Joint Photo Stream and Blog Post Summarization and Exploration

Gunhee Kim
Seoul National University
gunhee@snu.ac.kr

Seungwhan Moon
Carnegie Mellon University
seungwhm@cs.cmu.edu

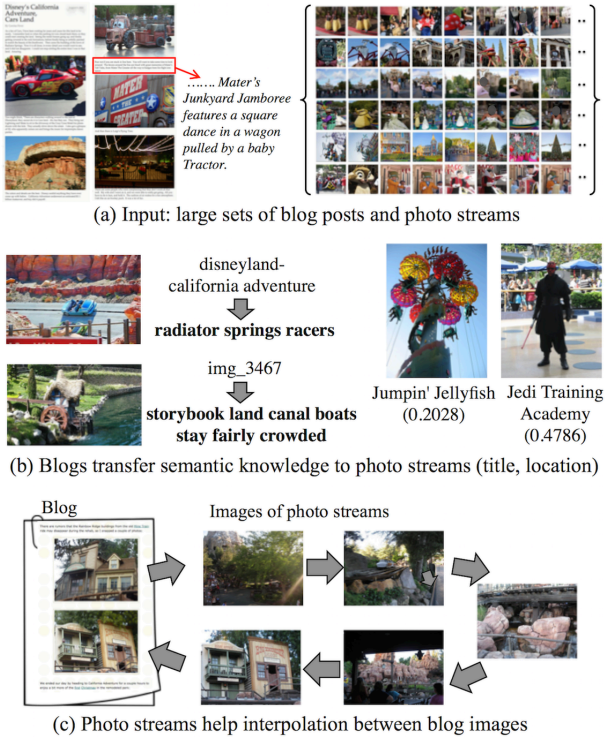
Leonid Sigal
Disney Research Pittsburgh
lsigal@disneyresearch.com

Abstract

We propose an approach that utilizes large collections of photo streams and blog posts, two of the most prevalent sources of data on the Web, for joint story-based summarization and exploration. Blogs consist of sequences of images and associated text; they portray events and experiences with concise sentences and representative images. We leverage blogs to help achieve story-based semantic summarization of collections of photo streams. In the opposite direction, blog posts can be enhanced with sets of photo streams by showing interpolations between consecutive images in the blogs. We formulate the problem of joint alignment from blogs to photo streams and photo stream summarization in a unified latent ranking SVM framework. We alternate between solving the two coupled latent SVM problems, by first fixing the summarization and solving for the alignment from blog images to photo streams and vice versa. On a newly collected large-scale Disneyland dataset of 10K blogs (120K associated images) and 6K photo streams (540K images), we demonstrate that blog posts and photo streams are mutually beneficial for summarization, exploration, semantic knowledge transfer, and photo interpolation.

1. Introduction

Photographs taken by general users can be regarded as personal statements of what stories they want to remember and tell about their experiences. Fig.1 shows one of the most evident examples, *visiting Disneyland*. In a single day, tens of thousands of people visit *Disneyland*, and many of them take large streams of photos about their special experiences with families or friends. In addition, some of the more enthusiastic visitors are also willing to write *travel blogs*, in which their personal stories unfold with itineraries, commentaries, impressions, and fun facts about the attractions. Most blogs include informative text along with the photos that users carefully choose as the most representative ones out of their large collections taken during their trip.



(a) Input: large sets of blog posts and photo streams

(b) Blogs transfer semantic knowledge to photo streams (title, location)

(c) Photo streams help interpolation between blog images

Figure 1. Motivation for joint summarization and exploration between large collections of photo streams and blog posts. (a) The input is two-fold: a set of photo streams and blog posts from *Disneyland*, which are captured by multiple users and at different times. (b) Blogs benefit photo stream summarization by transferring semantic knowledge: Examples are *automatic image titling* and attraction-based *image localization*. (c) Photo streams enhance blog posts by allowing interpolation between blog images. Two blog images of an attraction entrance used as a query, result in an illustration of what happens inside the attraction.

In this paper, as shown in Fig.1, we propose to take advantage of large collections of photo streams and blog posts in a *mutually-beneficial* way for the purpose of summarization and exploration. Blogs usually consist of sequences of images and associated text; they are written in a way of *storytelling* [16], by digesting key events with concise sentences and representative images. Thus, blog posts can help achieve a *story-based* semantic summarization of large-scale and ever-growing collections of photo streams

*This work has been done when Gunhee Kim was a postdoctoral researcher, and Seungwhan Moon was an intern at Disney Research.

that are often unstructured and associated with missing or inaccurate semantic labels. In the reverse direction, each blog benefits from a large set of photo streams, which can interpolate various photo paths between consecutive images in the blog. Each blog is written based on a single person's experience with a small number of selective images. Hence, the photo-path interpolation, achieved with photo streams, allows blog authors to explore alternative paths made by other visitors who follow a similar itinerary.

To implement joint summarization and exploration, we first collect a large set of photo streams and blog posts for an event of interest (*e.g*. *visiting Disneyland*) from the Web. We then jointly perform the two base tasks so that they help each other: (i) *alignment* between the images of blogs and photo streams, and (ii) *summarization* of photo streams. The alignment task discovers the correspondences from the blog photos to the images in the photo streams. The summarization task selects the most interesting and important images from photo streams with maximal coverage and minimal redundancy. Since blog photos are highly selective by users, encouraging photo stream summaries to have images that are closer matches to blog photos are likely to be more semantically meaningful. In the reverse, summaries of photo streams can make the alignment task faster and more focused. We formulate the alignment and the summarization as the optimization of two sets of latent ranking SVM problems [9, 28]. Hence, starting from initial summarization of photo streams, we alternate between solving one of the two tasks while conditioning on the output of the other.

