# Modeling and Analysis of Dynamic Behaviors of Web Image Collections

Gunhee Kim[1], Eric P. Xing[1], and Antonio Torralba[2]

[1] Carnegie Mellon University, Pittsburgh, PA 15213, USA
[2] Massachusetts Institute of Technology, Cambridge, MA 02139, USA
{gunhee,epxing}@cs.cmu.edu, torralba@csail.mit.edu

**Abstract.** *Can we model the temporal evolution of topics in Web image collections? If so, can we exploit the understanding of dynamics to solve novel visual problems or improve recognition performance?* These two challenging questions are the motivation for this work. We propose a nonparametric approach to modeling and analysis of topical evolution in image sets. A scalable and parallelizable sequential Monte Carlo based method is developed to construct the similarity network of a large-scale dataset that provides a base representation for wide ranges of dynamics analysis. In this paper, we provide several experimental results to support the usefulness of image dynamics with the datasets of 47 topics gathered from Flickr. First, we produce some interesting observations such as tracking of subtopic evolution and outbreak detection, which cannot be achieved with conventional image sets. Second, we also present the complementary benefits that the images can introduce over the associated text analysis. Finally, we show that the training using the *temporal association* significantly improves the recognition performance.

## 1 Introduction

This paper investigates the discovery and use of topical evolution in Web image collections. The images on the Web are rapidly growing, and it is obvious to assume that their topical patterns evolve over time. Topics may rise and fall in their popularity; sometimes they are split or merged to a new one; some of them are synchronized or mutually exclusive on the timeline. In Fig.1, we download *apple* images and their associated timestamps from Flickr, and measure the similarity changes with some canonical images of *apple*'s subtopics. As *Google trends* reveal the popularity variation of query terms in the search volumes, we can easily observe the affinity changes of each subtopic in the *apple* image set.

The main objectives of this work are as follows. First, we propose a nonparametric approach to modeling and analysis of temporal evolution of topics in Web image collections. Second, we show that understanding image dynamics is useful to solve novel problems such as subtopic outbreak detection and to improve classification performance using the *temporal association* that is inspired by studies in human vision [2,19,21]. Third, we present that the images can be a

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part V, LNCS 6315, pp. 85–98, 2010.

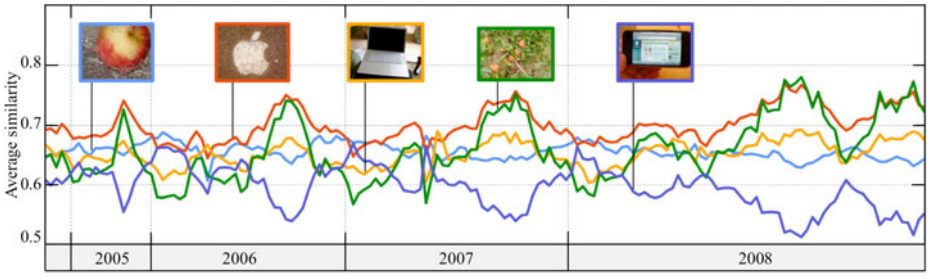


**Fig. 1.** The *Google trends*-like visualization of the subtopic evolution in the *apple* images from Flickr (*fruit*: blue, *logo*: red, *laptop*: orange, *tree*: green, *iphone*: purple). We choose the cluster center image of each subtopic, and measure the average similarity with the posterior (*i.e.* a set of weighted image samples) at each time step. The *fruit* subtopic is stable along the timeline whereas the *iphone* subtopic is highly fluctuated.

more reliable and delicate source of information to detect topical evolution than the texts.

Our approach is motivated by the recent success of the nonparametric methods [13,20] that are powered by large databases. Instead of using sophisticated parametric topic models [3,22], we represent the images with timestamps in the form of a *similarity network* [11], in which vertices are images and edges connect the temporally related and visually similar images. Thus, our approach is able to perform diverse dynamics analysis without solving complex inference problems. For example, a simple information-theoretic measure of the network can be used to detect subtopic outbreaks, which point out when the evolution speed is abruptly changed. The *temporal context* is also easily integrated with the classifier training in a framework of the Metropolis-Hastings algorithm.

The network generation is based on the sequential Monte Carlo (*i.e.* particle filtering) [1,9]. In the sequential Monte Carlo, the posterior (*i.e.* subtopic distribution) at a particular time step is represented by a set of weighted image samples. We track similar subtopics (*i.e.* clusters of images) in consecutive posteriors along the timeline, and create edges between them. The sampling based representation is quite powerful in our context. Since we deal with unordered natural images on the Web, any Gaussian or linearity assumption does not hold and multiple peaks of distributions are unavoidable. Another practical advantage is that we can easily control the tradeoff between accuracy and speed by managing the number of samples and parameters in the transition model. The proposed algorithm is easily parallelizable by running multiple sequential Monte Carlo trackers with different initialization and parameters. Our approach is also scalable and fast. The computation time is linear with the number of images.

For evaluation, we download more than 9M images of 47 topics from Flickr. Most standard datasets in computer vision research [7,18] have not yet considered the importance of temporal context. Recently, several datasets have introduced *spatial contexts* as fundamental cues to recognition [18], but the support for temporal context has still been largely ignored. Our experiments clearly show

that our modeling and analysis is practically useful and can be used to understand and simulate human-like visual experience from Web images.

### 1.1 Related Work

The temporal information is one of the most obvious features in video or auditory applications. Hence, here we review only the use of temporal cues for image analysis. The importance of temporal context has long been recognized in neuroscience research [2,19,21]. Wide range of research has supported that the *temporal association* (*i.e.* liking temporally close images) is an important mechanism to recognize objects and generalize visual representation. [21] tested several interesting experiments to show that temporally correlated multiple views can be easily linked to a single representation. [2] proposed a learning model for 3D object recognition by using the temporal continuity in image sequences.

In computer vision, [16] is one of the early studies that use temporal context in active object recognition. They used a POMDP framework for the modeling of temporal context to disambiguate the object hypotheses. [5] proposed a HMM-based temporal context model to solve scene classification problems. For the indoor-outdoor classification and the sunset detection, they showed that the temporal model outperformed the baseline content-based classifiers.

