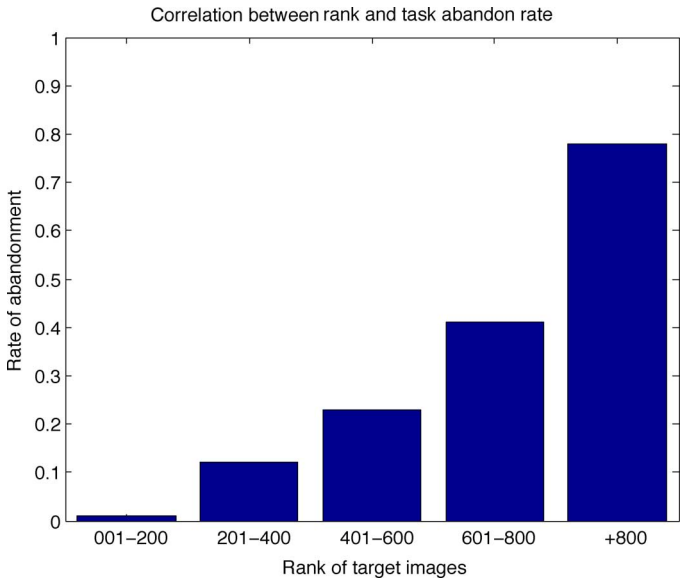


Fig. 17. Correlation between target-rank and task abandonment.

TABLE II
THE PERCENTAGE OF TASKS ABANDONED BY SUBJECTS RATE WHEN EACH IMAGE RETRIEVAL SYSTEM IS USED

Retrieval System	Google-L2	Query-specific
Abandonment Rate	12.8 %	9.4%

of the tasks were abandoned when the target image has the rank of more than 800.

VII. CONCLUSION

This work studies the feasibility and efficacy of learning query-specific distance functions for large-scale Web image search. We demonstrate that 1) co-click statistics derived from text-based search engine query logs can be used to predict how a person will compare images based on perceptual similarity, 2) one can learn query-specific distances from such sources of information, and the learned distances produce more accurate comparisons of images and reduce search abandonment-rate compare to the query-independent image distance function used by current Google image search.

There are two potential ways to improve the learning approaches proposed in this work. First, one way to improve query-specific distance functions is to allow related text-queries to "share" the learned distance functions. For example, one can first group the text queries into synsets, and learn *synset-specific* distance functions for each synset. One can derive synsets from an expert-knowledge database such as WordNet [20], or from the text and images associated such queries [34]. By allowing training data to be shared among related queries, synset-specific distances can be computed for less commonly used queries. Sharing distance functions also reduces the number of distance functions that need to be cached by the retrieval system.

Second, since our results have shown that query-specific distance functions can improve ranking accuracy in certain query categories (e.g., polysemy) more than others (e.g., animal), the ability to automatically select queries or query categories that are suitable for such distance functions would be beneficial. One possible approach is to measure the disagreement between the co-click statistics and the visual similarity produced by using un-weighted Euclidean distance, and use such disagreement as an indication of whether query-specific distance can be useful.

REFERENCES

[1] P. Borlund and P. Ingwersen, "The development of a method for the evaluation of interactive information retrieval systems," *J. Document.*, vol. 53, pp. 225–250, 1997.

[2] B. Carterette and R. Jones, "Evaluating search engines by modeling the relationship between relevance and clicks," in *Proc. Advances in Neural Inf. Process. Syst. (NIPS 2007)*.

[3] T. T. A. Combs and B. B. Bederson, "Does zooming improve image browsing?," in *Proc. 4th ACM Conf. Digital Libraries, DL '99*, New York, NY, USA, 1999, pp. 130–137.

[4] L. R. Conniss, A. J. Ashford, and M. E. Graham, Information Seeking Behaviour in Image Retrieval: Visor i Final Report. Technical Report Institute for Image Data Research, 2000, Library and Information Commission Research Report 95.

[5] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos, "The Bayesian image retrieval system, pichunter: Theory, implementation, and psychophysical experiments," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 20–37, 2000.

[6] I. J. Cox, M. L. Miller, S. M. Omohundro, and P. N. Yianilos, "Target testing and the pichunter Bayesian multimedia retrieval system," in *Proc. 3rd Int. Forum on Research and Technol. Advances in Digital Libraries*, Washington, DC, USA, 1996, p. 66.

