Web Image Prediction Using Multivariate Point Processes

Gunhee Kim
Computer Science Department
Carnegie Mellon University
Pittsburgh, 15213 PA
gunhee@cs.cmu.edu

Li Fei-Fei
Computer Science Department
Stanford University
Stanford, 94305 CA
feifeili@cs.stanford.edu

Eric P. Xing
Computer Science Department
Carnegie Mellon University
Pittsburgh, 15213 PA
epxing@cs.cmu.edu

ABSTRACT

In this paper, we investigate a problem of predicting *what images are likely to appear on the Web at a future time point*, given a query word and a database of historical image streams that potentiates learning of uploading patterns of previous user images and associated metadata. We address such a *Web photo prediction* problem at both a collective group level and an individual user level. We develop a predictive framework based on the multivariate point process, which employs a stochastic parametric model to solve the relations between image occurrence and the covariates that influence it, in a globally optimal, flexible, and scalable way. Using Flickr datasets of more than ten million images of 40 topics, our empirical results show that the proposed algorithm is more successful in predicting unseen Web images than other candidate methods, including reasoning on semantic meanings only, a state-of-art image retrieval method, and a generative topic model.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*

Keywords

Web image prediction, multivariate point processes, penalized Poisson regression, personalization

1. INTRODUCTION

The prevalence of digital cameras and smartphones has led to an explosion of pictures being uploaded and shared online, across websites, platforms and social networks. This phenomenon poses great challenges and opportunities in multimedia data mining research. In this paper, we address an interesting problem along this line – predicting likely images to appear on the Web at a future time point and retrieving images similar to them from the database, after learning patterns of previous user images and associated metadata.

Fig.1 shows our problem statement with an example of the query *world+cup*. Suppose that we have the image database downloaded from Flickr for the *world+cup* up to 12/31/2008[1]. Can we then estimate what would be the most likely pictures that are taken in a future query time[2] 6/6/2009, and retrieve images similar to them from the database? As Fig.1.(c) has shown, the pictures actually taken at 6/6/2009 and shared on Flickr are not necessarily about the best possible world cup pictures (if the definition of *best* is even possible). Instead, they are the pictures that not only reflect the semantic meaning of the keyword, but also people's intends at that given moment of time. Furthermore, if a user cue is supplemented, the image prediction becomes highly personalized as shown in Fig.1.(d), given that individual users have their own preferences and photo-taking styles.

The problem in this paper is closely related to one active area of research in information retrieval: *exploring the temporal dynamics of user behaviors on Web queries* [5, 16, 18, 22]. The popularity of queries and their best search results change over time as people's interests evolve. For example, in [18], it is reported that more than 7% of queries are the ones that do not actually contain a year, but the user implicitly formulate with a specific year in mind (*e.g.* miss universe, Olympics). Moreover, many of them are connected to the events that have occurred with predictable periodicity. This line of research aims to improve search relevance by identifying what search terms are sensitive to time, what documents should be retrieved to the query time, and what webpages are likely to be clicked by a user at a particular time point. However, much of previous work has targeted at the search of documents such as blogs and news archives by analyzing the query log data; modeling and predicting temporal dynamics of Web user images has yet received little attention, even though photos are another popular modality to share the information on the Web.

KDD'12, August 12–16, 2012, Beijing, China.
Copyright 2012 ACM 978-1-4503-1462-6 /12/08 ...$10.00.

[1] Strictly speaking, we address a varient of *image re-ranking*; we assume that a text-based image search engine (*e.g.* Flickr search engine) provides a large-scale pool of unordered Web images for a given query word. Then, our goal is to re-rank those images to be fit for a query time. Image re-ranking following a text-based image search is the de facto pipeline for major image search engines such as Google and Bing [4].

[2] We are interested in a future time point as a query rather than past or present because it is most interesting and challenging. For a query time in the past, we may trivially retrieve the images taken at that time from the database. However, if the query time is in future, we have to learn users' photo uploading patterns and extrapolate likely images to appear for the query time.

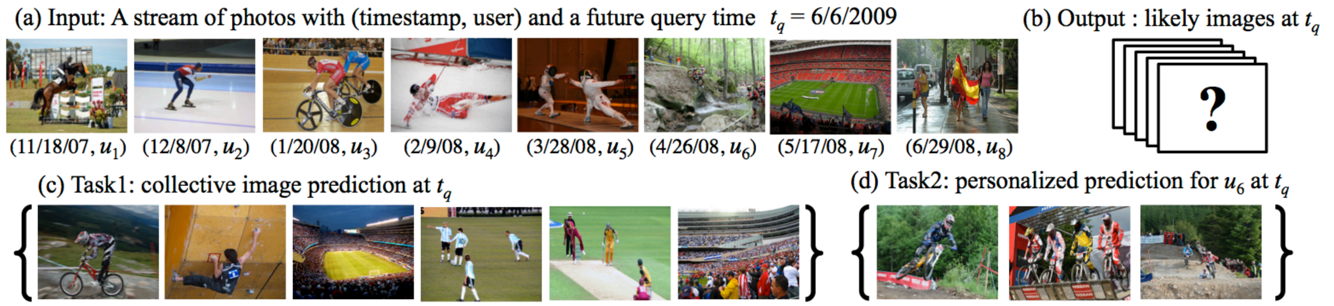

(a) Input: A stream of photos with (timestamp, user) and a future query time t_q = 6/6/2009
(b) Output : likely images at t_q
?
(11/18/07, u_1)
(12/8/07, u_2)
(1/20/08, u_3)
(2/9/08, u_4)
(3/28/08, u_5)
(4/26/08, u_6)
(5/17/08, u_7)
(6/29/08, u_8)
(c) Task1: collective image prediction at t_q
(d) Task2: personalized prediction for u_6 at t_q

Figure 1: (a) Given an image sequence of *world+cup*, can we guess what images are likely to appear at a future time t_q=6/6/2009? (c) Collective image prediction. The *world+cup* usually refers to the soccer event, so a soccer scene can be a reasonable guess. However, the actual Web images are diverse because they reflect different users' experiences and preferences. (d) Personalized image prediction for a user u_1. A user's unique angle of seeing the topic can make the prediction more focused. The images are sampled from Flickr at each time.

Consequently, one important application of our image prediction is *time and user sensitive Web image re-ranking.* Suppose that a user submits a query word of *world+cup* into *Google and Bing image search*, which then invariably retrieve redundant photos of soccer in the first page. Although the term *world+cup* usually refers to the international soccer event, it is also commonly used in other international sports and competitions (*e.g.* ski, skate, bicycle, or horse riding, as shown in Fig.1.(a)). Therefore, if the *world+cup* is submitted in winter by a user who likes skiing, it is more desirable to include ski world cup photos in the retrieved result. Our image prediction framework can enable the reranking of the retrieved images, so that various views of the query word are shown, according to who searches, and when the search takes place. With the majority of Web photos now coming from hundreds of millions of general users with different experiences and preferences, the contents of images that are associated even with the same keyword can be highly variable according to owners and temporal information.