As the Internet vision emerges as an active research area in computer vision, timing information starts to be used in the assistance of visual tasks. Surprisingly, however, the dynamics or temporal context for Web images has not yet been studied a great deal, contrary to the fact that the study of the dynamic behaviors of the texts on the Web has been one of active research areas in data mining and machine learning communities [3,22]. We briefly review some notable examples using timestamp meta-data for visual tasks. [6] developed an annotation method for personal photo collections, and the timestamps associated with the images were used for better correlation discovery between the images. [12] proposed a landmark classification for an extremely large dataset, and the temporal information was used for the constraints to remove misclassification. [17] also used the timestamp as an additional feature to develop an object and event retrieval system for online image communities. [10] presented a method to geolocate a sequence of images taken by a single individual. Temporal constraints from the sequence of images were used as a strong prior to improve the geolocation accuracy.

The main difference between their work and ours is that they considered the temporal information as additional meta-data or constraints to achieve their original goals (*i.e.* annotations in [6], classification and detection in [12,17], and the geolocation of images in [10]). However, our work considers the timestamps associated with images as a main research subject to uncover dynamic behaviors of Web images. To our best knowledge, there have been very few previous attempts to tackle this issue in computer vision research.

## 2 Network Construction by Sequential Monte Carlo

### 2.1 Image Description and Similarity Measure

Each image is represented by two types of descriptors, which are spatial pyramids of visual words [14] and HOG [4]. We use the codes provided by the authors of the papers. A dictionary of 200 visual words is formed by K-means to randomly selected SIFT descriptors [14]. A visual word is densely assigned to every pixel of an image by finding the nearest cluster center in the dictionary. Then visual words are binned using a two-level spatial pyramid. The oriented gradients are computed by Canny edge detection and Sobel mask [4]. The HOG descriptor is then discretized into 20 orientation bins in the range of [0°,180°]. Then the HOG descriptors are binned using a three-level spatial pyramid. The similarity measure between a pair of images is the cosine similarity, which is calculated by the dot product of a pair of $L_2$ normalized descriptors.

### 2.2 Problem Statement

The input of our algorithm is a set of images $\mathcal{I} = \{I_1, I_2, ..., I_N\}$ and associated tags of taken time $\mathcal{T} = \{T_1, T_2, ..., T_N\}$. The main goal is to generate an $N \times N$ sparse similarity network $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ by using the Sequential Monte Carlo (SMC) method. Each vertex in $\mathcal{V}$ is an image in the dataset. The edge set $\mathcal{E}$ is created between the images that are visually similar and temporally distant with a certain interval that is assigned by the *transition model* of the SMC tracker (Section 2.3). The weight set $\mathcal{W}$ is discovered by the similarity between descriptors of images (Section 2.1). For sparsity, each image is connected to its $k$-nearest neighbors with $k = a \log N$, where $a$ is a constant (*e.g.* $a$ =10).

### 2.3 Network Construction Using Sequential Monte Carlo

Algorithm 1 summarizes the proposed SMC based network construction. For better readability, we follow the notation of *condensation* algorithm [9]. The output of each iteration of the SMC is the conditional subtopic distribution (*i.e.* posterior) at every step, which is approximated by a set of images with relative importance denoted by $\{\boldsymbol{s}_t, \boldsymbol{\pi}_t\} = \{s_t^{(i)}, \pi_t^{(i)}, i = 1, \ldots, M\}$. Note that our SMC does not explicitly solve the *data association* during the tracking. In other words, we do not assign a subtopic membership to each image in $\boldsymbol{s}_t$. However, it can be easily obtained later by applying clustering to the subgraph of $\boldsymbol{s}_t$.

Fig.2 shows a downsampled example of a single iteration of the posterior estimation. At every iteration, the SMC generates a new posterior $\{\boldsymbol{s}_t, \boldsymbol{\pi}_t\}$ by running *transition*, *observation*, and *resampling*.

The image data are severely unbalanced on the timeline. (*e.g.* There are only a few images within a month in 2005 but a large number of images within even a week in 2008). Thus, in our experiments, we bin the timeline by the number of images instead of a fixed time interval. (*e.g.* The timeline may be binned by every 3000 images instead of a month). The function $\tau(T_i, m)$ is used to indicate the timestamp of the $m$-th image later from the image at $T_i$.

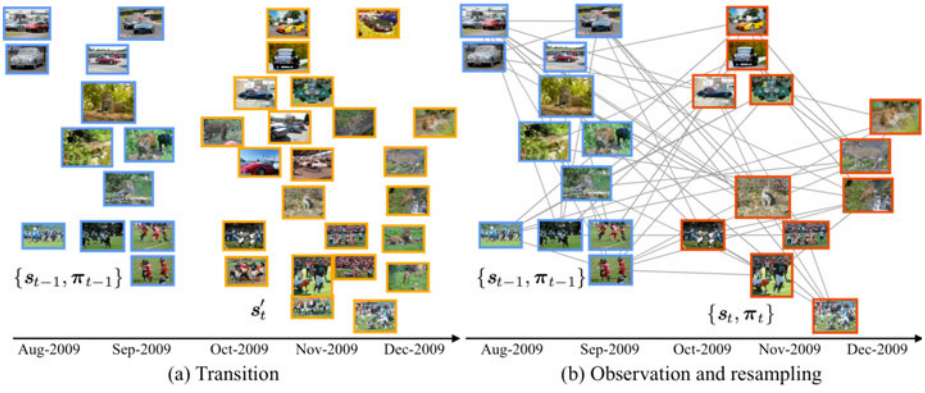


**Fig. 2.** An overview of the SMC based network construction for the *jaguar* topic. The subtopic distribution at each time step is represented by a set of weighted image samples (*i.e.* posterior) $\{\boldsymbol{s}_t, \boldsymbol{\pi}_t\}$. In this example, a posterior of the *jaguar* topic consists of image samples of *animal*, *cars*, and *football* subtopics. (a) The transition model generates new posterior candidates $\boldsymbol{s}'_t$ from $\boldsymbol{s}_{t-1}$. (b) The observation model discovers $\boldsymbol{\pi}'_t$ of $\boldsymbol{s}'_t$ and the resampling step computes $\{\boldsymbol{s}_t, \boldsymbol{\pi}_t\}$ from $\{\boldsymbol{s}'_t, \boldsymbol{\pi}'_t\}$. Finally, the network is constructed by similarity matching between two consecutive posteriors $\boldsymbol{s}_{t-1}$ and $\boldsymbol{s}_t$.

