Jointly Aligning and Segmenting Multiple Web Photo Streams for the Inference of Collective Photo Storylines

Gunhee Kim Eric P. Xing
School of Computer Science, Carnegie Mellon University
{gunhee,epxing}@cs.cmu.edu

Abstract

With an explosion of popularity of online photo sharing, we can trivially collect a huge number of photo streams for any interesting topics such as scuba diving as an outdoor recreational activity class. Obviously, the retrieved photo streams are neither aligned nor calibrated since they are taken in different temporal, spatial, and personal perspectives. However, at the same time, they are likely to share common storylines that consist of sequences of events and activities frequently recurred within the topic. In this paper, as a first technical step to detect such collective storylines, we propose an approach to jointly aligning and segmenting uncalibrated multiple photo streams. The alignment task discovers the matched images between different photo streams, and the image segmentation task parses each image into multiple meaningful regions to facilitate the image understanding. We close a loop between the two tasks so that solving one task helps enhance the performance of the other in a mutually rewarding way. To this end, we design a scalable message-passing based optimization framework to jointly achieve both tasks for the whole input image set at once. With evaluation on the new Flickr dataset of 15 outdoor activities that consist of 1.5 millions of images of 13 thousands of photo streams, our empirical results show that the proposed algorithms are more successful than other candidate methods for both tasks.

1. Introduction

As online sharing of personal photo streams is becoming popular, many of such photo streams often share overlapping contents. For example, one can easily download a huge number of photo streams associated with the query term *scuba+diving* from any photo sharing sites such as Flickr. The retrieved photo streams record various events and activities associated with *scuba+diving*, which are captured by different people from their unique experiences. Obviously, the photo streams are neither aligned nor calibrated since they are taken in different temporal, spatial, and personal perspectives. However, at the same time, they are likely to share *common storylines* consisting of sequences of events and activities repeatedly recurred across the *scuba+diving* photo streams (*e.g.* riding a boat, wearing equipment, underwater diving, and so on). The construction of such photo storylines can potentiate a variety of applications. For example, if a family decides to go to a scuba diving trip, they can make a plan by previewing what other people usually do. After the trip, they can also review the similarities and differences of their trip compared to others.

Figure 1. Motivation for jointly aligning and segmenting multiple photo streams with an example of three photo streams of *scuba+diving*. The input is any number of photo streams of a specific activity that are taken by various users at different time and places. The output is two-fold. (a) Photo stream alignment. The images of different photo streams are matched (as shown in the same colors). (b) Image cosegmentation. The shared regions in the aligned images are jointly segmented.

Therefore our challenging goal is to build such collective storylines from the photo streams of millions of users, and to discover the relations between the reconstructed storylines and photo streams of individual users. In this pa-

per, as a first technical step to achieve this ultimate goal, we propose a method to jointly perform *alignment* of multiple photo streams and *cosegmentation* of aligned images, as shown in Fig.1. In the alignment step, images of different photo sets are matched based on visual contents and associated meta-data. The alignment is a core task to build a big picture of storylines from a large number of fragmented photo streams of individual users. In the cosegmentation step, the aligned images are segmented together in order to facilitate image understanding such as pixel-level classification in the images. It is important to note that solving these two tasks are *mutually rewarding*. The main challenge of cosegmenting multiple photo streams is that the Web images by general users are too diverse to segment all at once. Jointly segmenting images with no commonality, which contradicts the basic assumption of cosegmentation, could be worse than individually segmenting each image. Therefore, the alignment step fills in the role of enabling grouping of images that share sufficient commonality, which provides a high-level clue for cosegmentation. Conversely, once we parse each image into multiple segments, image matching, a basic operation for the photo stream alignment can be improved. We can iterate these two steps in multiple rounds.

In our approach, photo stream alignment and image cosegmentation are achieved in a similar way. For the alignment, we first establish a sparse graph that connects similar photo streams to be aligned together as a Markov random field. Then, we perform belief propagation to jointly align all photo steams at once. Likewise, for image cosegmentation, we build a graph linking the coherent images that are beneficial to be segmented together, based on the output of the alignment step. Then, we perform cosegmentation of the entire image set all at once under the guidance of the graph by a message-passing style optimization.

For evaluation, we collect about 1.5 millions of images of 13 thousands of photo streams regarding 15 outdoor recreational activities from Flickr. Our experiments in Section 5 demonstrate that our approach outperforms other candidate methods on both photo stream alignment and image cosegmentation.

1.1. Previous work

While there has been little work on jointly aligning and segmenting multiple photo streams, the following two lines of research are remotely related to our work.

Cosegmentation: Our problem involves segmenting aligned photo streams together. It resembles the cosegmentation problem [1, 8, 10, 11, 15, 20], whose objective is to jointly segment recurring objects (or foregrounds) that are shared in multiple images. Our work is unique in several respects comparing to the large body of previous cosegmentation research. First, we focus on segmentation of unordered multiple Web photo streams. The cosegmentation of Flickr photo streams was discussed in [10], but it was applied to at most 20 images that are manually selected out of hundreds of pictures of a single Flickr photo stream. In contrast, here we can handle an arbitrary number of uncalibrated Web photo streams by closing the loop between segmentation and photo stream alignment. Second, in our experiments, we perform scalable segmentation with more than 100K images of 1K photo streams, which exceeds those of previous work by two orders of magnitude. To our knowledge, the largest dataset sizes in previous work are about 1K [10, 11].