On the technical aspect, we develop an image prediction algorithm using a multivariate point process, which is a stochastic process that consists of a series of random events occurring at points in time and space [6]. In our method, an observed image stream is viewed as an instance of the multivariate point process. Although this well-established statistical model has been employed for studying neural spiking activities [27], and for event detection in video [21], no attempt has been made for image retrieval or re-ranking so far. Nonetheless, we adapt it to offer several key advantages for large-scale image prediction as follows: (i) *Flexibility*: The image occurrence on the Web is correlated with a wide range of factors or covariates (*e.g.* season, time, user preference, and other external events). A parametric model can be easily set up to relate the image occurrence probability with any number of factors that influence it (section 3.3). (ii) *Optimality*: The sparse globally-optimal MLE solution is computed to identify only a small number of key factors and their relative weights (section 3.2). (iii) *Scalability*: The learning and prediction are performed in a linear time with respect to all parameters, including time steps and the number of covariates (section 3.4). (iv) *Prediction accuracies*: Our experiments on more than ten millions of Flickr images have demonstrated compelling results on both collective and personalized image forecast over various 40 topic keywords. Indeed we show that our approach outperforms other methods including a PageRank-based image retrieval [15] and a generative author-time topic model [23] (section 4).

1.1 Relations to Previous work

The problem of image prediction using large-scale Web photo collections remains an under-addressed topic in the image retrieval literature. Our work is remotely related to following four lines of research, but is significantly different on the task, utility and methodology. Due to vast volume of literatures on these topics, we introduce only some selected papers that are most closely related to our work.

Web content dynamics: This research aims at large-scale analyses to describe how the Web content changes over time. Most previous work [1, 28] has dealt with the textual content on the Web such as news articles and scientific libraries. In the image domain, the most related work to ours may be [15] in that both involve studying topic evolution in large-scale Flickr photos. However, the main tasks of [15] were subtopic outbreak detection and classification of noisy web images. They did not address the image prediction, which is our main task here. Also, they did not explore any issues regarding personalization, as done in this work.

Similar image retrieval: The image prediction problem is also related to similar image retrieval, a well-studied topic in computer vision [8, 20, 26]. They are related in a way that in both cases, given a query, relevant images are returned from the database. Yet, there are a number of key differences. Traditional similar image retrieval tends to focus solely on the semantic meaning of the query word and feature-wise image similarity, whereas our image prediction additionally emphasizes the temporal trends and user histories associated with the images.

Image based collaborative filtering: The goal of this research is to mine the trends of people's interests from community photos such as Flickr. Examples include the social trends in politics and market [13], and spatio-temporal events [24]. However, most existing work has used images as the source of information to infer other phenomena rather than taking themselves as a subject to be forecasted.

Leveraging Web photos to infer missing information: The final related work is on inferring missing information by leveraging a large-scale Web image corpus. Some notable examples include scene completion [12], geolocation estimation of a photo sequence [14], 3-D models of landmarks [25], semantic image hierarchy [17], and people matching [11]. However, future image occurrence has not been explored as missing information to be inferred.

1.2 Summary of Contributions

Departing from the literatures reviewed above, the main contributions of our work can be summarized as follows:

(1) We develop a method for collective and personalized image prediction. To the best of our knowledge, there have been few attempts so far on such prediction tasks using large-scale Web photos. Our work can be used in several interesting data mining applications, such as time and user based image suggestion and re-ranking.

(2) We design our algorithm using multivariate point processes. We are not aware of any prior instances of multivariate point process in image re-ranking applications; here we adapt this well-founded statistical model to address a number of key challenges of Web image prediction, including flexibility, optimality, scalability, and prediction accuracies.

2. PROBLEM STATEMENT

We define the image prediction as a variant of the time and user sensitive image re-ranking problem. As an input, an image database consists of Flickr photos in $[0, T)$ that are downloaded by a topic keyword, together with their metadata including timestamps and user IDs. Then, given a future time point $t_q > T$ in the form of (M/D/Y), we retrieve L number of the most likely images from the database. Actual Web images to be predicted at any days are usually hundreds or more in volume and extremely diverse in content. Therefore, we first predict the trends of image clusters, and sample multiple L images as output accordingly, in order to cover various aspects of the topic.

We address both *collective* and *personalized* image prediction. The former refers to a generic prediction for arbitrary individuals using all collected information; and the latter concerns a customized forecast for a particular individual u_q to be specified at test time. The personalized prediction focuses on an individual user's history, whereas the collective prediction deals with societally aggregated trends.

Our problem involves learning a model of the image occurrences with related factors or covariates, and then building a forecast algorithm to sample the likely images based on the learned model. Multivariate point processes are a unified statistical framework to solve these problems, which will be discussed in detail in the next section.

In this paper, we exploit three information modalities based on which a prediction is made: image description, user description, and timestamps at which photos are taken. For clarity, we explain bellow the preprocessing steps of the first two modalities, and the third one is self-explanatory.

Image Description: All images are clustered into M different groups, which we call as *visual clusters* in this paper. We first extract two types of features for each image, spatial pyramids of dense HSV SIFT and HOG[3]. Then, we construct a visual dictionary of M clusters (*e.g.* $M = 500$) for each topic by applying K-means to randomly selected $100K$ features. Finally, each image is assigned to the nearest visual cluster in the feature space.

User Description: Measuring user propensity is important in collaborative filtering [7] because a user's future behavior is likely to be correlated with those of users who are similar to her. Intuitively, each user can be represented by a set of images that she has posted. We first compute an M-dimensional histogram for each user where each bin represents the count of images belonging to the corresponding visual cluster. Instead of directly using the user descriptor, we perform the pLSI based user clustering proposed by Google News personalization [7]. In pLSI, the distribution of visual cluster v in a user u_i's images $(p(v|u_i))$ is given by the following generative model:

$$p(v|u_i) = \sum_{z \in \mathcal{Z}} p(v|z)p(z|u_i). \quad (1)$$

The latent variable $z \in \mathcal{Z}$ is assumed to represent the cluster of user propensity. Thus, $p(z|u_i)$ is proportional to the fractional membership of user i to cluster z. We denote $p(z|u_i)$ by $\mathbf{u}_i$, which is used as the descriptor of a user u_i. The $\mathbf{u}_i$ is an L-1 normalized $|\mathcal{Z}|$-dimensional vector (*i.e.* $|\mathcal{Z}| = 50$).

For simplicity and better exposition of our point process framework, we use relatively simple image and user descriptors, and assume that the images in the same visual cluster are interchangeable. However, it is straightforward to enhance our method by replacing them with richer descriptors (*e.g.* soft assignment of visual clusters) or by adding other types of information (*e.g.* text tags).