**Initialization.** The initialization samples the initial posterior $\boldsymbol{s}_0$ from the prior $p(\boldsymbol{x}_0)$ at $T_0$. $p(\boldsymbol{x}_0)$ is set by a Gaussian distribution $N(T_0, \tau^2(T_0, 2M/3))$ on the timeline, which means that $2M$ numbers of images around $T_0$ have nonzero probabilities to be selected as one of $\boldsymbol{s}_0$. The initial $\boldsymbol{\pi}_0$ is uniformly set to $1/M$.

**Transition Model.** The transition model generates posterior candidates $\boldsymbol{s}'_t$ rightward on the timeline from the previous $\{\boldsymbol{s}_{t-1}, \boldsymbol{\pi}_{t-1}\}$ (See Fig.2.(a) for an example). Each image $s^{(i)}_{t-1}$ in $\boldsymbol{s}_{t-1}$ recommends $m_i$ numbers of images that are similar to itself as candidates set $\boldsymbol{s}'_t$ for the next posterior. A more weighted image $s^{(i)}_{t-1}$ is able to recommend more images for $\boldsymbol{s}'_t$. ($\sum_i m_i = 2M$ and $m_i \propto \pi^{(i)}_{t-1}$). At this stage, we generate $2M$ candidates (*i.e.* $|\boldsymbol{s}'_t| = 2M$), and the observation and resampling steps reduce it to be $|\boldsymbol{s}_t| = M$ while computing weights $\boldsymbol{\pi}_t$.

Similarly to condensation algorithm [9], the transition consists of deterministic *drift* and stochastic *diffusion*. The *drift* describes the transition tendency of the overall $\boldsymbol{s}'_t$ (*i.e.* how far the $\boldsymbol{s}'_t$ is located from the $\boldsymbol{s}_{t-1}$ on the timeline). The *diffusion* assigns a random transition of an individual image. The *drift* and the *diffusion* are modeled by a Gaussian distribution $N(\mu_t, \sigma^2)$ and a Gamma distribution $\Gamma(\alpha, \beta)$, respectively. The final transition model is the product of these two distributions [8] in Eq.1. The asterisk of $P^{(i)*}_t(x)$ in Eq.1 means that it is not normalized. Renormalization is not required since we will use *importance sampling* to sample images on the timeline with the target distribution (See the next subsection with Fig.3 for the detail).

$$P^{(i)*}_t(x) = N(x; \mu_t, \sigma^2) \times \Gamma(x; \alpha^{(i)}_{t-1}, \beta^{(i)}_{t-1}) \tag{1}$$

**Algorithm 1.** The SMC based network generation

**Input:** (1) A set of images $\mathcal{I}$ sorted by timestamps $\mathcal{T}$. (2) Start time $T_0$ and end time $T_e$. (3) Posterior size $M$. (4) Parameters for *drift*: $(\triangle M_\mu, \sigma^2)$.
**Output:** Network $\boldsymbol{G}$
**Initialization**:
**1**: draw $s_0^{(i)} \sim N(T_0, \tau^2(T_0, 2M/3))$, $\pi_0^{(i)} = 1/M$ for $i = 1, \ldots, M$.
**while** $\mu_t < T_e$, ($\mu_0 = T_0$ and $\mu_t = \mu_{t-1} + \tau(\mu_{t-1}, \triangle M_\mu)$). **do**
[**Transition**]
**for all** $s_{t-1}^{(i)} \in \boldsymbol{s}_{t-1}$ with $\boldsymbol{x}^{(i)} = \emptyset$ **do**
**repeat**
**3**: draw $x \sim N(x; \mu_t, \sigma^2) \times \Gamma(x; \alpha_{t-1}^{(i)}, \beta_{t-1}^{(i)})$ ($\alpha_{t-1}^{(i)} \propto 1/\pi_{t-1}^{(i)}, \beta_{t-1}^{(i)} = \mu_t/\alpha_{t-1}^{(i)}$).
**4**: $\boldsymbol{x}^{(i)} \leftarrow x$ with probability of $w(s_{t-1}^{(i)}, x)$.
**until** $|\boldsymbol{x}^{(i)}| = m_i = 2M \times \pi_{t-1}^{(i)}$. Then, $\boldsymbol{s}'_t \leftarrow \boldsymbol{x}^{(i)}$.
**end for**
[**Observation**]
**4**: Compute self-similarity graph $\boldsymbol{W}_t$ of $\boldsymbol{s}'_t$. Row-normalize $\boldsymbol{W}_t$ to $\widetilde{\boldsymbol{W}}_t$.
**5**: Compute the stationary distribution $\boldsymbol{\pi}'_t$ by solving $\boldsymbol{\pi}'_t = \widetilde{\boldsymbol{W}}_t^T \boldsymbol{\pi}'_t$.
[**Resampling**]
**6**: Resample $\{\boldsymbol{s}_t, \boldsymbol{\pi}_t\}_{i=1}^M$ from $\{\boldsymbol{s}'_t, \boldsymbol{\pi}'_t\}$ by *systematic sampling* and normalize $\boldsymbol{\pi}_t$.
**7**: $\boldsymbol{G} \leftarrow \boldsymbol{W}_t(\boldsymbol{s}_t, \boldsymbol{s}_t)$, $\boldsymbol{W}_{t-1,t}(\boldsymbol{s}_{t-1}, \boldsymbol{s}_t)$, and then convert $\boldsymbol{G}$ into a $k$-NN graph.
**end while**

In sum, for each $s_{t-1}^{(i)}$, we sample an image $x$ using the distribution of Eq.1, which constrains the position of $x$ on the timeline. In addition, $x$ is required to be visually similar to its recommender. Thus, the sample $x$ is accepted with probability of $w(s_{t-1}^{(i)}, x)$, which is the cosine similarity between the descriptors of $s_{t-1}^{(i)}$ and $x$. This process is repeated until $m_i$ number of samples are accepted.