Large-scale image alignment: Image alignment has been one of fundamental tasks in a variety of computer vision problems. Recently, with the explosion of pictures available online, image alignment has become a key building block to solve various large-scale problems. Some notable examples include the reconstruction of 3D models of landmarks [19], the localization of tourists' photos [3], spatio-temporal reconstruction of time-varying 3D city models [17], and nonparametric object recognition and scene parsing [12]. However, their objectives of the image alignment are quite different from ours, which is to integrate with a subsequent image segmentation to infer common storylines of outdoor activities. As far as we know, [21] is one of the very few papers that involve the alignment of multiple photo streams. However, their algorithm was tested with relatively small datasets (*i.e.* 12 classes with less than 10 photo streams per class) compared to ours (*i.e.* 15 outdoor activities with 1K photo streams per activity). More importantly, they did not explore any sub-image level analysis; no image segmentation is performed.

1.2. Summary of Contributions

To conclude the introduction, we summarize the main contributions of this paper as follows.

(1) We propose an approach to jointly aligning and segmenting large-scale Web photo streams of different users. Compared to previous cosegmentation research, our approach can handle any number of uncalibrated photo streams. Compared to existing image alignment research, our work can widen its applicability for reconstructing collective storylines from multiple photo streams by closing the loop with cosegmentation in a mutually rewarding way.

(2) We propose large-scale alignment and cosegmentation algorithms that jointly work on the whole dataset by using message-passing based optimization. The algorithms are scalable; they run in a linear time with the number of photo streams and images, respectively.

(3) In experiments, we evaluate the proposed approach with our new Flickr dataset of 15 outdoor activities. Our largest experiments run on more than 100K images of 1K photo streams, which exceed those of previous work by orders of magnitude. We also show the superiority of our approach over other candidate methods for both tasks.

2. Approach

In this section, we describe the problem definition and the overview of our solution to the problem.

2.1. Problem Formulation

The input of our algorithm is the set of photo streams of a particular activity denoted by $\mathcal{P} = \{P^1, \cdots, P^L\}$, where L is the number of input photo streams. Each photo stream is a set of photos taken in sequence by a single photographer within a certain period of time, which is set to a single day in this paper. Without loss of generality, we assume that each photo stream is sorted by taken time. We also use $\mathcal{I} = \{I_1, \cdots, I_N\}$ to denote the whole image set without distinguishing the membership of photo streams. As a notation convention, we use superscripts to denote photo stream numbers and subscripts to denote image numbers.

Another input is related to the segmentation task; a user can provide the maximum number of foregrounds of interest per image K. Then, our algorithm automatically identifies K most dominant regions that are distinctive one another from the image and its aligned neighbors[1]. The background is defined as all the other regions that are not included in any of K foregrounds. For notational simplicity, we interchange the term background and foreground $K+1$.

The output of our algorithm is two-fold. The first output for the alignment is the set of correspondences between the images of different photo streams. If we represent each image as a vertex and each correspondence as an edge, the output can be summarized as an L-partite graph. The second output for the segmentation is assigning every pixel of each image to one of K foregrounds or background.

2.2. Overview of Algorithm

Our approach alternates between solving two target tasks, photo stream alignment and image cosegmentation. Given a large set of uncalibrated photo streams, we first build a nearest neighbor similarity graph that connects the photo streams to be aligned (see section 3.4). We formulate the alignment of the whole photo streams as an energy minimization problem, which can be solved by belief propagation on the graph. Its detailed procedure will be explained in section 3.3 and 3.4. As a result of the alignment, we can obtain the correspondences between the images of different photo streams, from which we establish an image graph connecting the similar images that are likely to share common foregrounds (see section 4.1). We perform large-scale cosegmentation for all images at once under the guidance of the image graph in a message-passing way, which will be discussed in section 4.2. The segmentation of images can enhance the similarity measurement between images, which subsequently contributes to a better photo stream alignment. This will be justified in section 3.2 with an intuitive example. Finally, we can return to the photo stream alignment step with the new segmentation-based image similarity.

[1] In segmentation literature, it is called an *unsupervised* setting. A user may provide some foreground examples in the form of bounding-boxes or pixel-wise annotations, which is called a *supervised* setting. In this paper, we focus on the unsupervised case because it is more challenging. Also, it is trivial to adapt our approach to the supervised setting.

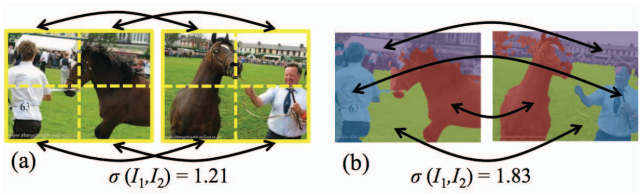


Figure 2. The benefit of segmentation for measuring image similarity. In this example, the same objects appear in different locations with different poses across the image pair. (a) When images are not yet segmented, we compute the image similarity from the spatial pyramid histograms on the whole images. (b) Once images are segmented, we find the best assignment between the segments of two images, and compute the mean of segment similarities.

3. Alignment of Photo Streams

We begin with our image description and similarity measure, and then discuss the proposed alignment algorithm.