3. POINT PROCESSES FOR WEB PHOTO STREAMS

3.1 Multivariate Point Processes

We employ a multivariate point process to model a stream of input images, as illustrated in Fig.2. Formally, a multivariate point process can be described by a counting process $\mathbf{N}(t) = (N^1(t), \cdots, N^M(t))^T$ where M is the number of visual clusters and $N^i(t)$ is the total number of observed images assigned to the$i$-th visual cluster in the interval $(0, t]$. Then, $N^i(t + \Delta) - N^i(t)$ represents the number of images in a small interval Δ. By letting $\Delta \to 0$, we obtain the *intensity function* $\lambda^i(t)$ (*i.e.* image occurrence rate) at t, which indicates the infinitesimal expected occurrence rate of the images of the i-th visual cluster at time t [6]:

$$\lambda^i(t) = \lim_{\Delta \to 0} \frac{P[N^i(t+\Delta) - N^i(t) = 1]}{\Delta}, \ i \in \{1, \ldots, M\}. \quad (2)$$

Data likelihood: Suppose that we partition the interval $(0, T]$ by a sufficiently large number K (*i.e.* $\Delta = T/K$) so that in each time bin Δ only one or zero image occurs. In other words, if we let $\Delta N^i_k = N^i_k - N^i_{k-1}$ be the number of images of the i-th visual cluster occurred between t_{k-1} and t_k, then ΔN^i_k can be zero or one. Here, t_k is the k-th interval ($t_k = k\Delta$). Now we can denote the sequence of images up to T by $N^i_{1:K} = (\Delta N^i_1, \cdots, \Delta N^i_K)$. This discretization induces that the log-likelihood function is represented by [27]

$$l(N^i_{1:K}|\boldsymbol{\theta}^i) = \sum_{k=1}^{K} \log(\lambda^i(t_k|\boldsymbol{\theta})\Delta)\Delta N^i_k - \sum_{k=1}^{K} \lambda^i(t_k|\boldsymbol{\theta})\Delta \quad (3)$$

where $\lambda^i(t_k|\boldsymbol{\theta})$ is the parametric form of the intensity function at k-th interval.

[3] We use the codes for dense SIFT and HOG available at http://www.vlfeat.org/ and http://www.robots.ox.ac.uk/∼vgg/software/, respectively.

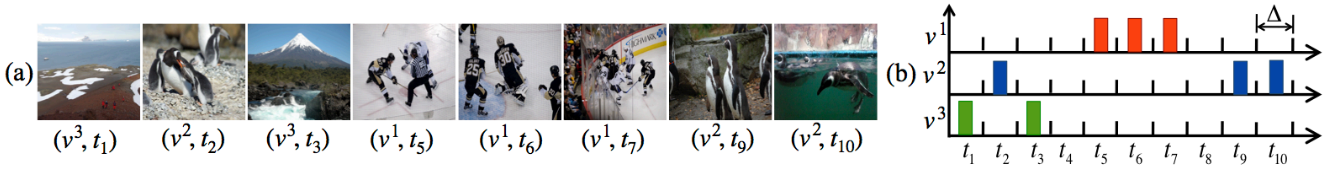


Figure 2: A multivariate point process for a short image stream of the *penguin*. (a) Each image is assigned to a visual cluster (v) up to M=3 and a timestamp (t). The visual clusters are {*ice hockey*, *animal penguin*, and *snowy mountain*}. (b) The image stream is modeled by a discrete-time trivariate point process according to visual clusters.

3.2 Regularized Generalized Linear Model

In order to connect the image occurrence with covariates, we model the intensity function as the exponential of a linear combination of functions f_j^i of the covariates x_k:

$$\log \lambda^i(t_k|\boldsymbol{\theta}^i) = \sum_{j=1}^{n} \theta_j^i f_j^i(x_1, \cdots, x_k),\ i \in \{1, \ldots, M\} \quad (4)$$

where $\boldsymbol{\theta}^i = (\theta_1^i, \cdots, \theta_n^i)$ is a vector of model parameters.

It is shown in [27] that the likelihood of a point process in Eq.(3) along with λ^i of Eq.(4) is identical to the likelihood of a *generalized linear model* (GLM) under a Poisson probability model and a log link function, which is also known as the *Poisson regression*.

L1 regularized likelihood: It is reasonable to assume that although numerous factors affect image occurrences, each visual cluster depends on only a small subset of them. Hence, it is important to detect a small number of strong covariates by encouraging a sparse estimator of $\boldsymbol{\theta}^i$, and we maximize the likelihood with l_1 penalty:

$$l_R(N^i_{1:K}|\boldsymbol{\theta}^i) = l(N^i_{1:K}|\boldsymbol{\theta}^i) - \mu \sum_{j=1}^{n} |\theta_j^i|. \quad (5)$$

We can efficiently solve the MLE solution to Eq.(5) (*i.e.* generalized linear models with mixed l_1/l_2 norm regularization) by using the cyclical coordinate descent in [10]. We use the regularized path to find the best regularization parameter μ; we perform a 10-fold cross validation procedure and choose μ that minimizes the mean cross-validated error.

Example: Here, we introduce a toy example to intuitively show how the proposed model predicts the image occurrence. For simplicity of the example, we assume that the intensity function is affected by only year and month covariates:

$$\log \lambda^i(t_k|\boldsymbol{\theta}^i) = \theta_0^i + \sum_{y=2003}^{2009} \theta_y^i I_y(t_k) + \sum_{m=1}^{12} \theta_m^i I_m(t_k) \quad (6)$$

where the parameter set comprises seven θ_y^i and twelve θ_m^i. $I_y(t_k)$ is an indicator function that is 1 if the year of t_k is y, and 0 otherwise (*e.g.* $I_y(t_k)$=1 if y=2009 and t_k= 03/01/2009). $I_m(t_k)$ is an indicator for the month.

Fig.3 shows the learned intensity functions $\lambda^i(t)$ of two visual clusters N^1 and N^2 with respect to years and months. Fig.3.(b) presents the observed image sequences. For both N^1 and N^2, the intensity functions increase every year (See Fig.3.(c)). The rates decrease in 2009 because the *shark* dataset was gathered up to mid 2009. Interestingly, N^1 and N^2 show different monthly behaviors (See Fig.3.(d)). The N^1 for *dolphins in the sea* has a higher intensity value (*i.e.* more frequently occurred) in summer, whereas the N^2 for the *ice hockey team* peaks around January. This observation is reasonable because sea tours are popular in summer and the ice hockey season takes place during winter.

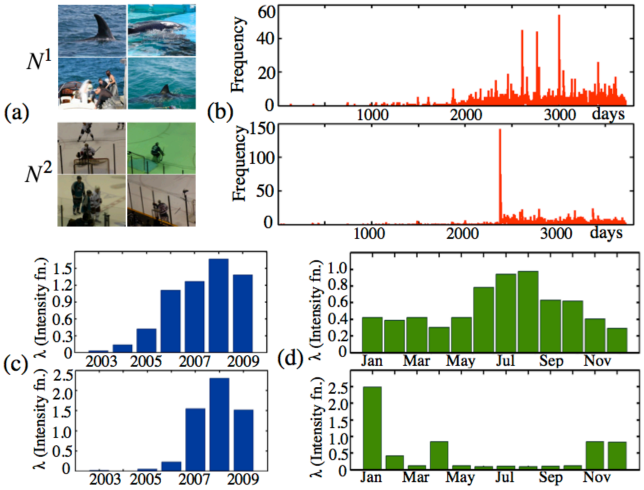


Figure 3: An example of the Poisson regression model for two visual clusters of the *Shark* topic: {*dolphins in the sea*, *ice hockey team*}. (a) Four images sampled from two visual clusters. (b) Observed occurrence data. (c)-(d) The estimated intensity functions for years and months. The N^1 (top) and N^2 (bottom) have different intensify functions peaked in summer and winter, respectively.