In Eq.1, the mean $\mu_t$ of $N(\mu_t, \sigma^2)$ is updated at every step as $\mu_t = \mu_{t-1} + \tau(\mu_{t-1}, \triangle M_\mu)$ where $\triangle M_\mu$ is the control parameter for the speed of the tracking. The higher $\triangle M_\mu$, the further $\boldsymbol{s}_t$ is located from $\boldsymbol{s}_{t-1}$ and the fewer the steps are executed until completion. The variance $\sigma^2$ of $N(\mu_t, \sigma^2)$ controls the spread of $\boldsymbol{s}_t$ along the timeline. A higher $\sigma^2$ results in a $\boldsymbol{s}_t$ that includes images with a longer time range.

A Gamma distribution $\Gamma(\alpha, \beta)$ is usually used to model the time required for $\alpha$ occurrences of events that follow a Poisson process with a constant rate $\beta$. In our interpretation, given an image stream, we assume that the occurrence of images of each subtopic follows the Poisson process with $\beta$. Then, $\Gamma(\alpha_{t-1}^{(i)}, \beta_{t-1}^{(i)})$ of Eq.1 indicates the time required for the next $\alpha$ images that have the same subtopic with $s_{t-1}^{(i)}$ in the image stream. Based on this intuition, the $\alpha_{t-1}^{(i)}$ for each $s_{t-1}^{(i)}$ is adjustively selected. A smaller $\alpha_{t-1}^{(i)}$ is chosen for the image $s_{t-1}^{(i)}$ with higher $\pi_{t-1}^{(i)}$ since the similar images to a more weighted $s_{t-1}^{(i)}$ are likely to occur more frequently in the dataset. The mean of Gamma distribution of each $s_{t-1}^{(i)}$ is aligned with the mean of the sample set $\mu_t$. Therefore, $\beta_{t-1}^{(i)} = \mu_t/\alpha_{t-1}^{(i)}$ since the mean of Gamma is $\alpha\beta$.

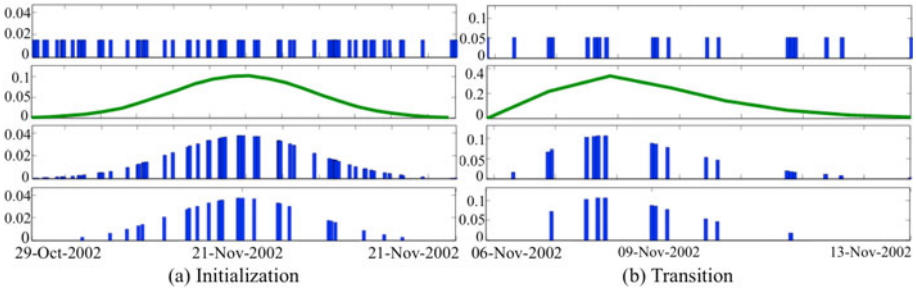


**Fig. 3.** An example of sampling images on the timeline during (a) the initialization and (b) the transition. From top to bottom: The first row shows the image distributions along the timeline. The images are regarded as the samples $(\{x^{(r)}\}_{r=1}^{R})$ from a proposal distribution $Q^*(x)$. They are equally weighted (*i.e.* $Q^*(x^{(r)}) = 1$). The second row shows the target distribution $P^*(x)$. (*e.g.* Gaussian in (a) and the product of Gaussian and Gamma in (b)). The third row shows the image samples weighted by $P^*(x^{(r)})/Q^*(x^{(r)})$. The fourth row shows the images chosen by *systematic sampling* [1].

The main reason to adopt the *product model* rather than the *mixture model* in Eq.1 is as follows. The *product model* only has a meaningful probability for an event when none of its component distribution has a low probability. (*i.e.* if one of two distributions has zero probability, their product does as well). It is useful in our application that the product with the Gaussian of the *drift* prevents the sampled images from severely spreading along the timeline by setting almost zero probability for the image outside the $3\sigma$ from $\mu_t$.

**Sampling Images with Target Distribution.** In the initialization and the transition, we sample a set of images on the timeline from a given target distribution $P^*(x)$. (*e.g.* Gaussian in the initialization and the product of Gaussian and Gamma in the transition). Fig.3 shows our sampling method, which can be viewed as an *importance sampling* [15]. The importance sampling is particularly useful for the transition model since there is no closed form of the product of Gaussian and Gamma distributions and its normalization is not straightforward.

**Observation Model.** The goal of the observation model is to generate weights $\boldsymbol{\pi}'_t$ for the $\boldsymbol{s}'_t$. First, the similarity matrix $\boldsymbol{W}_t$ of $\boldsymbol{s}'_t$ is obtained by computing pairwise cosine similarity of $\boldsymbol{s}'_t$. The $\boldsymbol{\pi}'_t$ is the stationary distribution of $\boldsymbol{W}_t$ by solving $\boldsymbol{\pi}'_t = \widetilde{\boldsymbol{W}}_t^T \boldsymbol{\pi}'_t$ where $\widetilde{\boldsymbol{W}}_t$ is row-normalized from $\boldsymbol{W}_t$ so that $\widetilde{w}_{ij} = w_{ij} / \sum_k w_{ik}$.

**Resampling.** The final posterior $\{\boldsymbol{s}_t, \boldsymbol{\pi}_t\} = \{s_t^{(i)}, \pi_t^{(i)}\}_{i=1}^{M}$ is resampled from $\{\boldsymbol{s}'_t, \boldsymbol{\pi}'_t\}$ by running the *systematic sampling* [1] on $\boldsymbol{\pi}''_t$. Then $\boldsymbol{\pi}_t$ is normalized so that their sum is one. The network $\boldsymbol{G}$ stores $\boldsymbol{W}_t(\boldsymbol{s}_t, \boldsymbol{s}_t)$ and the similarity matrix $\boldsymbol{W}_{t-1,t}(\boldsymbol{s}_{t-1}, \boldsymbol{s}_t)$ between two consecutive posteriors $\boldsymbol{s}_{t-1}$ and $\boldsymbol{s}_t$. As discussed in section 2.2, each vertex in $\boldsymbol{G}$ is connected to only its $k$-nearest neighbors.