3.1. Image Description

We use the dense feature extraction with vector quantization, which is one of standard methods in recent computer vision research. We densely extract two features from each image: HSV color SIFT and histogram of oriented edge (HOG) feature on a regular grid at steps of 4 and 8 pixels, respectively. Then, we 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 and segment descriptors, we build L_1 normalized spatial pyramid histograms to count the frequency of each visual word in multiple levels of regular grids.

3.2. Image Similarity Measure

It is vital to design a reliable similarity metric between images for an accurate alignment of photo streams. Our alternating approach is based on the assumption that the segmentation is helpful to enhance the measurement of image similarity. Fig.2 shows a typical example of such intuition where the same objects appear in different locations with different poses across the images. When images are not segmented yet, the image similarity is calculated from two-level spatial pyramid histograms on the whole images, which are not robust against location and pose variations. However, this issue can be largely alleviated even with an imperfect segmentation. Given the segment sets of two images I_1 and I_2, denoted by $\mathcal{F}_1$ and $\mathcal{F}_2$, we first solve the linear assignment problem (*i.e.* finding the best assignment

between the segments of two images), and then compute the mean of total similarity values as an image similarity metric. Formally, given a similarity metric between segments $\sigma_s : \mathcal{F}_1 \times \mathcal{F}_2 \rightarrow \mathsf{R}$, the image similarity σ is defined by

$$\sigma(I_1, I_2) = \max \left(\sum_{s \in \mathcal{F}_1} \sigma_s(s, f_s(s)) \right) / M \qquad (1)$$

where $f_s : \mathcal{F}_1 \rightarrow \mathcal{F}_2$ is a bijection and M is the number of segments. We use as σ_s the histogram intersection on the spatial pyramid histograms of the segments.

3.3. Pairwise Photo Stream Alignment

For a better understanding, our discussion starts from the alignment of a pair of photo streams P^1 and P^2. That is, the objective is to establish the correspondences between two photo streams through image matching. Our alignment objective is formulated based on the MRF energy function that has been applied to many computer vision problems such as deformable image matching [18] and SIFT flow [12]. Its strength lies in its flexibility to easily incorporate various energy terms related to alignment. It is of particular interest for our applications since we can leverage the terms regarding the meta-data associated with the images.

The goal of alignment is to find a matching $f : P^1 \rightarrow P^2 \cup \{\emptyset\}$ where $\emptyset$ is the null, meaning that if $f(p_i) = \emptyset$ for an image $p_i \in P^1$, p_i has no correspondence in P^2. Let $\widehat{p_i} \in P^2 \cup \{\emptyset\}$ denote the matched image to $p_i \in P^1$. The pairwise alignment is performed by minimizing the energy function as follows.

$$\begin{aligned} E(P^1, P^2) = &- \sum_{p_i \in P^1} \sigma(p_i, \widehat{p_i}) + \sum_{p_i \in P^1} \eta \min(|t(p_i) - t(\widehat{p_i})|, \tau) \\ &+ \sum_{(p_i, p_j) \in \Delta} \rho \, \sigma(p_i, p_j) \min(|t(\widehat{p_i}) - t(\widehat{p_j})|, \nu) \end{aligned} \qquad (2)$$

where τ and ν are the thresholds for truncated L_1 norms, and η and ρ are term weights. We let $t(p_i)$ be the timestamp of image p_i. The $\sigma(p_i, \widehat{p_i})$ is the image similarity between p_i and $\widehat{p_i}$. We let $\sigma(p_i, \emptyset) = 0$ and $t(\emptyset) = \infty$, which means that if $\min_{p_j \in P^2} \sigma(p_i, p_j) < \eta\tau + \rho\nu$, then p_i matches no image in P^2. The Δ contains the entire temporal neighborhood in a photo stream (*i.e.* $(p_i, p_j) \in \Delta$ means $|t(p_i) - t(p_j)| \leq \delta$). The first term accounts for the maximization of image similarity between the matched pairs, and the second term penalizes the time difference between the matched pairs. The third one is the smoothness term to encourage that the matched images to the neighbors in P^1 are also neighbors in P^2. This regularization is more strongly imposed for a pair of images that are more visually similar by weighting $\sigma(p_i, p_j)$. The optimization of Eq.(2) can be achieved by using the belief propagation [6, 12].

3.4. Multiple Photo Stream Alignment

We extend the pairwise alignment of Eq.(2) to that of an arbitrary number of photo streams $\mathcal{P}$. One naive approach may be to incrementally combine pairwise alignments starting from the most similar photo stream pair and progressing to the most distant one. However, this approach has two significant drawbacks [4]. First, it tends to be computationally intensive. Second, more importantly, this method does not treat all photo streams equally, which may lead to local minima according to the order of consideration.

To circumvent these issues, we jointly align all photo streams at once after constructing a graph between photo streams $\mathcal{G}_P = (\mathcal{P}, \mathcal{E}_P)$. For each photo stream $P^i \in \mathcal{P}$, we first find a set of photo streams that are sufficiently overlapped on timeline (*i.e.* the photo streams P^j such that (# of images of P^j within the time range of P^i)/ (total # of images P^j) $\geq \gamma$). Among them, we obtain K_P-nearest neighbors in terms of visual similarity, which is calculated by using the idea of Naive-Bayes Nearest-Neighbor [2] as follows. Given two photo streams P^i and P^j, for each image $p \in P^i$, we obtain the first nearest neighbor in P^j denoted by $\mathrm{NN}(p)$. Then, the similarity from P^i to P^j is computed by $\sum_{p \in P^i} \|\sigma(p, \mathrm{NN}(p))\|^2$. Finally, $\mathcal{E}_P$ includes all pairs of nearest neighbor photo streams.