For evaluation, we crawl the *Disneyland* dataset, consisting of about 540K images of 6K photo streams from FLICKR and 10K blog posts with 120K associated images from BLOGGER and WORDPRESS. Although we mainly discuss the proposed approach in the context of *Disneyland*, our approach can be extended to any problem domain, with little modification, because our NLP pre-processing (*e.g*. keyword extraction) is unsupervised. The only requirement is that the domain must have sufficient number of photo streams and blogs (*e.g*. tours of cities or museums).

In the experiments, we focus on showing that blog posts and photo streams are indeed mutually beneficial. First, we show that blog posts help achieve story-based semantic summarization of large sets of photo streams. In addition, we compare our algorithm with other candidate methods for the two tasks of *semantic knowledge transfer*. Since blogs consist of sequences of images and associated text, once alignment is complete, we can transfer the semantic knowledge associated with blog pictures to the aligned photo stream images, many of which have noisy or no semantic labels. Specifically, we show that blog posts improve the image localization accuracy (*i.e*. finding where photos were taken), and automatic image titling (*i.e*. creating descriptive titles for images). Second, we show that a large set of photo streams lead to better path interpolation between consecutive blog images. We quantitatively evaluate the performance of our approach for path interpolation via crowdsourcing, using Amazon Mechanical Turk.

Relation to previous work. In recent computer vision research, several studies have been conducted to organize and explore collections of unstructured tourists' photos. *Photo Tourism* [26] enables users to interactively browse large collections of geo-tagged photos of tourist attractions in a 3D space. In [13, 25], visitors' images are leveraged to create virtual tours of the world's landmarks on Google Maps. The 3D Wikipedia [21] establishes semantic navigation between the regions in the 3D landmark models and the Wikipedia text. In another line of work, [10, 11] addresses the problem of creating storyline graphs from Flickr image collections of outdoor activities. In comparison to these works, the key difference is that we explicitly leverage blog text and associated images to organize general users' photo streams. By doing so, we can achieve a semantically meaningful and *story*-based summary. Although 3D Wikipedia [21] also uses the documents in parallel with online visitors' photos, it focuses on a small number of well-structured text from Wikipedia. Instead, we use a large set of unstructured blog posts created by the general public.

Recently, there has been a focus on jointly leveraging visual and textual data to address challenging computer vision problems. Some notable problem domains that benefit from the synergistic interplay between the two complementary information sources include the generation of natural language descriptions from images [12, 17, 30] and videos [2, 6, 19] and joint detection and segmentation of scene and object types from images with textual information [4]. However, novel features of our work are two-fold. First, we use blog data as text sources, and second we aim at story-based exploration of photo streams and blogs.

Finally, in data mining research, there have been several previous papers to address the extraction of stories from web log data [3, 5, 18]. However, most of them are based on purely textual information. Among them, the work of [7] may be the most relevant to ours, because they also use blog pictures to discover popular landmarks in several big cities (*e.g*. Beijing, Sydney). However, [7] simply exploits the blogs as a repository of photos, whereas we close a loop for joint exploration between blog posts and photo streams. Therefore, we can perform several additional tasks, including semantic knowledge transfer from blogs to images and photo-path interpolation between blog photos.

Contributions. Our contributions are three-fold.

(1) To the best of our knowledge, our work is unique in jointly leveraging large sets of blog posts and photo streams for mutually-beneficial summarization and exploration. We show that blogs are useful for story-based summary of photo streams along with semantic knowledge transfer. At

the same time, a large set of photo streams help interpolate plausible image paths between any consecutive blog photos.

(2) We propose an approach for jointly solving alignment and summarization tasks in a unified ranking SVM framework. We alternatingly solve one of the two problems while conditioning on the solution of the other.

(3) For evaluation, we collect *Disneyland* dataset, consisting of 10K blog posts with 120K associated images, and 6K photo streams of 540K images. We demonstrate that blog posts and photo streams indeed help each other for summarization and exploration.

2. Input Data and Preprocessing

2.1. Photo streams and Blog Posts

The input of our approach is a set of photo streams $\mathcal{P} = \{P^1, \cdots, P^L\}$ and a set of blog posts $\mathcal{B} = \{B^1, \cdots, B^N\}$ for a topic of interest (*e.g.* Disneyland). Each photo stream is a set of images taken in sequence by a single user in a single day, denoted by $P^l = \{p_1^l, \cdots, p_{L^l}^l\}$. We assume that each image p_i^l is associated with timestamp t_i^l, based on which each photo stream is temporally sorted. Additionally, some photos may include GPS information g_i^l and text information s_i^l (*e.g.* titles and tags). Each blog post comprises a sequence of pairs of images and text blocks (denoted by $B^n = \{(I_1^n, T_1^n) \cdots, (I_{N^n}^n, T_{N^n}^n)\}$). We use $\mathcal{I}$ to denote all images of blogs. We discuss the details of our natural language processing for blog posts in Section 2.3.

Note that the used NLP techniques are unsupervised. One optional domain-specific input is the list of vocabularies for name entity extraction. Since we are particularly interested in location information, we add attraction and district names of Disneyland to the vocabulary list, referring to the official visitors' map. For other domains, the vocabulary list can be easily constructed from publicly available information, without any modification to the algorithm itself. Therefore, our approach is general and can be applied to any event or topic that has sufficient data.

2.2. Image Description

We use dense feature extraction with vector quantization, which is one of the standard methods in recent computer vision literature. We densely extract HSV color, SIFT and histogram of oriented edge (HOG) features on a regular grid for each image at steps of 4 and 8 pixels, respectively. We then form 300 visual words for each feature type by applying K-means to randomly selected descriptors. Finally, the nearest word is assigned to every node of the grid. As image or region descriptors, we build L_1 normalized spatial pyramid histograms to count the frequency of each visual word in three levels of regular grids. We define the image descriptor $\boldsymbol{v}$ by concatenating the two spatial pyramid histograms of color SIFT and HOG features.