## 3 Analysis and Results

### 3.1 Flickr Dataset

Table 1 summarizes 47 topics of our Flickr dataset. The topic name is identical to the query word. We downloaded all the images containing the query word. They are the images shown when a query word is typed in Flickr's search box without any option change. For the timestamp, we use the *date_taken* field of each image that Flickr provides.

We generate the similarity network of each topic by using the proposed SMC based tracking. The runtime is $O(NM)$ where $M$ is constant and $M \ll N$ (*i.e.* $1000 \leq M \leq 5000$ in our experiments). The network construction is so fast that, for example, it took about 4 hours for the *soccer* topic with $N = 1.1 \times 10^6$ and $M = 5,000$ in a matlab implementation on a single PC. The analysis of the network is also fast since most network analysis algorithms depend on the number of nonzero elements, which is $O(N \log N)$.

### 3.2 Evolution of Subtopics

Fig.4 shows the examples of the subtopic evolution of two topics, *big+ben* and *korean*. As we discussed in previous section, the SMC tracker generates the posterior sets $\{\boldsymbol{s}_0, \ldots, \boldsymbol{s}_e\}$. Five clusters in each posterior are discovered by applying spectral clustering to the subgraph $\boldsymbol{G}_t$ of each $\boldsymbol{s}_t$ in an unsupervised way. Obviously, the dynamic behavior is one of intrinsic properties of each topic. Some topics such as *big+ben* are stationary and coherent whereas others like *korean* are highly diverse and variant.

**Outbreak Detection of Subtopics.** The outbreak detection is important in Web mining since it reflects the change of information flows and people's interests. We perform the outbreak detection by calculating an information-theoretic measure of link statistics. Note that the consecutive posterior sets are

**Table 1.** 47 topics of our Flickr dataset. The numbers in parentheses indicate the numbers of downloaded images per topic. 9,751,651 images are gathered in total.

| | |
|---|---|
| Nation | *brazilian*(119,620), *jewish*(165,760), *korean*(254,386), *swedish*(94,390), *spanish*(322,085) |
| Place | *amazon*(160,008), *ballpark*(340,266), *big+ben*(131,545), *grandcanyon*(286,994), *pisa*(174,591), *wall+street*(177,181), *white+house*(241,353) |
| Animal | *butterfly+insect*(69,947), *cardinals*(177,884), *giraffe+zoo* (53,591), *jaguar*(122,615), *leopard*(121,061), *lobster*(144,596), *otter*(113,681), *parrot*(175,895), *penguin*(257,614), *rhino*(96,799), *shark*(345,606) |
| Object | *classic+car*(265,668), *keyboard*(118,911), *motorbike*(179,855), *pagoda*(128,019), *pedestrian*(112,116), *sunflower*(165,090), *television*(157,033) |
| Activity | *picnic*(652,539), *soccer*(1,153,969), *yacht*(225,508) |
| Abstract | *advertisement*(84,521), *economy*(61,593), *emotion*(119,899), *fine+art*(220,615), *horror*(157,977), *hurt*(141,249), *politics*(181,836) |
| Hot topic | *apple*(713,730), *earthquake*(65,375), *newspaper*(165,987), *simpson*(106,414), *starbucks*(169,728), *tornado*(117,161), *wireless*(139,390) |

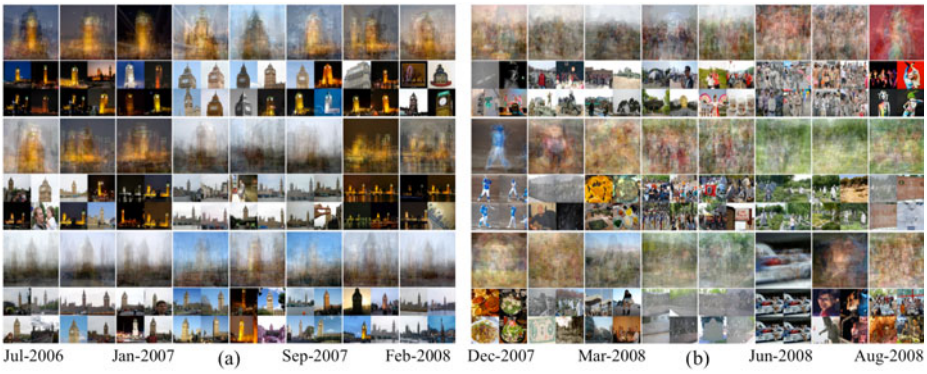


**Fig. 4.** Examples of subtopic evolution of *korean* and *big+ben* topics. Each column shows the clusters of each $\boldsymbol{s}_t$. From top to bottom, we show top three out of five clusters of each $\boldsymbol{s}_t$ with average images (the first row) and top-four highest ranked images in the cluster (the second row). The *big+ben* is relatively stationary and coherent whereas the *korean* topic is highly dynamic and contains diverse subtopics such as *sports*, *food*, *buildings*, *events*, and *Korean War Memorial Park*.

linked in our network. (*i.e.* $\boldsymbol{s}_{t-1}$ is connected to $\boldsymbol{s}_t$, which is linked to $\boldsymbol{s}_{t+1}$.) The basic idea of our outbreak detection is that if the subtopic distributions at step $t-1$ and $t+1$ are different each other, then the degree distribution of $\boldsymbol{s}_t$ to $\boldsymbol{s}_{t-1}$ ($\boldsymbol{f}_{t,t-1}$) and the degree distribution of $\boldsymbol{s}_t$ to $\boldsymbol{s}_{t+1}$ ($\boldsymbol{f}_{t,t+1}$) are dissimilar as well. For example, suppose that the dominant subtopic of $\boldsymbol{s}_{t-1}$ is *fruit apple* but the dominant one of $\boldsymbol{s}_{t+1}$ is *iphone*. Then, the degree of a *fruit apple* image $i$ in $\boldsymbol{s}_t$ has high $\boldsymbol{f}_{t,t-1}(i)$ but low $\boldsymbol{f}_{t,t+1}(i)$. On the other hand, an *iphone* image $j$ in $\boldsymbol{s}_t$ has high $\boldsymbol{f}_{t,t+1}(j)$ but low $\boldsymbol{f}_{t,t-1}(i)$. Both $\boldsymbol{f}_{t,t-1}$ and $\boldsymbol{f}_{t,t+1}$ are $|\boldsymbol{s}_t| \times 1$ histograms, each element of which is the sum of edge weights of a vertex in $\boldsymbol{s}_t$ with $\boldsymbol{s}_{t-1}$ and $\boldsymbol{s}_{t+1}$, respectively. In order to measure the difference between $\boldsymbol{f}_{t,t-1}$ and $\boldsymbol{f}_{t,t+1}$, we use *Kullback-Leibler* (KL) divergence in Eq.2.