The objective of multiple photo stream alignment reduces to find a matching $f : P^i \rightarrow P^j \cup \{\emptyset\}$ for all pairs $(P^i, P^j) \in \mathcal{E}_P$, which can be accomplished by minimizing

$$E = \sum_{(P^i, P^j) \in \mathcal{E}_P} E(P^i, P^j) \qquad (3)$$

where $E(P^i, P^j)$ is defined by Eq.(2). The optimization can be achieved by the belief propagation on the graph of photo streams $\mathcal{G}_P$, in such a way that we repeat a pairwise alignment of previous section by following the edges of $\mathcal{E}_P$ until convergence.

4. Large-Scale Cosegmentation

In this section, we explain our algorithm to construct an *image graph* and jointly segment the whole image set.

4.1. Building An Image Graph

For large-scale cosegmentation, we establish an *image graph* $\mathcal{G}_I = (\mathcal{I}, \mathcal{E}_C)$ where $\mathcal{I}$ is the set of images of all photo streams, and $\mathcal{E}_C$ is the set of edges that connect the images that share enough commonality to be segmented together. The edge set consists of two groups: $\mathcal{E}_C = \mathcal{E}_B \cup \mathcal{E}_W$ where $\mathcal{E}_B$ defines the edges between the images of different photo streams while $\mathcal{E}_W$ connects the images within the same photo stream. $\mathcal{E}_B$ is trivially obtained from the output of photo stream alignment; simply, all correspondences of image pairs are added to $\mathcal{E}_B$. $\mathcal{E}_W$ is useful for cosegmentation because the images in the same photo stream are

consecutively taken by the same camera, and thus they are likely to share common objects and scenes. In order to define $\mathcal{E}_W$, we find K_W-nearest neighbors for each image I_i among its temporal neighborhood in the same photo stream, which includes all images I such that $|t(I) - t(I_i)| \leq \delta$. In our experiments, δ is set to 2 hours.

4.2. Running Cosegmentation

We begin with some basic ingredients of our cosegmentation algorithm. We first oversegment every image of $\mathcal{I}$ by using the submodular image segmentation [11]. Let $\mathcal{S}_i$ denote the set of oversegments of image I_i. Then, the goal of segmentation reduces to finding an optimal disjoint partition $\mathcal{S}_i = \bigcup_{k=1}^{K+1} \mathcal{F}_i^k$ with $\mathcal{F}_i^k \cap \mathcal{F}_i^l = \emptyset$ if $k \neq l$, where $\mathcal{F}_i^k$ denotes the regions of foreground k in image I_i.

MFC algorithm: In our approach, we select the MFC [10] as our base cosegmentation algorithm, since it is scalable and has been successfully tested with Flickr user images. More specifically, we exploit two procedures of the MFC algorithm as our basic operations: *foreground modeling* and *region assignment* steps. The foreground models retain the appearance models of K foregrounds and the background. Formally, the k-th foreground model is defined as a parametric function $v^k : 2^{|\mathcal{S}_i|} \rightarrow \mathsf{R}$ that takes any subset $S \subset \mathcal{S}_i$ as input and returns its value to foreground k (*i.e.* how closely region S is relevant to foreground k). Each foreground model is learned from the regions that are allocated to the foreground after the region assignment step. Therefore, the foreground model can be accomplished by using any region classifiers or their combinations. In this paper, we use the Gaussian mixture model (GMM) on the RGB color and HSV SIFT spaces. Thus, $v^k(S)$ is defined as the mean log-likelihood of the descriptors of S to the k-th learned GMM model [14].

The role of the region assignment step is, given a set of learned foreground models $\{v^k\}_{k=1}^{K+1}$, to discover the optimal partition of $\mathcal{S}_i$ into $\{\mathcal{F}_i^k\}_{k=1}^{K+1}$ that maximizes the overall allocation values. We let c_i denote one such partition instance of image I_i. Generally, the set partition problem is NP-complete, but the region assignment of the MFC can solve it in a very efficient way by using combinatorial auction idea. We do not discuss its details, which can be found in [10]. Instead, we denote the region assignment procedure by $\{\mathcal{F}_i^k\}_{k=1}^{K+1} = \mathsf{RegAss}(\mathcal{S}_i, \{v^k\}_{k=1}^{K+1})$. In the following, we use the abbreviated notation of $\{v\}$ for $\{v^k\}_{k=1}^{K+1}$.

Message Passing based Cosegmentation: The basic idea of our large-scale cosegmentation is to iteratively perform foreground modeling and region assignment based on image graph $\mathcal{G}_I$. We view the image graph $\mathcal{G}_I$ as a MRF with hidden variables corresponding to the partition c_i of each image I_i. Consequently, we formulate the cosegmentation of whole image set $\mathcal{I}$ as the following energy maximization:

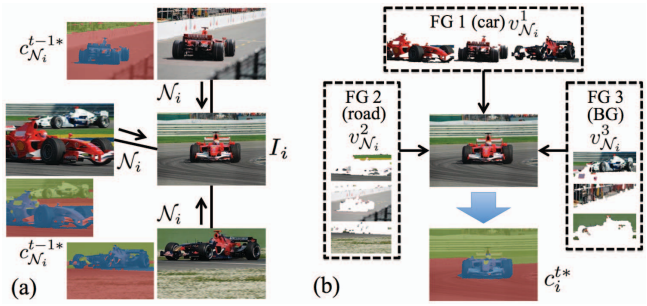


Figure 3. An intuition of our message-passing based cosegmentation at round t. (a) We show an image I_i to be segmented, and its three neighbors N_i in the image graph G_I. We also present color-coded partitions of best beliefs of N_i at $t-1$, denoted by $c^{t-1}_{N_i}$. (b) The message passing from N_i to I_i at round t ends up performing the region assignment for I_i by using the foreground models $\{v_{N_i}\}$ learned from $c^{t-1}_{N_i}$. As a result, we obtain the partition of the best belief of image I_i at t, denoted by c^t_i.

$$D(\mathcal{I}; \mathcal{G}_I) = \alpha \sum_{I_i \in \mathcal{I}} \psi\left(c_i; \{v\}\right) + \sum_{(I_i, \mathcal{N}_i) \in \mathcal{E}_C} \phi\left(c_i; \{v_{\mathcal{N}_i}\}\right) \quad (4)$$

where $\mathcal{N}_i$ denotes the neighborhood of image I_i in image graph $\mathcal{G}_I$, and α is a term weight. $\{v\}$ and $\{v_{\mathcal{N}_i}\}$ indicate the global and local foreground models, respectively. Both of them are implemented by the same region classifiers (*e.g.* GMM models). Only difference is the training data; $\{v_{\mathcal{N}_i}\}$ is learned from the regions of foregrounds only in $\mathcal{N}_i$, whereas $\{v\}$ is obtained without imposing such local restriction.

The objective of Eq.(4) consists of a unary term ψ and a pairwise term ϕ; it means that c_i is achieved by searching for the best partition not only for $\{v\}$ in the unary term ψ but also for $\{v_{\mathcal{N}_i}\}$ in the pairwise term ϕ. For a partition c_i of $\mathcal{S}_i$ into $\{\mathcal{F}_i\}$, the unary term is defined as the sum of assignment scores by $\{v\}$:

$$\psi\left(c_i; \{v\}\right) = \sum_{k=1}^{K+1} v^k(\mathcal{F}_i^k). \quad (5)$$

The pairwise term $\phi\left(c_i; \{v_{\mathcal{N}_i}\}\right)$ is defined as the exact same form of Eq.(5) only except replacing $\{v\}$ by $\{v_{\mathcal{N}_i}\}$.

Optionally, the unary term ψ can be reasonably ignored by setting α to 0, if it is hard to define a single set of globally applicable foreground models. For example, the *person* foregrounds are ubiquitous in all photo sets but their appearances can be severely varied in different photo sets. In this case, using only local models may be more robust.

Messages and beliefs: The energy maximization in Eq.(4) can be solved by the belief propagation, which proceeds by iteratively computing new *messages* for each edge in graph $\mathcal{G}_I$. Using the max-product algorithm (*i.e.* equivalently, the min-sum algorithm with negative log probabilities), the message from $\mathcal{N}_i$ to I_i at round t is defined by [6]

$$m^t_{\mathcal{N}_i \to I_i}(c_i) = \max_{c_{\mathcal{N}_i}} \Big(\phi\left(c_i; \{v_{\mathcal{N}_i}\}\right) + \psi\left(c_{\mathcal{N}_i}; \{v\}\right) + \sum_{s \in \mathcal{N}(\mathcal{N}_i) \setminus I_i} m^{t-1}_{s \to \mathcal{N}_i}(c_{\mathcal{N}_i}) \Big) \quad (6)$$

where $\mathcal{N}(\mathcal{N}_i) \setminus I_i$ denotes the neighbors of $\mathcal{N}_i$ except I_i. According to Eq.(6), the message computation involves the search for the best $c_{\mathcal{N}_i}$ (*i.e.* the partitions of neighbors) for every possible c_i. It results in an exponential explosion of the search space, which is largely unnecessary in practice. Therefore, we introduce an assumption that is reasonable for image cosegmentation as follows. *The best partitions $c_{\mathcal{N}_i}$ for the message $m^t_{\mathcal{N}_i \to I_i}(c_i)$ at round t is the same with those of the best beliefs of $\mathcal{N}_i$ at round $t-1$.*

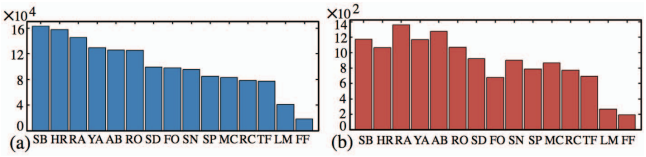
Fig.3 shows an intuitive example of how our message passing works with this assumption. Fig.3.(a) shows the image I_i to be segmented and its three neighbors $\mathcal{N}_i$ in image graph $\mathcal{G}_I$. We also illustrate the color-coded partitions of the best beliefs of $\mathcal{N}_i$ at round $t-1$, which are denoted by $c^{t-1*}_{\mathcal{N}_i}$. As shown in Fig.3.(b), when we compute the message $m^t_{\mathcal{N}_i \to I_i}(c_i)$, the assumption allows us to simply learn foreground models $\{v_{\mathcal{N}_i}\}$ from $c^{t-1*}_{\mathcal{N}_i}$ of Fig.3.(a), and to evaluate each possible c_i. By running $\{\mathcal{F}_i\} = \mathsf{RegAss}(\mathcal{S}_i, \{v_{\mathcal{N}_i}\})$, we can obtain the partition c^{t*}_i (*i.e.* the partition of the best belief of I_i at round t) as a result, which is also shown in Fig.3.(b).