The learned intensity functions can be used for a simple image prediction. For example, if the month of query time t_q is January, then $\lambda^2(t_q)(\approx 2.5) > \lambda^1(t_q)(\approx 0.4)$, and we can sample the images from N^2 six times more than from N^1.

3.3 A Composite Model of Intensity Functions

Now we introduce the full model of the intensity function λ^i that can be used in image prediction. Note that any probable factors can be flexibly included into the model without any performance loss, because our objective function in Eq.(5) encourages a sparse solution in which the weights of irrelevant covariates are zeros.

We assume that the occurrence of each visual cluster is affected by three types of covariates: (i) its own history, (ii) behaviors of other visual clusters, and (iii) external covariates. It leads to the following composite intensity function:

$$\lambda^i(t_k|\boldsymbol{\theta}^i) = \lambda_h^i(t_k|\boldsymbol{\theta}_h^i)\lambda_e^i(t_k|\boldsymbol{\theta}_e^i)\lambda_x^i(t_k|\boldsymbol{\theta}_x^i),\ i \in \{1, ..., M\} \quad (7)$$

where the λ_h^i, λ_e^i, and λ_x^i are called the components of intensity functions for history, correlation, and external covariates, respectively. They are described in Eq.(8)-(10) with an example of Fig.4. For brevity, we omit the superscript i in the following equations.

The first history component is modeled as a linear autore-

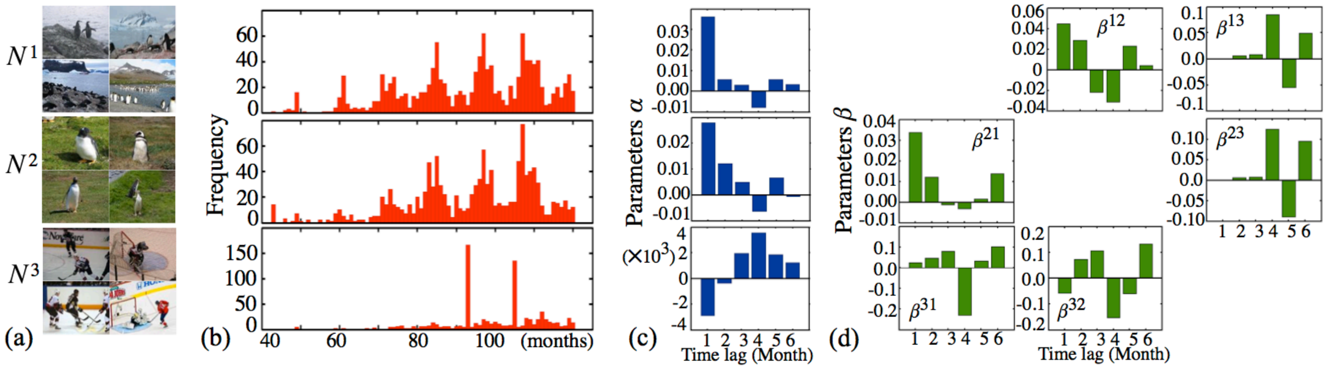


Figure 4: Examples of the *Penguin* topic for the learned parameters of history and correlation components. Visual clusters are {*penguins in landscape*, *penguins on grass*, *ice hockey team*}. (a) Four images sampled from each visual cluster. (b) Observed occurrence data. The N^1 and N^2 are strongly synchronized and periodically peaked in summer, whereas the N^3 has two high peaks in winter. (c)-(d) Learned parameters for history and correlation components, respectively.

gressive (AR) process of order P with $\boldsymbol{\theta}_h = \{\alpha_0, \cdots, \alpha_P\}$:

$$\log \lambda_h(t_k|\boldsymbol{\theta}_h) = \alpha_0 + \sum_{p=1}^{P} \alpha_p \Delta N\big(t_k - pd, t_k - (p-1)d\big). \quad (8)$$

$\Delta N(t_1, t_2)$ denotes the number of images during $[t_1, t_2)$, and d is the width of the time window (*e.g.* if Δ= 1 day and d=7, then $\Delta N(t_k - d, t_k)$ is the number of images occurred during previous one week from t). The history component reflects the dynamic behavior of a visual cluster. As shown in Fig.4.(c), the learned parameters of N^1 (top) and N^2 (middle) show the typical patterns for yearly periodic behaviors, whereas the parameters of N^3 (bottom) are biphasic, which indicates a bursty occurrence.

The second correlation component models the influence from the history of other visual clusters:

$$\log \lambda_e(t_k|\boldsymbol{\theta}_e) = \beta_0 + \sum_{\substack{c=1\\ c\neq i}}^{M} \sum_{q=1}^{R} \beta_j^{ic} \Delta N^c\big(t_k - qd, t_k - (q-1)d\big) \quad (9)$$

where the parameter $\boldsymbol{\theta}_e$ consists of $(M-1) \times R+1$ parameters of β in the full model. This correlation component is quite useful for the actual prediction in the Flickr dataset; we observe that there are strong correlations between visual clusters, and thus the existence or absence of a particular visual cluster gives a strong clue for others' prediction. The learned parameters β in Fig.4.(d) clearly present the correlations observed in Fig.4.(b). For example, the subfigures of β^{12} and β^{21} in Fig.4.(d) show that the occurrence of N^1 and N^2 are highly synchronized, whereas the subfigures of β^{13} and β^{23} illustrate the occurrence of N^3 precedes those of N^1 and N^2 by four months. For fast computation, instead of using the full pairwise model, we learn the correlations of each N^i with top K most frequent visual clusters.

The extrinsic component incorporates any types of factors that are likely to influence the image occurrence. In this paper, we use months and user descriptors as covariates:

$$\log \lambda_x(t_k|\boldsymbol{\theta}_x) = \gamma_0 + \sum_{m=1}^{12} \gamma_m g(t_k - m) + \sum_{z=1}^{Z} \gamma_z \mathbf{u}_{t_k\text{-}d:t_k}(z). \quad (10)$$

We use $g(t_k - m) \propto \exp(-\alpha(t_k - m)^2)$ for month covariates. The idea is that if an image occurs in June, some contributions are also given on nearby months like May and July, assuming that images are smoothly changed as time goes. The user covariate is the average of user preferences for the images in $[t_k - d, t_k)$. The $\mathbf{u}_{t\text{-}d:t}(z)$ is the mean of z-th elements of user descriptors for the images in $[t_k - d, t_k)$.