2.3. Natural Language Processing of Blog Posts

The text in a blog is usually highly correlated with the embedded images, and thus can be used as a rich source of semantic information. However, using blog text data in this way comes with several challenges: (i) it is hard to localize exactly *which part* of the text corresponds to each image, (ii) blog text is complex and is a challenge to understand even with state-of-the-art NLP algorithms, (iii) non-professional writing tends to contain a lot of lexical and syntactic errors. Further, we cannot expect sentences to describe all blog images or everything in those images.

Our goal here is to represent each blog post B^n by a sequence of images, associated meta-data and corresponding confidences: $\{(I_1^n, \mathbf{m}_1^n, \mathbf{v}_1^n) \cdots, (I_{N^n}^n, \mathbf{m}_{N^n}^n, \mathbf{v}_{N^n}^n)\}$, where $\mathbf{m}_i^n = \{\mathbf{l}_i^n, \mathbf{k}_i^n\}$, $\mathbf{l}_i^n$ is a list of name entities, $\mathbf{k}_i^n$ is a set of key phrases extracted from the text associated with image i in blog n, and $\mathbf{v}_i^n \in [0,1]$ is a vector of confidence scores of length $|\mathbf{l}_i^n| + |\mathbf{k}_i^n|$.

Named Entity Extraction. To extract named entities (*i.e.* mainly attraction names as locations in our setting) from the blog text, we use a linear chain CRF-based named entity recognizer (NER) [14, 27] trained on a standard NER training corpus (CoNLL 2003 [22]). We pick out only the location related entities, and find the closest matches to the venues of interest (*e.g.* Disney attractions). The confidence of a word tagged as one of the attractions is thus the posterior class probability of the NER labelling penalized by its Levenstein distance to the closest match.

Key Phrases Extraction. For key phrases extraction, we use an unsupervised statistical method called RAKE (Rapid Automatic Keywords Extraction) [20], which estimates key phrases by their word co-occurrence scores measured by term frequency and keyword adjacency [1].

Confidence of Photo Association. The key challenge in our work is to figure out how we associate the extracted text information (*e.g.* locations, key phrases) with blog images. Assuming that a text block closer to an image has a higher probability of belonging to the image, we employ a simple heuristic based on the image-to-text-distance. For example, the confidence score of an image belonging to a certain location can be calculated as a summation of the confidence scores of text blocks containing the name of the location, penalized by the distance of the image to that text block:

$$\mathbf{v}_i^n = \sum_{t \in T} \left\{ h_t(\mathbf{l}_i^n) - pen(t, I_i^n) \right\} \quad (1)$$

where $\mathbf{v}_i^n$ refers to the confidence score vector of an image I_i^n being associated with locations $\mathbf{l}_i^n$, T is a set of text blocks in the blog, $h_t(\mathbf{l}_i^n)$ is a confidence score of a text block $t \in T$ containing the locations $\mathbf{l}_i^n$, and $pen(\cdot)$ is a penalty function that degrades the association between a text block and an image based on the distance. We use

$pen(t, I_i^n) = d(t, I_i^n)/|T|$, where $d(t, I_i^n)$ is the index distance between a text block t and an image I_i^n where each text block and image is treated as one element in a sequence.

3. Approach

For joint exploration between blogs and photo streams, we solve the two subproblems: (i) alignment from blog images to photo streams, and (ii) summarization of photo streams. The alignment is achieved by building a bipartite image graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the vertex set consists of images of blogs and photo streams (*i.e.* $\mathcal{V} = \mathcal{I} \cup \mathcal{P}$), and the edge set $\mathcal{E}$ presents the correspondences between them. We denote the adjacency matrix by $\mathbf{W} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{P}|}$ where $|\mathcal{P}|$ is the number of images in all photo streams. Thus, the goal of alignment reduces to an estimate of $\mathbf{W}$. On the other hand, summarization aims to predict the best subset $S^l \subset P^l$ for each photo stream $P^l \in \mathcal{P}$. We use $\mathcal{S}$ to denote a set of summaries $\mathcal{S} = \{S^1, \cdots, S^L\}$. We now list our constraints for alignment (A1)–(A4) and summarization (S1)–(S2).

(A1) (*Sparsity*) We let $\mathbf{W}$ have a few nonzero elements. We retain only a small number of strong matches to avoid unnecessarily complex alignment where a blog image links too many images of photo streams.

(A2) (*Similarity*) If a blog image i is more similar to a photo stream image j than k, then $\mathbf{W}_{ij} > \mathbf{W}_{ik}$.

(A3) (*Summary*) We prefer to align blog images with the images in the summary $S^l \in \mathcal{S}$. If $j \in S^l$ and $k \notin S^l$, $\mathbf{W}_{ij} > \mathbf{W}_{ik}$ is encouraged.

(A4) (*Continuity*) Consecutive images in a blog are encouraged to match the images in the same photo stream, which makes easier the photo interpolation task.

(S1) (*Alignment*) Since blog images are representative and selective, the images of summary S^l should have as many alignment inlinks from blog images as possible. That is, $\mathbf{W}_{*S^l} \in \mathbb{R}^{|\mathcal{I}|\times|S^l|}$, a part of the adjacency matrix from all blog images to S^l, is non-sparse.

(S2) (*Coverage and Diversity*) The summary S^l should contain as few redundant images as possible, and not miss any important images of the photo streams.

As described in the above, computing $\mathbf{W}$ requires the result of photo stream summary $\mathcal{S}$ in (A3), and reversely computing summary $\mathcal{S}$ requires $\mathbf{W}$ in (S1). Therefore, we initialize $\mathcal{S}$, and then alternate between updating $\mathbf{W}$ and $\mathcal{S}$ until convergence. We formulate the alignment and summarization as two sets of ranking SVM problems with latent variables (*e.g.* [9, 15, 28]), The major strength of this formulation is its flexibility; one can easily add additional constraints, while using the exact same formulation and optimization, which can be efficiently solved via stochastic gradient descent. The details of alignment and summarization will be discussed in Section 3.1 and 3.2, respectively.