Consequently, the implementation of our message-passing based cosegmentation is straightforward; at every round, we iteratively segment each image I_i by using the learned foreground models from the partitioned regions of its neighbors $\mathcal{N}_i$ at previous round. Then, the segmented image I_i is subsequently used to learn the foreground models for its neighbors' segmentation. That is, we iteratively run foreground modeling and region assignment steps by following the edges of image graph $\mathcal{G}_I$.

Initialization: In order to proceed our iterative cosegmentation algorithm, we need initial image partitions as starting points of belief propagation. In the supervised scenario, we trivially begin from the labeled images. In an unsupervised setting, we apply the diversity ranking method of [11] to image graph $\mathcal{G}_I$ to discover a small number of central images and their neighbors. Then, the unsupervised version of MFC algorithm in [10] initially segments the images of each group, from which message passing begins.

4.3. Analysis of the Algorithm

The core procedures of our approach are the two belief propagation (BP) techniques for alignment and cosegmentation. The alignment BP works on the graph of photo streams while the cosegmentation BP runs on the image graph. Generally, the BP algorithm runs in $\mathcal{O}(T|\mathcal{E}|)$ where T is the number of iterations and $|\mathcal{E}|$ is the number of edges. Since we use only sparse KNN graphs where each vertex is connected to a constant number of neighbors, the alignment BP runs in $\mathcal{O}(TL)$ and the cosegmentation BP does in $\mathcal{O}(TN)$ where L and N are the number of photo streams and images, respectively. Moreover, the BP algorithm has been studied much for parallelization [7], which can further improve the speed of our algorithm. We summarize the pseudocode of our algorithm in supplementary material.



SB: *surfing+beach*, HR: *horse+riding*, RA: *rafting*, YA: *yacht*, AB: *air+ballooning*, RO: *rowing*, SD: *scuba+diving*, FO: *formula+one*, SN: *snowboarding*, SP: *safari+park*, MC: *mountain+camping*, RC: *rock+climbing*, TF: *tour+de+france*, LM: *london+marathon*, FF: *fly+fishing*.

Figure 4. Our Flickr datasets of 15 outdoor recreational activities. The number of images and photo streams are shown in (a) and (b), respectively. The dataset sizes are (1,514,976, 13,157) in total.

5. Experiments

We evaluate the proposed approach from two technical perspectives: photo stream alignment in section 5.1 and image cosegmentation in section 5.2. We present more details of experiments in supplementary material, including experimental design, application of baselines, and in-depth analysis of results. Our Matlab demo code is available at our webpage (`http://www.cs.cmu.edu/ gunhee`).

Flickr Dataset: Fig.4 summarizes our Flickr dataset that consists of 1,514,976 images of 13,157 photo streams for 15 outdoor recreational activity classes. Flickr is one of the best image sources to test our algorithm since a large number of photo streams of different users are freely available with rich associated meta-data. We use the class names as search keywords, and download all the photo streams that contain more than 50 images. We use all pictures of each photo stream without any filtering. For a quantitative segmentation evaluation, we manually annotate 100 images per class, from which we obtain approximate performance measures of algorithms. Although the labeled images are relatively few compared to dataset sizes, in practice the sampled annotation is widely adopted in standard large-scale benchmark datasets such as ImageNet [5].

5.1. Results on Alignment

Tasks: The performance of photo stream alignment is evaluated by a *temporal localization* task. It is inspired by the studies of geolocation estimation [3, 9], whose goal is to estimate the geolocations of individual pictures for a given sequence of a tourist's photos. We carry out our experiments similarly only except that the geolocation is replaced by the timestamp. We first randomly select 80% of photo streams of each class as training set and the others

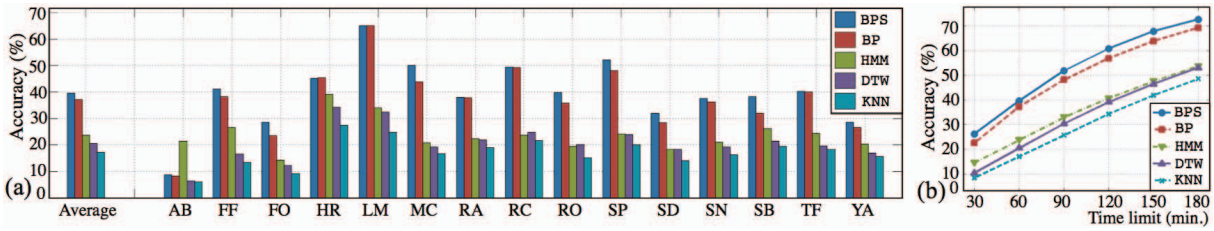


Figure 5. Comparison of temporal localization between our methods (BPS) and (BP) and the baselines (HMM), (DTW), and (KNN). In (a), we show the accuracies of all algorithms for 15 outdoor activity classes with = 60 minutes. In (b), we show the variation of average localization accuracies by changing time thresholds from 30 minutes to 180 minutes. The acronyms of activities are referred to Fig.4.

as test set. Then, the goal is to estimate the timestamps of all the images of the test photo streams by aligning them with training photo streams whose timestamps are known. Such temporal localization task is also important to achieve our ultimate goal, the picture-based storyline construction, which requires correctly locating each photo stream on the timeline to relate it with other photo streams.