$$D_{KL}(\boldsymbol{f}_{t,t+1} \parallel \boldsymbol{f}_{t,t-1}) = \sum_{i \in \boldsymbol{s}_t} \boldsymbol{f}_{t,t+1}(i) \log \frac{\boldsymbol{f}_{t,t+1}(i)}{\boldsymbol{f}_{t,t-1}(i)} \tag{2}$$

Fig.5.(a) shows an example of KL divergence changes along the 142 steps of *apple* tracking. The peaks of KL divergence indicate the radical subtopic changes from $\boldsymbol{s}_{t-1}$ to $\boldsymbol{s}_{t+1}$. We observed the highest peak at the step $t^* = 63$, where $\boldsymbol{s}_{t^*}$ is distributed in [May-2007, Jun-2007]. Fig.5.(b) represents ten subtopics of $\boldsymbol{s}_{t^*-1}$, $\boldsymbol{s}_{t^*}$, and $\boldsymbol{s}_{t^*+1}$, which are significantly different each other.

### 3.3 Comparison with Text Analysis

In this section, we empirically compare the image-based topic analysis with the text-based one. One may argue that the similar observations can be made from both images and the associated texts. However, our experiments show that the

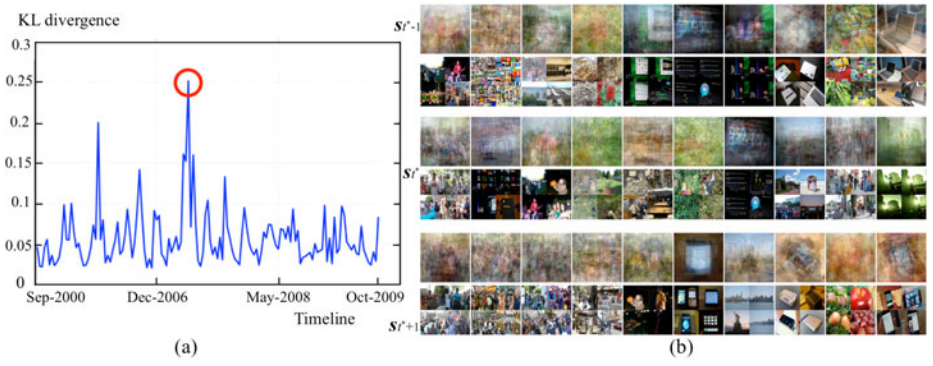


**Fig. 5.** The outbreak detection of subtopics. (a) The variation of KL divergences for the *apple* topic. The highest peak is observed at the step $t^*$=63 ([May-2007, Jun-2007] with the median of 11-Jun-2007). (b) The subtopic changes around the highest peak. Ten subtopics of $\boldsymbol{s}_{t^*-1}$, $\boldsymbol{s}_{t^*}$, and $\boldsymbol{s}_{t^*+1}$ are shown from top to bottom. In each set, the first row shows average images of top 15 images and the bottom row shows top four highest ranked ones in each subtopic. In $\boldsymbol{s}_{t^*-1}$ and $\boldsymbol{s}_{t^*}$, several subtopics about *Steve Jobs's presentation* are detected but disappear in $\boldsymbol{s}_{t^*+1}$. Rather, *crowds in street* (*i.e.* 1st $\sim$ 4th clusters) and *iphone* (*i.e.* 6,8,10-th clusters) newly emerge in $\boldsymbol{s}_{t^*+1}$.

associated texts do not overshadow the importance of information from the images. First of all, 13.70% of images in our dataset have no tags. It may be natural since the Flickr is oriented toward image sharing and thus text annotations are much less cared by users. In order to compare the dynamic behaviors detected from images and texts, we apply the outbreak detection method in previous section to both images and their associated tags. The only difference between them is the features: the spatial pyramids of SIFT and HOG for images and term frequency histograms for texts. Fig.6.(a) shows an example of outbreak detection using images and texts for the *grandcanyon* topic, which is one of the most stationary and coherent topics in our dataset (*i.e.* no matter when the images are taken, the majority of them are taken for the scene of the *Grand Canyon*). The image-based analysis is able to successfully detect its intrinsic stationary behavior. However, the text tags are highly fluctuated mainly because tags are subjectively assigned by different users with little consensus. This is a well-known noise source of the images from the Web image search, and our result can be its another supporting example from the dynamics view.

Another important advantage of image-based temporal analysis is that it conveys more delicate information that is hardly captured by text descriptions. Fig.6.(b) shows two typical examples about periodic updates of objects and events. For example, when a new *iphone* is released, the emergence of the *iphone* subtopic can be detected in the *apple* via both images and texts. However, the images can more intuitively reveal the upgraded appearance, new features, and visual context around the new event.