Initialization. We first obtain an initial set of summarizations of photo streams, denoted by $\mathcal{S}^{(0)}$, by applying K-means clustering on the image descriptors. We set $K = 0.3|P^l|$, where $|P^l|$ is the size of the photo stream. We use a relatively high K to select many images as an initial summary. As iterations proceed, $\mathcal{S}^{(t)}$ is updated to include a diverse set of canonical images while reducing redundancy.

3.1. Alignment between Blogs and Photo Streams

The objective of alignment is to find correspondences from blog images to photo streams (*i.e.* computing $\mathbf{W} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{P}|}$). We assume an initial $\mathcal{S}^{(o)}$ is available.

We design our alignment optimization based on the latent ranking SVM [8, 28], which minimizes a regularized margin-based loss, to satisfy the constraints (A1)–(A4) discussed previously. We solve the following optimization for each blog separately. We use $\mathbf{W}^n \in \mathbb{R}^{|\mathcal{I}^n|\times|P|}$ to denote a portion of the adjacency matrix for blog B^n.

$$\min_{\mathbf{W}^n, \xi} \frac{1}{2}\|\mathbf{W}^n\|_1 + \frac{\lambda_A}{M}\sum_{i=1}^{M}\xi_i \tag{2}$$
$$\text{s.t. } \forall(i,j,k) \in \mathcal{C}_d \cup \mathcal{C}_s\colon \mathbf{W}^n_{ij} - \mathbf{W}^n_{ik} \geq \Delta(\sigma_{ij}, \sigma_{ik}) - \xi_i$$
$$\forall(i,j,k) \in \mathcal{C}_c\colon \mathbf{W}^n_{ij} - \mathbf{W}^n_{ik} \geq \Delta(\#_{ij}, \#_{ik}) - \xi_i$$

where λ_A is a regularization parameter and $M = |\mathcal{C}_d| + |\mathcal{C}_s| + |\mathcal{C}_c|$ is the number of constraint instances. The objective of Eq.(2) uses ℓ_1-norm penalty instead of ℓ_2-norm to encourage the sparsity of $\mathbf{W}^n$ for the constraint (A1). The three constraint sets of $\mathcal{C}$, which encodes the (A2)–(A4), consist of a large number of image triplets (i, j, k) where i is sampled from blog B^n and j, k from photo streams.

First, $\mathcal{C}_d$ is the similarity constraints for (A2), penalizing the violation of $\mathbf{W}^n_{i,j} \leq \mathbf{W}^n_{j,k}$ when a blog image i is more similar to j than k. We use $\Delta(\sigma_{ij}, \sigma_{ik}) = |\sigma_{ij} - \sigma_{ik}|$ as the loss function, where σ_{ij} is the similarity between i and j.

Second, $\mathcal{C}_s$ is the summary constraint for (A3). We prefer matching with the images in the summary S. Thus, if $j \in \mathcal{S}$ and $k \notin \mathcal{S}$, then $\mathbf{W}^n_{i,j} \leq \mathbf{W}^n_{j,k}$ is penalized.

Finally, $\mathcal{C}_c$ is the continuity constraint for (A4), which boosts the consecutive images in a blog to match with the images in the same photo stream. Suppose $j \in P^l$ and $k \in P^m$. We define $\#_{ij} = \sum_{p\in P^l}(\sigma_{i-1,p} + \sigma_{i+1,p})/2|P^l|$, which indicates the mean of feature similarity between i's neighbors $(i-1,\ i+1)$ and P^l. With the loss function $\Delta(\#_{ij}, \#_{ik}) = |\#_{ij} - \#_{ik}|$, if i's neighbors are more similar to P^l than P^m, then $\mathbf{W}^n_{i,j} > \mathbf{W}^n_{j,k}$ is encouraged.

Constraint generation. The strength of Eq.(2) is the flexibility to easily incorporate various objectives related to alignment in a form of constraints. However, the size of $\mathcal{C}$ can be very large if we consider all possible combinations of triplets. For example, the largest size of $\mathcal{C}_d$ is $|B^n| \cdot \sum_{l=1}^{L}\binom{|P^l|}{2}$, where L is the number of photo streams.

Hence we generate $\mathcal{C}$ as follows. First, we find the K-nearest photo streams $\mathcal{N}(B^n)$ for each blog B^n, using the Hausdorff distance metric, and generate constraint samples only from $\mathcal{N}(B^n)$. We set $K = c \cdot \log(|\mathcal{P}|)$ with $c = 4$. We then fix the size of $\mathcal{C}$ (*e.g.* , 5∼10K per blog image) according to the allowed computational resource. For each blog image we sample triplet constraints using a weighted random sampling without replacement. For example, we generate $\mathcal{C}_s$ for blog image i by choosing one image from $j \in \mathcal{S}$ according to the weight of similarity σ_{ij} and the other from $k \notin \mathcal{S}$. We accept the triplet (i, j, k) only if $\sigma_{ij} > \sigma_{ik}$. We repeat adding triplets until the fixed set size is reached.

Optimization. We optimize Eq.(2) using an online stochastic coordinate descent algorithm in [23, 24]. Since the datasets of blogs and photo streams are large-scale and possibly ever-growing, it is beneficial to use the stochastic gradient descent formulation, which converges faster to a good solution than a batch algorithm. We present the detailed derivation and pseudocode in the supplementary.