In this paper, we introduce only three types of covariates for modeling of image occurrences, but one can freely add or remove functions of covariates according to the characteristics of image topics to be predicted unless they contradict the definition of Eq.(4). For example, other textual or social factors may be supplemented as covariates or AR functions can be replaced by a more general linear temporal model such as ARMA (AutoRegressive Moving-Average) model.

3.4 Learning and Prediction

The learning corresponds to obtain MLE solution $\boldsymbol{\theta}^{i*}$ of Eq.(5) from the observe image sequences $N^i_{i:K}$ by solving

$$\max_{\boldsymbol{\theta}} \left(\sum_{k=1}^{K} \log \lambda^i(t_k|\boldsymbol{\theta}) \Delta N^i_k - \sum_{k=1}^{K} \lambda^i(t_k|\boldsymbol{\theta}) - \mu \sum_{j=1}^{n} |\theta^i_j| \right) \quad (11)$$

where $\lambda^i(t_k|\boldsymbol{\theta})$ has the form of Eq.(7). We use the cyclical coordinate descent in [10].

In the prediction step, given a query time t_q, we first obtain the set of $\lambda^i(t_q|\boldsymbol{\theta}^{i*})$ for $i = \{1, \ldots, M\}$, which indicates the occurrence rates of each visual cluster for t_q. It is computed by gathering covariate values at t_q, and plugging them with $\boldsymbol{\theta}^*$ into the Eq.(7). The final output is L number of most likely images for t_q, which is sampled according to $\lambda^i(t_q|\theta^*)$. The images of visual clusters with higher $\lambda^i(t_q|\theta^*)$ are more likely to be chosen for t_q. We use the thinning algorithm [19], which is a rejection sampling to simulate new samples from intensity functions.

Our parametric model is scalable; The learning time is $O(M|T|J)$ where $|T|$ is the number of time steps (*e.g.* discretized by day) and J is the number of covariates. Our code written in Matlab takes about 30 minutes to learn the model for the 810K of *soccer* images with $M = 200$, $|T| = 1,500$, and $J = 118$. The complexity of prediction time is $O(MJ)$. In the same experiment, it takes far less than one second.

3.5 Personalization

Given a query user u_q, the idea of personalization is to weight more the history of pictures taken by u_q or similar users to u_q during learning. For collective prediction, one

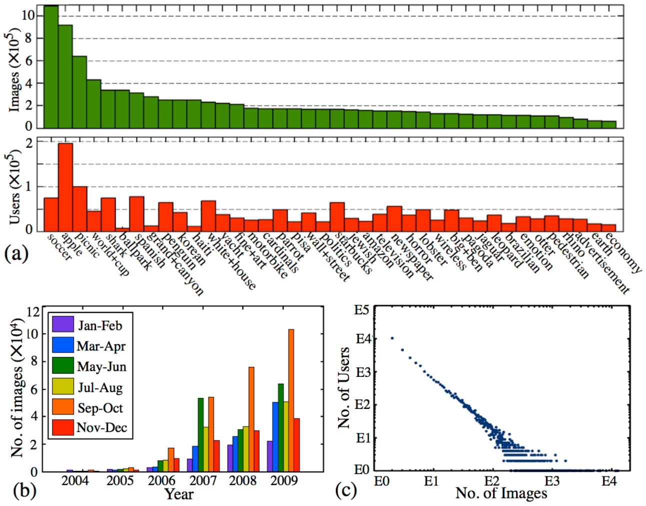


Figure 5: Flickr datasets. (a) The numbers of images and users for the 40 topics. (b) The seasonal distribution and (c) the log-log plot between the number of images and users for the *soccer* topic.

image occurrence is counted by equally one (*e.g.* the occurrence data in Fig.3.(b) simply count the number of images uploaded in each day). On the other hand, for personalized prediction, an image by u_q is weighted by a larger value so that the model fitting is more biased to the images of u_q. Likewise, the weight of an image occurrence can be assigned according to the similarity between its owner and u_q.

We implemented this idea by using the locally weighted learning framework [2], which is a form of lazy learning for a regression to adjust the weighting of data samples according to a query. More specifically, the weights of image data of any user u_x are assigned by $w_x{=}\sqrt{K(d(\mathbf{u}_q, \mathbf{u}_x))}$ where $\mathbf{u}_q$ is the user descriptor of a query user u_q, d is the distance function $d(\mathbf{u}_q, \mathbf{u}_x){=}(\mathbf{u}_q{-}\mathbf{u}_x)^2$, and K is the Gaussian kernel function $K(d) = \exp(-d^2/\sigma)$.

This approach is lazy learning in which the training is deferred until a query is available. If the number of users to be considered is very large, we can perform a user clustering method in collaborative filtering [7], and learn the prediction models offline for each group of users.

4. RESULTS

We evaluate the performance of our algorithm for collective and personalized image prediction using Flickr datasets. A simplified MATLAB demo code is available at our webpage `http://www.cs.cmu.edu/∼gunhee`.

4.1 Evaluation Setting

Datasets: Our dataset consists of 10,284,945 images of 40 topics from Flickr. Some topics are re-used from the datasets of [15] and others are newly downloaded. Both datasets are collected by the same protocol, in which the topic name is used as a search word and all queried images are downloaded without any filtering. For the timestamp, the *date_taken* field is used. Fig.5 summarizes some statistics of our Flickr dataset. Fig.5.(a) shows the numbers of images and users of all 40 topics, which are roughly classified into {*nations*, *places*, *animals*, *objects*, *activities*, *abstract*, *hot topics*}. Fig.5.(b) shows a seasonal variation in the *soccer* topic; the image uploading peaks in autumn but falls in winter. Fig.5.(c) is a log-log plot between the number of images (x-axis) and the frequencies of users (y-axis). The number of images per user follows Zipf's law in almost all topics. That is, a few users contribute the majority of images, and most users have only a small number of images.

Tasks: We first divide all image sets into training and test sets by time; training sets $\mathcal{I}_T$ consist of the images taken up to 12/31/2008 and test sets are the others. In the following experiments, $\mathcal{I}_T$ is used as the image database for retrieval and training data to learn the image occurrence patterns.

The collective image prediction is performed as follows; a topic name and a future time point t_q are given in a form of (M/D/Y) (*i.e.* t_q is a time point in 2009 or 2010). The images that are actually taken in [t_q±1 days] are the positive test set $\mathcal{I}_+$ to be estimated. The goal is to select L images $\mathcal{I}_e$ from $\mathcal{I}_T$ so that $\mathcal{I}_e$ and $\mathcal{I}_+$ are as similar as possible to each other. We set L to 200 for all our experiments, and the numbers of actual images at t_q ($|\mathcal{I}_+|$) are hundreds or thousands ($|\mathcal{I}_+| > L$ in almost all cases).

The personalized image prediction is tested similarly except that both a future time t_q and a query user u_q are specified at test time. The goal of the algorithm is to predict L likely images $\mathcal{I}_e$ for the user u_q at t_q. The actual images taken by u_q at t_q are the positive test set $\mathcal{I}_+$.