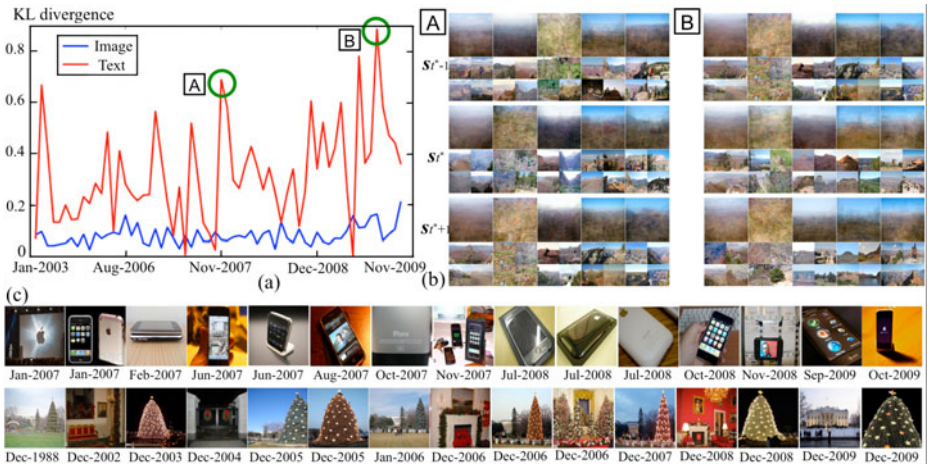


**Fig. 6.** The comparison between the topical analysis on the images and associated text tags. (a) The variation of KL divergences for the *grandcanyon* topic. The KL divergences of images are stationary along the timeline whereas those of texts are highly fluctuated. (b) The subtopic changes around the two highest peaks **A** (05-Nov-2007) and **B** (16-Aug-2009). Five subtopics of $s_{t^*-1}$, $s_{t^*}$, and $s_{t^*+1}$ are shown from top to bottom. Very little visual variation is observed between them. (c) 15 selected images tagged by *apple+new+iphone* (the first row) and *whitehouse+christmas* (the second row). They are sorted on the timeline.

### 3.4 Temporal Association for Classification

As pointed in neuroscience research [19,21], human perception tends to strongly connect temporally smoothed visual information. Inspired by these studies, we perform preliminary tests to see whether it holds in Web images as well; The subtopics that consistently appear along the timeline can be more closely related to the main topic rather than the ones that are observed for only a short period. For example, the *fruit apple* is likely to consistently exist in the *apple* image set, which may be a more representative subtopic of the *apple* rather than a specific model of an early *Mac* computer. In this experiment, we generate two training sets from the extremely noisy Flickr images and compare their classification performance; The first training set is constructed by choosing the images that are temporally and visually associated, and the other set is generated by the random selection without temporal context.

Since our similarity network links temporally close and visually similar images, dominant subtopics correspond to large clusters and their central images map to hub nodes in the graph. The stationary probability is a popular ranking measure, and thus the images with high stationary probabilities can be thought of temporally and visually strengthened images. However, the proposed network representation is incomplete in the sense that images are connected in an only local temporal space. In order to cope with this underlying uncertainty, we generate training sets by the Metropolis-Hasting (MH) algorithm.

We first compute the stationary probability $\boldsymbol{\pi}_G$ of the network $\boldsymbol{G}$. Since a general suggestion for a starting point in the MH is to begin around the modes of the distribution, we start from an image $\theta_o$ that has the highest $\boldsymbol{\pi}_G(\theta)$. From a current $\theta$ vertex, we sample a next candidate point $\theta^*$ from a proposal distribution $q(\theta_1, \theta_2)$ that is based on a random surfer model as shown in Eq.3; the candidate is chosen by following an outgoing edge of the $\theta$ with probability $\lambda$, but restarting it with probability $1 - \lambda$ according to the $\boldsymbol{\pi}_G$. A larger $\lambda$ weights more the local link structure of the network while a smaller $\lambda$ relies on $\boldsymbol{\pi}_G$ more. The new candidate is accepted with probability $\alpha$ in Eq.3 where $\widetilde{w}_{ij}$ is the element $(i, j)$ in the row-normalized adjacency matrix of $\boldsymbol{G}$. We repeat this process until the desired numbers of training samples are selected.

$$\alpha = \min\left(\frac{\boldsymbol{\pi}_G(\theta^*)q(\theta^*, \theta_{t-1})}{\boldsymbol{\pi}_G(\theta_{t-1})q(\theta_{t-1}, \theta^*)}, 1\right) \quad \text{where } q(i, j) = \lambda\widetilde{w}_{ij} + (1 - \lambda)\boldsymbol{\pi}_G(j) \tag{3}$$

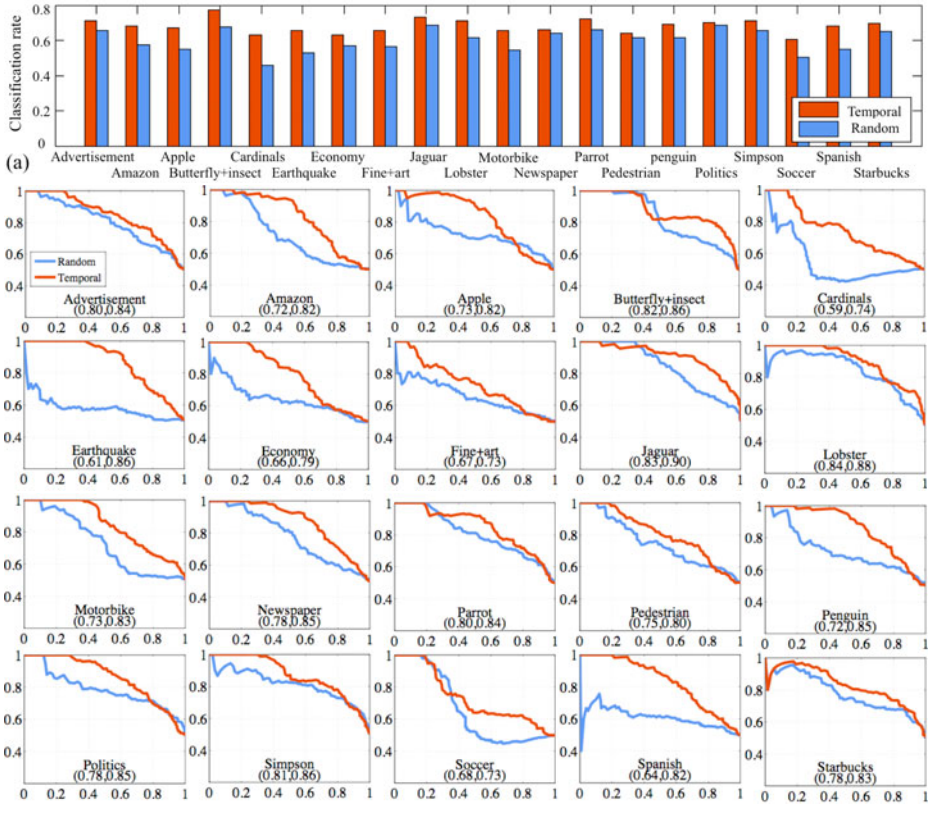


**Fig. 7.** Comparison of the binary classification performance between *Temporal* training and *Random* training. (a) Classification accuracies of selected 20 topics. (b) Corresponding Precision-Recall curves. The number $(n, m)$ underneath the topic name indicates the average precision of (*Random*, *Temporal*).