3.2. Photo Stream Summarization

For each photo stream P^l, the objective of summarization is to find the subset $S^{l*} = \text{argmax}_{S \subset P^l} \mathbf{s}^l_S$, in which $\mathbf{s}^l_S$ indicates a ranking score to any subset $S \subset P^l$. Although the size of all possible subsets is exponential (*i.e.* $2^{|P^l|}$), we will present a tractable approximate algorithm below, by limiting the number of subsets to explore. We compute the ranking score $\mathbf{s}^l$ based on the similar latent ranking SVM to satisfy the constraints (S1)–(S2) in Section 3.

$$\min_{\mathbf{s}^l, \xi} \frac{1}{2}\|\mathbf{s}^l\| + \frac{\lambda_S}{M}\sum_{i=1}^{M}\xi_i \qquad (3)$$
$$\text{s.t. } \forall (S^l_i, S^l_j) \in \mathcal{C}_p : \mathbf{s}^l_i - \mathbf{s}^l_j \geq \Delta_S(S^l_i, S^l_j) - \xi_i$$

where λ_S is a regularization parameter and $M = |\mathcal{C}_p|$. The loss function is $\Delta_S(S^l_i, S^l_j) = |\kappa(S^l_i, P^l) - \kappa(S^l_j, P^l)|$ for a subset pair (S^l_i, S^l_j) where

$$\kappa(S^l_i, P^l) = \sum_{p \in P^l} \max_{s \in S^l_i} \mathbf{q}_s \sigma(p, s) - \alpha|S^l_i| + \beta\nu(S^l_i). \qquad (4)$$

Partly inspired by the work on the summarization of online tourists' pictures [25], the function $\kappa(S^l_i, P^l)$ is defined as follows. The first term of Eq.(4) is a *weighted K-means objective* to boost the summary coverage in the (S2). The weight $\mathbf{q}_s$ encourages the summary to include the images with high correspondence inlinks from blogs as stated in (S1) (*i.e.* more an image i has inlinks, a higher $\mathbf{q}_i$ value is assigned). We compute the weight vector $\mathbf{q} \in \mathbb{R}^{|P^l| \times 1}$ as follows. We first build a similarity graph $\mathbf{S}^l$ between the images within P^l, in which consecutive images are connected as k-th order Markov chain ($k = 5$), and the weights are computed by feature similarity. Then we build an integrated similarity matrix $\mathbf{U} = [\mathbf{0}\ \mathbf{W}_{P^l}; \mathbf{W}^T_{P^l}\ \mathbf{S}^l]$ where $\mathbf{W}_{P^l} \in \mathbb{R}^{|\mathcal{I}| \times |P^l|}$ represents the similarity votes from all blog images to photo stream P^l. We compute PageRank vector $\mathbf{v}$ from $\mathbf{U}$, and the last $|P^l|$ dimensional part of $\mathbf{v}$ becomes $\mathbf{q}$. The second term of Eq.(4) penalizes the summary with too many images for brevity, and the third term is the *variance* of image descriptors of S^l_i to encourage the diversity. The parameters α and β can be tuned via cross validation.

As discussed earlier, the key difficulty in optimizing Eq.(3) is that the number of possible subsets $S \subset P^l$ is exponential. To cope with this, we generate constraints by using the greedy algorithm in Algorithm 1, as widely used in the structural SVM approaches for subset selection problems [15, 29]. The idea is that we select pairs of subsets as summary candidates for constraints, using a greedy algorithm, in which we iteratively add an image to a subset that allows the maximum increase of the objective. Since our subset selection can be regarded a special case of the budgeted max coverage problem, the greedy approach allows the $(1 - 1/e)$-approximation bound [25]. For optimization of Eq.(3), we use the similar online stochastic coordinate descent solver [23]. The only difference from Eq.(2) is ℓ_2-norm penalty, which is the same with conventional SVMs.

Algorithm 1: Greedy algorithm for constraint generation.

Input: (1) A photo streams P^l. (2) A size of subset K
Output: (1) A pair of subsets $(S_i, S_j) \in \mathcal{C}_p$.
1: Initialize $(S_i, S_j) \leftarrow (p_i, p_j) \in P^l$ by randomly sampling two images from P^l according to $\mathbf{q}$.
for $i = 1 \cdots K$ **do**
  2: $S_i \leftarrow \text{argmax}_{p \notin S_i} \kappa(S_i \cup \{p\}, P^l)$. Repeat for j.
  3: If $\kappa(S_i) < \kappa(S_j)$, swap S_i and S_j.

3.3. Interpolation

Once obtaining the results of alignment $\mathbf{W}$ and summarization $\mathcal{S}$, we perform the interpolation between consecutive blog images as follows. We first build the symmetric adjacency matrix $\mathbf{S}^l$ for each photo stream P^l, in which we represent a sequence of images in the photo stream by the first-order Markov chain with the edge weights of feature similarity. We then build a block-diagonal matrix $\mathbf{S} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ by combining S^l of all photo streams. We then create an integrated similarity matrix $\mathbf{V} = [\mathbf{0}\ \mathbf{W}; \mathbf{W}^T\ \mathbf{S}]$, which can be regarded as a big similarity matrix between all blog images and all photo streams.

Given a pair of blog images (I^n_i, I^n_j), we can apply Yen's algorithm [28] to $\mathbf{V}$ to generate K shortest paths between (I^n_i, I^n_j). As a final path, we only pick the images in the summary $\mathcal{S}$. Although the size of $\mathbf{V}$ is very large (*i.e.* $|\mathcal{I}| + |\mathcal{P}|$), the path planning is fast because $\mathbf{V}$ is extremely sparse and consecutive blog images are very close in the graph.

	blogspot	wordpress	typepad
All	15,257 (74,218)	37,240 (53,467)	594 (878)
Travelogue	5,152 (71,934)	4,815 (48,554)	108 (763)
Disney	3,270 (58,311)	2,350 (33,831)	28 (378)

Table 1. Statistics of blog posts and associated pictures in parentheses. Each post is labeled as one of *Travelogue*, *Disney* and *Junk*. The number of posts and associated pictures are (53,091, 128,563).