Baselines: For the alignment tests, we compare our algorithm with four baselines. As one of the simplest baselines, the (KNN) performs image matching by using only image similarity. We also choose two alternatives of image sequence alignment. The (HMM) is the hidden Markov model method that has been widely applied for localizing tourists' photo sets [3, 9]. The (DTW) is dynamic time warping, one of most popular algorithms for multiple sequence alignment [13]. Our algorithm is tested in two different ways, according to whether image segmentation is in a loop or not. The (BP) does not exploit the image segmentation output whereas the (BPS) is our fully geared approach. That is, this comparison can justify the usefulness of our alternating approach between alignment and segmentation.

Quantitative results: To compare the performances of algorithms, we use the similar evaluation metric to those of image geolocalization research [3, 9]. Given the estimated timestamps of all test images by each algorithm, we compute the percentage of images for which the estimated timestamps are within ϵ minutes of the groundtruths. Fig.5.(a) reports the accuracy rates of our algorithms and baselines across 15 activity classes with $\epsilon = 60$ minutes. The leftmost bar set is the average performance of 15 classes. Our algorithm significantly outperforms all the baselines in most classes. The average accuracy of our method (BPS) is 39.1%, which is notably higher than 23.7% of the best baseline (HMM). Fig.5.(b) compares the average accuracies of all algorithms according to different ϵ values from 30 to 180 minutes. In all ranges of ϵ, our (BPS) consistently outperforms the best baseline (HMM) by 17.1% points on average. Moreover, the accuracies of (BPS) is higher than those of (BP) by 3.6% points on average, which supports that segmentation can improve alignment.

Qualitative results: We present very preliminary results of storyline construction in supplementary material, which hints that our alignment works promisingly for this goal.

5.2. Results on Segmentation

Tasks: The task of image cosegmentation is to identify frequently recurring foregrounds in the image set. The accuracy is measured by the intersection-over-union metric $(GT_i \cap R_i)/(GT_i \cup R_i)$, where GT_i is the groundtruth of image i and R_i is the estimated regions by an algorithm. It is also a standard metric in PASCAL challenge. We compute the average values of this metric from all annotated images.

Baselines: We select three baselines of unsupervised segmentation methods that can discover multiple objects from a large-scale dataset (*i.e.* at least more than tens of thousands of images). The (LDA) [16] is an LDA-based unsupervised localization method, and the (COS) [11] is a state-of-art cosegmentation algorithm based on submodular optimization. We also test the MFC algorithm (MFC) without involving the alignment step; this comparison can quantify the contribution of alignment to cosegmentation. For (COS) and (MFC), we cluster the images into multiple subgroups by K-means on visual features, and apply the methods to each subgroup independently. We run our method and all the baselines in an unsupervised manner (*i.e.* without any seed labels) for a fair comparison. Since it is hard to know the best K beforehand (*e.g.* multiple foregrounds may exist in an image), we repeat each method by changing K from one to five, and report the best results.

Quantitative results: Fig.6 compares the segmentation performance between our method and the three baselines. In almost all classes, the accuracies of our algorithm (BP+MFC) are far better than those of the best baselines. Especially, our average accuracy is 43.5%, which is significantly higher than 34.3% of the best baseline (MFC), which indicates that our alignment step is more successful than simple clustering such as K-means for cosegmenting extremely diverse Web user images.

Segmentation examples: Fig.7 shows some selected examples of cosegmentation. We observe that the subjects and their appearances are severely variable even in the images that are collected with the same keyword. For example, in the *safari+park* class, tens of different animals occur, and in all classes, people are ubiquitously shown with different appearance, poses, and clothes. Moreover, a single class may include multiple other activities; for example,

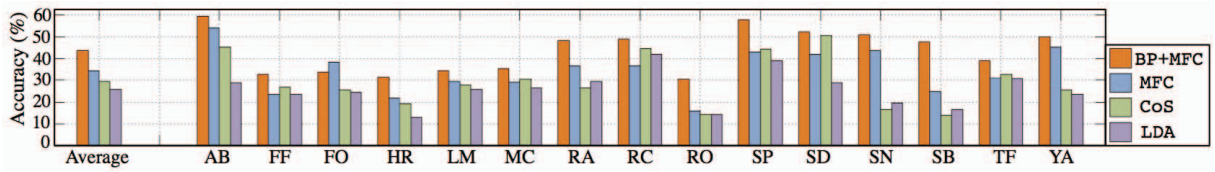


Figure 6. Cosegmentation accuracies between our method (BP+MFC) and the baselines (MFC), (COS), and (LDA) for 15 outdoor activities classes. The leftmost bar set shows the average accuracies. The acronyms of activities are referred to Fig.4.