For each topic, we randomly generate 20 t_q values and 20 (t_q, u_q) pairs as test cases of collective and personalized image forecast, respectively. A user is considered as u_q if she has a sufficiently large number of images in both training and test sets (*i.e.* at least 500 images in the training set and at least 100 images in the test set).

Baselines and Competitors: Since the Web image prediction is relatively novel, there are few existing methods to be compared. Hence, we come up with three alternatives for image prediction, and quantitatively compare them with our algorithm. Table 1 summarizes the baselines.

The (`SemIN`) [9] represents the prediction based on semantic meaning only. It is compared to show that the semantic meaning of a topic word is not enough to predict the user images on the Web. The (`RetPR`) [15] and (`TopAT`) [23][4] are the state-of-the-art methods for PageRank-based image retrieval and topic modeling for collaborative filtering.

In the personalized forecast, the locally weighted learning is also applied to all the other competitors except `SemIN`, which is a random sampling from the ImageNet dataset.

Evaluation Measures: We evaluated the performance of all algorithms by measuring the similarity between the estimated images $\mathcal{I}_e$ and actual images $\mathcal{I}_+$ at t_q. Due to lack of a perfect measure for image similarity, we calculate three popular metrics in image retrieval according to information levels: `L2`, `Tiny`, and average precision (`AP`), as shown in Table 2. `L2` is the most low-level metric by feature-wise comparison, and `AP` is the most high-level one based on classification ability. No single measure may be perfect, but one algorithm can be fairly said *better* if it constantly outperforms others in all three metrics. For `L2` and `Tiny`, we first find the one-to-one correspondences between estimated $\mathcal{I}_e$ and actual $\mathcal{I}_+$, and then calculate average distances. Since `L2` and `Tiny` are distance measures, a lower value means a better prediction, whereas in the `AP`, the higher is the better.

[4] We modified the toolbox at http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm.

Methods	Rationale	Description
Sampling from ImageNet (`SemIN`) [9]	Using semantic meaning only	ImageNet data are representative images of each topic cleaned by human annotation. We randomly sample images as predicted images for t_q.
PageRank-based prediction (`RetPR`) [15]	State-of-the-art retrieval algorithm	We first gather the images taken around the similar month to t_q from $\mathcal{I}_T$, and sample highly ranked images by using PageRank [3].
Author-Time Topic Model (`TopAT`) [23]	State-of-the-art topic modeling as collaborative filtering	We modify Author-Topic model [23] to jointly model image contents, users, and month data. We first estimate the subtopic distribution to know what images are popular in each month. We sample the images according to the subtopic distributions at the month of t_q.

Table 1: Summary of the three baselines used for quantitative comparison with our image prediction algorithm.

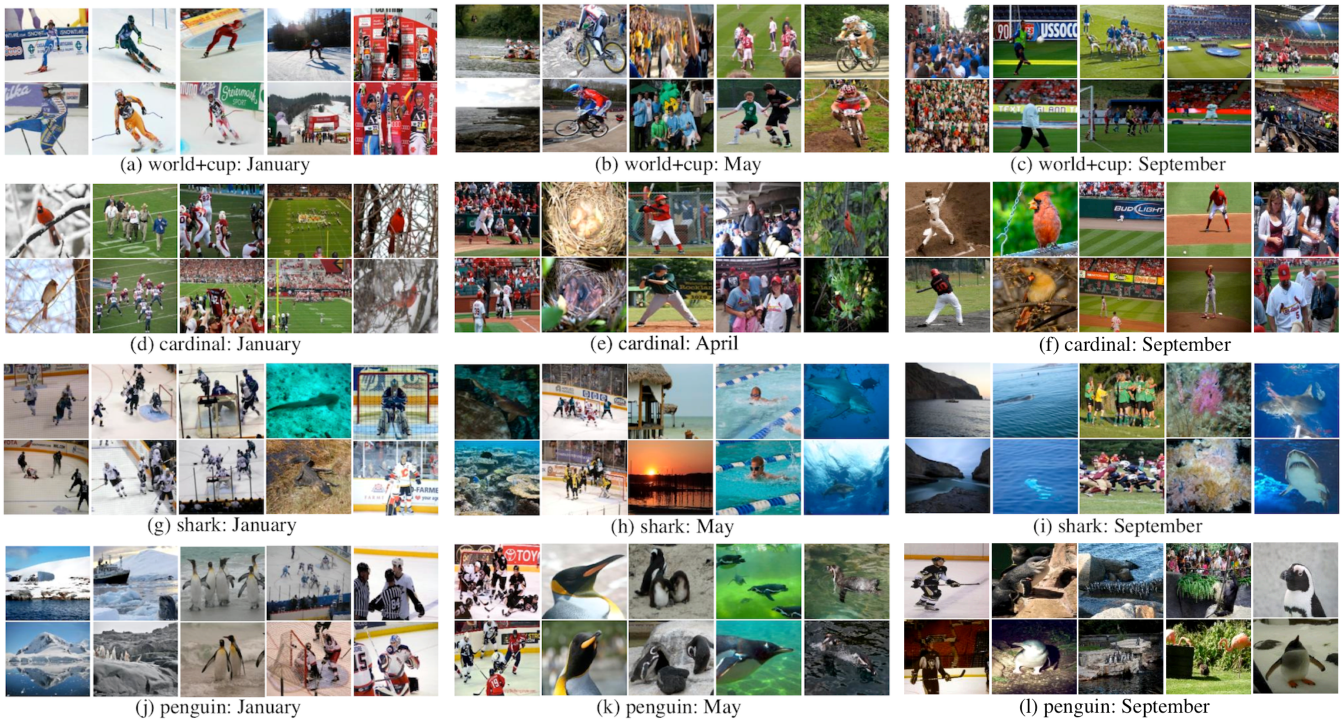

(a) world+cup: January
(b) world+cup: May
(c) world+cup: September
(d) cardinal: January
(e) cardinal: April
(f) cardinal: September
(g) shark: January
(h) shark: May
(i) shark: September
(j) penguin: January
(k) penguin: May
(l) penguin: September

Figure 7: Examples of collective image prediction for the topics of *world+cup* ((a)-(c)), *cardinal* ((d)-(f)), *shark* ((g)-(i)), and *penguin* ((j)-(l)) in some selected months. In all sets, we first find one-to-one correspondences between the estimated images $\mathcal{I}_e$ and the actual images $\mathcal{I}_+$ by the `L2` measure, and then sample five image pairs per month. The first row shows the estimated images by our method, and the second row depicts their matched actual images.

Metric	Description
`L2`	L2 distance between image descriptors (*i.e.* Spatial pyramids of denseSIFT and HOG features).
`tiny`	Inspired by [26], we first resize the images to 32×32 tiny color images, and compute SSD (the sum of squared differences between pixels of images).
`AP`	Let $\mathcal{I}_+$ be actual images at t_q (*i.e.* positive test data) and $\mathcal{I}_-$ be negative data by randomly selecting the same number images outside of $[t_q \pm 3$ months$]$. Each algorithm ranks top L images out of $(\mathcal{I}_+ \cup \mathcal{I}_-)$, from which average precisions are computed.