4. Experiments

We focus on showing that blogs and photo streams indeed help each other for summarization and exploration. We demonstrate the usefulness of blogs toward photo streams with two semantic knowledge transfer tasks: image localization (Section 4.2) and automatic image titling (Section 4.3). We then report the results of our photo stream summary using blogs (Section 4.4). Finally, we evaluate path interpolation between blog photos, by using a large set of photo streams (Section 4.5).

4.1. Datasets

Photo Stream Data. We download images from Flickr by querying keywords related to *Disneyland*. We then manually discard the photo streams that are not relevant to *Disneyland* or include less than 30 images. As a result, we collect 542,217 unique images of 6,026 valid photo streams. Each photo stream is a sequence of images taken by a single photographer within one day.

Blog Data. Table 1 summarizes our blog datasets. We first crawl 53,091 unique blog posts and 128,563 associated pictures from the three popular blog-publishing sites, `blogspot`, `wordpress`, and `typepad` by changing query terms from Google search. Then, the blogs are manually classified into three groups by the experts of the parks: *Travelogue*, *Disney* and *Junk*. The *Travelogue* label indicates the blog posts of our interest, which describes stories and events with multiple images in Disneyland. The *Disney* label is applied to the blog posts that are Disney-related but not travelogues, such as Disney films, cartoons, or merchandise. For experiments, we use the blogs of the *Travelogue*, whose size is 10,075 posts and 121,251 associated images.

4.2. Results of Image Localization

Task. We evaluate whether blogs can infuse semantic knowledge to the photo streams that are often contaminated with missing or noisy tags. Specifically, we evaluate that blogs can help solve the image localization task. *Disneyland* consists of multiple districts (*e.g.* *Tomorrowland*), each of which includes a set of attractions (*e.g.* *Astro Orbiter*, *Star Tours*). The task aims to find at which attraction a given image was taken. For groundtruth, we let expert labelers, with GPS information, to annotate 3,000 images of photo streams, out of which we randomly sample 2,000 images and perform localization tests. We repeat the test ten times.

Method	Top-1 Attr.	Top-5 Attr.	Top-1 Dist.
(JointRSVM)	**9.12%**	**22.83%**	**28.81%**
(KNN+KM)	7.34%	18.62%	24.27%
(DTW+KM)	4.05%	15.31%	21.03%
(VKNN)	5.16%	16.85%	22.12%
(VSVM)	4.63%	15.80%	20.63%
(Rand)	0.93%	4.63%	5.56%

Table 2. Image localization accuracies. We report top-1 and top-5 attraction accuracies, and top-1 district accuracies. Despite (JointRSVM) being only weakly supervised with blog text, it has twice the accuracy of fully supervised (VKNN) and (VSVM).

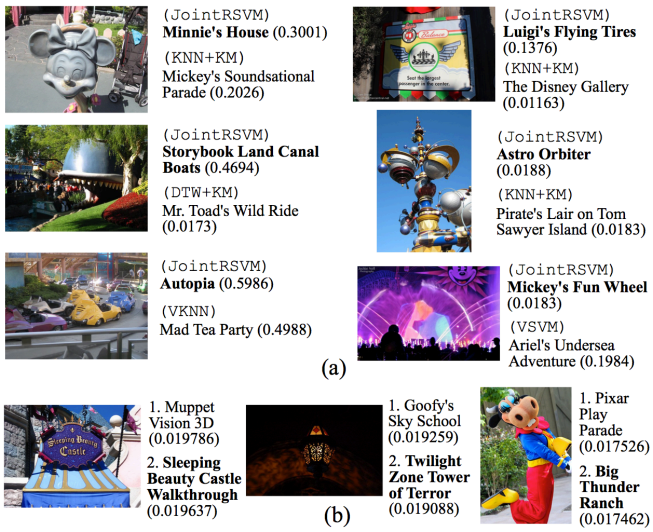


Figure 2. Examples of localization. (a) Comparison of results between our method and the best baseline in each case, with confidence values. (b) Typical near-miss failures. The correct answers (in bold font) are estimated as the second best ones. From left to right: (i) similar appearance of attractions, (ii) too dark images, and (iii) characters only weakly associated with location.

As location classes, we use 108 selected attractions and restaurants from 18 districts of the two Disney parks: *Disney California Adventure* and *Disneyland park*. We find the list of attractions from visitors' map. In supplementary we describe details of how we apply our method and baselines.

Baselines. We compare our method with four baselines. We test two vision-based methods, which are solely based on the visual contents of images. We download 100 top-ranked images from each of Google and Flickr by querying the 108 attraction names. Using these images as training data, we learn linear SVM and KNN classifiers, which are denoted by (VSVM) and (VKNN), respectively. This comparison can justify the usefulness of blog data. We also implement two baselines that use the same blog data but different algorithms. The (KNN+KM) uses the KNN search for alignment between blogs and photo streams, and K-means clustering for summarization of photo streams. The (DTW+KM) exploits DTW (dynamic time warping) for alignment and the same K-means clustering for summarization.

vs. Original title
64.2% (1605/2500)

# Votes	5	4	3	2	1	0
# Samples	156	92	94	60	55	43

Table 3. The results of image titling via AMT. Left: we show the percentage of turkers' responses that our prediction is better than original titles. Right: we present the number of test samples according to how many turkers voted for our titles. Each image is evaluated by five different turkers, and thus for 156 samples our titling is unanimously chosen by turkers.

We use (JointRSVM) to denote our full model. For the location keyword transfer from blogs to photo streams, we use named entity locations extracted using the method in Section 2.3. We also report chance performance (Rand) to show the difficulty of the localization task.