[7] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey, "An experimental comparison of click position-bias models," in *Proc. Int. Conf. Web Search and Web Data Mining, WSDM '08*, New York, NY, USA, 2008, pp. 87–94.

[8] R. Datta, D. Joshi, J. Li, and J. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surveys*, no. 2, 2008.

[9] J. Dean and S. Ghemawat, "MapReduce: A flexible data processing tool," *Commun. ACM*, vol. 53, no. 1, pp. 72–77, 2010.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR09*, 2009.

[11] A. Frome, Y. Singer, and J. Malik, "Image retrieval and classification using local distance functions," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2007.

[12] K. Grauman and T. Darrell, "The pyramid match kernel: Efficient learning with sets of features," *J. Mach. Learn. Res.*, vol. 8, pp. 725–760, Apr. 2006.

[13] G. Griffin, A. Holub, and P. Perona, Caltech-256 Object Category Dataset, California Institute of Technology, 2007, Tech. Rep. 7694.

[14] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl, "An algorithmic framework for performing collaborative filtering," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. and Develop. in Inf. Retrieval, SIGIR '99*, New York, NY, USA, 1999, pp. 230–237.

[15] S. C. H. Hoi, W. Liu, M. R. Lyu, and W. ying Ma, "Learning distance metrics with contextual constraints for image retrieval," in *Proc. Comput. Vision and Pattern Recognition*, 2006, pp. 2072–2078.

[16] V. Jain and M. Varma, "Learning to re-rank: Query-dependent image re-ranking using click data," in *Proc. Int. World Wide Web Conf.*, Mar. 2011.

[17] T. Joachims, "Optimizing search engines using clickthrough data," in *Proc. 8th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, KDD '02*, New York, NY, USA, 2002, pp. 133–142.

[18] T. Joachims, L. Granka, G. Inc, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay, "Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search," *ACM Trans. Inf. Sci. (TOIS 2007)*.

[19] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge Univ. Press, 2008.

[20] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, "Wordnet: An on-line lexical database," *Int. J. Lexicography*, vol. 3, pp. 235–244, 1990.

[21] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun, "Performance evaluation in content-based image retrieval: Overview and proposals," *Pattern Recognit. Lett.*, vol. 22, pp. 593–601, Apr. 2001.

[22] F. Radlinski, M. Kurup, and T. Joachims, "How does clickthrough data reflect retrieval quality?," in *Proc. 17th ACM Conf. Inf. and Knowledge Management, CIKM '08*, New York, NY, USA, 2008, pp. 43–52.

[23] S. Robertson, M. Vojnovic, and I. Weber, "Rethinking the ESP game," in *Proc. 27th Int. Conf. Extended Abstracts on Human Factors in Computing Systems, CHI EA '09*, New York, NY, USA, 2009, pp. 3937–3942.
[24] K. Rodden, W. Basalaj, D. Sinclair, and K. Wood, "Evaluating a visualisation of image similarity (poster abstract)," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. on Res. and Develop. in Inf. Retrieval, SIGIR '99*, New York, NY, USA, 1999, pp. 275–276.
[25] K. Rodden and K. R. Wood, "How do people manage their digital photographs?," in *Proc. SIGCHI Conf. on Human Factors in Comput. Syst., CHI '03*, New York, NY, USA, 2003, pp. 409–416.
[26] E. Rosch, "Natural categories," *Cognitive Psychol.*, vol. 7, pp. 573–605, 1973.
[27] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vision*, 2000.
[28] G. Salton, "The state of retrieval system evaluation," *Inf. Process. Manage.*, vol. 28, pp. 441–449, Mar. 1992.
[29] P. C. Saraiva, E. Silva de Moura, N. Ziviani, W. Meira, R. Fonseca, and B. Riberio-Neto, "Rank-preserving two-level caching for scalable search engines," in *Proc. 24th Annu. Int. ACM SIGIR Conf. on Res. and Develop. in Inf. Retrieval, SIGIR '01*, New York, NY, USA, 2001, pp. 51–58.
[30] M. Schultz and T. Joachims, "Learning a distance metric from relative comparisons," in *Proc. 16th Conf. Advances in Neural Inf. Process. Syst. (NIPS)*.
[31] N. V. Shirahatti and K. Barnard, "Evaluating image retrieval," in *Proc. 2005 IEEE Comput. Soc. Conf. on Comput. Vision and Pattern Recognition (CVPR'05)*, Washington, DC, USA, 2005, vol. 1, pp. 955–961.
[32] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, 2000.
[33] S. Shalev-Schwartz, Y. Singer, and N. Srebro, "Pegasos: Primal estimated sub-gradient solver for SVM," in *Proc. Int. Conf. Machine Learning*, 2007.
[34] D. Tsai, Y. Jing, Y. Liu, H. A. Rowley, S. Ioffe, and J. M. Rehg, "Large-scale image annotation using visual synset," in *Proc. Int. Conf. Comput. Vision*, 2011.
[35] S. Uchihashi and T. Kanade, "Content-free image retrieval by combinations of keywords and user feedbacks," in *Proc. Int. Conf. Image and Video Retrieval (CIVR)*, 2005.
[36] K. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. 18th Conf. Advances in Neural Inf. Processing Syst.*, , 2006, pp. 1437–1480.
[37] G. Wu and E. Y. Chang, "Formulating context-dependent similarity functions," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2005, pp. 725–734.
[38] E. Xing, A. Ng, M. Jordan, and S. Russel, "Distance metric learning, with applications to clustering with side-information," in *Proc. 15th Conf. Advances in Neural Inf. Process. Syst. (NIPS)*, 2002, pp. 450–459.
[39] L. Yang and R. Jin, Distance Metric Learning: A Comprehensive Survey. Tech. Rep., Michigan State Univ., Dept. Comput. Sci. Eng., 2006.
[40] X. Zhou and T. Huang, "Relevance feedback for image retrieval—A comprehensive survey," *Multimedia Syst.*, 2004.