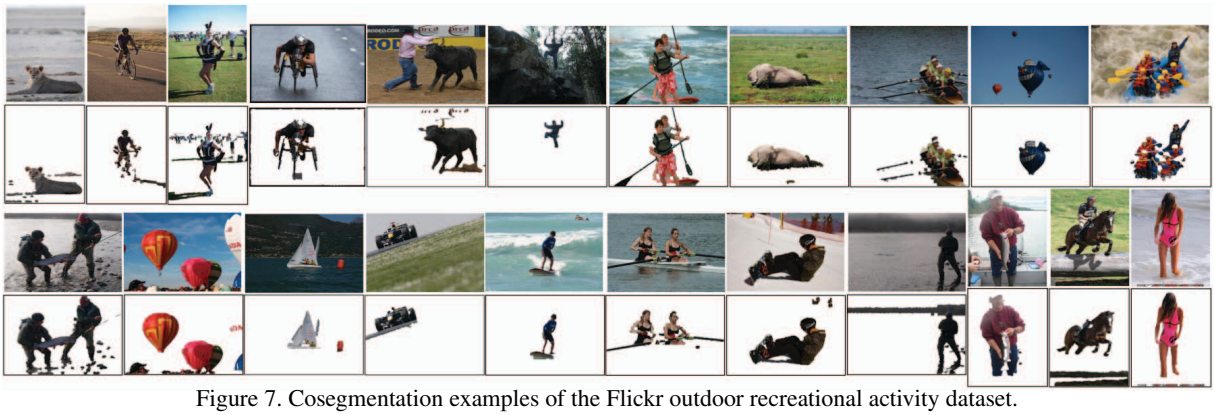


Figure 7. Cosegmentation examples of the Flickr outdoor recreational activity dataset.

the *mountain+camping* class contains the pictures of skiing, trekking, fishing, rock climbing, and hunting. Evidently, for the analysis of Web user images, it is extremely hard to pre-define the objects of interest and learn the classifiers beforehand. In contrast, our approach is greatly successful to quickly align a large-scale image set and segment out common regions in an unsupervised and bottom-up way, which can be a useful function for various Web applications.

6. Conclusion

We proposed a scalable approach to jointly aligning and segmenting multiple uncalibrated Web photo streams of different users. We demonstrated superior alignment and cosegmentation performance for the Flickr outdoor activity dataset over other candidate methods. The empirical results assured that our method can be a key component to achieve our ultimate goal: inferring collective photo storylines from Web images, which is a next direction of our future work.

Acknowledgement: This work is supported by NSF IIS-1115313 and AFOSR FA9550010247.

References

[1] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. Interactively Co-segmentating Topically Related Images with Intelligent Scribble Guidance. *IJCV*, 93:273–292, 2011. 2

[2] O. Boiman, E. Shechtman, and M. Irani. In Defense of Nearest-Neighbor Based Image Classification. In *CVPR*, 2008. 4

[3] C. Y. Chen and K. Grauman. Clues from the Beaten Path: Location Estimation with Bursty Sequences of Tourist Photos. In *CVPR*, 2011. 2, 6, 7

[4] D. Crandall, A. Owens, N. Snavely, and D. Huttenlocher. Discrete-Continuous Optimization for Large-Scale Structure from Motion. In *CVPR*, 2011. 4

[5] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-scale Hierarchical Image Database. In *CVPR*, 2009. 6

[6] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient Belief Propagation for Early Vision. *IJCV*, 70(1):41–54, 2006. 4, 5

[7] J. E. Gonzalez, Y. Low, and C. Guestrin. Residual Splash for Optimally Parallelizing Belief Propagation. In *AISTATS*, 2009. 6

[8] A. Joulin, F. Bach, and J. Ponce. Multi-Class Cosegmentation. In *CVPR*, 2012. 2

[9] E. Kalogerakis, O. Vesselova, J. Hays, A. A. Efros, and A. Hertzmann. Image sequence geolocation with human travel priors. In *ICCV*, 2009. 6, 7

[10] G. Kim and E. P. Xing. On Multiple Foreground Cosegmentation. In *CVPR*, 2012. 2, 5, 6

[11] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade. Distributed Cosegmentation via Submodular Optimization on Anisotropic Diffusion. In *ICCV*, 2011. 2, 5, 6, 7

[12] C. Liu, J. Yuen, and A. Torralba. Nonparametric Scene Parsing: Label Transfer via Dense Scene Alignment. In *CVPR*, 2009. 2, 4

[13] T. M. Rath and R. Manmatha. Word Image Matching Using Dynamic Time Warping. In *CVPR*, 2003. 7

[14] C. Rother, V. Kolmogorov, and A. Blake. GrabCut: Interactive Foreground Extraction using Iterated Graph Cuts. In *SIGGRAPH*, 2004. 5

[15] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of Image Pairs by Histogram Matching Incorporating a Global Constraint into MRFs. In *CVPR*, 2006. 2

[16] B. C. Russell, A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006. 7

[17] G. Schindler and F. Dellaert. Probabilistic Temporal Inference on Reconstructed 3D Scenes. In *CVPR*, 2010. 2

[18] A. Shekhovtsov, I. Kovtun, and V. Hlavac. Efficient MRF Deformation Model for Non-Rigid Image Matching. In *CVPR*, 2007. 4

[19] N. Snavely, I. Simon, M. Goesele, R. Szeliski, and S. M. Seitz. Scene Reconstruction and Visualization from Community Photo Collections. *Proc. IEEE*, 98(8):1370–1390, 2010. 2

[20] S. Vicente, C. Rother, and V. Kolmogorov. Object Cosegmentation. In *CVPR*, 2011. 2

[21] J. Yang, J. Luo, J. Yu, and T. Huang. Photo Stream Alignment for Collaborative Photo Collection and Sharing in Social Media. In *WSM*, 2011. 2