Results. Table 2 reports localization accuracies. The methods that utilize the blog text outperform the vision-based algorithms. Our method is the most accurate, despite being only weakly supervised, achieving almost twice the performance of fully supervised vision-based baselines of (VSVM) and (VKNN). The performance of (DTW+KM) is low because the image ordering in blogs and photo streams often does not match. We note that the task is challenging, with the chance of under 1%. In many cases, even for an human expert, the images are difficult to localize as content may match different locations (*e.g.* Mickey can be observed virtually at any location). Fig.2 shows examples of localization comparison with the best baselines in the top three rows, while the last row shows typical near-miss failures.

4.3. Results of Automatic Image Titling

Task. Image titling is the task of automatically generating a descriptive title of an image. The titles of online images are often missing or automatically assigned by cameras with meaningless codes (*e.g.* IMG1136.jpg). In this tests, we quantify how much blogs improve the existing titles of the photo streams via Amazon Mechanical Turk (AMT). We randomly sample 500 images out of photo streams, and generate titles by transferring semantic key phrases obtained in Section 2.3, over the alignment links discovered by our approach. We show to a turker a query image I_q, and original and estimated titles in a random order. We then ask turkers to choose which one is better for the image. We obtain answers from five different turkers for each query.

Results. Table 3 reports the results. Even considering a certain degree of unavoidable noisiness of AMT labels, our output is significantly preferred by AMT annotators. Our algorithm gains **64.2**% of votes, which justify our assumption that blogs help improve semantic understanding of possibly noisy Web images. Note, some test images have high-quality titles assigned by users, but not many.

Fig.3 illustrates examples of query images with actual and estimated titles. The top two rows present instances where our titles are better than the original, and the last row shows *failure* cases where the original titles were better.

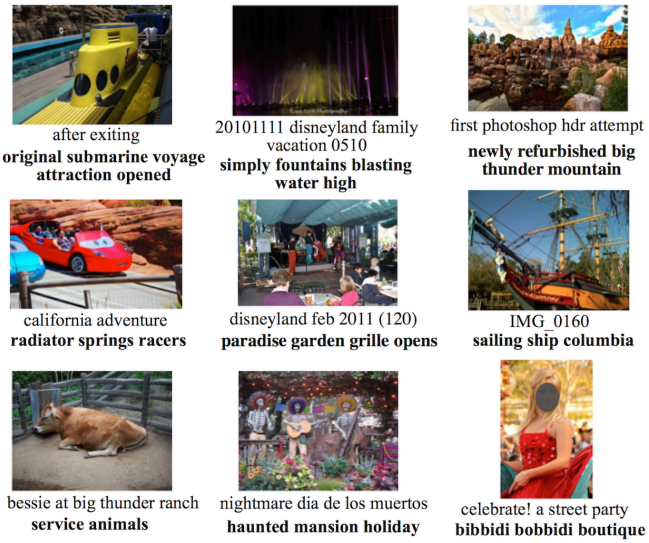


Figure 3. Examples of titling. Top two rows illustrate instances where our automatically generated titles (in bold) are better than originals (in plain). Our titles find and assign correct attraction names (*e.g. radiator springs racers*) or replace meaningless titles with richer and more meaningful ones (*e.g.* img_0160 to *sailing ship columbia*). The third row shows near-miss failure cases. In all the misses the original titles have high quality. From left to right: (i) our title is correct but less specific (*e.g.* the cow's name is *bessie*), (ii) skeletons in the image lead to our prediction toward the *haunted mansion*, (iii) *bibbidi bobbidi boutique* is an attraction where many princesses can be seen.

4.4. Results of Photo Stream Summary

Fig.4 shows qualitative comparison of summarization results for three selected photo streams. We show the top 6 images that are selected by our algorithm at two iterations ($t = 0$ and $t = 2$). Our algorithm converges within 2–3 iterations in most cases. We also show a simple (Unif) baseline that uniformly samples images. The results at $t = 0$ represent an initial content-based summary in which we apply K-means clustering to the visual descriptors of images. The (Unif) is at risk of selecting semantically meaningless images (*e.g.* , an image with plain water in the 3rd examples). The $\mathcal{S}^{(0)}$ may suffer from the limitation of using low-level features only. For example, if a photo stream is unstructured and includes many poorly-taken pictures, the summary can include such displeasing images. On the other hand, although our method uses the same low-level features, it can easily discover representative images thanks to similarity votes by the blog images that blog writers carefully choose with sufficient semantic intent and value.

4.5. Results of Interpolation between Blog Pictures

We now show quantitative and qualitative results of interpolation between blog images using a large set of photo streams. Groundtruth is unavailable; hence we perform user studies via AMT. We first randomly sample 300 pairs of

Figure 4. Qualitative comparison of photo stream summarization. We show the top six images of the summaries created by our method (at initialization and after 2 iterations) and the baseline (Unif). The results become more semantically meaningful after the two iterations.

Figure 5. Examples of the interpolations between blog images. The left pair of images illustrates consecutive blog query images, and the right photo sequences are the interpolation results by different algorithms.

consecutive images in the blogs as a test set denoted by $(I_q^1, I_q^2) \in \mathcal{I}_Q$. We run our algorithm and baselines to generate the most probable sequence of images between I_q^1 and I_q^2. On AMT we show I_q^1 and I_q^2 and then a pair of image sequences predicted by our method and one of the baselines. We ask a turkers to choose the most likely result. We obtain answers from five turkers for each query pair (I_q^1, I_q^2).

We compare our algorithm (JointRSVM) with two baselines that jointly use blog posts and photo streams: (KNN+KM), and (DTW+KM). Table 4 shows the results of pairwise AMT preference tests between our method and the two baselines. The number indicates the mean percentage of responses that choose our prediction as the most likely one to come between I_q^1 and I_q^2. Although the question is rather subjective, our algorithm significantly outperforms the baselines. Results of our method (JointRSVM) are preferred in **61.9**% and **66.5**% of test cases over (KNN+KM) and (DTW+KM) baselines.