We perform binary classification using the 128 nearest neighbor voting [20] in which we use the same descriptors and the cosine similarity in section 2.1. We generate the positive training set of each topic in two different ways; We sample 256 images by the MH method (called *Temporal* training) and randomly choose the same number of images (called *Random* training). For the negative training images, we randomly draw 256 images from the other topics of Flickr dataset. For the test sets, we downloaded 256 top-ranked images for each topic from Google Image Search by querying the same word in Table 1. The Google Image Search provides relatively clean images in the highest ranking. Since we would like to test whether the temporally associated samples are better generalization of the topic, the Google test sets are more suitable to our purpose than the images from the noisy Flickr dataset. In the binary classification test of each topic, the positive test images are the 256 Google images of the topic and the negative test images are 256 Google images that are randomly selected from the other topics. Note that in each run of experiment, only the positive training samples are different between *Temporal* and *Random* tests. The experiments are repeated ten times, and the mean scores are reported.

Fig.7 summarizes the comparison of recognition performance between *Temporal* and *Random* training. Fig.7.(a) shows the classification rates for the selected 20 topics. The accuracies of *Temporal* training are higher by 8.05% on average. Fig.7.(b) presents the corresponding precision-recall curves, which show that the *temporal association* significantly improves the confidence of classification. The *Temporal* training is usually better than the *Random* training in performance, but the improvement is limited in some topics; In highly variant topics (*e.g. advertisement* and *starbucks*), the temporal consistency is not easily captured. In stationary and coherent topics (*e.g. butterfly+insect* and *parrot*), the random sampling is also acceptable.

## 4 Discussion

We presented a nonparametric modeling and analysis approach to understand the dynamic behaviors of Web image collections. A sequential Monte Carlo based tracker is proposed to capture the subtopic evolution in the form of the similarity network of the image set. In order to show the usefulness of the image-based temporal topic modeling, we examined subtopic evolution tracking, subtopic outbreak detection, the comparison with the analysis on the associated texts, and the use of temporal association for recognition improvement. We believe that this line of research has not yet fully explored and various challenging problems still remain unsolved. In particular, more study on the temporal context for recognition may be promising.

**Acknowledgement.** This research is supported in part by funding from NSF IIS-0713379, DBI-0546594, Career Award, ONR N000140910758, DARPA NBCH 1080007 and Alfred P. Sloan Foundation awarded to Eric P. Xing, and NSF Career Award IIS 0747120 to Antonio Torralba.

## References


1. Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T.: A Tutorial on Particle Filters for On-line Non-linear/Non-Gaussian Bayesian Tracking. IEEE Trans. Signal Processing 50(2), 174–188 (2002)
2. Becker, S.: Implicit Learning in 3D Object Recognition: The Importance of Temporal Context. Neural Computation 11(2), 347–374 (1999)
3. Blei, D.M., Lafferty, J.D.: Dynamic Topic Models. In: ICML (2006)
4. Bosch, A., Zisserman, A., Munoz, X.: Image Classification using Random Forests and Ferns. In: ICCV (2007)
5. Boutell, M., Luo, J., Brown, C.: A Generalized Temporal Context Model for Classifying Image Collections. Multimedia Systems 11(1), 82–92 (2005)
6. Cao, L., Luo, J., Kautz, H., Huang, T.S.: Annotating Collections of Photos using Hierarchical Event and Scene Models. In: CVPR (2008)
7. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge, VOC 2010 Results (2010), `http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html`
8. Hinton, G.E.: Training Products of Experts by Minimizing Contrastive Divergence. Neural Computation 14(8), 1771–1800 (2002)
9. Isard, M., Blake, A.: CONDENSATION – Conditional Density Propagation for Visual Tracking. Int. J. Computer Vision 29(1), 5–28 (1998)
10. Kalogerakis, E., Vesselova, O., Hays, J., Efros, A., Hertzmann, A.: Image Sequence Geolocation with Human Travel Priors. In: ICCV (2009)
11. Kim, G., Torralba, A.: Unsupervised Detection of Regions of Interest using Iterative Link Analysis. In: NIPS (2009)
12. Li, Y., Crandall, D.J., Huttenlocher, D.P.: Landmark Classification in Large-scale Image Collections. In: ICCV (2009)
13. Liu, C., Yuen, J., Torralba, A.: Nonparametric Scene Parsing: Label Transfer via Dense Scene Alignment. In: CVPR (2009)
14. Liu, C., Yuen, J., Torralba, A., Sivic, J., Freeman, W.T.: SIFT Flow: Dense Correspondence across Different Scenes. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 28–42. Springer, Heidelberg (2008)
15. MacKay, D.: Information Theory, Inference and Learning Algorithms. Cambridge University Press, Cambridge (2002)
16. Paletta, L., Prantl, M., Pinz, A.: Learning Temporal Context in Active Object Recognition Using Bayesian Analysis. In: ICPR (2000)
17. Quack, T., Leibe, B., Gool, L.V.: World-scale Mining of Objects and Events from Community Photo Collections. In: CIVR (2008)
18. Russell, B.C., Torralba, A.: Building a Database of 3D Scenes from User Annotations. In: CVPR (2009)
19. Sinha, P., Balas, B., Ostrovsky, Y., Russell, R.: Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About. Proceedings of the IEEE 94(11), 1948–1962 (2006)
20. Torralba, A., Fergus, R., Freeman, W.T.: 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. IEEE PAMI 30(11), 1958–1970 (2008)
21. Wallis, G., Bulthöff, H.H.: Effects of Temporal Association on Recognition Memory. PNAS 98(8), 4800–4804 (2001)
22. Wang, X., McCallum, A.: Topics Over Time: a Non-Markov Continuous-Time Model of Topical Trends. In: KDD (2006)