Table 2: Summary of three evaluation metrics. For `L2` and `Tiny`, we first find the one-to-one correspondences between estimated $\mathcal{I}_e$ and actual $\mathcal{I}_+$, and then calculate average distances.

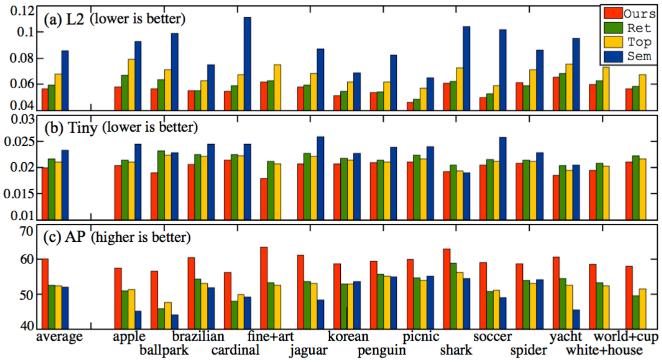


Figure 6: Quantitative comparison between our method and three baselines (`RetPG`[15], `TopAT`[23], `SemIN`[9]) for collective image prediction using three metrics in (a)-(c).

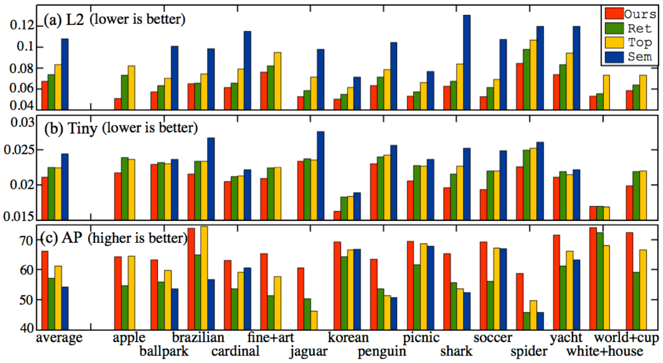


Figure 8: Quantitative comparison between our method and three baselines (`RetPG`[15], `TopAT`[23], `SemIN`[9]) for personalized image prediction using three metrics in (a)-(c).

4.2 Results of Collective Image Prediction

Fig.6 shows the quantitative comparison between our method and three baselines. In each figure, the leftmost bar set is the average performance of 40 topics, and the results of ten sampled topics follow. Our algorithm significantly outperformed all the competitors in most measures. In the average performance, the `L2`(`tiny`) measure of our method is smaller by 5.1(8.8)%, 17.1(6.5)%, and 34.2(15.5)% over the `RetPR`, `TopAT`, and `SemIN`, respectively. Our `AP` is also higher than the best of baselines by 8 %. Among the baselines, the `RetPR` was the best, and the `SemIN` was the worst.

Fig.7 shows several examples of collective prediction for the topics of *world+cup*, *cardinals*, *shark*, and the *penguin* in several months. Each figure is obtained as follows. We first find the one-to-one correspondences between the estimated $\mathcal{I}_e$ and the actual $\mathcal{I}_+$ by the `L2` measure. Out of L matched pairs (L=200), we only sample five matches and show the predicted images by our algorithm in the first row, and the matched actual images in the second row. The five samples may be too small compared to 200 matched pairs, but here our goal is qualitative analysis. We already show the quantitative superiority of our algorithm over three baselines in Fig.6.

The term *world+cup* is commonly used in different sports and competitions (*e.g.* soccer, ski, skating, cycling, horse riding, and even cocktail competitions), and the popular sports in the images are changed according to the query times. The topic *cardinal* is also used in different meanings, including a bird, a baseball team, and an American football team. The football images are frequent from fall to winter, whereas the baseball images are dominant from spring to fall. They agree with the scheduled seasons of corresponding sports. The bird images are also varied according to the query times. The popular backgrounds of bird images are snowy fields in winter and leafy trees in summer. The images about eggs and baby cardinals also appear in summer. Similar observations can be made as well in the *shark* and *penguin* topics, as shown in Fig.7.(g)-(l).

This observation concludes that indeed the Web image collections are extremely diverse, but they follow some patterns that can be predictable. Specifically, our predictive model works well for polysemous topics that show strong annual or periodic trends, and is promisingly applicable to image suggestion or re-ranking.

4.3 Results of Personalized Image Prediction

Fig.8 shows the quantitative results of personalized image prediction. In the average performance, our method is far better than all the baselines. The personalized prediction is more accurate than the collective forecast, because knowing the user at query time provides a strong clue to predict the images.

Fig.9 delivers a clear evidence for the importance of personalization in image prediction tasks. Even with the same keyword, users show various preferences. As shown in Fig.9.(a)-(c), the meanings of the term *fine+art* are differently recognized according to the users, such as paintings, classes, and photography. Other examples in Fig.9, including *Brazilian*, *apple*, and *picnic*, also show that this personal variation in user photo sets is quite common.

This observation presents that our method can be employed in the image search where a query word has a board concept, which can be varied according to people's thoughts and interests. This personalized search has been widely studied in textual information retrieval, but our analysis also reveals that images can convey more delicate information about user preferences that are hardly captured by text descriptions (*e.g.* what paintings does the user like? How does the presentation look like?).

5. CONCLUSION

We studied the collective and personalized image prediction tasks, as a time and user sensitive Web image re-ranking using large-scale Flickr images. The multivariate point process model was successfully tailored to achieve the flexibility, optimality, scalability, and prediction accuracies. As a promising direction of future research, it is interesting to incorporate other meta data surrounding Flickr photos such as comments or favs for better forecast.

6. ACKNOWLEDGMENTS

This work is supported by Google, ONR N000140910758, AFOSR FA9550010247, NSF IIS-0713379, and NSF DBI-0640543. L.F-F is supported by grant from NSF-IIS-1115313.

7. REFERENCES

[1] A. Ahmed, Q. Ho, C. H. Teo, J. Eisenstein, A. J. Smola, and E. P. Xing. Online Inference for the Infinite Topic-Cluster Model: Storylines from Streaming Text. In *AISTAT*, 2011.

[2] C. G. Atkeson, A. W. Moore, and S. Schaal. Locally Weighted Learning. *AI Review*, 11(1):11–73, 1997.

[3] S. Brin and L. Page. The Anatomy of a Large-scale Hypertextual Web Search Engine. In *WWW*, 1998.