Fig.5 shows examples of the predicted image sequences by our algorithm and two baselines. On the left of each set, we show two *query* blog images; we show the estimated images by our method and two baselines in the rows. Since the query blog images and the retrieved images are disjoint, each algorithm can only return similar (but not identical) images at best, from the photo streams of other users. In most cases, our interpolations are more coherent and representative of the query images (*Darth Vader* in bottom right).

vs. (KNN+KM)	# Votes	5	4	3	2	1	0
61.9% (929/1500)	# Samples	23	91	99	69	15	3
vs. (DTW+KM)	# Votes	5	4	3	2	1	0
66.5% (997/1500)	# Samples	51	90	94	50	15	0

Table 4. The results of blog photo interpolation via AMT pairwise preference tests between our method and two baselines. The numbers indicates the percentage of responses that our prediction is more likely to occur between two query blog photos.

Experimental results clearly show that we can leverage photo streams to densely fill in detail among sparsely selected blog images. For example, in blogs one or two images may be chosen for a given attraction. They may be representative snapshots, but fail to capture the progression of shows or experiences that our interpolation can fill in.

5. Conclusion

We proposed an approach that takes advantage of large collections of photo streams and blog posts, for joint story-based summarization and exploration. To achieve this, we design alternating optimization over two latent Ranking SVM problems for alignment and summarization. On a newly collected large-scale dataset of blogs and Flickr photo streams for Disneyland, we showed that blogs and photo streams are mutually beneficial for a variety of tasks.

References

[1] A. Aizawa. An Information-Theoretic Perspective of TF-IDF Measures. *Info. Proc. Manag.*, 39(1):45–65, 2003. 3

[2] P. Das, C. Xu, R. F. Doell, and J. J. Corso. A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching. In *CVPR*, 2013. 2

[3] K. El-Arini, G. Veda, D. Shahaf, and C. Guestrin. Turning Down the Noise in the Blogosphere. In *KDD*, 2009. 2

[4] S. Fidler, A. Sharma, and R. Urtasun. A Sentence is Worth a Thousand Pixels. In *CVPR*, 2013. 2

[5] A. S. Gordon and R. Swanson. Identifying Personal Stories in Millions of Weblog Entries. In *ICWSM Data Challenge Workshop*, 2009. 2

[6] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-shot Recognition. In *ICCV*, 2013. 2

[7] R. Ji, X. Xie, H. Yao, and W.-Y. Ma. Mining City Landmarks from Blogs by Graph Modeling. In *ACMMM*, 2009. 2

[8] T. Joachims. Training Linear SVMs in Linear Time. In *KDD*, 2006. 4

[9] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-Plane Training of Structural SVMs. *Mach Learn*, 77:27–59, 2009. 2, 4

[10] G. Kim and E. P. Xing. Joint Summarization of Large-scale Collections of Web Images and Videos for Storyline Reconstruction. In *CVPR*, 2014. 2

[11] G. Kim and E. P. Xing. Reconstructing Storyline Graphs for Image Recommendation from Web Community Photos. In *CVPR*, 2014. 2

[12] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby Talk: Understanding and Generating Image Descriptions. In *CVPR*, 2011. 2

[13] A. Kushal, B. Self, Y. Furukawa, D. Gallup, C. Hernandez, B. Curless, and S. M. Seitz. Photo Tours. In *3DIMPVT*, 2012. 2

[14] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *ICML*, 2001. 3

[15] L. Li, K. Zhou, G.-R. Xue, H. Zha, and Y. Yu. Enhancing Diversity, Coverage and Balance for Summarization through Structure Learning. In *WWW*, 2009. 4, 5

[16] B. A. Nardi, D. J. Schiano, M. Gumbrecht, and L. Swartz. Why We Blog. *CACM*, 47(12):41–46, 2004. 1

[17] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *NIPS*, 2011. 2

[18] A. Qamra, B. Tseng, and E. Y. Chang. Mining Blog Stories Using Community-based and Temporal Clustering. In *CIKM*, 2006. 2

[19] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating Video Content to Natural Language Descriptions. In *ICCV*, 2013. 2

[20] S. Rose, D. Engel, N. Cramer, and W. Cowley. Automatic Keyword Extraction from Individual Documents. In *Text Mining: Applications and Theory*, pages 1–20. John Wiley and Sons, Ltd, 2010. 3

[21] B. C. Russell, R. Martin-Brualla, D. J. Butler, S. M. Seitz, and L. Zettlemoyer. 3D Wikipedia: Using Online Text to Automatically Label and Navigate Reconstructed Geometry. In *SIGGRAPH Asia*, 2013. 2

[22] E. F. T. K. Sang and F. D. Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *CONLL*, 2003. 3

[23] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. In *ICML*, 2007. 5

[24] S. Shalev-Shwartz and A. Tewari. Stochastic Methods for L1 Regularized Loss Minimization. In *ICML*, 2009. 5

[25] I. Simon, N. Snavely, and S. M. Seitz. Scene Summarization for Online Image Collections. In *ICCV*, 2007. 2, 5

[26] N. Snavely, S. M. Seitz, and R. Szeliski. Photo Tourism: Exploring Photo Collections in 3D. In *SIGGRAPH*, 2006. 2

[27] C. Sutton and A. McCallum. Piecewise Pseudolikelihood for Efficient Training of Conditional Random Fields. *ICML*, 2007. 3

[28] C.-N. J. Yu and T. Joachims. Learning Structural SVMs with Latent Variables. In *ICML*, 2009. 2, 4, 5

[29] Y. Yue and T. Joachims. Predicting Diverse Subsets Using Structural SVMs. In *ICML*, 2008. 5

[30] C. L. Zitnick, D. Parikh, and L. Vanderwende. Learning the Visual Interpretation of Sentences. In *ICCV*, 2013. 2