Yushi Jing received his B.S., M.S., and Ph.D. in computer science from Georgia Institute of Technology. Currently, he works as a senior research scientist at Google Research. As one of the founding members of the computer vision research group, he worked on multiple projects related to visual-search, image recognition, large-scale data mining and visualization. He is the recipient of best-student-paper-award at International Conference on Machine Learning (ICML) and his work on Google VisualRank was covered by New York Times in 2008. Outside of research, he enjoys building software tools to help non-profits and NGOs, with experiences in the United Nations and Google.org.

Michele Covell received her BS in electrical engineering from the University of Michigan and M.S. and Ph.D. from MIT in signal processing. She joined SRI International, in the area of active acoustic-noise control, and then Interval Research Corporation, where her research covered a wide range of topics in audio, image, and video processing. In 2000, she joined a startup (YesVideo) and worked on faster-than-real-time video analysis. She moved to the Mobile Streaming Media group in HP Labs, as a key contributor in streaming-video services in 3G telephony networks. This work is listed as one of the top-40 accomplishments from HP Labs' 40-year history. She moved to Google, in the research group, in 2005, where she focused for several years on large-scale audio and video fingerprinting, identification, and retrieval. For this work, she received two Google awards—one for innovation and one for financial impact. More recently, she has been working in image and texture analysis and in large-scale graph labeling problems.

David Tsai received the BE degree in Electrical Engineering from Tsinghua University in 2009 and the M.E. degree in Computer Science from Georgia Institute of Technology in 2012, advised by Professor James M. Rehg. He also worked part time in Google on various projects including large scale image search, image annotation, visual browsing and video object segmentation from 2010 to 2012. He is currently on leave from the Ph.D. program in Georgia Tech. His research interests include large scale computer vision, machine learning and multimedia. In 2010 he received the BMVC Best Student Paper Award. In 2012 he received Best Paper Award in ECCV Workshop on Web-scale Vision and Social Media.

James M. Rehg is a Professor in the School of Interactive Computing at the Georgia Institute of Technology, where he is the Director of the Center for Behavior Imaging, co-Director of the Computational Perception Lab, and Associate Director of Research in the Center for Robotics and Intelligent Machines. He received his Ph.D. from CMU in 1995 and worked at the Cambridge Research Lab of DEC (and then Compaq) from 1995–2001, where he managed the computer vision research group. He received the National Science Foundation (NSF) CAREER award in 2001, and the Raytheon Faculty Fellowship from Georgia Tech in 2005. He and his students have received a number of best paper awards, including best student paper awards at ICML 2005 and BMVC 2010. Dr. Rehg is active in the organizing committees of the major conferences in computer vision, most-recently serving as the Program co-Chair for ACCV 2012. His research interests include Computer Vision, Robotics, and Machine Learning. He is a member of the IEEE and ACM.