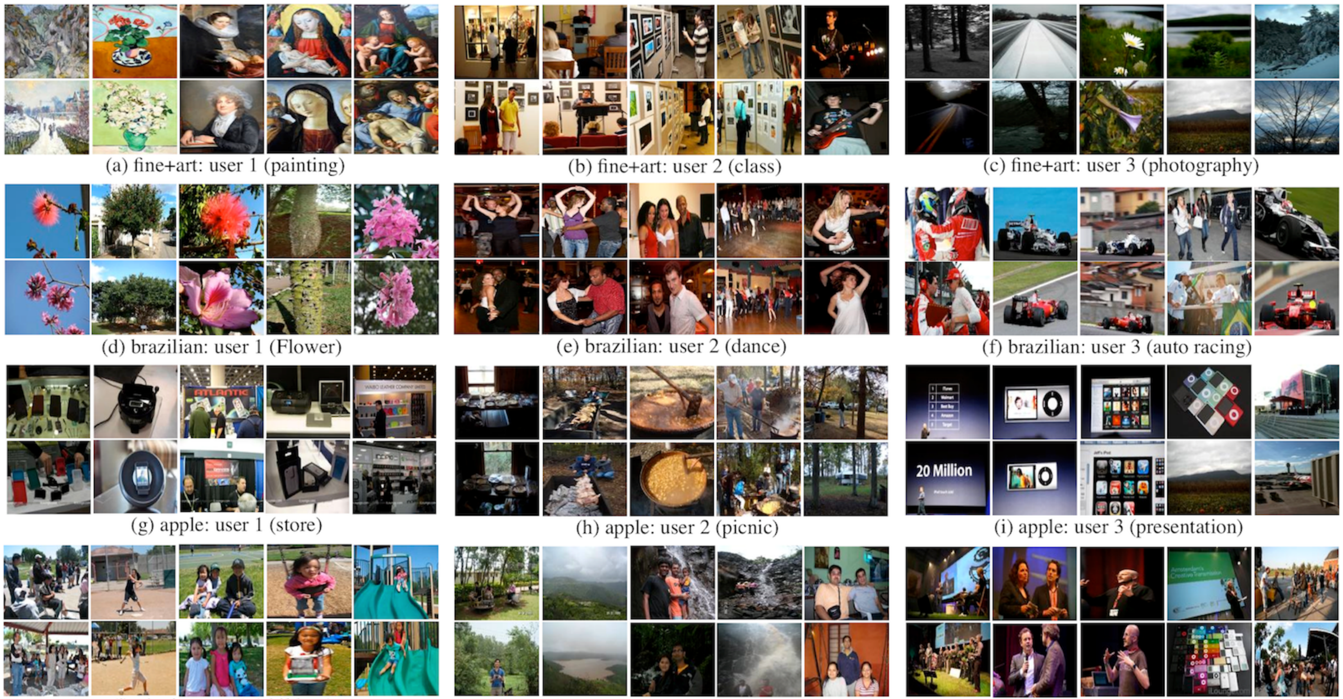
[4] J. Cui, F. Wen, and X. Tang. Real Time Google and Live Image Search Re-ranking. In *MM*, 2008.

[5] W. Dakka, L. Gravano, and P. G. Ipeirotis. Answering General Time-Sensitive Queries. In *CIKM*, 2008.

[6] D. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes.* Springer, 2003.

[7] A. S. Das, M. Datar, , A. Garg, and S. Rajaram. Google News Personalization: Scalable Online Collaborative Filtering. In *WWW*, 2007.

[8] J. Deng, A. C. Berg, and L. Fei-Fei. Hierarchical Semantic Indexing for Large Scale Image Retrieval. In *CVPR*, 2011.



(a) fine+art: user 1 (painting)
(b) fine+art: user 2 (class)
(c) fine+art: user 3 (photography)
(d) brazilian: user 1 (Flower)
(e) brazilian: user 2 (dance)
(f) brazilian: user 3 (auto racing)
(g) apple: user 1 (store)
(h) apple: user 2 (picnic)
(i) apple: user 3 (presentation)
(j) picnic: user 1 (children)
(k) picnic: user 2 (mountain)
(l) picnic: user 3 (presentation)

Figure 9: Examples of personalized image prediction for the topics of *fine+art* ((a)-(c)), *Brazilian* ((d)-(f)), *apple* ((g)-(i)) and *picnic* ((j)-(l)) for some selected users. Similarly to Fig.7, in all sets, the first row shows the predicted images, and the second row is matched actual images. The themes of images are largely varied according to users.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[10] J. Friedman, T. Hastie, and R. Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Statistical Software*, 33:1–22, 2010.

[11] R. Garg, D. Ramanan, S. Seitz, and N. Snavely. Where's Waldo: Matching People in Images of Crowds. In *CVPR*, 2011.

[12] J. Hays and A. A. Efros. Scene Completion Using Millions of Photographs. *SIGGRAPH*, 26(3), 2007.

[13] X. Jin, A. Gallagher, L. Cao, J. Luo, and J. Han. The Wisdom of Social Multimedia: Using Flickr for Prediction and Forecast. In *ACM MM*, 2010.

[14] E. Kalogerakis, O. Vesselova, J. Hays, A. A. Efros, and A. Hertzmann. Image Sequence Geolocation with Human Travel Priors. In *ICCV*, 2009.

[15] G. Kim, E. P. Xing, and A. Torralba. Modeling and Analysis of Dynamic Behaviors of Web Image Collections. In *ECCV*, 2010.

[16] A. Kulkarni, J. Teevan, K. M. Svore, and S. T. Dumais. Understanding Temporal Query Dynamics. In *WSDM*, 2011.

[17] L.-J. Li, C. Wang, Y. Lim, D. Blei, and L. Fei-Fei. Building and Using a Semantivisual Image Hierarchy. In *CVPR*, 2010.

[18] D. Metzler, R. Jones, F. Peng, and R. Zhang. Understanding Temporal Query Dynamics. In *SIGIR*, 2009.

[19] Y. Ogata. On Lewis' Simulation Method for Point Processes. *IEEE T. Information Theory*, 27(1):23–31, 1981.

[20] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases. In *CVPR*, 2008.

[21] K. Prabhakar, S. Oh, P. Wang, G. Abowd, and J. M. Rehg. Temporal Causality and the Analysis of Interactions in Video. In *CVPR*, 2010.

[22] K. Radinsky, K. Svore, S. Dumais, J. Teevan, A. Bocharov, and E. Horvitz. Modeling and Predicting Behavioral Dynamics on the Web. In *WWW*, 2012.

[23] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The Author-Topic Model for Authors and Documents. In *UAI*, 2004.

[24] V. K. Singh, M. Gao, and R. Jain. Social Pixels: Genesis and Evaluation. In *ACM MM*, 2010.

[25] N. Snavely, I. Simon, M. Goesele, R. Szeliski, and S. M. Seitz. Scene Reconstruction and Visualization From Community Photo Collections. *Proceedings of the IEEE*, 98(8), 2010.

[26] A. Torralba and W. T. F. Rob Fergus. 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. *IEEE PAMI*, 30:1958–1970, 2008.

[27] W. Truccolo, U. T. Eden, M. R. Fellows, J. P. Donoghue, and E. N. Brown. A Point Process Framework for Relating Neural Spiking Activity to Spiking History, Neural Ensemble, and Extrinsic Covariate Effects. *J. Neurophysiol*, 93(2):1074–1089, 2005.

[28] X. Wang and A. McCallum. Topics Over Time: a Non-Markov Continuous-Time Model of Topical Trends. In *KDD*, 2006.